THÈSE DE DOCTORAT
DE SORBONNE UNIVERSITÉ

Spécialité : Physique

École doctorale n°564 : Physique en Île-de-France

réalisée

au Laboratoire de Physique Théorique de la Matière Condensée

sous la direction de Maria BARBI

présentée par

Antony LESAGE

pour obtenir le grade de :

DOCTEUR DE SORBONNE UNIVERSITÉ

Sujet de la thèse :

Inférence bayésienne des paramètres structuraux de la chromatine de la drosophile sur des images super-résolues de domaines épigénétiques

---

soutenue le 20 septembre 2019

devant le jury composé de :

M.	Diego CATTONI	Rapporteur
M.	John MARKO	Rapporteur
Mme	Isabelle CALLEBAUT	Examinatrice
Mme	Karine DUBRANA	Examinatrice
M.	Cédric VAILLANT	Examinateur
Mme	Maria BARBI	Directrice de thèse
M.	Vincent DAHIREL	Invité

Remerciements

Qu'importe l'issue du chemin quand
seul compte le chemin parcouru.

David Le Breton

La thèse est un moment charnière autant sur le plan scientifique que personnel, empli de doutes et jalonné de succès, où « qu'importe l'issue [...] quand seul compte le chemin parcouru ».

Je tiens tout particulièrement à exprimer ma gratitude envers ceux qui m'ont épaulé tout au long de ce chemin.

À Maria BARBI pour m'avoir ouvert les portes de son équipe pour y faire ma thèse, dans la continuité de mon stage de M2, et de l'avoir dirigée.

À Vincent DAHIREL pour m'avoir encadré durant ma thèse et de m'offrir l'opportunité de poursuivre en post-doc.

Sans oublier à Jean-Marc VICTOR pour sa contribution scientifique, nos innombrables conversations et, non des moindres, qui a su supporter mon humour au quotidien.

Je m'adresse conjointement à vous trois pour vous dire : merci ! Merci pour tous ses moments passés ensemble : des pauses café-chocolat à la collaboration à Venise. Merci de m'avoir fait confiance et pour votre optimisme à toute épreuve.

Je tiens à remercier Diego CATTONI et John MARKO pour avoir accepté d'être rapporteur et pour leurs critiques constructives concernant le manuscrit ; Isabelle CALLEBAUT, Karine DUBRANA et Cédric VAILLANT pour avoir accepté d'être examinateur et pour leur bienveillance.

À mes compagnons de thèse : Léopold CARRON, Mélody MERLE, Mathias CASIULIS, Gregory PAGE pour leur convivialité.

À Chloé DURET pour son soutient inconditionnel durant ma dernière année de thèse et notamment la phase de rédaction.

À ma famille pour m'avoir soutenu jusqu'à la fin de mes études.

À mes amis de longue date Geoffrey MONET et Adrien POULENARD.

À mon amie Mathilde AMIOT pour nos interminables conversations.

Enfin, à tous ceux que j'aurais pu oublier dans cette page de remerciements.

Table des matières

Introduction **v**

1 Chromatine de la drosophile **1**
1.1 Chromatine . . . 2
1.2 Nucléosome . . . 3
1.3 Épigénétique . . . 4
1.4 Chromatine de la drosophile . . . 6
1.4.1 Approche structurale . . . 6
1.4.2 Approche architecturale . . . 7
1.5 Imagerie des domaines épigénétiques . . . 9
1.5.1 Microscopie super-résolue . . . 9
1.5.2 Les données analysées . . . 10
1.5.3 Polyploïdie . . . 12
1.6 Problématique . . . 13

2 Rayon de giration **15**
2.1 Polymère . . . 15
2.1.1 Copolymère . . . 16
2.1.2 Homopolymère . . . 18
2.2 Faisceau . . . 19
2.2.1 Système d'homopolymères de même taille . . . 19
2.2.2 Faisceau d'homopolymères . . . 20
2.3 Distribution du rayon de giration d'un faisceau . . . 22
2.4 Distribution d'un faisceau de chromosomes . . . 22
2.4.1 Les chromosomes appariés forment un faisceau . . . 23
2.4.2 Estimation de l'extension transversale du faisceau . . . 24
2.5 Modèle de faisceau gaussien . . . 27
2.6 Modèle de faisceau maxwellien . . . 28
2.6.1 Faisceau maxwellien cylindrique . . . 29
2.6.2 Disque . . . 30
2.6.3 Cylindre . . . 30
2.7 En bref . . . 31

3 Physique des polymères **33**
3.1 Chaine idéale . . . 33
3.1.1 Distance bout à bout . . . 34
3.1.2 Rayon de giration . . . 35
3.2 Chaine de Kuhn . . . 35
3.3 Chaine persistante . . . 36
3.3.1 Distance bout à bout . . . 37
3.3.2 Rayon de giration . . . 38
3.4 Chaine réelle . . . 38
3.5 Énergie libre d'une marche auto-évitante . . . 39
3.5.1 Classes de conformations . . . 40
3.5.2 Raccordements . . . 42
3.5.3 Variable d'échelle . . . 42
3.5.4 Synthèse et généralisation . . . 43
3.5.5 Correction logarithmique pour les conformations globulaires . . . 44
3.5.6 Conclusion . . . 45

4 Marche auto-évitante attractive **47**
4.1 Rappels de physique statistique . . . 48
4.1.1 Micro-état ou conformation . . . 48
4.1.2 Entropie . . . 48
4.1.3 Ensemble microcanonique . . . 49
4.1.4 Ensemble canonique . . . 50
4.1.5 Fonction génératrice des moments et des cumulants . . . 52
4.1.6 Macro-état de densité fixée . . . 53
4.1.7 En bref . . . 54
4.2 Approche phénoménologique de VICTOR (état de l'art) . . . 54
4.2.1 Lois de puissance des cumulants . . . 54
4.2.2 Conclusion . . . 55

5 Résultats numériques **57**
5.1 Simulation . . . 57
5.1.1 Théorème d'ergodicité . . . 57
5.1.2 Algorithme de Metropolis . . . 58
5.1.3 Serpent rampant . . . 63
5.1.4 Données simulées . . . 63
5.1.5 Confinement . . . 65
5.2 Inférence des paramètres . . . 67
5.2.1 Inférence bayésienne . . . 67
5.2.2 Échantillonnage de l'espace des paramètres . . . 68
5.2.3 Écriture de la log-vraisemblance . . . 68
5.2.4 Détail du calcul numérique . . . 69
5.3 Comparaison aux simulations et correction du modèle . . . 70
5.3.1 Première confrontation . . . 70
5.3.2 Terme manquant de surface . . . 70
5.3.3 Interpolation des fonctions a_i . . . 72

5.3.4 Corrélations des paramètres du modèle 74
5.4 Conclusion 76

6 Analyse des expériences **77**
6.1 Expérience 77
6.1.1 Simple anaylse des données expérimentales 78
6.1.2 Modèle convolué 80
6.1.3 Résultats 81
6.2 Conclusion 84
6.2.1 L'attraction rapproche les domaines inactifs et réprimés des conditions critiques 84
6.2.2 Géométrie du faisceau 84
6.2.3 Les paramètres obtenus soutiennent la thèse des faibles longueurs de persistance 85
6.2.4 La dépendance des paramètres aux couleurs épigénétiques indique une structure spéciale pour les domaines inactifs 86
6.2.5 Recherche d'une base moléculaire pour expliquer les paramètres énergétiques inférés 86
6.2.6 La comparaison avec des expériences sur une solution de nucléosomes dans les noyaux révèle des caractéristiques de criticité et un rôle clé pour l'interaction nucléosome-nucléosome 87

7 Conclusion générale **89**
7.1 Principaux résultats 90
7.2 Améliorations 90
7.3 Limitations 90
7.4 Perspectives à court terme 91
7.5 Perspectives à moyen terme 91

A Faisceau brownien libre **93**
A.1 Pont brownien libre 93
A.2 Faisceau dense de ponts browniens libres 94
A.3 Algorithme de simulation d'un pont brownien sur réseau 95

B Quadrature Tanh-Sinh **97**
B.1 Méthode des trapèzes 98
B.2 Décroissance en exponentielle double 98
B.3 Transformation en exponentielle double 99
B.4 Algorithme pour la méthode des trapèzes 100
B.4.1 Troncature 101
B.4.2 Récurrence 101
B.4.3 Cas pathologique du dépassement de capacité 102
B.4.4 Critère de convergence 105
B.4.5 Heuristique de t_0 105
B.4.6 Synthèse 106
B.5 Extension du catalogue de changements de variable 107
B.5.1 Exp-Exp 107

B.5.2 Exp 110
B.5.3 Sinh 111
B.5.4 Bibliothèque *dequad* 111
B.6 Tables récapitulatives 113

Bibliographie **115**

Introduction

Récemment, de grandes avancées ont permis de mieux caractériser l'architecture et la dynamique des génomes. Au cœur de ce problème, un domaine particulier suscite aujourd'hui un intérêt croissant dans plusieurs communautés scientifiques : l'épigénétique. C'est l'étude du vaste ensemble des modifications biochimiques de l'ADN et des protéines architecturales qui dirigent le destin cellulaire sans en altérer l'information génétique.

Des domaines fonctionnels localisés dans les chromosomes ont été mis en évidence. Chez la drosophile, ces domaines fonctionnels sont biochimiquement définis par des marques épigénétiques, ce qui suggère que l'arrangement spatial peut être le « chainon manquant » entre l'épigénétique et l'activité génétique.

Depuis peu, la microscopie de super-résolution (STORM) a permis d'imager la structure de ces domaines avec une résolution sans précédent. Elle a, en particulier, donné accès à la distribution des rayons de giration pour des domaines de différentes longueurs et associés à des états d'activité transcriptionnelle différents : actif, inactif ou réprimé [1]. Les lois d'échelle observées nécessitaient le développement d'un cadre théorique pour les interpréter de façon cohérente, à la lumière de la physique des polymères.

Ce travail de thèse a consisté à contribuer à la modélisation physique de la chromatine, ainsi qu'à mettre en place une méthodologie d'analyse pour extraire les paramètres structuraux des domaines fonctionnels, à partir de ces données de super-résolution.
En corollaire, les résultats de ce travail pourraient, dans une moindre mesure, permettre de trouver la bonne modélisation à adopter pour comprendre l'organisation de la chromatine chez la drosophile.

Ce manuscrit de thèse se décompose comme suit.
Au chapitre 1 sont introduites les données issues de la microscopie super-résolue sur lesquelles s'appuie la thèse, notamment les distributions du rayon de giration des domaines fonctionnels.
Au chapitre 2 sont définis successivement le rayon de giration d'un polymère et d'un faisceau de polymères, en vue de modéliser l'appariement des chromosomes homologues de la drosophile.
La conformation d'un polymère est modélisée comme une marche auto-évitante attractive.
Ainsi, au chapitre 3 est établie, à partir de la synthèse de la littérature, l'énergie libre d'une marche auto-évitante comme base de départ.
Et, au chapitre 4 est construite l'énergie libre d'une marche auto-évitante attractive, capable de réaliser la transition de phase *coil-globule* des polymères. De cette construc-

tion émerge un paramétrage de l'énergie libre par des coefficients, hors d'atteinte par la théorie, qui encodent toute la phénoménologie de cette transition.
Au chapitre 5 sont réalisées des simulations de marches auto-évitantes attractives et est adaptée une méthode d'inférence bayésienne en vue de déterminer les coefficients paramétrant l'énergie libre.
Enfin, au chapitre 6, sont assemblés les résultats des chapitres précédents pour aboutir à la distribution du rayon de giration d'un faisceau de chromosomes et sont ainsi analysées les données de microscopie super-résolue de Boettiger et al. [1].

Chapitre 1

Chromatine de la drosophile

L'acide désoxyribonucléique, ou ADN, est un macromolécule qui contient l'information génétique qui permet à un organisme de se développer, fonctionner et de se reproduire. L'ADN est un enchainement de nucléotides formés à partir d'adénine (A), thymine (T), guanine (G) ou de cytosine (C), briques élémentaires de l'information génétique. Illustré à la Figure 1.1, l'ADN est un appariement de deux brins complémentaires et adopte une structure en double hélice, connue depuis la publication historique de Watson et Crick [2] de 1953. Chaque base d'un brin s'associe à sa base complémentaire de l'autre brin formant une paire de bases (bp), unité structurale de l'ADN.

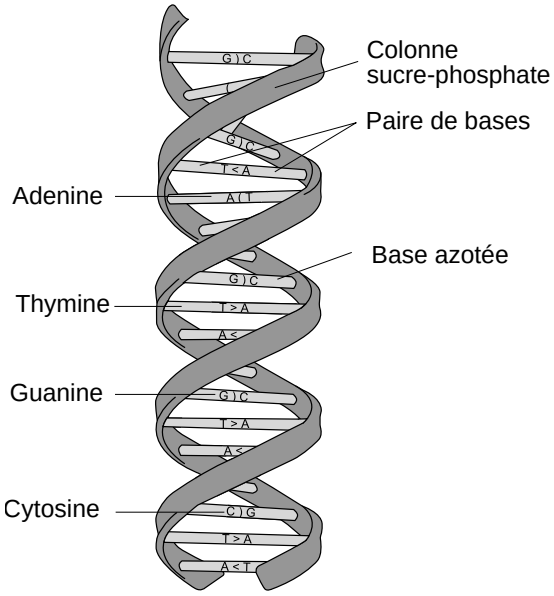

Figure 1.1 – Schéma de la structure en double hélice de l'ADN [3].

Les organismes vivants se divisent en trois catégories : les eucaryotes, les archées et les bactéries. Les premiers voient leur matériel génétique séparé du reste de la cellule par une membrane (nucléaire) délimitant le noyau, tandis que celui des archées et des bactéries cohabite avec le reste de la cellule. Le terme « eucaryote » vient du grec ancien ευ et κάρυον signifiant respectivement « vrai » et « noyau » : possédant un « vrai noyau », par opposition historiquement aux « procaryotes » qui n'en ont pas.

Le schéma de la Figure 1.2 présente la structure type d'une cellule eucaryote et permet d'en identifier le noyau.

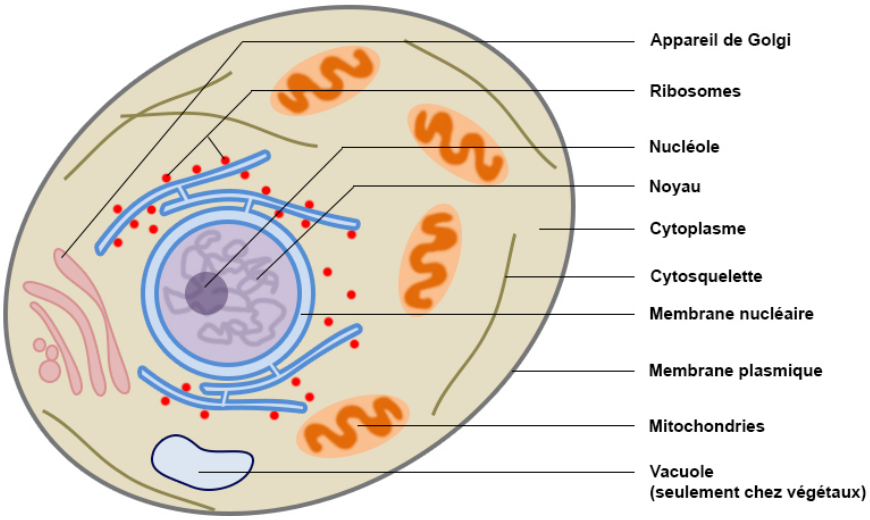

FIGURE 1.2 – Schéma d'une cellule eucaryote [4].

1.1 Chromatine

Historiquement, c'est le biologiste allemand WALTHER FLEMMING qui est le premier à avoir observé et décrit le contenu du noyau cellulaire (eucaryote). Il constate que le noyau est rempli d'une substance qu'il appelle *chromatine* pour son affinité avec les colorants.

> ...in view of its refractile nature, its reactions, and above all its affinity to dyes, is a substance which I have named chromatin. [5]

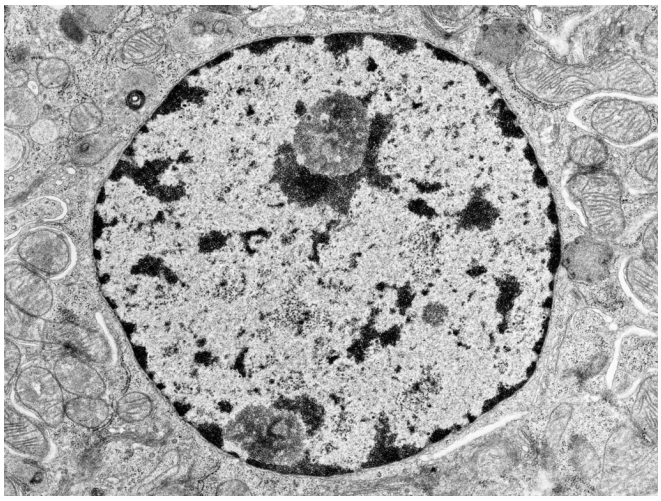

FIGURE 1.3 – Euchromatine et hétérochromatine [6].

La chromatine est la structure au sein de laquelle se trouve l'ADN dans le noyau des cellules (eucaryotes). Elle existe sous deux formes, que l'on peut voir à la Figure 1.3,

qui reflètent le niveau d'activité de la cellule :

1. L'hétérochromatine se présente sous la forme de petites régions irrégulières, foncées et éparpillées dans le noyau, le plus souvent accumulées à proximité de l'enveloppe nucléaire.
2. L'euchromatine correspond aux régions plus claires et répandues dans tout le noyau. Elle ne se colore pas facilement.

Aujourd'hui, on sait que l'euchromatine est prévalente dans les cellules qui sont actives dans la transcription de plusieurs de leurs gènes, tandis que l'hétérochromatine est plus abondante dans les cellules qui sont moins actives ou non actives [6].

1.2 Nucléosome

Depuis Walther Flemming, on sait que l'ADN s'associe à des protéines architecturales, appelées histones, pour former un complexe que l'on appelle *nucléosome*. Chaque nucléosome enroule de 146 à 147 paires de bases (bp) d'ADN. Les nucléosomes s'enchainent constituant la *fibre* de chromatine. Ils sont présents et très conservés chez tous les eucaryotes.

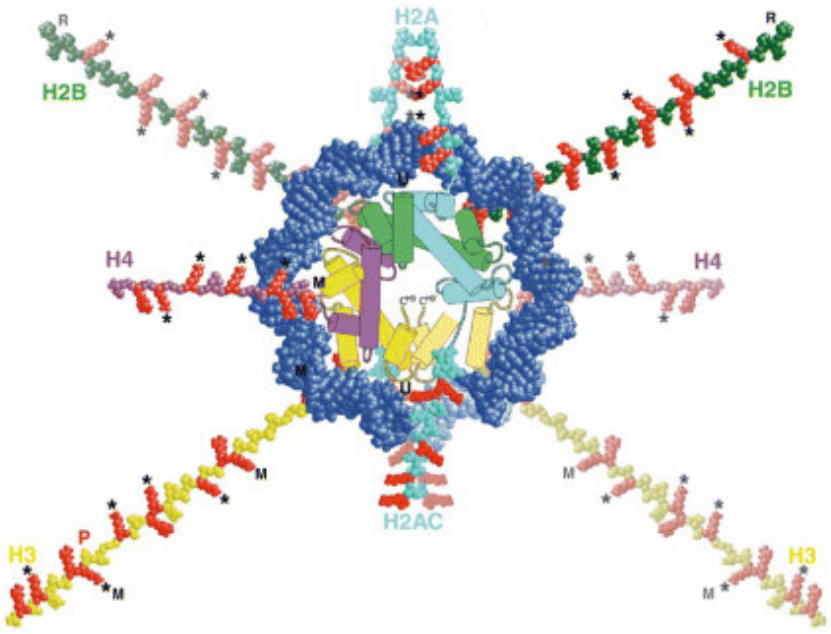

Figure 1.4 – Nucléosome [7].

Conformément à la Figure 1.4, l'ADN (en bleu foncé) s'enroule autour d'un cœur d'histones (en bleu clair, en vert, en violet et en jaune) formant le nucléosome. Les histones H2A (en bleu clair), H2B (en vert), H4 (en violet) et H3 (en jaune) possèdent des queues riches en acides aminés basiques, notamment des lysines (en rouge). Ces queues n'ont pas de structure secondaire, *i.e.* elles ne se replient pas de façon structurée en hélice α ou en feuillet β.

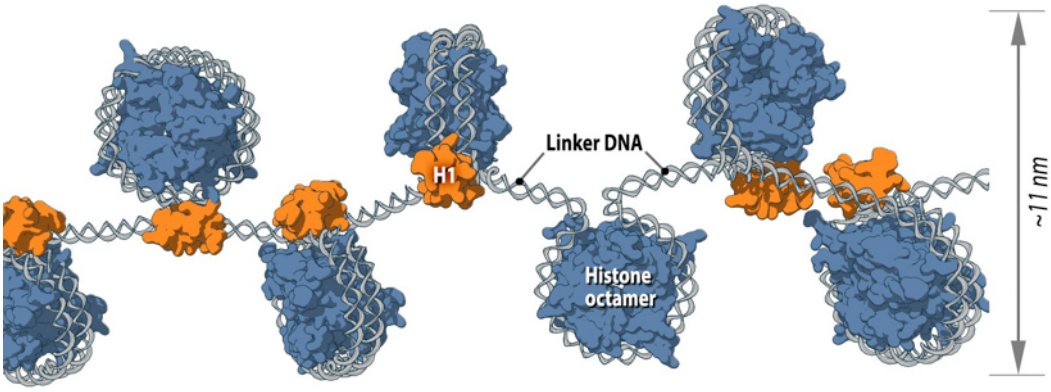

FIGURE 1.5 – Fibre de chromatine [8].

Conformément à la Figure 1.5, les histones se positionnent de manière ordonnée le long de l'ADN. Deux nucléosomes sont régulièrement espacés par un ADN de liaison d'une longueur de l'ordre de 50 bp. Cette régularité permet aux nucléosomes de s'agencer en une *fibre* dont l'architecture peut, en principe, s'adapter aux contraintes dues à la régulation de l'expression génétique.

1.3 Épigénétique

Les lysines présentes sur les queues des nucléosomes sont le lieu de modifications biochimiques qui n'altèrent pas l'information génétique encodée. C'est pourquoi, on parle de modifications *épigénétiques*, littéralement à la surface des gènes.

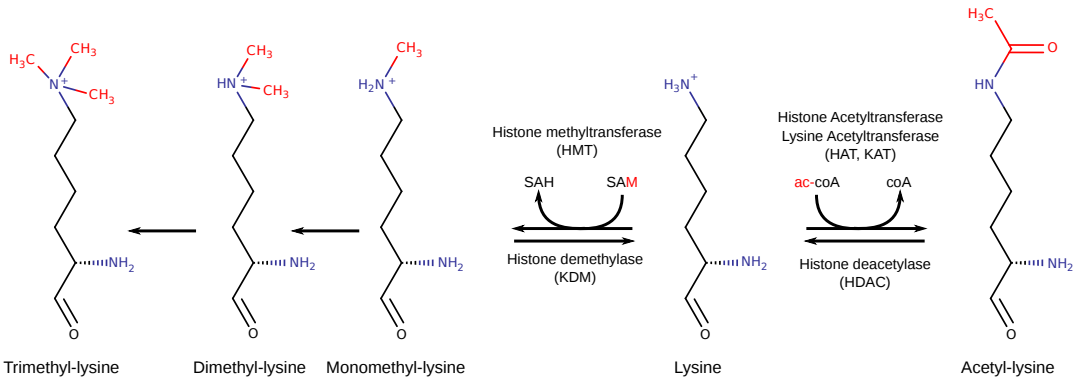

FIGURE 1.6 – Modifications post-transcriptionnelles des lysines [9].

Conformément à la Figure 1.6, l'extrémité réactive de la lysine est chargée positivement. Elle peut être méthylée jusqu'à trois fois, conservant sa charge, ou acétylée, lui faisant perdre sa charge.

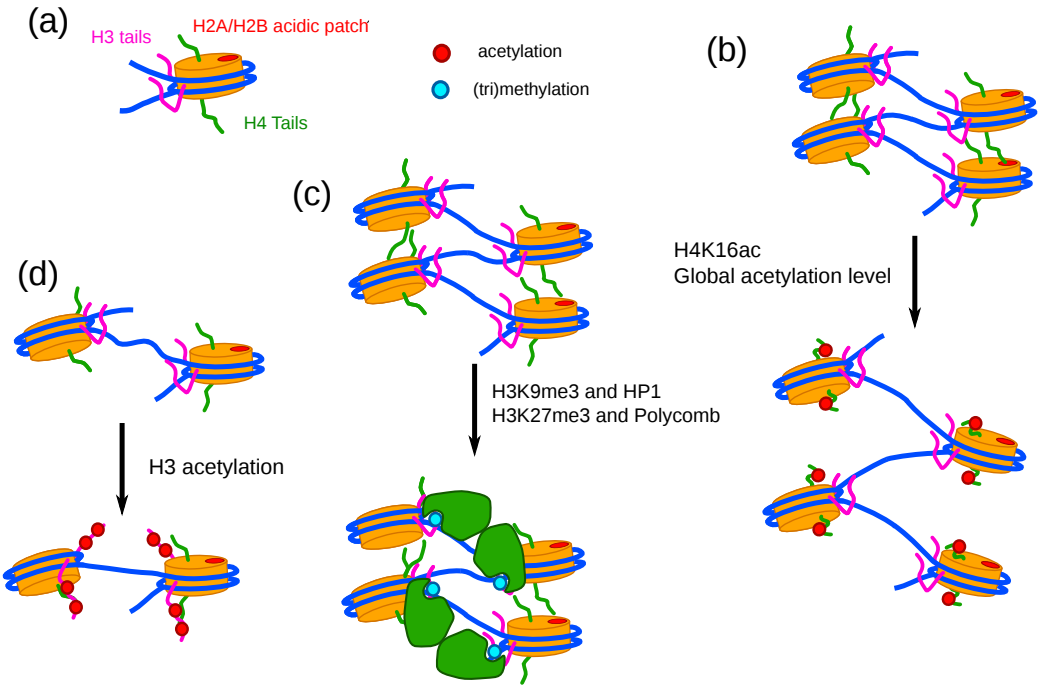

FIGURE 1.7 – Réarrangements des nucléosomes [9].

Bien qu'elles n'affectent pas la séquence génétique, ces modifications épigénétiques modulent à la fois l'expression des gènes marqués, ainsi que l'agencement des nucléosomes entre eux. On parle aussi de *marques* épigénétiques. Comme l'illustre la Figure 1.7, le cœur d'histones possède une partie acide chargée négativement (a) sur laquelle peut venir se fixer une lysine d'une queue d'un histone voisin (b), conférant ainsi une structure tassée à la fibre non modifiée.
En acétylant les lysines, elles perdent leur charge et ainsi leur capacité à se fixer aux autres nucléosomes. Il en découle une décondensation de la fibre de nucléosomes (b, en bas).
De plus, les triméthylations peuvent être reconnues par des protéines, notamment les complexes répresseurs Polycomb-1 (PRC1) qui viennent réprimer les gènes marqués et condenser la fibre de nucléosomes.

Chez les organismes eucaryotes, il apparait alors une forte corrélation entre marquages épigénétiques, structure de la fibre de nucléosomes et régulation de l'expression des gènes.

1.4 Chromatine de la drosophile

Les eucaryotes vont d'unicellulaires, telle que la levure, aux mammifères en passant par la drosophile. Ils ont tous une organisation nucléaire différente, de complexité croissante [9].
Chez la levure, les chromosomes sont en permanence ancrés par leur centromère à la membrane nucléaire et s'organisent en brosse de polymères [10, 11].
Chez la drosophile, les chromosomes sont partitionnés en domaines épigénétiques qui se replient sur eux-mêmes [12].
Chez les mammifères, par dessus les domaines épigénétiques s'ajoutent d'autres mécanismes de régulation telles que les boucles à extrusion [13, 14].

On s'intéresse, ici, à l'organisation nucléaire de la drosophile qui est riche du point de vue de la physique. L'enjeu à terme est de pouvoir comprendre graduellement la complexité de l'organisation nucléaire des mammifères (et notamment de l'être humain) par l'intermédiaire d'un système plus simple telle que la drosophile.

La chromatine se structure à différentes échelles. Je vais présenter la vision historique, puis revenir sur l'approche moderne de l'architecture de la chromatine.

1.4.1 Approche structurale

L'existence *in vitro* d'une fibre de nucléosomes, dite de 30 nm [15], a montré que la chromatine pouvait se structurer et se replier de manière régulière sur plusieurs ordres de grandeur, potentiellement, jusqu'à l'échelle du chromosome tout entier (*cf.* Figure 1.8).

Cependant, l'existence *in vivo* d'une telle fibre n'a jamais été prouvée [16]. La chromatine est une fibre, dite de 11 nm, très désordonnée, dont on sait peu de choses sur son organisation spatiale à plus grandes échelles.

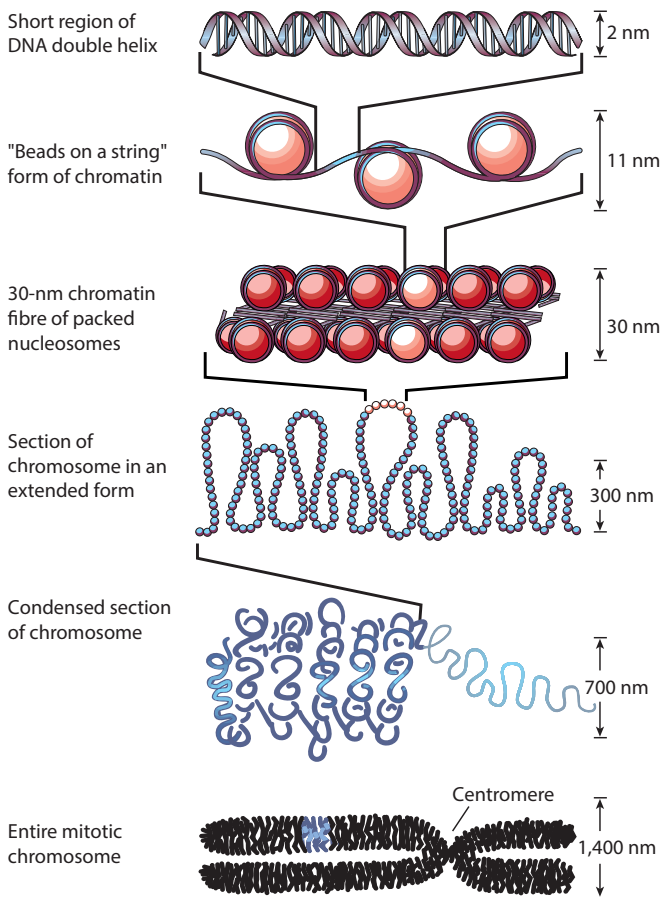

FIGURE 1.8 – Structuration de la chromatine sur plusieurs échelles [17].

1.4.2 Approche architecturale

À l'échelle du noyau (*cf.* Figure 1.9), les chromosomes ne se mélangent pas et restent invariablement confinés dans leur *territoire chromosomique*. Chaque chromosome est partitionné en domaines épigénétiques formant des blocs continus de chromatine qui se replient physiquement sur eux-mêmes.

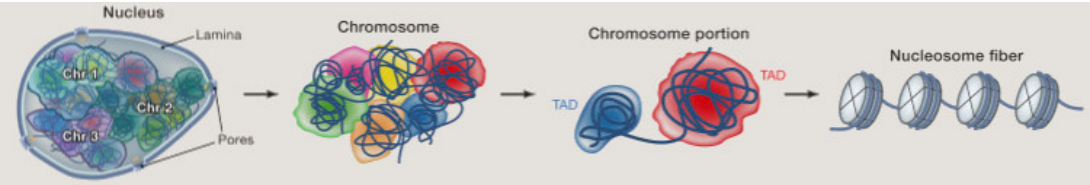

FIGURE 1.9 – Architecture de la chromatine, modifiée de SEXTON et CAVALLI [18].

De telles structures sont mises en évidence expérimentalement à l'aide de cartes de contact (*cf.* Figure 1.10), permettant de décrire les contacts que chaque région du génome établit avec les autres. On les obtient avec des techniques de capture conformationnelle de chromosomes (3C, 5C, Hi-C [19–21]). On y reporte le nombre de fois

qu'un site génomique i a été en contact avec un site j, donnant après normalisation une carte de probabilité de contact.

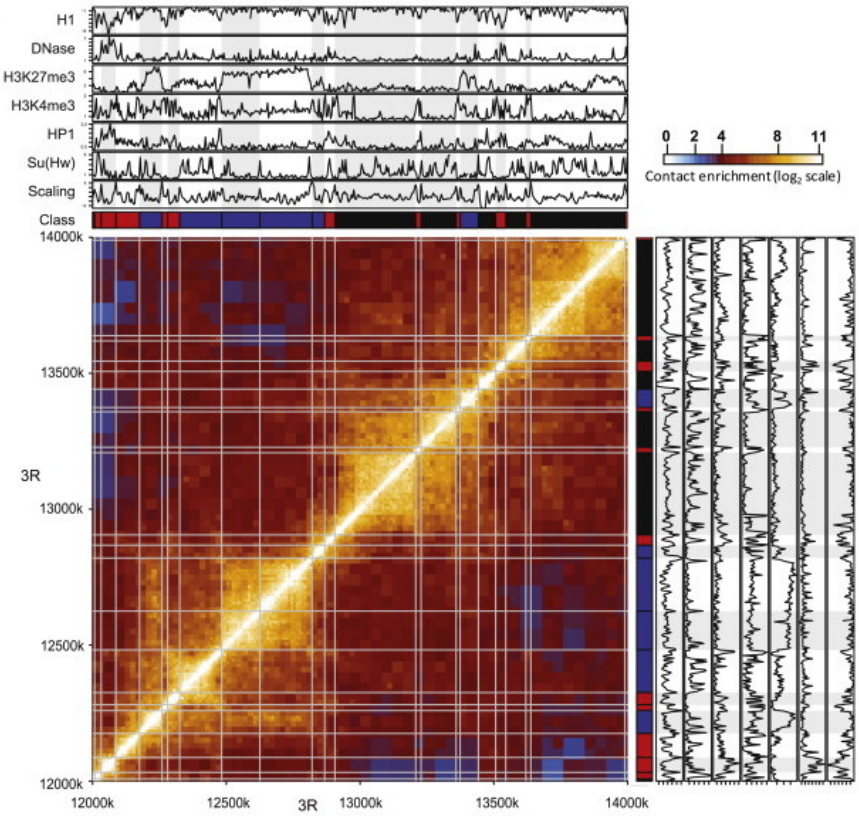

FIGURE 1.10 – Les carrés jaunes sur la diagonale de la matrice de contact représentent les domaines physiques. Ils sont corrélés aux domaines épigénétiques (rouges, noirs ou bleus) [12].

FILION et al. [22] ont effectué une *analyse en composantes principales* sur 53 protéines liées à la chromatine et sur 4 modifications des histones, qui permet d'identifier 5 combinaisons de marques épigénétiques de la chromatine, appelées *couleurs* de la chromatine (*cf.* Figure 1.11).

Les couleurs s'organisent en domaines épigénétiques continus qui forment une partition des chromosomes et tendent à se replier sur eux-mêmes pour former des structures tridimensionnelles, appelées domaines *physiques* (*cf.* Figure 1.10).

Par ailleurs, les couleurs sont corrélées à l'état d'activité transcriptionnelle des gènes au sein des domaines.

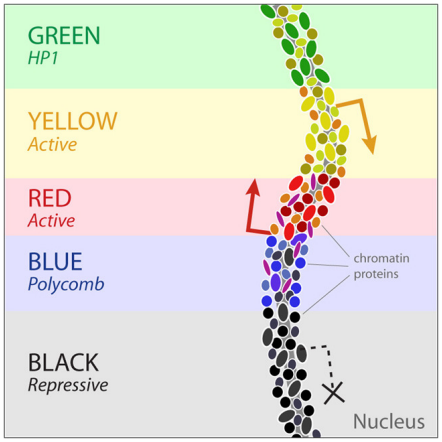

FIGURE 1.11 – Les 5 couleurs de la chromatine de FILION et al. [22].

Les domaines au sein de la chromatine *rouge* et *jaune* sont activement transcrits, tandis que les gènes se trouvant dans la chromatine *bleue* sont réprimés par les protéines du groupe Polycomb. Les domaines de la chromatine *noire* sont, quant à eux, inactifs, *i.e.* dans un état latent.

La disposition des couleurs sur les chromosomes caractérise un type cellulaire. C'est la mise en place de ces marques épigénétiques qui est responsable de la différentiation cellulaire.

1.5 Imagerie des domaines épigénétiques

Jusqu'à l'apparition des techniques de microscopie super-résolue, la seule façon de visualiser la structure des domaines épigénétiques était par le biais de méthodes chimiques (Hi-C), destructives donnant la structure moyenne sur une population de plusieurs millions de cellules.

1.5.1 Microscopie super-résolue

La structure des domaines épigénétiques est longtemps restée hors d'atteinte par les techniques physiques d'imagerie. Conformément à la Figure 1.12, d'un côté, la cristallographie aux rayons X est limitée par la structure désordonnée de la fibre de chromatine de 11 nm et ne permet pas de remonter plus haut. De l'autre côté, le pouvoir de résolution des instruments optiques est intrinsèquement limité par la diffraction et ne permet pas de descendre plus bas que le chromosome.

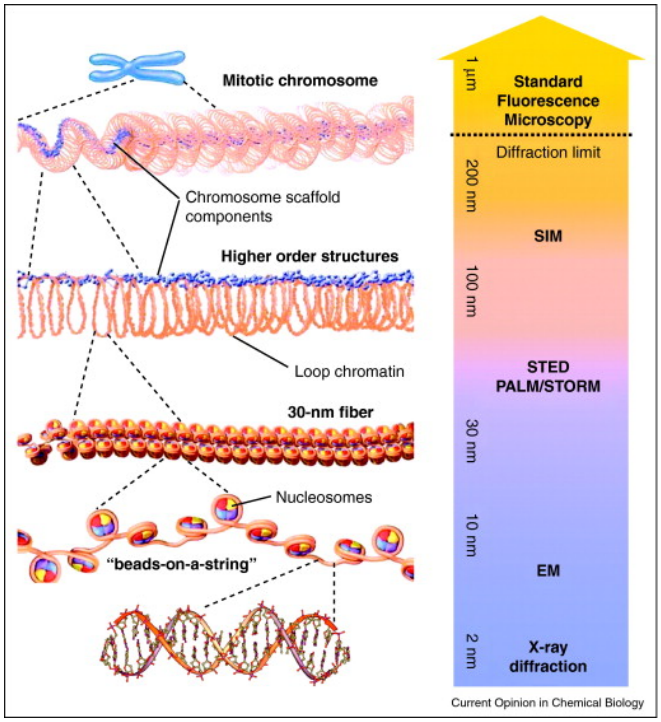

FIGURE 1.12 – Limites de résolution expérimentale [23].

La microscopie super-résolue permet de descendre en dessous de la limite de résolution optique et d'imager la structure d'un domaine épigénétique dans une seule cellule. Ce qui permet non plus d'accéder à la structure moyenne, mais à la structure instantannée des domaines épigénétiques.

Une image super-résolue a une résolution de l'ordre de quelques dizaines de nanomètres, bien inférieure à la limite de résolution optique. Elle est obtenue grâce à différentes techniques, en particulier des méthodes de traitement d'images.

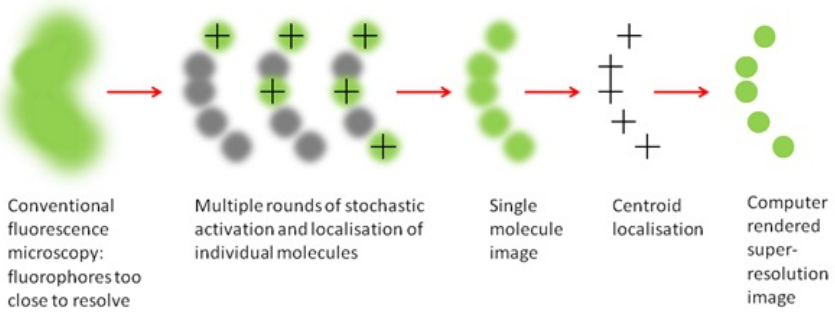

FIGURE 1.13 – Principe du STORM [24].

Prenons l'exemple de la microscopie de reconstruction optique stochastique, ou *Stochastic Optical Reconstruction Microscopy* (STORM) en anglais, dont le principe est illustré à la Figure 1.13. Le principe consiste à échantillonner aléatoirement l'image en faisant scintiller des sondes fluorescentes accrochées à l'objet à imager. Ainsi, on peut séparer le centre des taches lumineuses produites par les sondes et reconstruire l'image complète en superposant les clichés obtenus. Ce qui n'aurait pas été possible si les sondes avaient été toutes illuminées en même temps.

1.5.2 Les données analysées

En 2016, BOETTIGER et al. [1] publient des images super-résolues de domaines épigénétiques de la drosophile dans trois couleurs épigénétiques différentes. Ces couleurs sont caractérisées par des marquages épigénétiques spécifiques et corrélés à un état d'activité de transcription des gènes : actif, inactif ou réprimé.

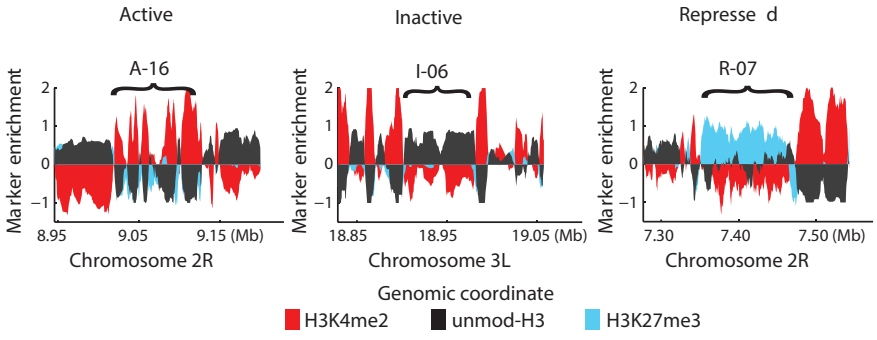

FIGURE 1.14 – Enrichissement en marqueurs [1].

Les domaines sont labellisés grâce à la méthode de séquençage par immunoprecipitation de chromatine ou *Chromatin ImmunoPrecipitation Sequencing* (ChIP-Seq) en anglais, qui permet de mesurer l'enrichissement des domaines en H3K4me2, unmod-H3 et H3K27me3 (*cf.* Figure 1.14). Si l'un de ces marqueurs prévaut sur les autres alors le domaine sera dit respectivement actif, inactif ou réprimé. Les marquages correspondent :

— H3K4me2 à la diméthylation (me2) de la lysine 4 (K4) de l'histone 3 (H3).
— unmod-H3 à l'absence de modification biochimique (unmod) de l'histone 3 (H3).
— H3K27me3 à la triméthylation (me3) de la lysine 27 (K27) de l'histone 3 (H3).

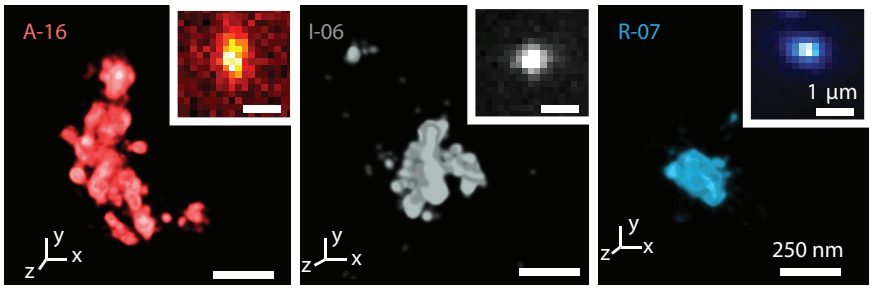

FIGURE 1.15 – Domaines épigénétiques super-résolus [1].

À l'aide des images tridimensionnelles super-résolues (*cf.* Figure 1.15), les auteurs ont réalisé des mesures de rayon de giration d'un domaine épigénétique donné dans une population de cellules Kc167. Le rayon de giration représente la taille caractéristique de l'objet dans l'espace. Il sera défini au chapitre 2.

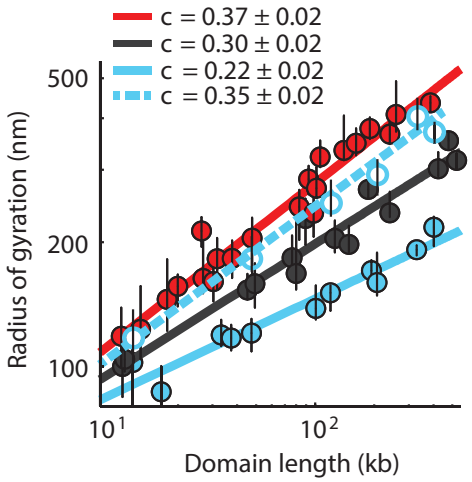

FIGURE 1.16 – Lois de puissance du rayon de giration [1].

En reportant le rayon de giration médian (en nm) obtenu pour les domaines d'un même état épigénétique en fonction de leur longueur (en kb [1]), on observe Figure 1.16, en échelle logarithmique, un comportement en loi de puissance différent pour les trois états. Ce qui semble raisonnable. En effet, l'architecture des domaines joue un rôle

clé dans la régulation des gènes en leur sein. On peut donc s'attendre à ce que l'architecture spécifique des trois états d'activité se reflète dans la manière dont le rayon de giration croit avec la longueur du domaine. Par exemple, les domaines réprimés sont revêtus d'une marque H3K27me3 qui va recruter une protéine PRC1 qui va venir compacter la fibre (*cf.* Figure 1.7).
Du point de vue du physicien, le marquage épigénétique définit des interactions spécifiques au sein d'un domaine. On observe bien, ici, que l'évolution du rayon de giration en fonction de la taille des domaines dépend de la nature des interactions. Pour une longueur donnée, un domaine réprimé sera plus compact (aura un rayon de giration plus petit) qu'un domaine actif ou inactif. Comme on le verra au chapitre 3, cette évolution en lois de puissance, *i.e.* droites en échelle logarithmique, a une justification simple en physique des polymères.

1.5.3 Polyploïdie

Une donnée supplémentaire importante est que les cellules de la ligne cellulaire Kc167, utilisée pour les mesures de rayon de giration chez Boettiger et al. [1], sont *tétraploïdes*, *i.e.* elles ont 4 paires de chromosomes. Alors qu'en temps normal, les cellules de la drosophile sont diploïdes, *i.e.* elles ont 2 paires de chromosomes. La polyploïde (*cf.* Figure 1.17) renvoie au nombre de chromosomes homologues dans le noyau d'une cellule. Ainsi à titre d'exemple, chez l'Homme, où l'on a 23 paires de chromosomes : nos cellules sont diploïdes [2].

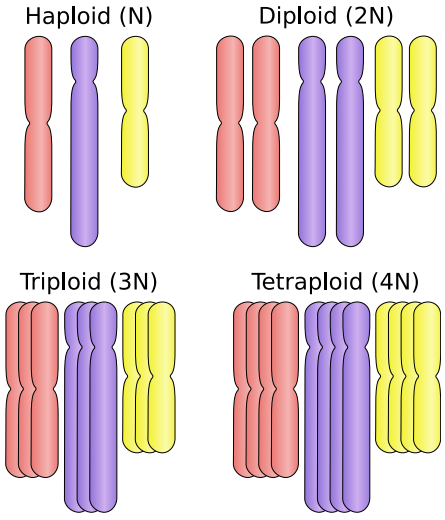

Figure 1.17 – Polyploïdie [25].

Chez la drosophile, un autre fait important est que les chromosomes homologues sont régulièrement appariés, à environ 80% [26, 27], formant un *faisceau* de chromosomes.

1. Kilobase noté kb pour 1000 bp.
2. À l'exception de certains tissus qui peuvent être tétraploïdes.

1.6 Problématique

L'objectif de ma thèse est de mettre en place une méthode d'analyse des images de microscopie super-résolue des domaines épigénétiques. Ces images permettent de remonter à la distribution du rayon de giration de chaque domaine épigénétique, qui contient toutes leurs propriétés statistiques. Il apparait, alors, nécessaire de les exploiter pour fournir une analyse plus riche que celle des médianes proposée par Boettiger et al. [1].

Chapitre 2

Rayon de giration

On rappelle que Boettiger et al. [1] ont mesuré les distributions du rayon de giration des domaines épigénétiques de la lignée cellulaire tétraploïde Kc167. Ces distributions encodent toutes les propriétés statistiques des domaines, c'est pourquoi on cherche à les modéliser.
On doit également tenir compte du fait que les chromosomes homologues de la drosophile sont appariés à environ 80% [26, 27], formant un *faisceau* de chromosomes. L'un des enjeux de ce chapitre sera donc de décrire le rayon de giration d'un faisceau de polymères et d'en exhiber l'expression de sa distribution.

Dans toute la suite, $\boldsymbol{r}$ désignera le vecteur position parcourant l'espace accessible.

2.1 Polymère

Le mot « polymère » vient du grec ancien πολύς, *polús* et μέρος, *méros* qui signifient respectivement « beaucoup » et « partie ». Un polymère est une molécule constituée d'un grand nombre d'unités structurales de base, identiques ou semblables, appélées monomères.

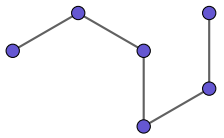

Figure 2.1 – Polymère.

Les monomères sont rattachés le long d'une chaine pouvant se déployer de façon différente dans l'espace. On appelle conformation d'un polymère, ou encore conformère, la disposition spatiale de ses différents monomères.

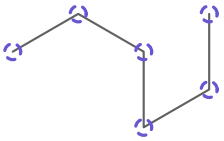

Figure 2.2 – Conformère, *i.e.* la chaine d'un polymère privé de ses monomères.

Si tous les monomères sont identiques, on parle alors d'homopolymère, sinon on parle d'hétéropolymère ou encore de copolymère.

En physique, le polymère est souvent réduit à son conformère, donnant la formule du rayon de giration

$$R^2 = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{G}_i - \boldsymbol{G})^2 \tag{2.1}$$

où N désigne le nombre de monomères, $\boldsymbol{G}_i$ leur barycentre, *i.e.* position, et $\boldsymbol{G} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{G}_i$ le barycentre du polymère.

2.1.1 Copolymère

On souhaite déterminer le rayon de giration d'un polymère dans un cas plus général où les monomères ne seraient plus réduits à des points. Le point de départ est d'introduire la distribution de masse, ρ_i, du i-ème monomère :

$$\int \rho_i(\boldsymbol{r})\, d\boldsymbol{r} = m_i.$$

Ensuite, on définit la fonction indicatrice du i-ème monomère

$$\delta_i(\boldsymbol{r}) = \frac{1}{m_i}\rho_i(\boldsymbol{r})$$

qui décrit sa géométrie à l'aide de la distribution de masse ρ_i. L'indicatrice δ_i correspond à la fraction du monomère présent dans le volume d'intégration et est normée :

$$\int \delta_i(\boldsymbol{r})\, d\boldsymbol{r} = 1. \tag{2.2}$$

Ce qui signifie que le monomère est entièrement inclus dans le volume accessible. À présent, on peut aisément définir le barycentre $\boldsymbol{G}_i$ et le rayon de giration R_i du monomère

$$\boldsymbol{G}_i = \int \boldsymbol{r}\, \delta_i(\boldsymbol{r})\, d\boldsymbol{r} \tag{2.3}$$

$$R_i^2 = \int (\boldsymbol{r} - \boldsymbol{G}_i)^2 \delta_i(\boldsymbol{r})\, d\boldsymbol{r} \tag{2.4}$$

comme étant la moyenne et l'écart type de la distribution δ_i.

La distribution de masse d'un polymère à N monomères,

$$P(\boldsymbol{r}) = \sum_{i=1}^{N}\rho_i(\boldsymbol{r})\,,$$

s'écrit simplement comme la somme des distributions de masse de ses monomères. Elle vérifie la propriété intuitive que la masse M du polymère est la somme des masses m_i

des monomères :

$$\int P(\boldsymbol{r})\, d\boldsymbol{r} = M$$
$$\sum_{i=1}^{N} \int \rho_i(\boldsymbol{r})\, d\boldsymbol{r} = M$$
$$\sum_{i=1}^{N} m_i = M \tag{2.5}$$

De manière analogue, on définit la fonction indicatrice d'un polymère

$$\Delta(\boldsymbol{r}) = \frac{1}{M} P(\boldsymbol{r})$$

$$\Delta(\boldsymbol{r}) = \sum_{i=1}^{N} \frac{m_i}{M} \delta_i(\boldsymbol{r}) \tag{2.6}$$

comme somme pondérée des indicatrices δ_i des monomères. Elle correspond à la fraction du polymère présent dans le volume d'intégration et est également normée :

$$\int \Delta(\boldsymbol{r})\, d\boldsymbol{r} = 1.$$

On vérifie que le barycentre du polymère, défini par

$$\boldsymbol{G} = \int \boldsymbol{r} \Delta(\boldsymbol{r})\, d\boldsymbol{r}$$
$$\boldsymbol{G} = \sum_{i=1}^{N} \frac{m_i}{M} \int \boldsymbol{r} \delta_i(\boldsymbol{r})\, d\boldsymbol{r}$$
$$\boldsymbol{G} = \sum_{i=1}^{N} \frac{m_i}{M} \boldsymbol{G}_i$$

est bien le barycentre des barycentres $\boldsymbol{G}_i$ des monomères.

Maintenant que nous avons défini l'indicatrice du polymère et vérifié qu'elle permettait de retrouver un résultat cohérent sur le centre de masse $\boldsymbol{G}$ du polymère, on va expliciter son rayon de giration

$$R^2 = \int (\boldsymbol{r} - \boldsymbol{G})^2 \Delta(\boldsymbol{r})\, d\boldsymbol{r}\,.$$

Pour ce faire, on utilise la définition de l'indicatrice Δ (2.6) du polymère

$$R^2 = \sum_{i=1}^{N} \frac{m_i}{M} \int (\boldsymbol{r} - \boldsymbol{G})^2 \delta_i(\boldsymbol{r})\, d\boldsymbol{r}\,,$$

et on recentre la variable $\boldsymbol{r}$ dans chaque intégrale autour de son barycentre $\boldsymbol{G}_i$

$$R^2 = \sum_{i=1}^{N} \frac{m_i}{M} \int ((\boldsymbol{r} - \boldsymbol{G}_i) + (\boldsymbol{G}_i - \boldsymbol{G}))^2 \delta_i(\boldsymbol{r})\, d\boldsymbol{r}\,.$$

Puis, on développe l'identité remarquable $((\boldsymbol{r}-\boldsymbol{G}_i)+(\boldsymbol{G}_i-\boldsymbol{G}))^2$ et on factorise par tout ce qui ne dépend pas de la variable d'intégration $\boldsymbol{r}$ sous le signe intégral

$$\begin{aligned} R^2 =& \sum_{i=1}^{N} \frac{m_i}{M} \cdot \underbrace{\int (\boldsymbol{r}-\boldsymbol{G}_i)^2 \delta_i(\boldsymbol{r})\, d\boldsymbol{r}}_{=R_i^2} \\ &+ \sum_{i=1}^{N} \frac{m_i}{M} (\boldsymbol{G}_i-\boldsymbol{G})^2 \cdot \underbrace{\int \delta_i(\boldsymbol{r})\, d\boldsymbol{r}}_{=1} \\ &+ 2\sum_{i=1}^{N} \frac{m_i}{M} (\boldsymbol{G}_i-\boldsymbol{G}) \cdot \underbrace{\int (\boldsymbol{r}-\boldsymbol{G}_i)\delta_i(\boldsymbol{r})\, d\boldsymbol{r}}_{=\mathbf{0}}. \end{aligned}$$

On reconnait successivement le rayon de giration du i-ème monomère (2.4), la norme de sa distribution (2.2) et son barycentre (2.3). Ainsi, on aboutit à la formule du rayon de giration d'un copolymère

$$R^2 = \underbrace{\sum_{i=1}^{N} \frac{m_i}{M} R_i^2}_{\text{monomères}} + \underbrace{\sum_{i=1}^{N} \frac{m_i}{M} (\boldsymbol{G}_i-\boldsymbol{G})^2}_{\text{conformère}}$$

où l'on peut identifier deux contributions : l'une venant directement des monomères, l'autre venant du conformère.

2.1.2 Homopolymère

Dans le cas d'un homopolymère, tous les monomères ont la même masse $m_i = m$, ce qui permet de réécrire la somme (2.5) comme $Nm = M$. En d'autres termes, partout où apparait la somme $\sum_{i=1}^{N} \frac{m_i}{M}$ dans le cas d'un copolymère, on peut la remplacer par $\frac{1}{N}\sum_{i=1}^{N}$. On obtient alors la formule du rayon de giration d'un polymère dont les monomères ont une masse identique

$$R^2 = \frac{1}{N}\sum_{i=1}^{N} R_i^2 + \frac{1}{N}\sum_{i=1}^{N} (\boldsymbol{G}_i-\boldsymbol{G})^2. \tag{2.7}$$

Par ailleurs, si les distributions de masse des monomères sont toutes identiques à une translation près, $\rho_i(\boldsymbol{r}) = \rho(\boldsymbol{r}-\boldsymbol{G}_i)$, alors leurs indicatrices sont également identiques à une translation près : $\delta_i(\boldsymbol{r}) = \delta(\boldsymbol{r}-\boldsymbol{G}_i)$. De ce fait, les monomères ont tous le même rayon de giration $R_i = R_{\text{mono}}$. Ainsi, on aboutit à la formule du rayon de giration d'un homopolymère

$$R^2 = R_{\text{mono}}^2 + \frac{1}{N}\sum_{i=1}^{N} (\boldsymbol{G}_i-\boldsymbol{G})^2. \tag{2.8}$$

On prend le soin de vérifier la cohérence de la formule pour le cas de monomères ponctuels. Leur distribution s'écrit $\delta_i(\boldsymbol{r}) = \delta_D(\boldsymbol{r}-\boldsymbol{G}_i)$ à l'aide de la distribution

ponctuelle de Dirac δ_D qui a les deux propriétés notables suivantes :

$$\int \delta_D(\boldsymbol{r})\, d\boldsymbol{r} = 1$$
$$\int f(\boldsymbol{r})\delta_D(\boldsymbol{r} - \boldsymbol{r}_0)\, d\boldsymbol{r} = f(\boldsymbol{r}_0).$$

En utilisant cette distribution dans l'équation (2.4), on obtient $R^2_{\text{mono}} = 0$. Il faut comprendre par là que le volume d'un monomère ponctuel est nul. On retrouve bien que le rayon de giration d'un polymère (2.7), dont les monomères sont ponctuels ($R_i^2 = 0$), est égal au rayon de giration de son conformère (2.1).

2.2 Faisceau

Dans cette partie, on cherche à exprimer le rayon de giration d'un faisceau de polymères. Par faisceau de polymères, on entend assemblage de polymères liés ensemble et de ce fait possédant une conformation similaire. Pour ce faire, on procède par étapes en exprimant d'abord, le plus généralement possible, le rayon de giration d'un système d'homopolymères, puis celui d'un faisceau d'homopolymères.

2.2.1 Système d'homopolymères de même taille

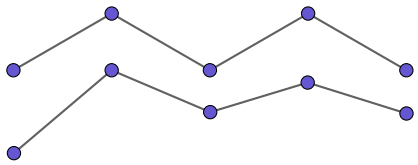

FIGURE 2.3 – Système d'homopolymères.

En suivant la même démarche, on introduit l'indicatrice du k-ième polymère Δ_k. Ce qui nous permet de définir son barycentre $\boldsymbol{G}_k$ et son rayon de giration R_k^2. On introduit ensuite l'indicatrice du système à n polymères

$$\Delta(\boldsymbol{r}) = \frac{1}{n}\sum_{k=1}^{n} \Delta_k(\boldsymbol{r})$$

qui donne le barycentre du système

$$\boldsymbol{G} = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{G}_k$$

et aboutit au rayon de giration du système de polymères

$$R^2 = \underbrace{\frac{1}{n}\sum_{k=1}^{n} R_k^2}_{\text{polymères}} + \underbrace{\frac{1}{n}\sum_{k=1}^{n} (\boldsymbol{G}_k - \boldsymbol{G})^2}_{\text{poly-conformère}} \tag{2.9}$$

où l'on peut identifier deux contributions : l'une venant des polymères, l'autre venant de la disposition des barycentres des polymères, que l'on appelle, ici, poly-conformère.

On note la ressemblance de l'équation (2.9) avec l'équation (2.7), qui tient au fait que la formule du rayon de giration d'un système, au sens large, s'écrira toujours de la même façon. Il y aura un terme venant de la distribution spatiale des constituants du système et l'autre venant du rayon de giration moyen des constituants. Dans le cas de l'équation (2.7), le système est l'ensemble des monomères. Alors que dans l'équation (2.9), le système est l'ensemble des polymères.

Une dernière chose à noter, c'est que le terme R_k^2 correspond au rayon de giration du k-ième homopolymère. Il est donc donné par l'équation (2.8)

$$R_k^2 = R_{\text{mono}}^2 + \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{G}_{ik} - \boldsymbol{G}_k)^2$$

où l'on prend le soin d'introduire des indices k sur chacun des termes. Ainsi le barycentre des monomères et celui des polymères s'écrivent respectivement $\boldsymbol{G}_{ik}$ et $\boldsymbol{G}_k$. En somme, le rayon de giration d'un système d'homopolymères de même taille s'écrit

$$R^2 = \underbrace{R_{\text{mono}}^2}_{\text{monomères}} + \underbrace{\frac{1}{n}\sum_{k=1}^{n}\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{G}_{ik} - \boldsymbol{G}_k)^2}_{\text{polymères}} + \underbrace{\frac{1}{n}\sum_{k=1}^{n}(\boldsymbol{G}_k - \boldsymbol{G})^2}_{\text{poly-conformère}}. \tag{2.10}$$

2.2.2 Faisceau d'homopolymères

Dans cette partie, je vais introduire la notion de faisceau de polymères qui permet de sa ramener à un polymère effectif possédant une certaine épaisseur, en décomposant non plus le système (longitudinalement) en polymères, mais (transversalement) en paquets de monomères que je vais appeler macromères.

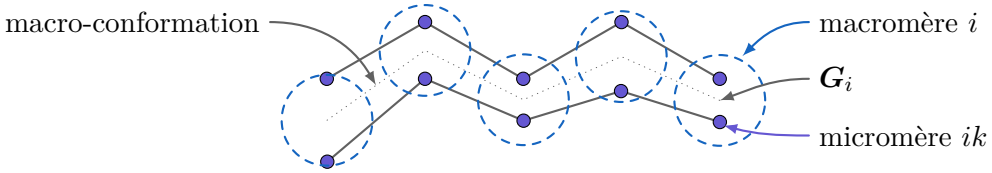

FIGURE 2.4 – Faisceau d'homopolymères.

Un faisceau est un système de polymères dont les monomères homologues, *i.e.* de même abscisse curviligne i, se retrouvent proches spatialement et dont l'orientation des liens est fortement corrélée. Je vais appeler *micromère* les monomères des polymères constituants le faisceau. Le faisceau en lui-même est assimilable à un polymère dont :

1. Les monomères sont formés des micromères homologues et sont appelés macromères.
2. La conformation est décrite par les barycentres des macromères $\boldsymbol{G}_i$ et est appelée macro-conformation.

Le rayon de giration d'un faisceau à N macromères s'écrit

$$R^2 = \underbrace{\frac{1}{N}\sum_{i=1}^{N} R_i^2}_{\text{macromères}} + \underbrace{\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{G}_i - \boldsymbol{G})^2}_{\text{macro-conformère}}$$

où R_i^2 désigne le rayon de giration du i-ème macromère

$$R_i^2 = \underbrace{R_{\text{micro}}^2}_{\text{micromère}} + \underbrace{\frac{1}{n}\sum_{k=1}^{n}(\boldsymbol{G}_{ik} - \boldsymbol{G}_i)^2}_{\text{micro-conformère}}.$$

Ce dernier se décompose en deux termes : l'un venant des micromères, l'autre venant de la conformation des micromères homologues au sein du macromère. Je parle de micro-conformation. En somme, le rayon de giration d'un faisceau s'écrit

$$R^2 = \underbrace{R_{\text{micro}}^2}_{\text{micromères}} + \underbrace{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{n}\sum_{k=1}^{n}(\boldsymbol{G}_{ik} - \boldsymbol{G}_i)^2}_{\text{micro-conformères}} + \underbrace{\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{G}_i - \boldsymbol{G})^2}_{\text{macro-conformère}}. \tag{2.11}$$

À présent, si on met en regard l'équation (2.10) avec l'équation (2.11), en différenciant minutieusement les indices i, k et ik, on peut définir plus précisément la macro-conformation. On rappelle que :

— L'indice i est relatif au i-ème macromère.
— L'indice k est relatif au k-ème homopolymère.
— Les indices ik sont relatifs au i-ème micromère au sein du k-ème homopolymère.
— Le nombre d'homopolymères au sein du faisceau est noté n.
— Le nombre de micromères au sein d'un homopolymère est noté N.

Puisque la macro-conformation du faisceau est la conformation moyenne de ses homopolymères constitutifs, on a l'égalité des rayons de giration suivante :

$$\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{G}_i - \boldsymbol{G})^2 = \frac{1}{n}\sum_{k=1}^{n}\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{G}_{ik} - \boldsymbol{G}_k)^2. \tag{2.12}$$

Il ne reste qu'à identifier le rayon de giration du poly-conformère à celui du micro-conformère moyen :

$$\frac{1}{n}\sum_{k=1}^{n}(\boldsymbol{G}_k - \boldsymbol{G})^2 = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n}\sum_{k=1}^{n}(\boldsymbol{G}_{ik} - \boldsymbol{G}_i)^2. \tag{2.13}$$

En définitive, on peut écrire le rayon de giration d'un faisceau sous sa forme la plus simple :

$$R^2 = R_{\text{micro}}^2 + \frac{1}{n}\sum_{k=1}^{n}(\boldsymbol{G}_k - \boldsymbol{G})^2 + \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{G}_i - \boldsymbol{G})^2. \tag{2.14}$$

Puisque l'on est dans un faisceau, par définition, les conformations des homopolymères constitutifs sont très fortement corrélées :

$$\frac{1}{n}\sum_{k=1}^{n}\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{G}_{ik} - \boldsymbol{G}_k)^2 \approx \left.\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{G}_{ik} - \boldsymbol{G}_k)^2\right|_{k=1}.$$

Il vient alors que la macro-conformation (2.12) est assimilable à la conformation d'un *seul* homopolymère :

$$\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{G}_i - \boldsymbol{G})^2 \approx \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{G}_{ik} - \boldsymbol{G}_k)^2\Big|_{k=1} . \tag{2.15}$$

Du point de vue du rayon de giration, le fait d'avoir un faisceau, c'est comme n'avoir qu'un seul polymère avec une certaine épaisseur (2.13).

2.3 Distribution du rayon de giration d'un faisceau

On vient d'expliciter la formule du rayon de giration d'un faisceau de polymères. À présent, on souhaite exhiber une formule pour la *distribution* du rayon de giration à partir de l'expression précédemment obtenue pour le rayon de giration.

On peut réécrire l'équation (2.14) du rayon de giration d'un faisceau de polymères comme somme de variables aléatoires

$$R^2 = R^2_{\text{micro}} + R^2_\sigma + R^2_N \tag{2.16}$$

où R^2_{micro} est la variable aléatoire à valeur certaine décrivant les micromères dont la distribution s'écrit

$$p_{\text{micro}}(r^2) = \delta_D(r^2 - R^2_{\text{micro}}) , \tag{2.17}$$

R^2_σ est la variable aléatoire décrivant le poly-conformère, *i.e.* la largeur du faisceau et R^2_N est la variable aléatoire décrivant le macro-conformère, *i.e.* la conformation d'un seul polymère. La première hypothèse que l'on va formuler est que ces trois variables aléatoires sont indépendantes. Ce qui signifie que micromères, poly-conformères et macro-conformères sont décorrélés. C'est vrai dans la mesure où :

1. Le volume exclu des micromères ne va pas changer fondamentalement ni les micro-conformations, ni la macro-conformation : R^2_σ et R^2_N sont indépendant de R^2_{micro}.
2. Au sein du faisceau, l'orientation des liens homologues étant fortement corrélée, les micro-conformères sont dans un plan orthogonal aux liens homologues : R^2_σ et R^2_N sont orthogonaux.

La conséquence immédiate est que la distribution de probabilité p de R^2 s'écrit comme la convoluée des distributions de probabilité des variables R^2_{micro}, R^2_σ et R^2_N

$$p = p_{\text{micro}} * p_\sigma * p_N .$$

La distribution du rayon de giration d'un faisceau de polymères se décompose en distributions indépendantes. On va pouvoir traiter séparément la distribution de la largeur du faisceau p_σ et la distribution du conformère p_N.

2.4 Distribution d'un faisceau de chromosomes

Les méthodes pour obtenir les distributions p_N et p_σ sont ici très différentes. En effet, la construction de la distribution p_N repose à la fois sur la théorie des polymères,

ainsi que sur des simulations. Elle sera l'objet des chapitres suivants. Tandis que dans ce chapitre, on va analyser des distributions expérimentales du rayon de giration de faisceau de chromosomes en vue d'obtenir l'expression de p_σ. Mais avant, on va vérifier qu'un faisceau de chromosomes se comporte bien comme un faisceau de polymères tel qu' on l'a précédemment défini.

2.4.1 Les chromosomes appariés forment un faisceau

On formule, ici, la deuxième hypothèse qu'un faisceau de chromosomes appariés se comporte comme un faisceau d'homopolymères, *i.e.* qui vérifie la propriété (2.15).

On appuie notre hypothèse sur les données de Hi-M de Gizzi et al. [28]. La capture conformationnelle des chromosomes ou *Microscopy-Based Chromosome Conformation Capture* en anglais, abrégée en *Hi-M* est une technique expérimentale qui permet de mesurer des cartes de distance au sein des chromosomes et notamment les cartes de distances intra-domaine topologique de la Figure 2.5. On peut y voir l'influence de l'appariement des chromosomes homologues, chez la drosophile, sur la structure interne des domaines topologiques. Les auteurs placent les mêmes marqueurs sur les mêmes sites génomiques des chromosomes homologues. Ils classent les chromosomes comme *appariés* lorsque toutes les marques ne sont détectées qu'une seule et unique fois. Autrement dit, il suffit qu'un seul marqueur soit détecté plus d'une fois pour que les chromosomes soient classés comme *non appariés*. Donc, qu'ils soient complètement ou que partiellement non appariés, les chromosomes seront classés comme non appariés. Les auteurs notent que la structure des domaines est qualitativement la même qu'on soit dans un chromosome apparié ou non. On en conclut que l'appariement des chromosomes n'altère pas leur conformation.

Remarque. *On peut formuler une critique quant à l'exactitude de cette conclusion. Pour que la comparaison soit pertinente, il faudrait qu'elle soit faite entre des chromosomes complètement appariés et des chromosomes complètement non appariés. Ici, on est en train de dire que, en moyenne, la structure des chromosomes partiellement appariés est semblable à la structure des chromosomes complètement appariés.*

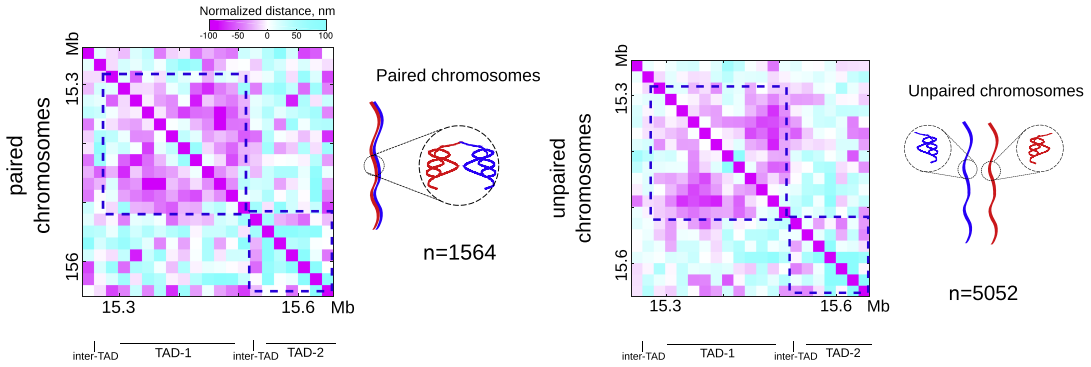

Figure 2.5 – Tirée de l'article de Gizzi et al. [28]. Cartes de distance Hi-M normalisées de chromosomes homologues appariés (à gauche) ou non appariés (à droite). La représentation schématique indique les chromosomes homologues appariés et non appariés. Nombre de noyaux examinés : $n = 1564$ (appariés), $n = 5052$ (non appariés).

De plus, pour une analyse plus quantitative, Gizzi et al. calculent les distances moyennes des paires $\langle |\boldsymbol{r}_i - \boldsymbol{r}_j| \rangle$ et obtiennent, à la Figure 2.6, un coefficient de corrélation de 0.91 entre les chromosomes appariés et non appariés. Puisque les distances des paires sont très fortement corrélées, l'approximation (2.15) est vérifiée. On peut alors en conclure qu'un faisceau de chromosomes se comporte bien comme un faisceau d'homopolymères précédemment défini.

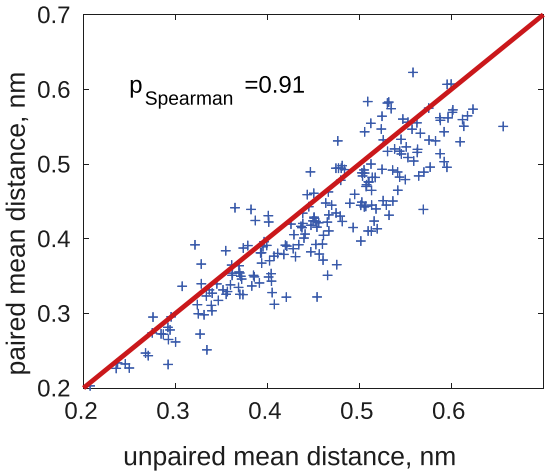

Figure 2.6 – Tirée de l'article de Gizzi et al. [28]. Distances moyennes des paires au sein des chromosomes appariés par rapport aux distances moyennes des paires au sein des chromosomes non appariés, représentées par des croix bleues. La ligne rouge représente une pente égale à 1. On remarque que les chromosomes non appariés sont très légèrement plus gros que les chromosomes appariés. Le coefficient de corrélation de Pearson vaut $0,91$.

2.4.2 Estimation de l'extension transversale du faisceau

Avant de s'intéresser à la distribution de l'extension transversale du faisceau p_σ, on souhaite obtenir une estimation de son extension moyenne $\langle R_\sigma^2 \rangle$.

Les travaux de Cattoni et al. [29] en 3D-SIM permettent d'estimer la taille du faisceau à $\langle R_\sigma^2 \rangle \sim (100\,\mathrm{nm})^2$, chez des cellules tétraploïdes. En effet, en ne marquant qu'un seul site génomique, les auteurs n'observent qu'un seul spot dans 75% des cas. Ce qui signifie que les quatre chromosomes sont à une distance inférieure à la résolution expérimentale qui vaut ici $\sim 100\,\mathrm{nm}$. Leurs mesures suggèrent que les chromosomes, dans ce type cellulaire, se comportent plutôt comme un unique polymère plus épais.

Pour une modélisation plus quantitative, il nous faudrait pouvoir accéder à la distribution de R_σ^2. C'est pourquoi, on formule la troisième hypothèse que dans un faisceau constitué de polymères courts, le rayon de giration est principalement dominé par son épaisseur plutôt que par sa longueur :

$$R_\sigma^2 \gg R_N^2. \tag{2.18}$$

Il en résulte que la distribution du rayon de giration du faisceau se résume à celle de son épaisseur :

$$p \approx p_{\mathrm{micro}} * p_\sigma.$$

Ce qui signifie que dans les données expérimentales de Boettiger et al. [1], la distribution observée du rayon de giration des domaines les plus courts serait celle de l'épaisseur du faisceau.

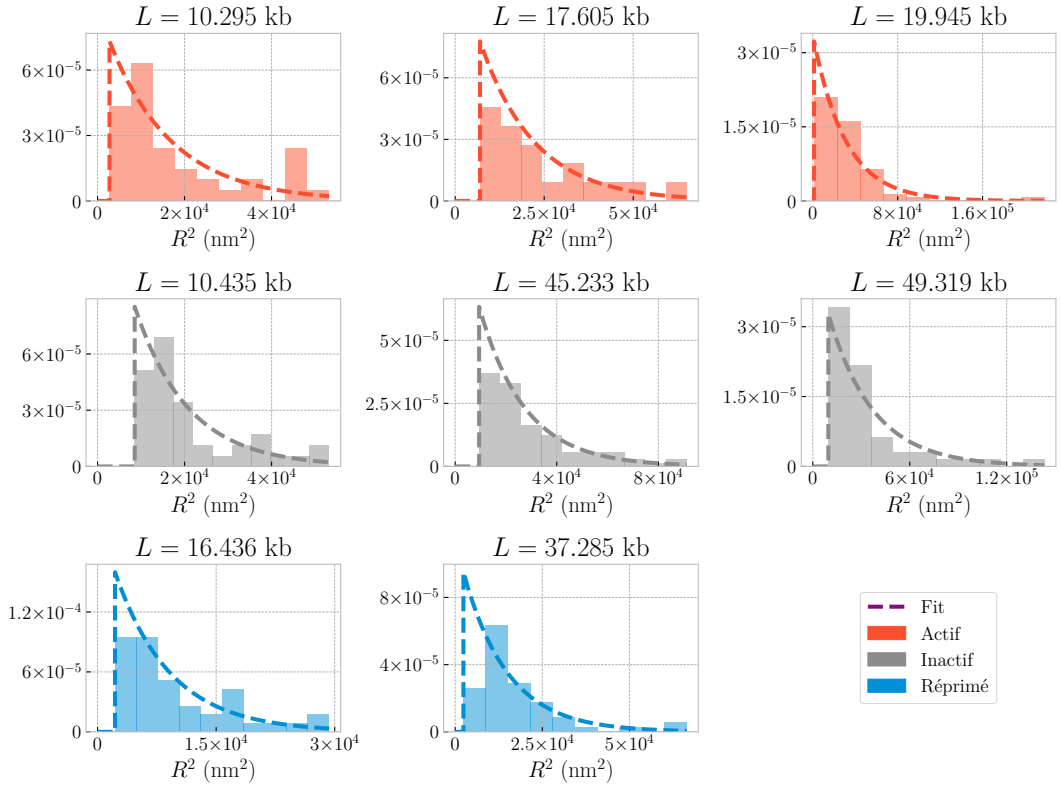

Figure 2.7 – Distributions du rayon de giration des plus petits domaines. Le *fit* est réalisé sur les paramètres λ et μ de la fonction $x \mapsto f_\lambda(x-\mu)$. Sur l'ensemble de ces domaines, on trouve en moyenne $\lambda^{-1/2} \approx 128\,\text{nm}$.

L'allure (Figure 2.7) des distributions des domaines épigénétiques (actifs, inactifs et réprimés) les plus courts suggère que le rayons de giration r^2 suit une loi exponentielle de paramètre λ :

$$f_\lambda(x) = \lambda e^{-\lambda x}. \tag{2.19}$$

J'ai ajusté la loi exponentielle de paramètres (λ, μ), tels que $x \mapsto f_\lambda(x-\mu)$, d'espérance $1/\lambda + \mu$ et d'écart type $1/\lambda$. On obtient les paramètres (λ, μ) optimaux en égalant moyenne et écart type des données (M et E) à la moyenne et à l'écart type de la distribution exponentielle :

$$\begin{cases} M &= \frac{1}{\lambda} + \mu \\ E &= \frac{1}{\lambda} \end{cases} \iff \begin{cases} M - E &= \mu \\ E &= \frac{1}{\lambda} \end{cases}$$

On trouve par cette simple analyse que la largeur du faisceau vaut en moyenne $\lambda^{-1/2} \approx 128\,\text{nm}$ sur les domaines étudiés. Ce qui est compatible avec l'estimation de Cattoni et al. [29] qui donne une taille de faisceau de l'ordre de $\sim 100\,\text{nm}$.

Fort de cette constatation, il vient immédiatement que la distribution radiale p_σ est une loi exponentielle de paramètre λ :

$$p_\sigma(r^2) = f_\lambda(r^2). \tag{2.20}$$

En combinant les équations (2.17) et (2.20), on obtient pour $r^2 \geqslant R^2_{\text{micro}}$ la distribution

$$p_{\text{micro}} * p_\sigma(r^2) = f_\lambda(r^2 - R^2_{\text{micro}})$$

qui est la description la plus simple que l'on puisse obtenir à partir des données expérimentales. La distribution f_λ décrit la distribution radiale du faisceau dont l'épaisseur caractéristique est $\lambda^{-1/2}$. On aboutit à la distribution complète du rayon de giration

$$p(r^2) = \int_0^{r^2 - R^2_{\text{micro}}} f_\lambda(r^2 - R^2_{\text{micro}} - s^2) \cdot p_N(s^2)\, ds^2 \tag{2.21}$$

qui sera la base du travail qui suit.

Hypothèses fondamentales

Voici un récapitulatif des hypothèses fondamentales qui ont été formulées pour aboutir à l'équation (2.21) :

1. Le rayon de giration du faisceau $R^2 = R^2_{\text{micro}} + R^2_\sigma + R^2_N$ se décompose en somme de variables aléatoires indépendantes, dont la distribution est $p = p_{\text{micro}} * p_\sigma * p_N$.
2. La distribution du rayon de giration du macro-conformère p_N est identique à celle du rayon de giration du conformère d'*un* des polymères du faisceau.
3. La distribution radiale du faisceau $p_\sigma(r^2) = f_\lambda(r^2)$ suit une loi exponentielle de paramètre λ, dont l'épaisseur caractéristique est $\lambda^{-1/2}$.

Justification *a posteriori* de la dernière hypothèse

Afin de justifier l'hypothèse (2.18), je vais invoquer des résultats des prochains chapitres de cette thèse. Comme on le verra dans le chapitre suivant, le rayon de giration moyen de la conformation d'un polymère suit la loi de puissance

$$\langle R^2_N \rangle \sim K^2_{\text{nm}} N^{2\nu}$$

où K_{nm} est la longueur de Kuhn du polymère en nanomètre, N le nombre de segments de Kuhn et $\nu = 0.588$ est l'exposant de Flory. La longueur de Kuhn de la chromatine est de l'ordre de $K_{\text{nm}} \sim 50\,\text{nm}$ et pour les polymères les plus courts $N^\nu \sim 1$. Il vient alors que

$$\frac{\langle R^2_N \rangle}{\langle R^2_\sigma \rangle} \sim \frac{1}{4}.$$

Ce qui signifie, en effet, que chez les polymères les plus courts, le rayon de giration de la largeur du faisceau prédomine sur la longueur des polymères constitutifs. Et donc que leur distribution observée pour ces domaines est bien la distribution radiale p_σ.

2.5 Modèle de faisceau gaussien

Je propose un modèle simple pour expliquer la forme exponentielle de la distribution du rayon de giration, des petits domaines, obtenue dans les données expérimentales de la Figure 2.7.

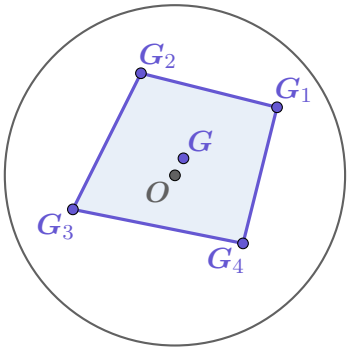

FIGURE 2.8 – Schéma d'un poly-conformère ($n = 4$).

On souhaite modéliser le rayon de giration du poly-conformère de l'équation (2.14)

$$R_\sigma^2 = \frac{1}{n}\sum_{k=1}^{n}(\boldsymbol{G}_k - \boldsymbol{G})^2 \tag{2.22}$$

où les $\boldsymbol{G}_k$ sont les barycentres des polymères constitutifs du faisceau et $\boldsymbol{G}$ est le barycentre du faisceau défini comme

$$\boldsymbol{G} = \frac{1}{n}\sum_{k=1}^{n}\boldsymbol{G}_k. \tag{2.23}$$

Le plus simple consiste à tirer le vecteur $\boldsymbol{G}_k$ selon une gaussienne centrée à d dimensions et de variance σ^2 :

$$\boldsymbol{G}_k \sim \mathcal{N}^d\left(0, \sigma^2\right).$$

Il en résulte que la somme (2.22) est distribuée selon la loi « khi-deux » $(\sigma^2/n)\chi^2_\ell$ à $\ell = d(n-1)$ degrés de liberté et de variance σ^2/n. Le préfacteur σ^2/n est juste une notation pour indiquer la variance de la distribution, il n'est pas à prendre au sens multiplicatif strict. Il y a $n-1$ vecteurs $\boldsymbol{G}_k$, à d dimensions, indépendants à cause de l'équation (2.23) qui impose une contrainte sur un vecteur. La loi χ^2 peut se mettre sous la forme générale

$$\frac{1}{2\lambda}\chi^2_\ell(x) = \frac{(\lambda x)^{\ell/2-1}}{\Gamma\left(\ell/2\right)} f_\lambda(x)$$

où l'on peut facilement reconnaitre la loi exponentielle f_λ de paramètre λ (2.19). Dans le cas où il y a deux degrés de liberté ($\ell = 2$), on retrouve exactement la loi exponentielle $\frac{1}{2\lambda}\chi^2_2 = f_\lambda$. Sinon, $\frac{1}{2\lambda}\chi^2_\ell$ est équivalent à f_λ dans la limite $\lambda x \to +\infty$ et à un polynôme en $\lambda x \to 0$. En somme la distribution de R_σ^2 se comporte comme une loi exponentielle aux grands r^2 et s'écrit

$$p_\sigma(r^2) = (\sigma^2/n)\chi^2_\ell(r^2)$$

avec $\lambda = n/(2\sigma^2)$ et $\ell = d(n-1)$.

Une somme de variables aléatoires gaussiennes élevées au carré suffit à retrouver la décroissance exponentielle des distributions du rayon de giration observée dans les données.

Dans l'absolu, rien n'oblige le poly-conformère à être ni purement à deux dimensions, ni parfaitement symétrique à trois dimensions. On pourrait très bien imaginer un modèle où les $\boldsymbol{G}_k$ seraient tirés selon une gaussienne à trois dimensions et asymétrique dans une direction : $\boldsymbol{G}_k \sim \mathcal{N}^2\left(0,\sigma^2\right)\mathcal{N}\left(0,\sigma_Z^2\right)$ avec $\sigma^2 \neq \sigma_Z^2$. On peut écrire la distribution à l'aide des fonctions caractéristiques, *i.e.* leurs transformées de Fourier. On rappelle que la fonction caractéristique de la distribution $\frac{1}{2\lambda}\chi_\ell^2$ est

$$\varphi_\ell(t) = \left(1 - i\frac{t}{\lambda}\right)^{-\ell/2} ,$$

et que la fonction caractéristique de la somme de variables indépendantes est le produit des fonctions caractéristiques :

$$\varphi(t) = \left(1 - i\frac{2\sigma^2}{n}t\right)^{1-n}\left(1 - i\frac{2\sigma_Z^2}{n}t\right)^{\frac{1}{2}(1-n)} .$$

La distribution se retrouve à l'aide de la transformée de Fourier inverse :

$$\psi(x) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-itx}\varphi(t)\,dt\,.$$

La distribution s'écrit finalement :

$$p_\sigma(r^2) = \psi(r^2).$$

Remarque. *Une autre approche de la modélisation du faisceau qui a été infructueuse est présentée dans l'Annexe A. Elle se fonde sur l'idée que le faisceau peut être vu comme une succession de ponts browniens à travers des points de contrôles qui décrivent la conformation du faisceau.*

2.6 Modèle de faisceau maxwellien

On va discuter d'un raffinement de la description obtenue précédemment. Pour pouvoir donner une interprétation géométrique au paramètre λ, il faut comprendre que la variable aléatoire R_σ^2 décrit l'épaisseur radiale du faisceau de polymères. On formule la quatrième hypothèse que R_σ suit une distribution de Maxwell à deux dimensions, d'où le nom de faisceau maxwellien. Pour reconstruire la distribution de Maxwell, on décompose le rayon de giration R_σ^2 comme $R_\sigma^2 = X^2 + Y^2$ où les variables X et Y sont distribuer selon la gaussienne de variance σ^2

$$g(x,y) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right).$$

On utilise la méthode de la transformation inverse pour exprimer la distribution dans les coordonnées « polaires » (r^2, θ)

$$p_\sigma(r^2, \theta)\, dr^2 d\theta = g(x, y)\, dxdy$$

$$p_\sigma(r^2, \theta) = g(x, y) \underbrace{\left| \frac{dxdy}{dr^2 d\theta} \right|}_{=1/2}$$

$$p_\sigma(r^2, \theta) = \frac{1}{4\pi\sigma^2} \exp\left(-\frac{r^2}{2\sigma^2} \right)$$

où $\left| \frac{dxdy}{dr^2 d\theta} \right|$ désigne le jacobien du changement de variable $(x, y) \to (r^2, \theta)$. Puis on considère la loi marginale selon la coordonnée radiale r^2 :

$$\begin{aligned} p_\sigma(r^2) &= \frac{1}{2\sigma^2} \exp\left(-\frac{r^2}{2\sigma^2} \right) \cdot \underbrace{\frac{1}{2\pi} \int_0^{2\pi} d\theta}_{=1} \\ p_\sigma(r^2) &= \frac{1}{2\sigma^2} \exp\left(-\frac{r^2}{2\sigma^2} \right). \end{aligned} \tag{2.24}$$

On constate que l'on retrouve une distribution exponentielle f_λ de paramètre $\lambda = \frac{1}{2\sigma^2}$.

Remarque. *Pour obtenir l'expression de la distribution de Maxwell à deux dimensions, il faut utiliser le changement de variable polaire $(x, y) \to (r, \theta)$ dont le jacobien vaut $\left| \frac{dxdy}{drd\theta} \right| = r$. En passant, on remarque que l'on peut calculer simplement le jacobien de l'équation (2.24) : $\left| \frac{dxdy}{dr^2 d\theta} \right| = \left| \frac{dr}{dr^2} \right| \cdot \left| \frac{dxdy}{drd\theta} \right| = \frac{1}{2r} \cdot r = \frac{1}{2}$.*

En résumé, on vient de démontrer que si une distribution $p(r)$ est une maxwellienne à deux dimensions et de variance σ^2, alors $p(r^2)$ est une distribution exponentielle f_λ de paramètre $\lambda = 1/2\sigma^2$.

2.6.1 Faisceau maxwellien cylindrique

On peut aller encore un pas plus loin dans l'interprétation, en supposant que les micromères sont des segments et que les micro-conformères sont des disques. Il en résulte que les macromères sont des cylindres de même longueur que les micromères et de même rayon que les micro-conformères.

Segment

Le rayon de giration d'un segment de longueur L s'écrit $R^2 = \frac{1}{12} L^2$. En effet, si on introduit l'indicatrice du segment

$$\delta_{\text{seg}}(\boldsymbol{r}) = \begin{cases} 1/L & \text{si } \boldsymbol{r} \text{ appartient au segment} \\ 0 & \text{sinon} \end{cases}$$

elle est normée, centrée en **0**. Le rayon de giration s'écrit par parité de l'intégrande

$$R^2 = \int \boldsymbol{r}^2 \delta_{\text{seg}}(\boldsymbol{r})\, d\boldsymbol{r}$$
$$R^2 = \frac{2}{L} \int_0^{L/2} x^2\, dx$$
$$R^2 = \frac{1}{12} L^2.$$

Il suffit alors de poser $R^2_{\text{micro}} = \frac{1}{12} L^2$.

Remarque. *On verra dans la section suivante que le segment joue un rôle fondamental dans la modélisation d'un polymère, au travers du segment de Kuhn.*

2.6.2 Disque

Le rayon de giration d'un disque de rayon R_D s'écrit $R^2 = \frac{1}{2} R_D^2$. En effet, si on introduit l'indicatrice du disque de surface $S = \pi R_D^2$

$$\delta_{\text{dis}}(\boldsymbol{r}) = \begin{cases} 1/S & \text{si } \boldsymbol{r} \text{ appartient au disque} \\ 0 & \text{sinon} \end{cases}$$

elle est normée, centrée en **0**. Le rayon de giration s'écrit en coordonnées polaires

$$R^2 = \int \boldsymbol{r}^2 \delta_{\text{dis}}(\boldsymbol{r})\, d\boldsymbol{r}$$
$$R^2 = \frac{1}{S} \int_0^{R_D} r^2 \cdot r\, dr d\theta$$
$$R^2 = \frac{1}{2} R_D^2.$$

Il suffit alors de poser $R^2_\sigma = \frac{1}{2} R_D^2$. Maintenant, si on suppose que le rayon R_D est distribué selon une maxwellienne de variance σ^2, on obtient la distribution

$$p_\sigma(r^2) = 2 f_\lambda(2r^2) = f_{2\lambda}(r^2)$$

de paramètre $2\lambda = 1/\sigma^2$, dont l'épaisseur caractéristique est $\sigma = 1/\sqrt{2\lambda}$. On notera la propriété de la distribution exponentielle

$$2 f_\lambda(2x) = 2\lambda e^{-2\lambda x} = f_{2\lambda}(x).$$

2.6.3 Cylindre

Le rayon de giration d'un cylindre de longueur L et de rayon R_D s'écrit comme la somme du rayon de giration du segment de même longueur et du disque de même rayon $R^2 = \frac{1}{12} L^2 + \frac{1}{2} R_D^2$. En posant $R^2_{\text{micro}} + R^2_\sigma = \frac{1}{12} L^2 + \frac{1}{2} R_D^2$, on obtient la distribution

$$p_{\text{micro}} * p_\sigma(r^2) = f_{2\lambda}(r^2 - \frac{1}{12} L^2)$$

ce qui aboutit, en remplaçant dans l'équation (2.21), à la distribution complète du rayon de giration

$$p(r^2) = \int_0^{r^2-\frac{1}{12}L^2} f_{2\lambda}(r^2 - \frac{1}{12}L^2 - s^2) \cdot p_N(s^2)\, ds^2$$

dont l'épaisseur caractéristique du faisceau est $\sigma = 1/\sqrt{2\lambda}$, le rayon moyen des cylindres.

En somme, un faisceau maxwellien cylindrique est un polymère dont les monomères sont des cylindres de rayon distribué selon une maxwellienne.

2.7 En bref

On vient de montrer que le rayon de giration d'un faisceau de polymère est donné par l'équation (2.14). On réécrit cette équation comme une somme de variable aléatoire (2.16), ce qui donne génériquement la distribution du rayon de giration d'un faisceau de polymères (2.21). De plus, on constate dans les données de Boettiger et al. [1] que la distribution transversale du rayon de giration du faisceau est une loi exponentielle (2.20).

Chapitre 3

Physique des polymères

Dans ce chapitre, ainsi que le suivant, on va s'intéresser de plus près à la physique des polymères, à l'équilibre thermodynamique, pour tenter d'avoir la meilleure description possible de la distribution de probabilité p_N du rayon de giration d'un polymère à N monomères.

Dans ce chapitre plus spécifiquement, je vais présenter les modèles classiques de polymères, notamment la notion fondamentale de longueur de Kuhn et les lois d'échelles décrivant l'évolution de la taille caractéristique d'un polymère en fonction de son nombre de monomères.

Les sections 3.1 à 3.3 sont inspirées du deuxième chapitre de l'ouvrage « Introduction to Biopolymer Physics » de VAN DER MAAREL [30].

3.1 Chaine idéale

Un polymère est une chaine à $N+1$ monomères dont les centres de masse sont repérés par les vecteurs $\boldsymbol{G}_i$. On parle de chaine idéale, ou encore de chaine librement jointe, lorsque les N liens $\boldsymbol{\ell}_i = \boldsymbol{G}_i - \boldsymbol{G}_{i-1}$ décrivent une marche aléatoire. Les liens ont tous la même longueur $\|\boldsymbol{\ell}_i\| = \ell$, n'interagissent pas entre eux et leurs orientations sont décorrélées

$$\langle \boldsymbol{\ell}_i \cdot \boldsymbol{\ell}_j \rangle = \ell^2 \delta_{ij}$$

où δ_{ij} désigne le symbole de Kronecker défini comme

$$\delta_{ij} = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{sinon} \end{cases}.$$

Une grandeur invariante de la chaine est sa longueur de contour L défini par

$$L = \sum_{i=1}^{N} \|\boldsymbol{\ell}_i\| = N\ell\,,$$

la somme des longueurs des liens.

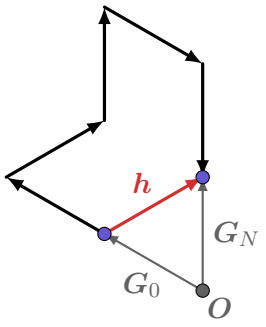

FIGURE 3.1 – Schéma d'une marche aléatoire à N liens, $\boldsymbol{\ell}_i = \boldsymbol{G}_i - \boldsymbol{G}_{i-1}$ et de vecteur bout à bout $\boldsymbol{h} = \boldsymbol{G}_N - \boldsymbol{G}_0$.

3.1.1 Distance bout à bout

Une façon possible de mesurer la taille de la chaine est sa distance bout à bout quadratique moyenne. On introduit, donc, la distance bout à bout $\boldsymbol{h}$

$$\boldsymbol{h} = \boldsymbol{G}_N - \boldsymbol{G}_0 = \sum_{i=1}^{N} \boldsymbol{\ell}_i$$

avec $\boldsymbol{G}_0$ et $\boldsymbol{G}_N$ les deux extrémités. La distance bout à bout quadratique moyenne s'écrit alors :

$$\begin{aligned}
\langle h^2 \rangle &= \langle \boldsymbol{h} \cdot \boldsymbol{h} \rangle \\
\langle h^2 \rangle &= \left\langle \left(\sum_{i=1}^{N} \boldsymbol{\ell}_i \right) \cdot \left(\sum_{j=1}^{N} \boldsymbol{\ell}_j \right) \right\rangle \\
\langle h^2 \rangle &= \sum_{i=1}^{N} \langle \boldsymbol{\ell}_i \cdot \boldsymbol{\ell}_i \rangle + 2 \sum_{i<j} \langle \boldsymbol{\ell}_i \cdot \boldsymbol{\ell}_j \rangle.
\end{aligned}$$

De manière générale, on peut écrire les corrélations des orientations des liens

$$\begin{aligned}
\langle \boldsymbol{\ell}_i \cdot \boldsymbol{\ell}_i \rangle &= \ell^2 \\
\langle \boldsymbol{\ell}_i \cdot \boldsymbol{\ell}_i \rangle &= \ell^2 \langle \cos \theta_{ij} \rangle
\end{aligned}$$

où θ_{ij} est l'angle entre le lien i et le lien j. La distance bout à bout quadratique moyenne s'écrit, donc, en toute généralité :

$$\langle h^2 \rangle = N\ell^2 \left(1 + \frac{2}{N} \sum_{i<j} \langle \cos \theta_{ij} \rangle \right). \tag{3.1}$$

Dans le cas d'une chaine idéale, on a $\langle \boldsymbol{\ell}_i \cdot \boldsymbol{\ell}_j \rangle = 0$, soit $\langle \cos \theta_{ij} \rangle = 0$. Ce qui aboutit à la formule de la distance bout à bout quadratique d'une chaine idéale à N segments :

$$\langle h^2 \rangle = N\ell^2.$$

3.1.2 Rayon de giration

Une autre façon de mesurer la taille de la chaine, introduite dans le chapitre 2, est le rayon de giration. On peut exprimer le rayon de giration d'une chaine à N monomères en fonction de la distance bout à bout h_{ij} entre les monomères i et j :

$$\langle R^2 \rangle = \frac{1}{N^2} \sum_{i<j} \langle h_{ij}^2 \rangle.$$

Dans le cas d'une chaine idéale, tous les domaines de taille $n \leqslant N$ ont les mêmes propriétés quelque soit leur position dans la chaine. L'absence d'interaction entrainant l'absence d'effet de bord, $\langle h_{ij}^2 \rangle$ est la distance bout à bout quadratique moyenne d'un polymère à $|j-i|$ segments. Ce qui permet d'écrire :

$$\begin{aligned}
\langle R^2 \rangle &= \frac{\ell^2}{N^2} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} |j-i| \\
\langle R^2 \rangle &= \frac{\ell^2}{N^2} \frac{(N^2-1)N}{6} \\
\langle R^2 \rangle &= \frac{1}{6} \left(1 - \frac{1}{N^2}\right) N\ell^2.
\end{aligned}$$

Grâce à l'équivalence $1 - \frac{1}{N^2} \sim 1$ pour $N \gg 1$, on aboutit à l'expression asymptotique du rayon de giration :

$$\langle R^2 \rangle \sim \frac{1}{6} N\ell^2. \tag{3.2}$$

Remarque. *On note que le rayon de giration $\langle R^2 \rangle$ est proportionnel au nombre de monomères, tandis que le rayon bout à bout $\langle h^2 \rangle$ est proportionnel au nombre de segments. Donc, pour un polymère à N monomères, on a précisément $\langle h^2 \rangle = (N-1)\ell^2$.*

3.2 Chaine de Kuhn

On vient de voir, dans le cas d'une chaine idéale, que la corrélation de l'orientation entre les différents segments est nulle, $\langle \boldsymbol{\ell}_i \cdot \boldsymbol{\ell}_j \rangle = 0$. Dans le cas d'une chaine de Kuhn, l'orientation entre segments est corrélée le long de la chaine, à courte portée uniquement. Cette corrélation permet de rendre compte de la rigidité de la chaine. La distance bout à bout quadratique moyenne s'écrit en toute généralité comme à l'équation (3.1) pour N segments, où il nous faut discuter le terme $\sum_{i<j} \langle \cos\theta_{ij} \rangle = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \langle \cos\theta_{ij} \rangle$ contenant les corrélations.

Le modèle de la chaine de Kuhn repose sur deux hypothèses :

1. Les corrélations ont une portée limitée à s monomères, *i.e.* $\langle \cos\theta_{ij} \rangle = 0$ pour $|j-i| > s$. La portée est courte devant le nombre de segments : $s \ll N$.
2. La chaine est homogène et les effets de bords sont négligeables, *i.e.* les corrélations ne dépendent que de la distance $|j-i|$ qui sépare deux segments, même aux bords. On peut alors écrire $\langle \cos\theta_{ij} \rangle = \langle \cos\theta_{1j} \rangle$.

La première hypothèse restreint la somme sur j uniquement aux termes inclus dans la portée s de l'interaction, en éliminant les termes nuls :

$$\sum_{j=i+1}^{N} \langle \cos\theta_{ij} \rangle = \sum_{i<j<i+s} \langle \cos\theta_{ij} \rangle.$$

Tandis que la seconde hypothèse rend les corrélations proportionnelles asymptotiquement au nombre N de segments :

$$\sum_{i=1}^{N-1} \sum_{i<j<i+s} \langle \cos\theta_{1j} \rangle = (N-1) \sum_{j=2}^{s} \langle \cos\theta_{1j} \rangle \sim N \sum_{j=2}^{s} \langle \cos\theta_{1j} \rangle. \tag{3.3}$$

En injectant les corrélations (3.3) à courte portée dans l'équation (3.1), on aboutit à la distance bout à bout quadratique moyenne de la chaine de Kuhn :

$$\langle h^2 \rangle \sim \alpha \cdot N\ell^2.$$

On retrouve le même résultat que pour la chaine idéale à un facteur multiplicatif près :

$$\alpha = 1 + 2 \sum_{j=2}^{s} \langle \cos\theta_{1j} \rangle.$$

On peut s'affranchir des corrélations à courte portée en renormalisant la longueur des segments de la chaine et retrouver le comportement d'une chaine idéale

$$\langle h^2 \rangle \sim N_k \ell_K^2 \,, \tag{3.4}$$

de $N_K = N/\alpha$ segments de longueur $\ell_K = \alpha \cdot \ell$, tout en préservant la longueur de contour initiale :

$$L = N\ell = N_K \ell_K. \tag{3.5}$$

La longueur $\ell_K = \alpha \cdot \ell$ est appelée longueur de Kuhn et définit la taille d'un segment de Kuhn.

En d'autres termes, une chaine possédant des corrélations à courte portée est équivalente (de même longueur L et de même taille $\langle h^2 \rangle$) à une chaine idéale de segments de Kuhn, dont les orientations sont par définition décorrélées. On vient d'introduire un concept fondamental de la physique des polymères : la longueur de Kuhn, qui est la longueur à partir de laquelle on perd les effets des corrélations à courte portée.

3.3 Chaine persistante

La chaine persistante, plus communément appelée *worm-like chain* en anglais, est la variante continue de la chaine de Kuhn. On l'obtient en prenant la limite jointe d'une infinité de segments de taille infinitésimale $(N, \ell) \to (+\infty, 0)$, tout en préservant la longueur de contour $L = N\ell$.

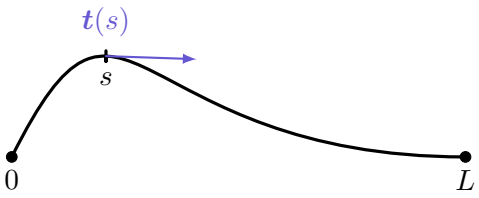

FIGURE 3.2 – Schéma d'une chaine persistante et de son vecteur tangent unitaire $\boldsymbol{t}(s)$ d'abscisse curviligne s.

3.3.1 Distance bout à bout

La distance bout à bout de la chaine continue s'exprime comme

$$\boldsymbol{h} = \int_0^L \boldsymbol{t}(s)\, ds\,,$$

où $\boldsymbol{t}(s) = \frac{\partial}{\partial s}\boldsymbol{r}(s)$ est le vecteur tangent unitaire au contour. L'orientation de la chaine est corrélée à courte portée, *i.e.* corrélation à décroissance exponentielle, traduisant une certaine rigidité :

$$\langle \boldsymbol{t}(s) \cdot \boldsymbol{t}(s') \rangle = \langle \boldsymbol{t}(s - s') \cdot \boldsymbol{t}(0) \rangle = e^{-|s-s'|/L_p}$$

La corrélation est invariante par translation et décroît exponentiellement avec une longueur caractéristique L_p que l'on appelle longueur de persistance de la chaine. La distance bout à bout quadratique moyenne d'une chaine persistante s'écrit

$$\begin{aligned} \langle h^2 \rangle &= \int_0^L ds \int_0^L ds' \langle \boldsymbol{t}(s) \cdot \boldsymbol{t}(s') \rangle \\ \langle h^2 \rangle &= 2L_p^2 \left(\frac{L}{L_p} - 1 + e^{-L/L_p} \right). \end{aligned} \tag{3.6}$$

On peut s'intéresser au cas limite de faible persistance $L_p \ll L$ d'une chaine continue, où l'on obtient :

$$\langle h^2 \rangle \sim 2L_p L.$$

En identifiant au résultat (3.4) obtenu pour une chaine de Kuhn, $\langle h^2 \rangle \sim N_k \ell_K^2 = \ell_K L$, on trouve que la longueur de Kuhn égale le double de la longueur de persistance :

$$\ell_K = 2L_p. \tag{3.7}$$

Par définition, puisque les segments de Kuhn sont exempts de corrélation, pour une chaine persistante, on perd toute trace de corrélation au bout de deux fois sa longueur de persistance.

3.3.2 Rayon de giration

Brièvement, dans une démarche analogue donnant le rayon de giration (3.2) d'une chaine idéale, on écrit celui d'une chaine persistante

$$\langle R^2 \rangle = \frac{1}{L^2} \int_0^L ds \int_s^L ds' \langle h^2(s'-s) \rangle \,,$$

où $\langle h^2(s'-s) \rangle$ désigne la distance bout à bout quadratique moyenne entre les points d'abscisse curviligne s et s'. Toute sous-partie de la chaine suit la loi (3.6) avec une longueur $s'-s$, on a alors

$$\langle R^2 \rangle = 2\frac{L_p^2}{L^2} \int_0^L ds \int_s^L ds' \left(\frac{s'-s}{L_p} - 1 + e^{-(s'-s)/L_p} \right)$$
$$\langle R^2 \rangle = \frac{1}{3}\frac{L_p^4}{L^2} \left[\frac{L^3}{L_p^3} - 3\frac{L^2}{L_p^2} + 6\frac{L}{L_p} - 6\left(1 - e^{-L/L_p}\right) \right] ,$$

qui dans la limite de faible persistance $L_p \ll L$ donne

$$\langle R^2 \rangle = \frac{1}{3} L_p L = \frac{1}{6} L \ell_K = \frac{1}{6} N_K \ell_K^2 .$$

On retrouve le comportement (3.2) d'une chaine idéale avec l'identité (3.7) liant la longueur de persistance à la longueur de Kuhn. Le modèle de la chaine persistante est équivalent au modèle de la chaine de Kuhn dans la limite de faible persistance $L_p \ll L$.

Le modèle de la chaine de Kuhn prédit que la taille caractéristique de la chaine croit comme la racine carrée du nombre de segments, $\langle h^2 \rangle^{1/2} \sim \langle R^2 \rangle^{1/2} \sim N^{1/2}$. Par ailleurs, la corrélation de l'orientation locale des segments le long de la chaine n'a pas d'influence sur les propriétés de la chaine à plus longue distance. Elle rend uniquement compte de la rigidité de la chaine à courte échelle. De plus, ce modèle ignore les interactions de volume exclu des voisins les plus proches spatialement, empêchant essentiellement la chaine de se croiser.

3.4 Chaine réelle

Les modèles que l'on a considérés jusqu'à présent font fi des interactions de volume exclu de segments spatialement proches. Pour une chaine réelle, il faut tenir compte à la fois des interactions de la chaine avec elle-même, l'empêchant notamment de se passer au travers, ainsi qu'avec le solvant dans lequel elle se trouve. On peut identifier trois classes de conformations pour une chaine, plus ou moins probables, en fonction des conditions physiques :

1. Les conformations *stretch* : dans le cas où le polymère est contraint par les extrémités ou dans la limite d'une tige rigide $L_p \gg L$.
2. Les conformations *coil* : dans le cas dit « bon solvant », le polymère enfle afin de maximiser la surface de contact avec le solvant.

3. Les conformations *globule* : dans le cas dit « mauvais solvant », le polymère s'effondre sur lui-même, formant une boule, afin de minimiser la surface de contact avec le solvant.

Elles sont clairement établies expérimentalement [31–33].

Pour une chaine réelle, l'évolution de sa taille caractéristique en fonction du nombre de monomères se généralise sous la forme

$$\langle R^2 \rangle^{1/2} \sim N^{\psi} \,, \tag{3.8}$$

grâce à une approche type champ moyen [34, 35]. L'équation (3.8) est ce que l'on appelle une loi de puissance où l'exposant ψ, *i.e.* la puissance, discrimine les conformations par sa valeur. Les conformations *stretch* décrivent une marche auto-évitante étirée et ont un exposant $\psi = 1$, elle se comporte comme une tige et leur taille $\langle R^2 \rangle^{1/2}$ croit linéairement avec leur nombre N de monomères. Les conformations *coil* décrivent une marche auto-évitante et ont un exposant $\psi = \nu = 0.588 \approx 3/5$, plus grand que dans le cas d'une marche aléatoire, traduisant leur gonflement dû au volume exclu de la chaine. L'exposant ν est l'exposant de Flory. Enfin, les conformations *globule* ont un exposant $\psi = 1/3$, *i.e.* l'inverse de la dimension spatiale, traduisant l'isotropie des conformations.

	Stretch	*Coil*	*Globule*
ψ	1	$0.588 \approx 3/5$	$1/3$

TABLE 3.1 – Récapitulatif des exposants des classes de conformations ($d = 3$). [35, 36]

Dans l'absolu, il est difficile de déterminer la nature d'une classe de conformations *stretch*, *coil* ou *globule*, par la simple évaluation de sa taille caractéristique $\langle R^2 \rangle$ ou $\langle h^2 \rangle$. En revanche, regarder comment cette taille dépend du nombre de monomères (3.8) permet de lever le voile sur la nature de la conformation, grâce à la valeur de l'exposant ψ.

3.5 Énergie libre d'une marche auto-évitante

Pour modéliser la conformation d'une chaine réelle, dans un premier temps, il faut tenir compte des interactions répulsives, dites de volume exclu, empêchant la chaine de se croiser avec elle-même. Puis dans un second temps, il faut tenir compte des interactions attractives lorsque deux monomères entrent en contact, traduisant l'affinité des monomères avec le solvant.

On souhaite établir l'expression de l'énergie libre d'une chaine réelle. Pour modéliser l'auto-évitement, dans l'énergie, de la chaine avec elle-même, deux choix s'offrent à nous :

1. Partir d'une marche aléatoire, dont on connait l'énergie libre [37], et ajouter une interaction répulsive pour modéliser à la fois le volume exclu, mais aussi pour empêcher les liens de se croiser. Le premier pouvant être modéliser avec un potentiel de sphère dure. Le second, du fait des interactions à longues portées, est un problème, pour le moment, insoluble exactement.

2. Partir d'une marche auto-évitante, dont on ne connait pas l'énergie libre, mais dont on peut étudier le comportement au travers de simulations.

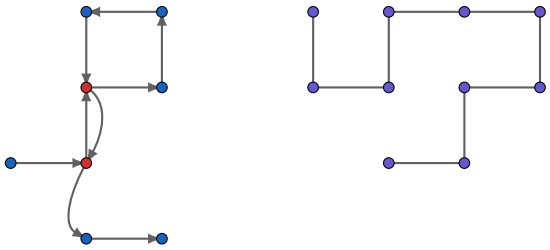

FIGURE 3.3 – À gauche : une marche aléatoire, à droite : une marche auto-évitante.

Le choix qui est fait, ici, est d'étudier la phénoménologie des lois d'échelle d'une marche auto-évitante pour remonter à l'expression de son énergie libre. C'est la démarche initiée par LHUILLIER [38], puis reprise par VICTOR et al. [39–42] que l'on va développer dans cette section.

3.5.1 Classes de conformations

Les conformations se répartissent en trois classes, *i.e.* sous-ensembles, (*stretch*, *coil*, *globule*) en fonction de leur rayon de giration. En taille finie N, les frontière entre ces trois classes sont plutôt floues. L'ambiguïté vient du fait qu'il existe un *crossover*, *i.e.* un hybride, en taille finie, où l'on peut passer continûment de la classe *coil* à la classe *globule*. Cependant, l'ambiguïté est levée dans la limite thermodynamique ($N \to +\infty$), car cette transition est discontinue.

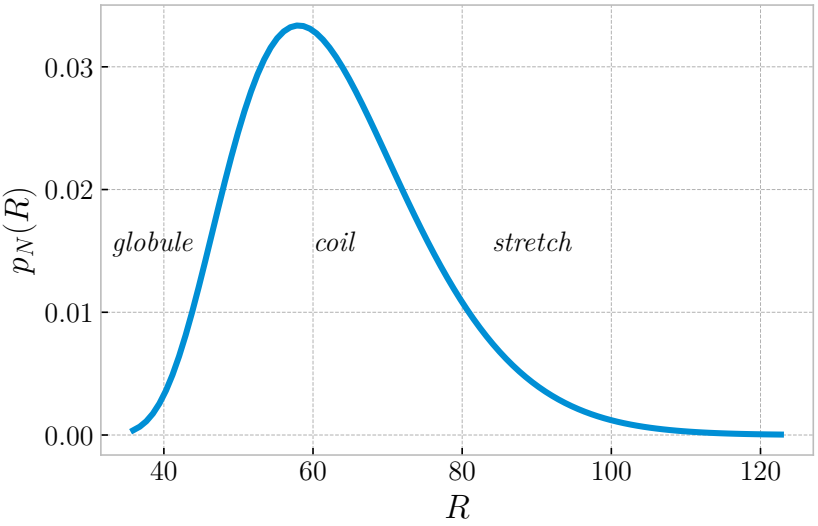

FIGURE 3.4 – Répartition des conformations en fonction de leur rayon de giration pour $N = 5012$.

Du fait que les conformations *coil* soient intermédiaire aux conformations *stretch* et *globule*, cela signifie que :

— Une conformation *stretch* peut être vue comme un *coil* étiré.

— Une conformation *globule* dilué peut être vue comme un *coil* recroquevillé.

Stretch

Les conformations étirées sont invariantes par translation et peuvent être vues comme une succession de parties indépendantes dû à l'absence de contrainte entre les segments liée au volume exclu. Ce qui implique que l'énergie libre d'une telle classe de conformations

$$\beta F_S = N\mathcal{S}(\lambda)$$

est extensive et dépend de sa densité linéique $\lambda = N/R$ au travers d'une fonction $\mathcal{S}$. On suppose que cette fonction $\mathcal{S}$ est une loi de puissance

$$\mathcal{S}(\lambda) = A_S \cdot \lambda^s$$

de préfacteur A_S et de puissance s.

Globule

Les conformations globulaires sont homogènes, ce qui signifie que la densité locale en monomères ne dépend pas de la position au sein des conformations et coïncide avec la densité globale $\rho = N/R^d$, où d est la dimension de l'espace. Ce qui implique de l'énergie libre d'une telle classe de conformations

$$\beta F_G = N\mathcal{G}(\rho)$$

est extensive et dépend de sa densité volumique $\rho = N/R^d$ au travers d'une fonction $\mathcal{G}$. Dans la limite de faible densité ($\rho \to 0$), on suppose que cette fonction $\mathcal{G}$ est une loi de puissance

$$\mathcal{G}(\rho) = A_G \cdot \rho^g$$

de préfacteur A_G et de puissance g.

Coil

Les conformations torsadées sont invariantes d'échelle, de dimension fractale $1/\nu$. Ce qui implique que l'énergie libre d'une telle classe de conformations

$$\beta F_C = \mathcal{C}(\phi)$$

ne dépend que de la variable d'échelle $\phi = N/R^{1/\nu}$ au travers d'une fonction $\mathcal{C}$, où $\nu \approx 0.588$ est l'exposant de Flory. Puisqu'une conformation *coil* peut passer continûment d'une conformation étirée à une conformation globulaire à basse densité, on suppose que cette fonction $\mathcal{C}$ est une somme de lois de puissance

$$\mathcal{C}(\phi) = A_S \cdot \phi^{c_s} + A_G \cdot \phi^{c_g}$$

de préfacteurs A_S, A_G et de puissances c_s, c_g. Dans chacune des deux limites, alternativement, l'un des deux termes va prédominer sur l'autre, permettant le raccordement avec la classe correspondante tout en assurant *et* la continuité *et* la bonne loi d'échelle.

3.5.2 Raccordements

Puisque ces trois familles ne sont que des facettes d'une marche auto-évitante, leur énergie libre ne devrait être que des cas limites d'une *même* énergie libre sous-jacente, où il serait possible de passer continûment de *stretch* à *coil*, puis de *coil* à *globule*. On peut retrouver en partie cette énergie libre en faisant des raccordements par continuité des lois de puissance de la classe *coil* avec la classe *stretch*, puis avec la classe *globule*.

Stretch-Coil

Commençons par étudier une conformation à la limite entre *stretch* et *coil*. Puisque l'on est à la limite, il devrait être équivalant d'utiliser l'énergie libre de la classe *stretch* ou *coil*. Plus précisément, l'expression des deux énergies libres doit coïncider :

$$A_S \cdot N \left(\frac{N}{R}\right)^s = A_S \cdot \left(\frac{N}{R^{1/\nu}}\right)^{c_s}$$

Pour cela, il est nécessaire d'égaler les exposants de N et de R, de part et d'autre, donnant lieu au système :

$$\begin{cases} 1+s &= c_s \\ \nu s &= c_s \end{cases} \Longrightarrow \begin{cases} c_s &= -\nu/(1-\nu) \\ s &= -1/(1-\nu) \end{cases}$$

En introduisant l'exposant de Fisher-Pincus [35]

$$\delta = \frac{1}{1-\nu},$$

on peut réécrire les exposants comme $c_s = -\nu\delta$ et $s = -\delta$.

Globule-Coil

À présent, intéressons-nous à la limite entre un *globule* à basse densité ($\rho \to 0$) et un *coil*. Puisque l'on est à la limite, il devrait être équivalant d'utiliser l'énergie libre de la classe *globule* ou *coil*. Plus précisément, l'expression des deux énergies libres doit coïncider :

$$A_G \cdot N \left(\frac{N}{R^d}\right)^g = A_C \cdot \left(\frac{N}{R^{1/\nu}}\right)^{c_g}$$

Pour cela, il est nécessaire d'égaler les exposants de N et de R, de part et d'autre, donnant lieu au système :

$$\begin{cases} 1+g &= c_g \\ \nu d g &= c_g \end{cases} \Longrightarrow \begin{cases} c_g &= \nu d/(\nu d-1) \\ g &= 1/(\nu d-1) \end{cases}$$

3.5.3 Variable d'échelle

Il apparait une variable d'échelle naturelle dans le *globule* :

$$t = \rho^g$$

qui peut s'interpréter comme une densité « renormalisée » compte tenu de la nature auto-évitante de la conformation, avec $g = 1/(\nu d - 1)$.

Le fait le plus surprenant, c'est que l'on peut, également, réécrire l'énergie libre d'un *stretch* en terme de cette nouvelle variable d'échelle. Pour cela, on cherche sous quelles puissances de N et de t on peut l'écrire :

$$N^x t^y = N^x \left(\frac{N}{R^d}\right)^{gy} = N \left(\frac{N}{R}\right)^{-\delta}$$

Ce qui se traduit par le système suivant :

$$\begin{cases} x + gy &= 1 - \delta \\ dgy &= -\delta \end{cases} \implies x = y = -q$$

où q est la notation utilisée par Imbert et al. [42] pour

$$q = \frac{\nu - 1/d}{\nu d - 1}$$

3.5.4 Synthèse et généralisation

Avant d'aller plus loin, il est nécessaire d'effectuer une synthèse. En somme, pour que l'énergie libre de la classe *coil* puisse s'interpoler avec celle de la classe *stretch* ou de la classe *globule*, elle doit s'écrire comme la somme des énergies libres de chacun des deux états :

$$\beta F_C(t) = \beta F_S(t) + \beta F_G(t)$$

Cette énergie libre est une fonction de la variable d'échelle $t = \rho^{1/(\nu d-1)}$. Elle interpole les deux termes qui prédominent dans des limites opposées. À faible densité ($t \ll 1$), on sera dans la classe *stretch*, dont l'énergie libre s'écrit :

$$\beta F_S(t) = A_S \cdot (Nt)^{-q}$$

alors qu'à plus haute densité ($t \gg 1$), on sera dans la classe *globule*, dont l'énergie libre s'écrit :

$$\beta F_G(t) = N\mathcal{G}(t)$$

Notons que pour faire le raccordement *globule-coil*, on s'est intéressé à cette dernière fonction que dans la limite de basse densité, qui se traduit par $t \to 0$. En somme, on ne s'est intéressé qu'à son développement limité au premier ordre, $\mathcal{G}(t) = A_G \cdot t + \mathcal{O}(t)$. Cependant Imbert et al. [41, 42] ont montré que il était nécessaire d'étendre le développement de la fonction à l'ordre 2

$$\mathcal{G}(t) = A_1 \cdot t + A_2 \cdot t^2 + \cdots$$

pour pouvoir reproduire correctement des résultats issus de simulations. Nous garderons cette forme par la suite.

3.5.5 Correction logarithmique pour les conformations globulaires

On sait que la (densité de) probabilité d'une classe de conformations de densité t est proportionnelle à l'exponentielle de l'énergie libre

$$p_N(t) \propto e^{-\beta F_C(t)}$$

Cependant, on peut montrer qu'il est nécessaire d'introduire une correction logarithmique (additive) en s'intéressant au nombre de marches auto-évitantes de taille N. L'expression donnée à l'équation (I.21) du DE GENNES [35], résultante d'études numériques, s'écrit

$$\mathcal{N} = \Lambda \mu^N N^{\gamma-1} \tag{3.9}$$

où :

— Λ est une constante dépendante du réseau.
— μ est la connectivité du réseau.
— γ est l'exposant d'« amélioration ».

L'exposant d'amélioration sert, comme son nom le suggére, à améliorer la prédiction du nombre de marches auto-évitantes dans l'équation (3.9) par rapport au nombre de marches aléatoires qui s'écrit quant à lui

$$\mathcal{N}_{RW} = \mu_{RW}^N$$

où pour un réseau cubique, on a $\mu_{RW} = 6$, alors que $\mu \approx 4.46$ [35]. De même, sur un réseau cubique l'exposant d'amélioration vaut $\gamma \approx 7/6$ [35].

Pour prendre en compte cette correction, il faut se placer dans la classe globulaire où le nombre de conformations à une densité t donnée s'écrit de la même façon qu'en (3.9) [43]

$$\mathcal{N}_G(t) = \Lambda(t)\mu(t)^N N^{\gamma_g - 1}$$

où :

— $\mu(t)$ est la connectivité d'un *globule* de paramètre t.
— γ_g est l'exposant d'amélioration pour les conformations globulaires.
— $\Lambda(t)$ est une constante dépendant à la fois du réseau et de la densité.

À haute densité ($t \gg 1$), la probabilité d'observer une conformation (globulaire) de densité t s'écrit

$$p_N(t) = \frac{\mathcal{N}_G(t)}{\mathcal{N}} = \frac{\Lambda(t)}{\Lambda} \left(\frac{\mu(t)}{\mu}\right)^N N^{\gamma_g - \gamma}$$

On sait par ailleurs que la probabilité doit dépendre de N comme $p_N \propto \exp(-N\mathcal{G}(t))$, ce qui permet l'identification

$$\frac{\mu(t)}{\mu} = e^{-\mathcal{G}(t)}$$

Enfin, IMBERT et al. [42] ont conjecturé que le terme

$$\frac{\Lambda(t)}{\Lambda} \propto t^{-c}$$

suivait également une loi de puissance avec un exposant $-c$. Ils ont également montré que cet exposant devait satisfaire la relation

$$c = 1 + \gamma - \gamma_g$$

et ont mesuré sa valeur $c \approx 1.13$ sur des simulations.

3.5.6 Conclusion

On vient de montrer que l'énergie libre d'une marche auto-évitante s'écrit phénoménologiquement en fonction de la densité t comme

$$\beta F(t) = -\frac{1}{k_B} S_0(t) = A_1 \cdot Nt + A_2 \cdot Nt^2 + A_3 \cdot (Nt)^{-q} + c \ln(Nt) \tag{3.10}$$

où le terme en A_3 vient de l'énergie libre des conformations *stretch-coil* alors que le reste vient de l'énergie libre des conformations globulaires, notamment la correction logarithmique. Cette énergie libre est purement entropique, car on a considéré des marches auto-évitantes sans énergie d'interaction entre monomères ($U = 0$), d'où $\beta F = -S_0/k_B$. Avant de pouvoir achever l'expression de l'énergie libre d'une marche auto-évitante attractive, il nous faut revenir sur quelques bases de physique statistique.

Chapitre 4

Marche auto-évitante attractive

Au chapitre précédent, on a obtenu l'énergie libre d'une marche auto-évitante. On souhaite à présent l'étendre au cas d'une marche marche auto-évitante *attractive*.

Le simple ajout d'un paramètre $-\varepsilon$ d'interaction attractive par contact permet de retrouver la transition *coil-globule* des polymères. En effet, un polymère auto-attractif de taille infinie effectue une transition *coil-globule* (du premier ordre) à une température critique Θ ou de façon équivalente à une énergie critique $\varepsilon_\Theta = k_B\Theta \approx 0.27$ [44–46].

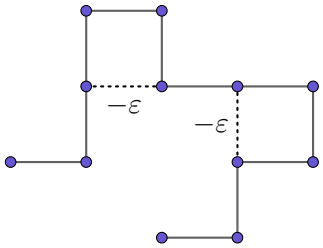

FIGURE 4.1 – Marche auto-évitante attractive avec une énergie $-\varepsilon$ par contact.

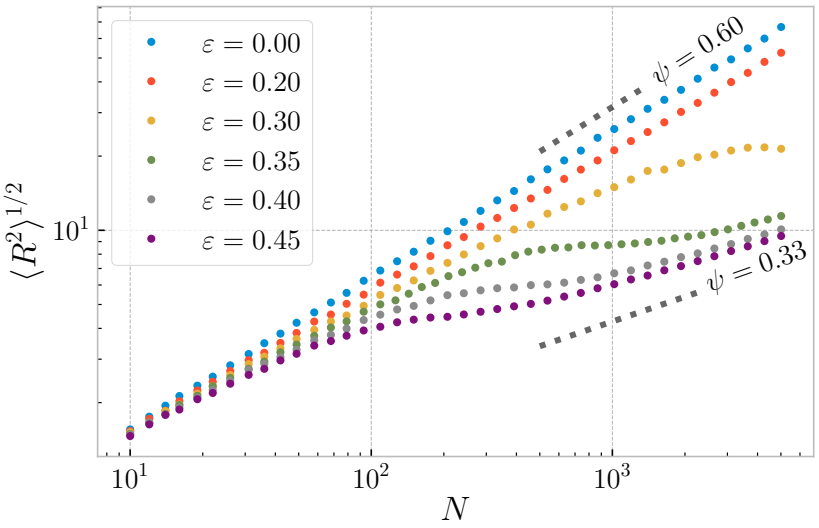

FIGURE 4.2 – Transition *coil-globule*.

Si l'énergie d'interaction est trop faible ($\varepsilon < \varepsilon_\Theta$), alors le polymère adopte une conformation de la classe *coil*. C'est la répulsion qui prédomine.
Tandis que si l'énergie d'interaction est suffisamment forte ($\varepsilon > \varepsilon_\Theta$), alors le polymère s'effondre sur lui-même et adopte une conformation de la classe *globule*. C'est l'attraction qui prédomine.

La relativement faible valeur de l'énergie de transition ($\varepsilon_\Theta < 1$) vient de l'existence d'effets coopératifs entre les monomères. Du fait qu'il y ait moins de monomères en taille *finie*, il faut alors plus d'énergie par contact pour stabiliser les conformations globulaires. Ainsi, l'énergie de transition ε_{Θ_N} dépend du nombre de monomères N et est d'autant plus élevée que le polymère est petit ($\varepsilon_{\Theta_N} > \varepsilon_\Theta$). C'est-à-dire que si on se place à une énergie suffisamment grande ($\varepsilon > \varepsilon_\Theta$), alors les petits polymères seront encore *coil* ($\psi \approx 3/5$), tandis que les plus grands seront déjà *globule* ($\psi = 1/3$).

Malgré une certain nombre d'études [41, 42, 44], il manquait encore une formulation de l'énergie libre capable de reproduire, raisonnablement, le comportement d'une marche aléatoire auto-évitante attractive au voisinage du *crossover*. Dans ce chapitre, je vais développer une approche théorique pour obtenir une première expression de l'énergie libre, à partir d'argument de lois d'échelle. Cette expression fera intervenir des paramètres issus des lois d'échelle qui seront des fonctions, uniquement, du paramètre d'interaction ε. Ces paramètres encoderont toute la phénoménologique de la transition *coil-globule* et ne pourront, cependant, être déterminer qu'au travers d'une étude numérique au chapitre suivant.

4.1 Rappels de physique statistique

Dans cette partie, je vais m'employer à faire des rappels succincts de physique statistique qui vont me servir de fondement pour appuyer l'approche phénoménologique employée afin d'obtenir l'expression complète de l'énergie libre d'une marche auto-évitante attractive.

4.1.1 Micro-état ou conformation

En physique statistique, on appelle micro-état la configuration microscopique d'un système. Dans le cas d'un polymère à N monomères dont la conformation est assimilable à une marche auto-évitante, sa configuration microscopique est donnée par sa conformation $\mathscr{C}$.

4.1.2 Entropie

À chaque conformation $\mathscr{C}$, on peut lui associer une probabilité $p_{\mathscr{C}}$. L'ensemble de ces probabilités satisfait la propriété de normalisation

$$\sum_{\mathscr{C}} p_{\mathscr{C}} = 1 \tag{4.1}$$

et permet de construire l'entropie du système :

$$S = -k_B \sum_{\mathscr{C}} p_{\mathscr{C}} \ln p_{\mathscr{C}}$$

où k_B désigne la consante de Boltzmann. L'interprétation de l'entropie est quelque peu délicate. Il faut tout d'abord introduire la notion de macro-état qui est un état macroscopique caractérisé par des variables d'état (typiquement l'énergie E, le volume V et le nombre de particules N). Le macro-état est l'ensemble des micro-états « compatibles » avec les variables d'état observées. Plus un macro-état a de micro-états « compatibles », moins on a d'information sur son état microscopique instantanée. L'entropie quantifie le manque d'information sur le micro-état instantané du système étant donné un macro-état (à l'équilibre thermodynamique).

Par exemple, si on prend un macro-état où il n'y a qu'un seul micro-état compatible, appelons p_0 sa probabilité. Elle vaut $p_0 = 1$, car toutes les autres micro-états i ont une probabilité $p_i = 0$. On a alors $p_0 \ln p_0 = 0$ et pour les autres $p_i \ln p_i = 0$, d'où $S = 0$. L'entropie d'un macro-état parfaitement connu, *i.e.* n'ayant qu'un seul micro-état, a une entropie nulle. Inversement, on peut montrer que si tous les micro-états accessibles sont équiprobables, alors l'entropie est maximale.

4.1.3 Ensemble microcanonique

Pour savoir quelle probabilité attribuer à une conformation $\mathscr{C}$, *i.e.* à un micro-état, il faut d'abord savoir quelles sont les variables d'état que l'on doit considérer. Elles diffèrent en fonction des conditions physiques dans lesquelles se trouve le système et pondèrent différemment les micro-états. On parle plus généralement d'ensemble statistique pour la façon dont on attribue les probabilités aux micro-états.

Lorsque le système physique est isolé, le triplet de variables d'état : énergie, volume et nombre de particules (E, V, N) est stationnaire à l'équilibre thermodynamique et de ce fait caractérise un macro-état. Le postulat fondamental de la physique statistique stipule que tous les micro-états accessibles d'un système isolé et à l'équilibre sont équiprobables. La probabilité d'une conformation s'écrit alors

$$p_{\mathscr{C}} = \frac{1}{\Omega(E, V, N)}$$

où $\Omega(E, V, N)$ désigne, en toute généralité, le nombre de micro-états ayant l'énergie E à dE près, pour un système ayant accès au volume V et possédant N particules.

Remarque. *Dans toute la suite, on va considérer que le volume accessible est infini $V \to +\infty$ pour éviter les effets de confinement et pour l'élider dans les notations.*

Dans le cas d'une marche auto-évitante (attractive ou non), les conformations ont une probabilité

$$p_{\mathscr{C}} = \frac{1}{\Omega(E, N)}$$

où $\Omega(E, N)$ désigne, plus spécifiquement, le nombre de marche auto-évitante d'énergie E et de taille N. Il est à noter que N désigne non pas le nombre de particules présentes dans le système, mais bien le nombre de monomères constitutifs. En effet, on ne considère ici que la statistique d'une chaine unique de taille N.

Entropie

Dans le cas particulier de l'ensemble microcanonique, *i.e.* d'une marche auto-évitante isolée, on peut écrire

$$S(E,N) = k_B \ln \Omega(E,N) \underbrace{\sum_{\mathscr{C}} p_{\mathscr{C}}}_{=1}$$

en prenant le soin de ne substituer que le terme en $\ln p_{\mathscr{C}}$ et en reconnaissant la normalisation de la probabilité (4.1). On aboutit à l'expression de l'entropie microcanonique

$$S(E,N) = k_B \ln \Omega(E,N) \tag{4.2}$$

Dans le cas particulier d'une marche auto-évitante sans interaction ($E = 0$) de taille N, on écrit l'entropie

$$S_0(N) = S(E=0,N) = k_B \ln \Omega_0(N) \tag{4.3}$$

où $\Omega_0(N) = \Omega(E=0,N)$ est le nombre de marches auto-évitantes de N monomères sans interaction.

4.1.4 Ensemble canonique

L'ensemble canonique désigne l'ensemble des probabilités des micro-états d'un sous-système en contact avec un thermostat, où le système {sous-système + thermostat} est isolé. Le triplet de variables d'état est : la température, le volume et le nombre de particules (T,V,N). On montre qu'à chaque micro-état $\mathscr{C}$, on peut lui associer une énergie $E_{\mathscr{C}}$ qui permet d'exprimer sa probabilité

$$p_{\mathscr{C}} = \frac{1}{Z} e^{-\beta E_{\mathscr{C}}} \tag{4.4}$$

où $\beta = (k_B T)^{-1}$ et Z apparait comme une constante pour assurer la normalisation des probabilités (4.1).

Dans notre cas, on s'intéresse à un polymère dont les monomères ont une interaction attractive $-J$ par contact qui permet d'exprimer l'énergie d'une conformation

$$E_{\mathscr{C}} = -J m_{\mathscr{C}}$$

où $m_{\mathscr{C}}$ est le nombre de contacts de la conformation $\mathscr{C}$. La conformation d'un tel polymère est modélisable comme une marche auto-évitante attractive.

Remarque. *Pour une chaine, N désigne le nombre de monomères.*

Fonction de partition

La constante de normalisation Z est appelée fonction de partition canonique et est définie par

$$Z = \sum_{\mathscr{C}} e^{-\beta E_{\mathscr{C}}} \tag{4.5}$$

Cette fonction est fondamentale, car elle encode toutes les propriétés thermodynamiques du système à l'équilibre. C'est-à-dire que l'on peut exprimer toutes les variables d'état du système en fonction de Z.

Remarque. *Dans le cas d'une marche auto-évitante ($\varepsilon = 0$), ensembles microcanonique et canonique coïncident ($Z = \Omega_0$).*

Énergie interne

L'énergie interne $U = \langle E \rangle$ est la moyenne macroscopique de l'ensemble des énergies microscopiques $E_{\mathscr{C}}$:

$$\langle E \rangle = \sum_{\mathscr{C}} p_{\mathscr{C}} E_{\mathscr{C}} \tag{4.6}$$

On peut exprimer l'énergie interne $\langle E \rangle$ en fonction de la fonction de partition Z, en utilisant la définition de la probabilité (4.4) et de la fonction de partition canonique (4.5) :

$$\langle E \rangle = \frac{1}{Z} \sum_{\mathscr{C}} E_{\mathscr{C}} e^{-\beta E_{\mathscr{C}}}$$

$$\langle E \rangle = -\frac{1}{Z} \frac{\partial}{\partial \beta} \underbrace{\sum_{\mathscr{C}} e^{-\beta E_{\mathscr{C}}}}_{=Z}$$

L'énergie interne s'exprime, donc, en fonction de la fonction de partition Z comme :

$$\langle E \rangle = -\frac{\partial}{\partial \beta} \ln Z$$

Entropie

Dans l'ensemble canonique, en utilisant la même démarche que dans l'ensemble microcanonique (4.2), l'entropie s'exprime

$$S = k_B \sum_{\mathscr{C}} p_{\mathscr{C}} \left(\ln Z + \beta E_{\mathscr{C}} \right)$$

$$S = k_B \ln Z \underbrace{\sum_{\mathscr{C}} p_{\mathscr{C}}}_{=1} + k_B \beta \underbrace{\sum_{\mathscr{C}} p_{\mathscr{C}} E_{\mathscr{C}}}_{=\langle E \rangle}$$

en utilisant successivement : la définition de la probabilité canonique (4.4), la normalisation de la probabilité (4.1) et la définition de l'énergie interne (4.6). On aboutit à l'expression de l'entropie canonique :

$$S = k_B \ln Z + k_B \beta \langle E \rangle \tag{4.7}$$

Énergie libre

En rapprochant la définition thermodynamique de l'énergie libre $F = U - TS$ de l'entropie canonique (4.7), on obtient :

$$F = \langle E \rangle - TS = -\frac{1}{\beta} \ln Z \tag{4.8}$$

que l'on peut écrire de manière équivalente $Z = e^{-\beta F}$. On constate que l'énergie libre est en bijection avec la fonction de partition, ce qui signifie qu'il est équivalant de travailler avec Z ou avec F. Par conséquent, l'énergie libre encode, également, toutes les propriétés thermodynamiques du système à l'équilibre.

4.1.5 Fonction génératrice des moments et des cumulants

Cette partie est inspirée du chapitre « Series expansions » de l'ouvrage de KARDAR [47], traitant des développements en haute température. On y aborde un point formel, d'une part, pour justifier l'écriture de l'énergie libre d'une marche auto-évitante attractive. D'autre part, pour montrer qu'elle peut se construire théoriquement à partir des propriétés statistiques d'une marche auto-évitante *sans attraction*.
On procède par analyse-synthèse.

Analyse

On introduit le paramètre d'attraction

$$\varepsilon = \beta J = \frac{J}{k_B T}$$

qui est l'énergie d'interaction par contact, par unité de $k_B T$ et qui permet de réécrire la fonction de partition comme

$$Z(\varepsilon) = \sum_{\mathscr{C}} e^{\varepsilon m_{\mathscr{C}}}, \quad \text{avec} \quad -\beta E_{\mathscr{C}} = \varepsilon m_{\mathscr{C}} \tag{4.9}$$

On développe l'exponentielle en série entière, puis on distribue la somme sur les conformations sur chacun des termes du développement

$$Z(\varepsilon) = \sum_{\mathscr{C}} \left(1 + \varepsilon m_{\mathscr{C}} + \frac{\varepsilon^2}{2} m_{\mathscr{C}}^2 + \cdots \right)$$
$$Z(\varepsilon) = \Omega_0 + \varepsilon \sum_{\mathscr{C}} m_{\mathscr{C}} + \frac{\varepsilon^2}{2} \sum_{\mathscr{C}} m_{\mathscr{C}}^2 + \cdots$$

où $\sum_{\mathscr{C}} 1 = \Omega_0$ est le nombre de marches auto-évitantes de taille N de l'équation (4.3). En factorisant par Ω_0, on fait apparaitre le développement sur les moments de la distribution du nombre de contacts au sein d'une marche auto-évitante

$$Z(\varepsilon) = \Omega_0 \left(1 + \langle m \rangle_0 \cdot \varepsilon + \langle m^2 \rangle_0 \cdot \frac{\varepsilon^2}{2} + \cdots \right)$$

où $\langle \square \rangle_0 = \frac{1}{\Omega_0} \sum_{\mathscr{C}} \square_{\mathscr{C}}$ désigne la moyenne dans l'ensemble canonique à $\varepsilon = 0$ de $\square$.

Remarque. *La fonction Z/Ω_0 est la fonction génératrice des moments de la distribution du nombre de contacts d'une marche auto-évitante ($\varepsilon = 0$).*

Synthèse

On factorise la fonction de partition par le nombre de marches auto-évitantes Ω_0

$$Z(\varepsilon) = \Omega_0 \left(\frac{1}{\Omega_0} \sum_{\mathscr{C}} \exp(\varepsilon m_{\mathscr{C}}) \right)$$
$$Z(\varepsilon) = \Omega_0 \langle e^{\varepsilon m} \rangle_0$$

À présent, on peut exprimer l'énergie libre en fonction du paramètre d'attraction ε, en prenant soin de reconnaitre l'entropie d'une marche auto-évitante S_0 de l'équation (4.3)

$$\begin{aligned} -\beta F(\varepsilon) &= \ln Z(\varepsilon) \\ -\beta F(\varepsilon) &= \frac{1}{k_B} S_0 + K(\varepsilon) \end{aligned} \tag{4.10}$$

On introduit K la fonction génératrice des cumulants de la distribution du nombre de contacts d'une marche auto-évitante définie comme

$$\begin{aligned} K(\varepsilon) &= \ln \langle e^{\varepsilon m} \rangle_0 \\ K(\varepsilon) &= \sum_{n=1}^{+\infty} \kappa_n \frac{\varepsilon^n}{n!} \end{aligned} \tag{4.11}$$

où κ_n désignent les cumulants de la distribution du nombre de contacts. Ils sont indépendants de ε.

Remarque. *La fonction génératrice K des cumulants s'annule en $\varepsilon = 0$.*

Il est remarquable que l'énergie libre d'une marche auto-évitante attractive puisse être écrite en terme de l'entropie S_0 et de la fonction génératrice des cumulants K de la distribution du nombre de contacts, toutes deux d'une marche auto-évitante ($\varepsilon = 0$).

On peut exprimer la fonction génératrice K en fonction de l'énergie interne U d'une marche auto-évitante attractive, en identifiant l'expression $F = U - TS$ à l'équation (4.10)

$$\begin{aligned} F(\varepsilon) &= U - TS \\ F(\varepsilon) &= -\frac{1}{\beta} K(\varepsilon) - TS_0 \end{aligned}$$

on obtient

$$K(\varepsilon) = \frac{1}{k_B}(S(\varepsilon) - S_0) - \beta U(\varepsilon)$$

On peut constater que K incorpore à la fois l'énergie interne U et la différence d'entropie que génère l'ajout du paramètre d'attraction ε par rapport à la marche auto-évitante.

4.1.6 Macro-état de densité fixée

Pour faire le lien avec l'expression de l'énergie libre obtenue à l'équation (3.10), on souhaite écrire l'énergie libre d'une marche auto-évitante attractive en fonction de sa densité t. C'est pourquoi on s'intéresse, ici, au macro-état de densité t (à dt près), *i.e.* à l'ensemble des conformations $\mathscr{C}_t$ de densité t (à dt près). La fonction de partition du macro-état de densité t s'écrit

$$Z(t, \varepsilon) = \sum_{\mathscr{C}_t} e^{-\beta E_{\mathscr{C}_t}}$$

On retrouve la fonction de partition globale en sommant sur toutes les densités. Par cette définition, il vient immédiatement que l'énergie libre en fonction de t s'écrit

$$-\beta F(t,\varepsilon) = \frac{1}{k_B} S_0(t) + K(t,\varepsilon)$$

où $S_0(t)$ et $K(\varepsilon,t)$ désignent, respectivement, l'entropie et la fonction génératrice des cumulants de la distribution du nombre de contacts pour des conformations de densité t dans l'ensemble canonique à $\varepsilon = 0$. La fonction génératrice s'écrit alors comme

$$K(t,\varepsilon) = \sum_{n=1}^{+\infty} \kappa_n(t) \frac{\varepsilon^n}{n!}$$

où $\kappa_n(t)$ sont les cumulants pour une densité t donnée.

Remarque. *En toute rigueur, le cumulant $\kappa_n(t)$ est une fonction de la densité t, du volume V et du nombre de monomères N.*

4.1.7 En bref

On vient de montrer que l'énergie libre d'une marche auto-évitante *attractive* peut s'écrire comme un développement (4.10) en basse énergie autour de l'entropie d'une marche auto-évitante. Ce développement ne fait intervenir que des fonctions propres à la marche auto-évitante : l'entropie (4.3) et la fonction génératrice (4.11) des cumultants de la distribution du nombre de contacts. En théorie, si on connait toutes les propriétés statistiques d'une marche auto-évitante, on accède alors automatiquement à l'énergie libre d'une marche auto-évitante *attractive*.

4.2 Approche phénoménologique de Victor (état de l'art)

Les marches auto-évitantes attractives possèdent une physique très riche. Grâce à l'introduction d'un paramètre d'attraction ε, la chaine peut réaliser une transition de phase *coil-globule* [34, 35, 48, 49] où les effets de taille finie sont prépondérants, car en pratique les polymères sont toujours très loin de la limite thermodynamique ($N \to +\infty$). Cette transition est remarquable, car dite tricritique [35, 44]. Cependant, cet aspect ne sera pas considéré ici.

4.2.1 Lois de puissance des cumulants

Bien que la fonction génératrice $K(t,\varepsilon)$ ait été construite *ab initio*, *i.e.* à partir des principes premiers de la physique statistique, son écriture demeure incomplète. À ce stade de la construction de l'énergie libre, il est nécessaire d'expliciter la dépendance à la fois en t et en N de la fonction génératrice $K(t,\varepsilon)$. Cependant, l'écriture exacte des cumulants $\kappa_n(t)$ demeure hors d'atteinte.

C'est pourquoi, on se base sur les travaux numériques d'Imbert et Victor [41] pour obtenir une expression approchée de $K(t,\varepsilon)$. Il s'avère que leur expression, en lois de puissance, des deux premiers cumultants est proche de l'écriture de l'entropie obtenue à l'équation (3.10). On fait, donc, l'hypothèse que les cumulants $\kappa_n(t)$ ont

la même dépendance en N et en t que l'entropie d'une marche auto-évitante, ce qui permet de réécrire la fonction génératrice

$$\begin{aligned} K(t,\varepsilon) &= \sum_{n=1}^{+\infty} \kappa_n(t)\frac{\varepsilon^n}{n!} \\ K(t,\varepsilon) &= b_1(\varepsilon)\cdot Nt + b_2(\varepsilon)\cdot Nt^2 + b_3(\varepsilon)\cdot (Nt)^{-q} \end{aligned} \tag{4.12}$$

où b_1, b_2, b_3 sont des fonctions continues de ε.

Remarque. *Les fonctions b_1, b_2, b_3 s'annulent en $\varepsilon = 0$.*

4.2.2 Conclusion

En injectant l'entropie obtenue à l'équation (3.10) pour une marche auto-évitante dans l'équation (4.10), on aboutit à l'équation de l'énergie libre d'une marche auto-évitante attractive

$$\beta F(t,\varepsilon) = a_1(\varepsilon)\cdot Nt + a_2(\varepsilon)\cdot Nt^2 + a_3(\varepsilon)\cdot (Nt)^{-q} + c\ln Nt \tag{4.13}$$

avec $a_i(\varepsilon) = A_i - b_i(\varepsilon)$ qui par transitivité sont, aussi, des fonctions continues de ε.

L'énergie libre d'une marche auto-évitante attractive obtenue à l'équation (4.13) incorpore et généralise celle de l'équation (3.10) d'une marche auto-évitante. Si on regarde la définition des fonctions a_i, les constantes A_i viennent de l'entropie S_0 d'une marche auto-évitante (3.10), alors que les fonctions b_i sont issues du développement (4.12) en cumulants et correspondent aux interactions attractives.

Le début du développement de l'énergie libre $a_1(\varepsilon)\cdot Nt + a_2(\varepsilon)\cdot Nt^2$ rappelle le développement en série du viriel (à basse densité) avec une densité renormalisé $t = \rho^{1/(\nu d-1)}$. Les coefficients a_1 et a_2 jouent le rôle, respectivement, de deuxième et troisième coefficients du viriel. Plus particulièrement, le signe du coefficient a_1 décrit la prédominance des interactions répulsives ($a_1(\varepsilon) > 0$) ou attractives ($a_1(\varepsilon) < 0$). Lorsqu'il s'annule, il marque le point Θ, *i.e.* l'écrantage des interactions répulsives de volume exclu par les interactions attractives de contacts.

Nous verrons au chapitre 5 que l'expression obtenue n'est pas complète et qu'il sera nécessaire de revenir sur l'hypothèse (4.12) pour y ajouter l'énergie de surface.

On vient d'aboutir à l'expression de l'énergie libre d'une marche auto-évitante attractive où l'on a explicité la taille N et incorporé le paramètre d'attraction ε par l'intermédiaire des fonctions a_i, qui elles sont indépendantes de N. Ces fonctions sont, *a priori*, universelles et contiennent toute la phénoménologie de la transition *coil-globule*. Cependant, on ne peut en dégager une expression sur la base d'arguments théoriques. C'est pourquoi, je propose d'inférer les valeurs des fonctions sur la base de simulations, puis d'en effectuer une interpolation.

Chapitre 5

Résultats numériques

Dans le modèle que l'on vient de mettre en place dans le chapitre précédent, on a réussi à séparer la dépendance de l'énergie libre βF en la taille N et en le paramètre d'attraction ε au travers des fonctions a_1, a_2, a_3. C'est pourquoi on va pouvoir dès à présent écrire formellement l'énergie libre $\beta F_N(t|\varepsilon)$.

L'objectif de ce chapitre est de compléter la construction du modèle grâce à une approche phénoménologique. On procède en deux temps. D'abord, on infère les paramètres du modèle grâce aux simulations, où les conformations ont pu être échantillonnées pour en déduire les distributions du rayon de giration à différents couples (ε, N). Puis, on vérifie la capacité du modèle à reproduire les données d'expériences numériques « exacts » d'une marche auto-évitante attractive. On verra que le modèle ne sera capable de reproduire les simulations qu'en y incluant l'énergie de surface des conformations globulaires.

5.1 Simulation

Dans cette partie, je vais présenter l'algorithme du « serpent rampant » qui est une variante de l'algorithme de Metropolis [50]. Avant de présenter successivement ceux deux algorithmes, ainsi que les données obtenues, je vais revenir sur les fondements de l'algorithme de Metropolis, *i.e.* le théorème d'ergodicité.

5.1.1 Théorème d'ergodicité

En physique statistique, on doit distinguer deux types de moyenne. Tout d'abord, il y a la moyenne statistique d'une observable A

$$\langle A \rangle = \sum_{\mathscr{C}} p_{\mathscr{C}} A_{\mathscr{C}} \tag{5.1}$$

que l'on a vu dans le chapitre précédent. C'est la moyenne sur l'ensemble des configurations à l'équilibre permises par les contraintes extérieures (volume, nombre de particules, ...). Elle est pratique à manier pour les calculs. En revanche, d'un point de vue expérimental, il est pratique d'utiliser la moyenne temporelle d'une observable A

le long d'une trajectoire, d'une durée τ, dans l'espace des conformations

$$\overline{A}(\tau) = \frac{1}{\tau} \int_0^\tau A(\mathscr{C}(t))dt$$

où $\mathscr{C}(t)$ représente l'évolution temporelle de la conformation selon une dynamique à l'équilibre. Une hypothèse très générale de la physique statistique, et souvent (mais pas toujours) réalisée, est que la moyenne temporelle sur une réalisation de la dynamique d'un système à l'équilibre coïncide avec la moyenne statistique sur les conformations permises

$$\langle A \rangle = \lim_{\tau \to +\infty} \overline{A}(\tau)$$

c'est l'hypothèse d'ergodicité. L'interprétation de cette hypothèse est qu'au bout d'une durée infinie, un système à l'équilibre finit par explorer toutes les conformations $\mathscr{C}$ permises avec la probabilité $p_{\mathscr{C}}$.

Numériquement, on a plutôt une évolution discrète des conformations $\mathscr{C}_k \mapsto \mathscr{C}_{k+1}$. L'échantillonnage de la valeur moyenne d'une observable A s'écrit dans ce cas

$$\overline{A}_n = \frac{1}{n} \sum_{k=1}^{N} A_{\mathscr{C}_k}$$

où n est le nombre d'échantillons. Ainsi, l'hypothèse d'ergodicité pour une évolution temporelle discrète devient

$$\langle A \rangle = \lim_{n \to +\infty} \overline{A}_n$$

5.1.2 Algorithme de Metropolis

Le calcul de la valeur moyenne (5.1) d'une observable A à l'équilibre peut être problématique si le nombre de conformations $\mathscr{C}$ est très grand ou s'il est difficile de les paramétriser. De plus, dans le cas de l'ensemble canonique, il faut calculer la fonction de partition $Z = \sum_{\mathscr{C}} e^{-\beta E_{\mathscr{C}}}$ en amont, pour déterminer les probabilités $p_{\mathscr{C}} = \frac{1}{Z} e^{-\beta E_{\mathscr{C}}}$. Ce qui nécessite, là aussi, de sommer sur toutes les conformations $\mathscr{C}$. L'algorithme de Metropolis remplace la somme explicite (5.1) sur toutes les conformations $\mathscr{C}$ par la somme sur un échantillon de n conformations $\mathscr{C}_k$ générées par un processus aléatoire selon la distribution à l'équilibre $p_{\mathscr{C}}$

$$\langle A \rangle \approx \frac{1}{n} \sum_{k=1}^{n} A_{\mathscr{C}_k}$$

Ce processus aléatoire doit être ergodique, *i.e.* capable d'explorer toutes les conformations $\mathscr{C}$ avec la probabilité $p_{\mathscr{C}}$.

L'algorithme utilise une chaine de Markov comme processus pour sauter aléatoirement de conformations c en conformations c' proches dans l'espace des conformations, avec une probabilité $\pi(c', c)$. Cette probabilité se factorise comme $\pi(c', c) = \pi(c'|c)\pi(c)$, où $\pi(c)$ est la probabilité que le processus visite la conformation c et $\pi(c'|c)$ est le taux de transition de la conformation c vers c'. Ces taux de transition sont illustrés à la Figure 5.1.

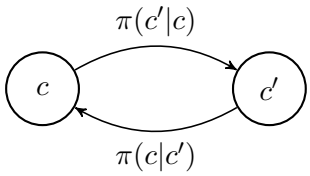

FIGURE 5.1 – Chaine de Markov.

La chaine de Markov explore dynamiquement l'espace des conformations et cette dynamique obéit à l'équation maîtresse

$$\underbrace{\pi(c,k+1)-\pi(c,k)}_{\text{variation temporelle}} = \underbrace{\sum_{c'\neq c}\pi(c|c')\pi(c',k)}_{\text{flux entrant}} - \underbrace{\sum_{c'\neq c}\pi(c'|c)\pi(c,k)}_{\text{flux sortant}}$$

décrivant l'évolution temporelle des probabilités du processus aléatoire, où $\pi(c,k)$ est la probabilité que la conformation c soit visitée à l'itération k.

Pour que le processus reproduise la distribution à l'équilibre $\pi(c,k)=\pi(c)=p_c$, il est nécessaire que π soit stationnaire, *i.e.* indépendant de k :

$$\pi(c,k+1)-\pi(c,k)=\sum_{c'\neq c}\left(\pi(c|c')\pi(c',k)-\pi(c'|c)\pi(c,k)\right)=0$$

Pour cela, il suffit que le processus vérifie le bilan détaillé

$$\pi(c|c')\pi(c',k)-\pi(c'|c)\pi(c,k)=0 \tag{5.2}$$

i.e. que chacun des termes de la somme soit également nul. Ce que l'on peut réécrire de façon plus pragmatique

$$\frac{\pi(c')}{\pi(c)}=\frac{\pi(c'|c)}{\pi(c|c')}$$

avec $\pi(c,k)=\pi(c)$, puisqu'elle est à présent stationnaire.

L'algorithme de Metropolis élémentarise la transition en deux étapes. D'abord, la proposition de la transition de c vers c' avec une probabilité $g(c'|c)$. Puis, l'acceptation de cette transition avec la probabilité $A(c',c)$. Le taux de transition se décompose alors

$$\pi(c'|c)=g(c'|c)A(c',c)$$

Ce qui permet de transcrire l'équation (5.2) du bilan détaillé en taux d'acceptation

$$\frac{A(c',c)}{A(c,c')}=\frac{g(c|c')\pi(c')}{g(c'|c)\pi(c)}$$

Pour ce taux d'acceptation, si la conformation c' est moins probable que la conformation c, alors elle est acceptée avec le taux $A(c',c)=g(c|c')\pi(c')/g(c'|c)\pi(c)$. Inversement, si la conformation c', est plus probable que la c, elle est automatiquement acceptée $A(c,c')=1$. On peut, donc, écrire la probabilité d'acceptation sous la forme synthétique

$$A(c',c)=\max\left(1,\frac{g(c|c')\pi(c')}{g(c'|c)\pi(c)}\right)$$

En pratique, il est utile de s'assurer qu'il est aussi probable que le processus propose la conformation c' venant de c, que la conformation c venant de c', $g(c|c') = g(c'|c)$, simplifiant la probabilité d'acceptation

$$A(c', c) = \max\left(1, \frac{\pi(c')}{\pi(c)}\right) \tag{5.3}$$

Enfin, dans l'ensemble canonique, la distribution à l'équilibre s'écrit $\pi(c) \propto \exp(-\beta E_c)$, ce qui permet de déterminer la probabilité d'acceptation sans faire intervenir la fonction de partition Z

$$A(c', c) = \max\left(1, e^{-\beta(E_{c'} - E_c)}\right)$$

Pour une marche auto-évitante attractive, on rappelle que l'énergie E_c d'une formation c s'écrit $-\beta E_c = \varepsilon m_c$, comme à l'équation (4.9), où m_c est le nombre de contacts au sein de la conformation c et $-\varepsilon$ est l'énergie par contacts, en unités de $k_B T$. Finalement, pour une marche auto-évitante attractive la probabilité d'acceptation (5.3) s'écrit :

$$A(c', c) = \max\left(1, e^{\varepsilon(m_{c'} - m_c)}\right)$$

Implémentation de l'algorithme de Metropolis

```rust
use num_traits::Float;
use rand::distributions::Distribution;
use rand::distributions::Standard;
use rand::Rng;

/// Metropolis algorithm.
pub trait Metropolis {
    type State;
    type Step;
    type Energy: Float;

    /// Proposes a new `State`.
    fn propose(&self, rng: &mut impl Rng) -> Option<Self::State>;

    /// Updates to the next `State` and produces the corresponding `Step`.
    fn accept(&mut self, state: Option<Self::State>) -> Self::Step;

    /// Computes the current `Energy`.
    fn energy(&self) -> Self::Energy;

    /// Computes the new `Energy`.
    fn new_energy(&self, state: &Self::State) -> Self::Energy;

    /// Determines the next `State` by following the Metropolis algorithm.
    fn metropolis(&self, rng: &mut impl Rng) -> Option<Self::State>
    where
        Standard: Distribution<Self::Energy>,
    {
        let state = self.propose(rng)?;
        let delta = self.energy() - self.new_energy(&state);

        Some(state).filter(|_| {
            delta.is_sign_positive() || rng.gen() < delta.exp()
        })
    }

    /// Produces an infinite `Iterator` over `Metropolis::Step`
    /// with the given `Rng`.
    fn metropolis_iter<R: Rng>(self, rng: R) -> MetropolisIter<Self, R>
    where
        Self: Sized,
    {
        MetropolisIter { inner: self, rng }
    }
}
```

Extrait 5.1: Metropolis.

D'un point de vue plus technique, j'ai articulé l'implémentation de l'algorithme de Metropolis en Rust, un langage développé par la fondation Mozilla, autour du concept de `trait` et d'`Iterator`. Un `trait` indique au compilateur Rust la fonctionnalité, *i.e.* un ensemble des fonctions, qu'un type [1] particulier a et peut partager avec d'autres types. Les `trait`s sont similaires aux *interfaces* présentes dans d'autres langages de programmation. L'Extrait 5.1 montre la déclaration du `trait` `Metropolis`. Pour qu'un type puisse implémenter le `trait` `Metropolis`, il doit fournir un implémentation des fonctions suivantes :

- `propose` : génère une conformation à partir de la conformation actuelle
- `accept` : met à jour la conformation si la nouvelle est acceptée, sinon laisse la conformation actuelle inchangée, et monitore l'état actuel via `Self::Step`
- `energy` : calcule l'énergie de la conformation actuelle
- `new_energy` : calcule l'énergie de la nouvelle conformation

Si le type en question fournit une implémentation de ces fonctions, il pourra être transformé, via la méthode `metropolis_iter`, en un itérateur réalisant les itérations de l'algorithme. Cette transformation est permise grâce à une implémentation automatique du `trait Iterator` pour tout type implémentant le `trait` `Metropolis` présent dans l'Extrait 5.2.

```rust
/// Metropolis iterator.
pub struct MetropolisIter<M, R> {
    inner: M,
    rng: R,
}

impl<M, R> Iterator for MetropolisIter<M, R>
where
    M: Metropolis,
    R: Rng,
    Standard: Distribution<M::Energy>,
{
    type Item = M::Step;

    fn next(&mut self) -> Option<Self::Item> {
        let state = self.inner.metropolis(&mut self.rng);
        let step = self.inner.accept(state);

        Some(step)
    }
}
```

Extrait 5.2: MetropolisIter.

1. Un *type* définit la nature des valeurs que peut prendre une donnée, ainsi que les opérateurs qui peuvent lui être appliqués.

5.1.3 Serpent rampant

Jusqu'à maintenant, on a décrit l'algorithme de Metropolis dans les grandes lignes, sans pour autant expliciter la procédure permettant de générer une conformation auto-évitante à partir d'une autre. Il existe plusieurs algorithmes et mon choix s'est porté sur l'algorithme de Wall et Mandel [51] dit du serpent rampant, ou *slithering snake* en anglais, qui reproduit le mouvement de reptation d'une chaine. Les étapes clés de l'algorithme sont illustrées à la Figure 5.2 et énumérées ci-dessous :

1. Partir d'une conformation auto-évitante.
2. Tirer aléatoirement l'orientation (tête-queue).
3. Retirer la queue de la conformation.
4. Proposer une nouvelle tête auto-évitante, sinon rejet de la nouvelle conformation.
5. Mettre à jour le nombre de contacts (décompter les contacts autour de l'ancienne queue et compter les contacts autour de la nouvelle tête).
6. Calculer la différence d'énergie entre les deux conformations.
7. Accepter la nouvelle conformation selon l'algorithme de Metropolis, sinon reprendre l'ancienne conformation.

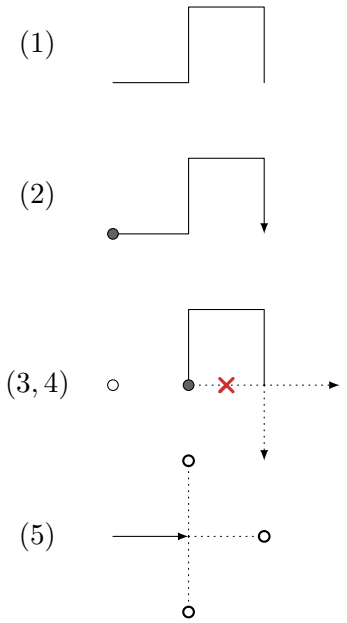

Figure 5.2 – Principales étapes de l'algorithme du serpent rampant.

Techniquement, la vérification de la présence d'une partie de la chaine à une position donnée, aux étapes (4) et (5), se fait efficiemment grâce à une table de hachage, plus précisément à un `HashSet`. À chaque monomère est associé une valeur, appelée clé de hachage, qui ne dépend que de ses coordonnées dans l'espace. Lorsqu'une nouvelle tête est produite, on vérifie que sa clé n'est pas déjà utilisée, sinon cela signifie que la tête est en collision avec une autre partie de la chaine. On utilise, également, cette technique pour déterminer le nombre de contacts autour de chaque extrémité, à partir du nombre de plus proches voisins. Il faut faire attention qu'à chaque extrémité, il y a systématiquement *un* plus proche voisin à ne pas comptabiliser qui est, en fait, le lien avec le reste de la chaine. La fonction de comptage du nombre de contacts s'écrit donc : `contacts = neighbors - 1`.

5.1.4 Données simulées

Grâce à l'algorithme de Metropolis, j'ai échantillonné des conformations à (ε, N) fixés qui m'ont permis d'obtenir les distributions du rayon de giration pour tous les

couples de paramètres de :

— ε allant par pas de 0.05 de 0.00 à 0.10, puis par pas de 0.01 de 0.10 à 0.46,
— N uniformément réparties en 40 valeurs en échelle logarithmique de 10 à 5012,

dont un extrait est donné à la Figure 5.3.

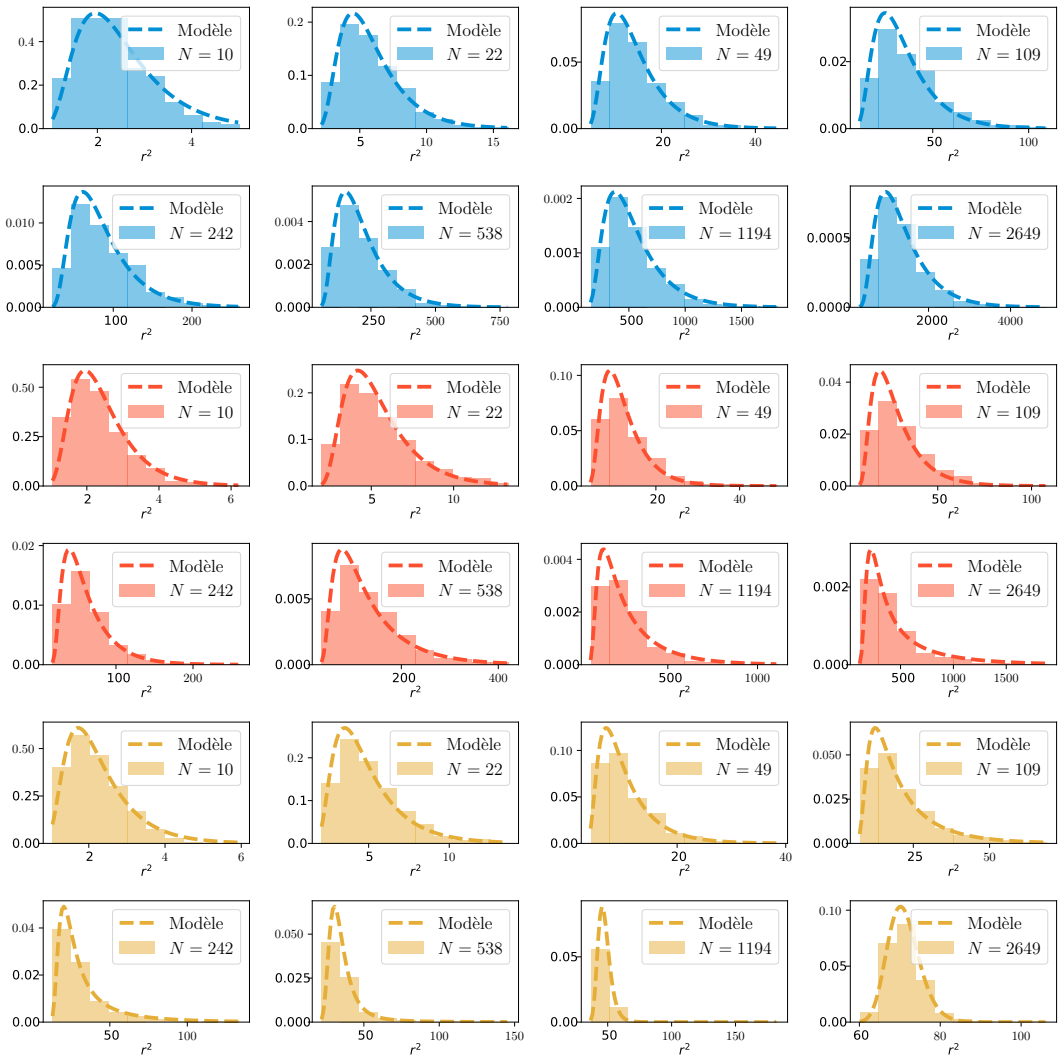

FIGURE 5.3 – Extrait des distributions de rayons de giration à différents couples (ε, N), notamment $\varepsilon = 0.20$ (en bleu), $\varepsilon = 0.30$ (en rouge) et $\varepsilon = 0.40$ (en jaune).

Pour chaque couple (ε, N), 1024 conformations indépendantes ont été échantillonnées tous les temps de décorrélation τ, avec une phase de thermalisation préalable de $64\,\tau$. Le temps de décorrélation a été conjecturé à $\tau = \frac{1}{2}N^2$ et permet, effectivement, d'obtenir des conformations décorrélées sur la plage de ε explorée.

5.1.5 Confinement

Les éléments de programmation ci-décrits permettent d'échantillonner des chaines auto-évitantes dans un espace non borné, mais des contraintes géométriques peuvent, très facilement, y être intégrées. Par exemple, je présente dans l'Extrait 5.4 une extension du `trait` `Metropolis` pour effectuer une marche auto-évitante confinée. Tout type implémentant le `trait` `Metropolis` peut être transformé en un processus confiné `Confined` via la méthode `confined` où l'utilisateur doit fournir une fonction `is_inside`, retournant un booléen, vérifiant que la nouvelle conformation est bien contenue dans le volume de confinement. L'exemple d'utilisation présenté dans l'Extrait 5.3 a permis de simuler une marche auto-évitante entre deux plans infinis pour la publication de Socol et al. [52]. Dans cette publication, l'équipe d'Aurélien Bancaud a étudié la dynamique de molécules individuelles d'ADN transportées dans un flux de fluide viscoélastique avec une force électrophorétique opposée à l'aide d'un microscope à fluorescence. Pour soutenir les observations, j'ai mesuré dans mes simulations la distribution du rayon de giration du petit et du grand axe de l'ellipsoïde d'inertie. Ce travail n'étant pas dans la ligne principale de mon projet de thèse, je ne rentrerais pas dans les détails ici. Pour plus d'information, l'article [52] est joint avec la thèse.

```
1 Isaw::<i32>::new(size, epsilon)
2     .confined(|state| 0 <= state.head.z && state.head.z < h)
3     .metropolis_iter(rng)
4     // ...
```

Extrait 5.3: Exemple d'une marche auto-évitante attractive confinée entre deux plans infinis distant de `h` dans la direction z.

```rust
/// Metropolis algorithm.
pub trait Metropolis {
    // ...

    /// Produces a confined `Metropolis`.
    fn confined<B>(self, is_inside: B) -> Confined<Self, B>
    where
        Self: Sized,
        B: Fn(&Self::State) -> bool,
    {
        Confined {
            inner: self,
            is_inside,
        }
    }
}

/// Confined Metropolis.
pub struct Confined<M, B> {
    inner: M,
    is_inside: B,
}

impl<M, B> Metropolis for Confined<M, B>
where
    M: Metropolis,
    B: Fn(&M::State) -> bool,
{
    type State = M::State;
    type Step = M::Step;
    type Energy = M::Energy;

    fn new_state(&self, rng: &mut impl Rng) -> Option<Self::State> {
        self.inner.propose(rng).filter(&self.is_inside)
    }

    // ...
}
```

Extrait 5.4: Confined Metropolis.

5.2 Inférence des paramètres

Les simulations décrites ci-dessus m'ont permis d'obtenir un échantillonnage de conformations couvrant une large plage de tailles N et d'énergie ε. C'est grâce à la comparaison de ces données avec le modèle théorique que j'ai pu déterminer la dépendance en ε des paramètres du modèle, et ainsi le définir complètement (au moins sur la plage en ε considérée).

Dans cette partie, je vais présenter la méthode employée pour inférer les valeurs $a_i(\varepsilon)$ sur les distributions de rayon de giration produites par l'algorithme du serpent rampant. Elle repose sur l'inférence bayésienne qui permet de trouver les paramètres optimaux d'un modèle étant donné un échantillon de données.

5.2.1 Inférence bayésienne

En supposant qu'un échantillon $\boldsymbol{x} = (x_1, \dots, x_n)$ de taille n dont les x_i suivent une loi de densité de probabilité $p_{\boldsymbol{\theta}}$ de paramètres $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots)$, *a priori* inconnus, l'inférence bayésienne permet de remonter aux paramètres $\boldsymbol{\theta}^*$ optimaux qui décrivent au mieux l'échantillon $\boldsymbol{x}$.

Pour cela, on considère la probabilité jointe $\mathcal{P}(\boldsymbol{x}, \boldsymbol{\theta})$ d'observer un échantillon $\boldsymbol{x}$ pour un modèle $p_{\boldsymbol{\theta}}$ de paramètres $\boldsymbol{\theta}$. Cette probabilité se factorise symétriquement en $\boldsymbol{x}$ et en $\boldsymbol{\theta}$ grâce aux probabilités conditionnelles

$$\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{x})\mathcal{P}(\boldsymbol{x}) = \mathcal{P}(\boldsymbol{x}, \boldsymbol{\theta}) = \mathcal{P}(\boldsymbol{x}|\boldsymbol{\theta})\mathcal{P}(\boldsymbol{\theta}) \tag{5.4}$$

où :

— $\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{x})$ est la probabilité d'avoir les paramètres $\boldsymbol{\theta}$ étant donné l'échantillon $\boldsymbol{x}$.
— $\mathcal{P}(\boldsymbol{x})$ est, formellement, la probabilité d'avoir l'échantillon $\boldsymbol{x}$.
— $\mathcal{P}(\boldsymbol{x}|\boldsymbol{\theta})$ est la probabilité d'avoir l'échantillon $\boldsymbol{x}$ étant donnés les paramètres $\boldsymbol{\theta}$.
— $\mathcal{P}(\boldsymbol{\theta})$ est la probabilité, *a priori*, des paramètres $\boldsymbol{\theta}$.

En réarrangeant les termes de l'équation (5.4), on obtient le théorème de Bayes

$$\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{\mathcal{P}(\boldsymbol{x}|\boldsymbol{\theta})\mathcal{P}(\boldsymbol{\theta})}{\mathcal{P}(\boldsymbol{x})} \tag{5.5}$$

où $\mathcal{P}(\boldsymbol{x}|\boldsymbol{\theta})$ est la vraisemblance des données $\boldsymbol{x}$ qui s'exprime en fonction de la densité de probabilité $p_{\boldsymbol{\theta}}$ comme

$$\mathcal{P}(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p_{\boldsymbol{\theta}}(x_i)\, dx \tag{5.6}$$

On souhaite obtenir la distribution des paramètres $\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{x})$. Elle est définie pour un échantillon $\boldsymbol{x}$ donné et s'écrit, grâce au théorème de Bayes (5.5), comme

$$\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{x}) \propto \mathcal{P}(\boldsymbol{x}|\boldsymbol{\theta})\mathcal{P}(\boldsymbol{\theta})$$

à une constante multiplicative $\mathcal{P}(\boldsymbol{x})$ près. On peut omettre cette constante, en toute généralité, sans que ça ait d'incidence sur le problème d'optimisation, puisque indépendante de $\boldsymbol{\theta}$. Dans le cas particulier où l'on n'a aucune information *a priori* sur la distribution des paramètres, $\mathcal{P}(\boldsymbol{\theta})$ est uniforme, *i.e.* est une constante que l'on pourra

omettre. On parle alors d'inférence bayésienne « naïve », où $\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{x})$ est assimilable à la vraisemblance $\mathcal{P}(\boldsymbol{x}|\boldsymbol{\theta})$.

En pratique, la vraisemblance (5.6) est un produit de termes, souvent plus petits que un, faisant tendre le produit vers zéro. Numériquement pour palier à ce problème, on considère la log-vraisemblance définie comme le logarithme de la vraisemblance

$$\mathscr{L}(\boldsymbol{\theta}) = \ln \mathcal{P}(\boldsymbol{\theta}|\boldsymbol{x}) = \ln \mathcal{P}(\boldsymbol{x}|\boldsymbol{\theta}) + \ln \mathcal{P}(\boldsymbol{\theta})$$

qui a la vertu de transformer le problème d'optimisation d'un produit en optimisation d'une somme. Le terme $\ln \mathcal{P}(\boldsymbol{\theta})$ permet de borner l'espace des paramètres explorable.

Le problème d'optimisation est le suivant : on cherche les paramètres $\boldsymbol{\theta}^*$ qui maximisent la log-vraisemblance $\mathscr{L}$:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathscr{L}(\boldsymbol{\theta})$$

ou, de façon équivalente, qui minimisent la log-vraisemblance négative $-\mathscr{L}$.

Du point de vue du physicien, il est intéressant de noter que ce problème de minimisation de $-\mathscr{L}$ peut se ramener à un problème de minimisation de l'énergie libre, dans le cas où les probabilités s'écrivent $p_{\boldsymbol{\theta}}(t) = Z^{-1}(\boldsymbol{\theta}) \exp(-\beta F_{\boldsymbol{\theta}}(t))$ et t est l'observable. Dans ce cas, la log-vraisemblance négative moyenne s'écrit

$$-\frac{1}{n}\mathscr{L}(\boldsymbol{\theta}) = \overline{\beta F_{\boldsymbol{\theta}}(t)} - \beta F_{\boldsymbol{\theta}}$$

avec $\overline{\beta F_{\boldsymbol{\theta}}(t)} = n^{-1} \sum_{i=1}^{n} \beta F_{\boldsymbol{\theta}}(t_i)$ la moyenne sur l'échantillon (de taille n) de l'énergie libre exprimée en t et $\ln Z(\boldsymbol{\theta}) = -\beta F_{\boldsymbol{\theta}}$ l'énergie libre canonique obtenue à l'équation (4.8). En l'absence de la distribution *a priori* des paramètres, il parait clairement que minimiser $-\mathscr{L}$ équivaut, dans ce cas, à minimiser la différence entre la moyenne sur l'échantillon de l'énergie libre avec l'énergie libre canonique.

5.2.2 Échantillonnage de l'espace des paramètres

Pour extraire le plus d'information possible de l'échantillon $\boldsymbol{x}$ à notre disposition, on souhaite non seulement trouver les paramètres optimaux $\boldsymbol{\theta}^*$, mais aussi inférer la distribution de probabilité des paramètres au voisinage de $\boldsymbol{\theta}^*$. Pour cela, on utilise un algorithme de type Metropolis avec une propriété d'invariance affine [53], ce qui augmente significativement les performances de l'algorithme. Par ailleurs, il existe une implémentation [54], en Python, parallélisable que j'ai adapté, pour mes besoins, en Rust.

5.2.3 Écriture de la log-vraisemblance

Pour l'analyse des données de simulation, l'observable est le rayon de giration r^2 et le modèle est $p_N(t|\boldsymbol{\theta}) = Z^{-1}(\boldsymbol{\theta}) \exp(-\beta F_N(t|\boldsymbol{\theta}))$ de l'équation (4.13), où $\boldsymbol{\theta} = (a_1, a_2, a_3)$.

La première chose à effectuer est de traduire le modèle de la variable t à la variable r^2. Pour cela, on utilise la méthode de la transformation inverse

$$p_N(r^2|\boldsymbol{\theta})\, dr^2 = p(t|\boldsymbol{\theta})\, dt$$
$$p_N(r^2|\boldsymbol{\theta}) = p(t|\boldsymbol{\theta}) \left| \frac{dt}{dr^2} \right|$$

où $\left|\frac{dt}{dr^2}\right| = \frac{d}{2} g N^g (r^2)^{-dg/2-1}$ désigne le jacobien du changement de variable $t \to r^2$, avec $t = (N/r^d)^g$, $g = (\nu d - 1)^{-1}$ et $d = 3$. On aboutit à l'expression de l'énergie libre exprimée en r^2

$$\beta F_N(r^2|\boldsymbol{\theta}) = \beta F_N(t(r^2)|\boldsymbol{\theta}) + \left(\frac{3}{2}g + 1\right) \ln r^2 + \text{constante} \tag{5.7}$$

L'énergie libre est définie à une constante près qui n'a pas d'influence sur la suite.

Puisqu'on souhaite inférer $a_i(\varepsilon)$ sur l'ensemble des tailles N simulées à ε fixé, on introduit la vraisemblance des données r^2 simulées pour un N donné :

$$-\frac{1}{n}\mathscr{L}_N(\boldsymbol{\theta}) = \overline{\beta F_N(r^2|\boldsymbol{\theta})} + \ln Z_N(\boldsymbol{\theta})$$

où $Z_N(\boldsymbol{\theta})$ est, ici, la fonction de partition calculée avec l'énergie libre (5.7) exprimée en r^2. Puis, on introduit la vraisemblance de l'ensemble des données à ε fixé

$$\mathscr{L}(\boldsymbol{\theta}) = \sum_N \mathscr{L}_N(\boldsymbol{\theta}) + \ln \mathcal{P}(\boldsymbol{\theta}) \tag{5.8}$$

qui sera la fonction à optimiser.

5.2.4 Détail du calcul numérique

Numériquement, la fonction $\exp(-\beta F_N(r^2|\boldsymbol{\theta}))$ peut prendre des valeurs plus grandes que ce qu'est capable de conserver en mémoire un ordinateur. C'est pourquoi, on introduit l'énergie libre réduite $\beta f(r^2|\boldsymbol{\theta})$ définie comme

$$-\beta f(r^2|\boldsymbol{\theta}) = \max_{r^2} \beta F_N(r^2|\boldsymbol{\theta}) - \beta F_N(r^2|\boldsymbol{\theta})$$

telle que $-\beta f(r^2|\boldsymbol{\theta}) \leqslant 0 \implies e^{-\beta f(r^2|\boldsymbol{\theta})} \leqslant 1$ et vérifiant

$$p_N(r^2|\boldsymbol{\theta}) = \frac{e^{-\beta F_N(r^2|\boldsymbol{\theta})}}{Z_N(\boldsymbol{\theta})} = \frac{e^{-\beta f(r^2|\boldsymbol{\theta})}}{z_N(\boldsymbol{\theta})}$$

avec la fonction de partition réduite

$$z_N(\boldsymbol{\theta}) = \int_0^{+\infty} e^{-\beta f_N(r^2|\boldsymbol{\theta})}\, dr^2$$

qui se calcule numériquement et correspond au cas Exp de la Table B.3. En effet, l'intégrale est impropre en $+\infty$ et l'intégrande a des propriétés de décroissance exponentielle à la fois en 0 et en $+\infty$, ce qui nécessite une technique d'intégration numérique adaptée que j'ai étoffé en me basant sur des publications non exhaustives. Elle est détaillée dans l'Annexe B.

Finalement pour l'optimisation numérique de la fonction (5.8), on utilise l'énergie libre et la fonction de partition réduites dans l'expression de la log-vraisemblance négative partielle

$$-\frac{1}{n}\mathscr{L}_N(\boldsymbol{\theta}) = \overline{\beta f_N(r^2|\boldsymbol{\theta})} + \ln z_N(\boldsymbol{\theta})$$

conduisant au même résultat, sans les problèmes de dépassement de capacité de la mémoire de l'ordinateur.

Remarque. *La maximisation de l'énergie libre est réalisée avec l'algorithme de* Nelder *et* Mead *[55] que j'ai également dû réimplémenter en Rust.*

5.3 Comparaison aux simulations et correction du modèle

Dans cette partie, on infère les paramètres de notre modèle conformément à la section 5.2, puis on vérifie que ses prédictions reproduisent les données des simulations : dans un premier temps, médianes et moyennes des distributions échantillonnées, puis dans un second temps, les distributions elles-mêmes. Comme on le verra, cet ajustement nous conduira également à introduire un terme d'énergie supplémentaire, rendant compte des effets de tension de surface dans les configurations globulaires, ainsi qu'à proposer une forme analytique inédite pour ce terme.

5.3.1 Première confrontation

On ne présente, ici, qu'une confrontation limitée des prédictions du modèle aux données simulées. À la Figure 5.4, la simple comparaison avec les médianes prédites en fonction de la taille N, en échelle logarithmique, pour $\varepsilon = 0.35$, suffit à se rendre compte que le modèle est incorrect, ou du moins incomplet. En effet, on retrouve approximativement les comportements asymptotiques attendus à la fois pour les médianes et pour les moyennes, même si la prédiction est en dessous pour la partie gauche de la courbe et au dessus pour la partie droite.

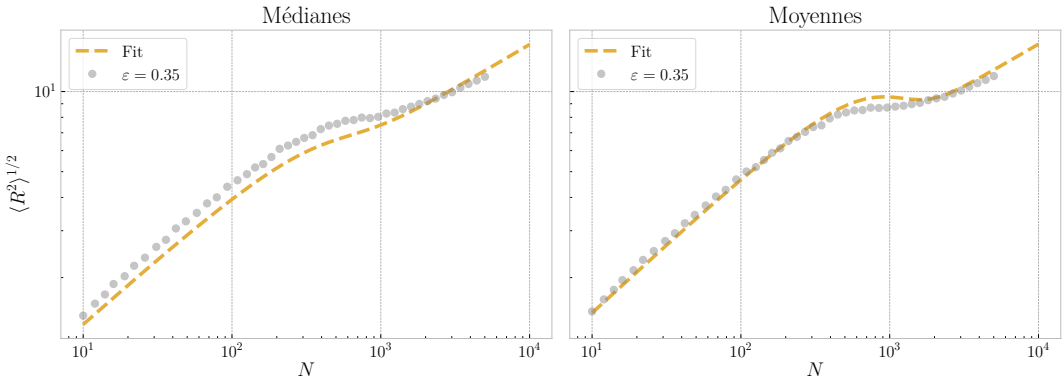

FIGURE 5.4 – Confrontation des prédictions du modèle aux distributions échantillonnées, en médianes et en moyennes.

5.3.2 Terme manquant de surface

Jusqu'à maintenant, on a omis la tension de surface qui est non négligeable en taille finie [56]. La tension de surface apparait dans les conformations globulaires et vient du fait que les monomères ont plus d'affinité entre eux qu'avec le solvant. Les monomères minimisent leur énergie lorsqu'ils sont entourés d'autres monomères. Il en résulte une pénalité énergétique pour les monomères se trouvant à l'interface avec le solvant, qui est proportionnelle à la surface de la conformation. Par ailleurs, RISSANOU et al. [56] ont montré sur des simulations que la surface des conformations globulaires avait une épaisseur valant au minimum la moitié du rayon de la conformation. Ce qui est non négligeable et signifie que le développement en t proposé à l'équation (4.12) de la fonction génératrice des cumulants est incomplet.

J'ai essayé plusieurs formes possibles pour l'énergie de surface, d'abord j'ai puisé dans la littérature, puis j'en ai construite une sur la base d'arguments d'échelle.

Dans la littérature, l'énergie de surface est, en partie, décrite chez Lifshitz et al. [57], puis complétée par Khokhlov [58] et reprise dans l'ouvrage *Statistical Physics of Macromolecules* [59]. Khokhlov l'écrit

$$\beta F_\sigma \sim R^{d-1}\tau^2$$

où R est le rayon de giration, d la dimension et $\tau = (T - T_\theta)/T_\theta$ est l'écart relatif de la température T à la température T_θ de la transition de phase *coil-globule*. On peut réécrire R^{d-1} en fonction de la taille N et de la densité ρ comme

$$R^{d-1} = N^{(d-1)/d}\rho^{-(d-1)/d}$$

Or dans la théorie de Khokhlov [58] $\tau \sim \rho$, ce qui permet d'écrire le terme de surface comme

$$\beta F_\sigma \sim N^{(d-1)/d}\rho^{(d+1)/d} \tag{5.9}$$

Les arguments pour aboutir à cette expression sont un peu obscurs et reposent sur des lois d'échelle qui sont vraies en dimension critique supérieure ($d \geqslant 4$) où notamment la variable d'échelle n'est plus la densité renormalisée $t = \rho^{1/(\nu d-1)}$, mais simplement la densité ρ.

En incorporant l'énergie de surface avec un nouveau paramètre d'ajustement, on constate à la Figure 5.5 que le terme proposé par Khokhlov (courbe rouge) ne permet pas, non plus, de reproduire les données de nos simulations.

C'est pourquoi nous proposons une énergie de surface dont la forme est déduite, une nouvelle fois, par des arguments d'échelle. Comme l'a montré Imbert et al. [42], $\hat{t} = N^{1/2}t$ est la variable d'échelle qui décrit le *crossover*, *i.e.* le plateau, de la transition *coil-globule* en taille finie. Ce qui signifie qu'au *crossover*, (*i*) l'énergie libre est une fonction de la variable d'échelle $\hat{t}$. De plus, Owczarek et al. [43] ont montré que (*ii*) l'énergie de surface dépend de la taille comme $N^{(d-1)/d}$. L'énergie de surface doit donc satisfaire ces deux conditions, on écrit (*i*) $\hat{t}^\sigma = N^{\sigma/2}t^\sigma$, où σ est l'exposant que l'on cherche, et satisfait (*ii*) $\sigma/2 = (d-1)/d$, soit $\sigma = 2(d-1)/d$. L'énergie de surface est donc de la forme

$$\beta F_\sigma \sim (Nt^2)^{(d-1)/d} \tag{5.10}$$

Ainsi en tenant compte de l'énergie de surface, le développement (4.12) en t de la fonction génératrice des cumulants s'écrit

$$K(t,\varepsilon) = b_1(\varepsilon)\cdot Nt + b_2(\varepsilon)\cdot Nt^2 + b_3(\varepsilon)(Nt)^{-q} + b_4(\varepsilon)\cdot(Nt^2)^{2/3}$$

Par conséquent, l'énergie libre d'une marche auto-évitante attractive devient

$$\beta F_N(t|\varepsilon) = a_1(\varepsilon)\cdot Nt + a_2(\varepsilon)\cdot Nt^2 + a_3(\varepsilon)\cdot(Nt)^{-q} + a_4(\varepsilon)\cdot(Nt^2)^{2/3} + c\ln Nt \tag{5.11}$$

avec l'introduction d'une nouvelle fonction $a_4 = -b_4$ à inférer sur les simulations.

En confrontant les prédictions du modèle corrigé aux données de nos simulations, on constate qu'il reproduit correctement non seulement médianes et moyennes à la Figure 5.5, mais aussi toutes les distributions, dont un extrait est donné à la Figure 5.3.

Remarque. *Il est intéressant de noter quatre points.*
Premièrement, la fonction a_4 n'apparait pas pour les marches auto-évitantes ($A_4 = 0$) et s'écrit donc $a_4(\varepsilon) = -b_4(\varepsilon)$.
Deuxièmement, l'énergie de surface n'affecte que les conformations globulaires ($t \gg 1$) du fait qu'elle s'exprime en puissance de la variable Nt^2.
Troisièmement, l'expression de l'énergie de surface de Khokhlov *à l'équation (5.9) est similaire à la notre à l'équation (5.10) pour $d = 3$. À ceci près qu'elle n'a pas la même variable d'échelle.*
Quatrièmement, il serait opportun de tester la robustesse des prédictions du modèle pour d'autres dimensions $d \neq 3$ et conforter l'expression de notre énergie de surface qui prédit une autre dépendance en d.

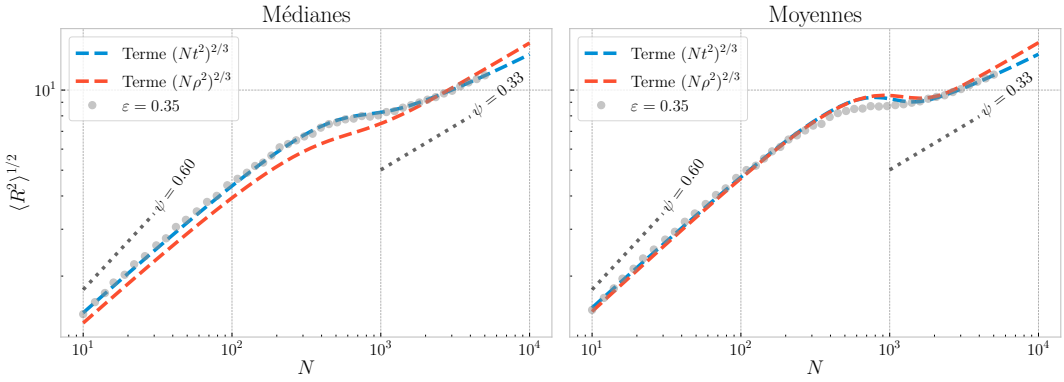

Figure 5.5 – Comparaisons de deux termes d'énergie de surface : le notre $(Nt^2)^{2/3}$ (courbe bleue) et celui de Khokhlov $(N\rho^2)^{2/3}$ (courbe rouge), en médianes et en moyennes.

5.3.3 Interpolation des fonctions a_i

Maintenant que nous avons une expression correcte de l'énergie libre, nous pouvons inférer les valeurs des fonctions a_i pour chaque valeur ε et en faire l'interpolation.

On réitère l'inférence des paramètres, conformément à la section 5.2, pour chaque valeur de ε simulée en utilisant l'énergie libre de l'équation (5.11). À la Figure 5.6, on reporte sur des graphiques les valeurs $a_i(\varepsilon)$ inférées en fonction de ε. Première constatation, les valeurs $a_i(\varepsilon)$ semblent décrire une fonction continue de ε. Pour les quatre graphiques, on observe un changement notable de comportement au voisinage d'une valeur critique $\varepsilon_c \approx 0.29$. Les fonctions a_2 et a_4 présentent également un changement de comportement vers une seconde valeur critique $\varepsilon'_c \approx 0.17$.

On a appliqué une régression polynomiale de degré 3 de la forme

$$a_i(\varepsilon) = c_3\varepsilon^3 + c_2\varepsilon^2 + c_1\varepsilon + c_0$$

de part et d'autre de chaque valeur critique ε_c observée, à l'exception de a_4 où seule la valeur critique $\varepsilon_c \approx 0.29$ a été prise en compte, le changement de comportement autour

$\varepsilon_c \approx 0.17$ étant assez régulier pour qu'il puisse être capturé par un seul polynôme d'ordre 3.

Cette procédure est suffisante pour reproduire les courbes $a_i(\varepsilon)$ avec une bonne approximation. Les coefficients des différents polynômes sont résumés dans la Table 5.1 et sont valides pour l'intervalle de ε allant de 0.00 à 0.46.

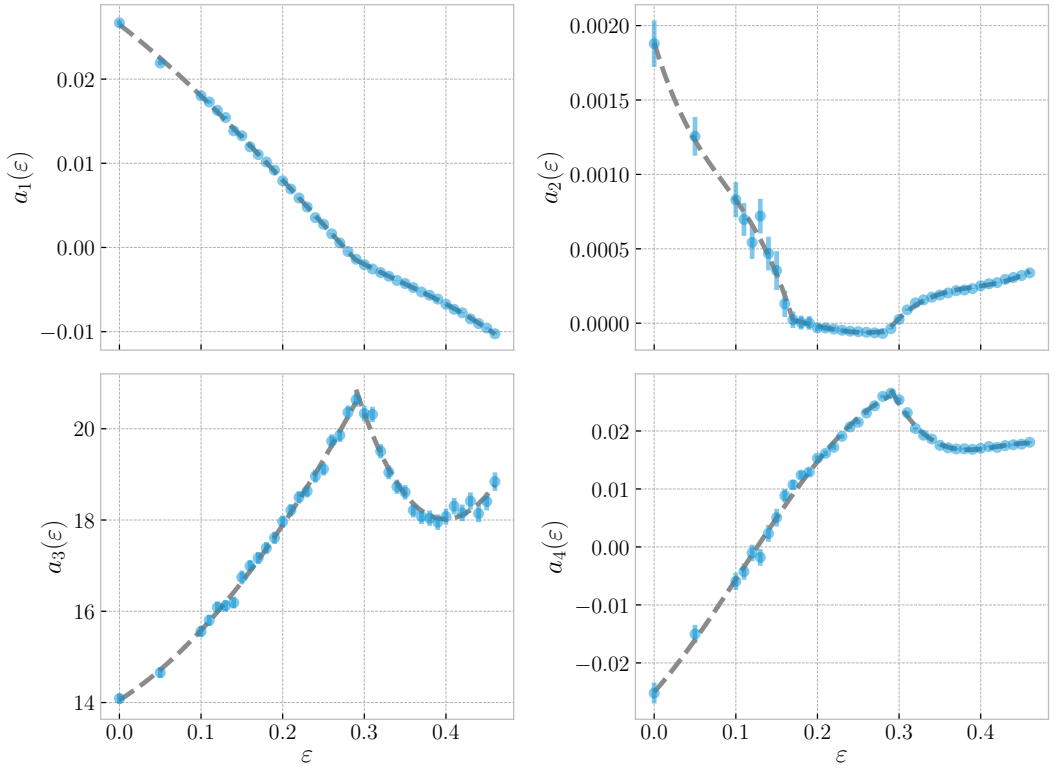

FIGURE 5.6 – Inférence des valeurs $a_i(\varepsilon)$.

a_i	ε_c	Partie	c_0	c_1	c_2	c_3
a_1	0.2890	Gauche	0.02648	-0.07413	-0.1218	0.1520
		Droite	0.03470	-0.2465	0.5839	-0.5668
a_2	0.1699 0.2862	Gauche	0.001885	-0.01808	0.1227	-0.4757
		Centrale	-0.00006015	0.004102	-0.03171	0.06115
		Droite	-0.01034	0.07912	-0.1991	0.1687
a_3	0.2930	Gauche	14.06	11.22	41.45	-8.246
		Droite	66.67	-280.6	489.7	-230.8
a_4	0.2933	Gauche	-0.02502	0.1581	0.5143	-1.559
		Droite	0.4306	-2.988	7.142	-5.647

TABLE 5.1 – Coefficients des régressions polynomiales.

5.3.4 Corrélations des paramètres du modèle

On s'intéresse, ici, à l'étude plus fine de l'inférence des paramètres du modèle pour deux régimes différents : $\varepsilon < \varepsilon_\theta$ (*coil*) et $\varepsilon > \varepsilon_\theta$ (*globule*).

Les figures 5.7 et 5.8 ont été réalisées grâce à « corner.py : Scatterplot matrices in Python » [60]. Elles représentent les corrélations et les lois marginales des paramètres pour $\varepsilon = 0.00$ (*coil*) et $\varepsilon = 0.40$ (*globule*), respectivement. Ces deux valeurs de ε ont été choisies à titre d'exemple représentatif pour les deux régimes.

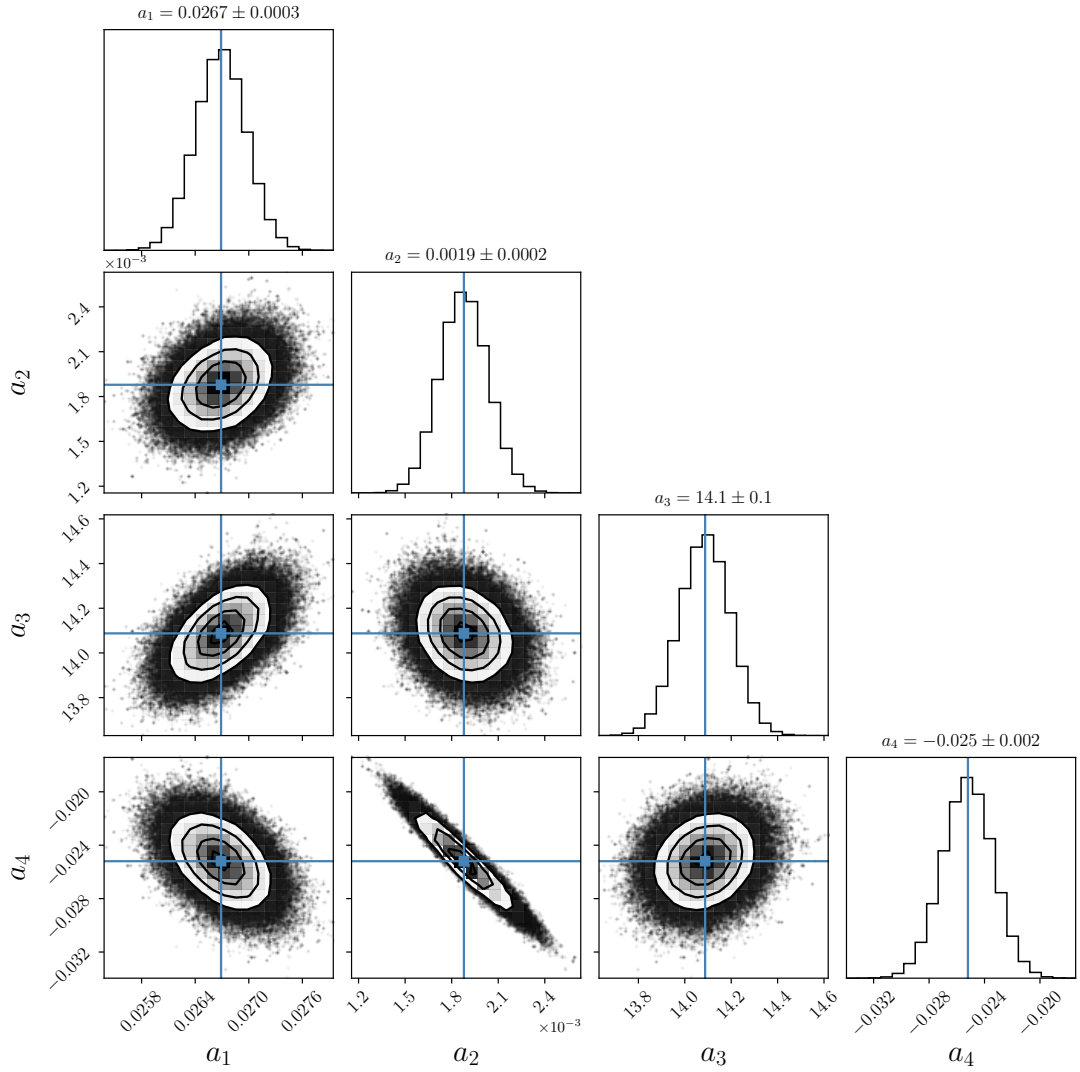

FIGURE 5.7 – Vue en coupe de la distribution des paramètres pour $\varepsilon = 0.00$. Les projections 2D montrent les corrélations entre deux paramètres, tandis que les projections 1D représentent les lois marginales des paramètres. Chaque point représente un point visité dans l'espace des paramètres, où les régions de plus forte densité en points sont les régions des paramètres les plus probables. La valeur retenue pour chaque paramètre est sa valeur médiane.

À la Figure 5.7 pour $\varepsilon = 0.00$, on peut observer une corrélation entre les paramètres a_2 et a_4. Les paramètres sont anti-corrélés et se compensent numériquement en ordre de grandeur. Cette corrélation vient du fait que l'on se trouve dans le régime purement *coil* du modèle où les paramètres sont superflus ($a_2 = a_4 = 0$). L'optimisation satisfait ainsi la contrainte numérique : $a_2 \cdot Nt^2 + a_4 \cdot (Nt^2)^{2/3} \approx 0$.

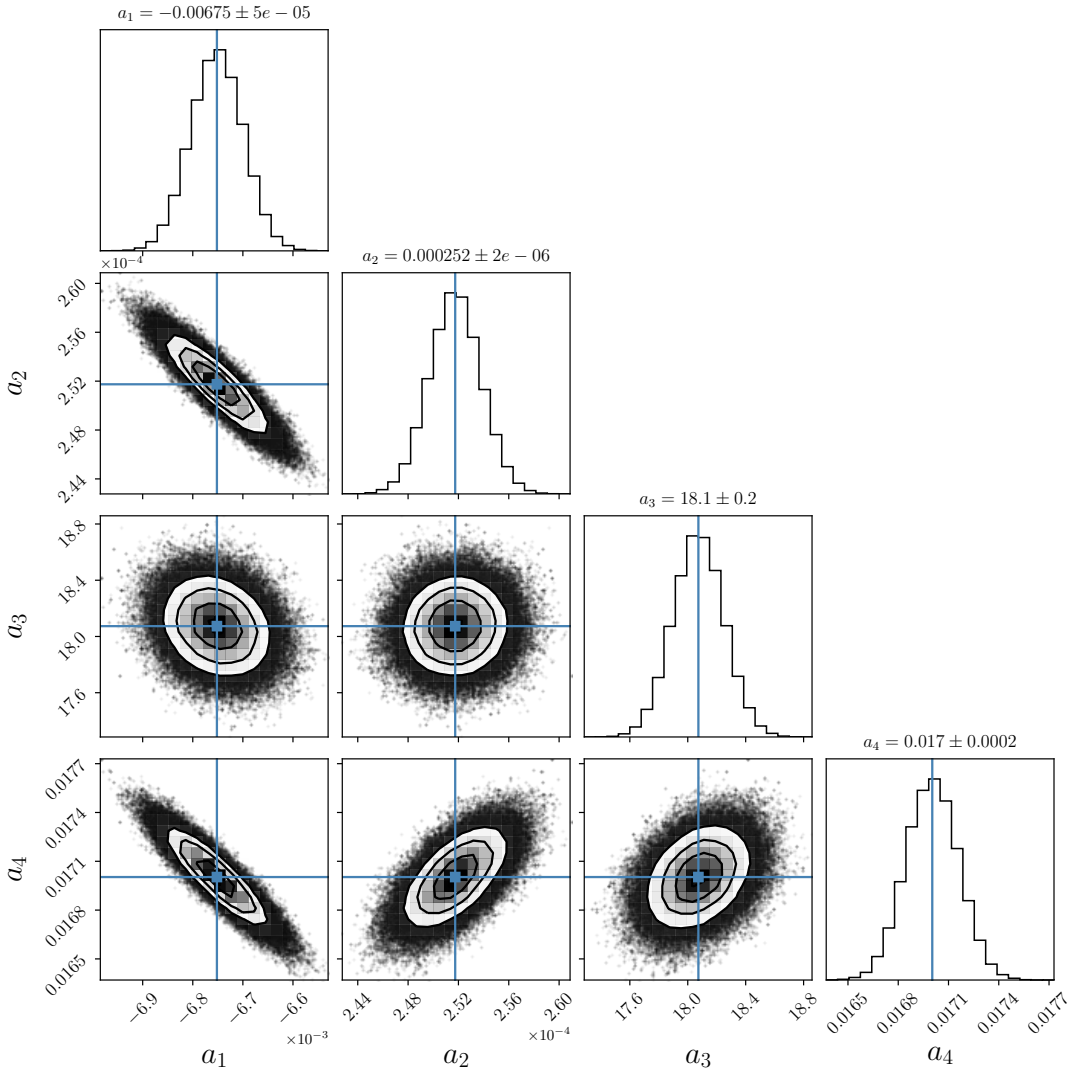

FIGURE 5.8 – Vue en coupe de la distribution des paramètres pour $\varepsilon = 0.40$. Les projections 2D montrent les corrélations entre deux paramètres, tandis que les projections 1D représentent les lois marginales des paramètres. Chaque point représente un point visité dans l'espace des paramètres, où les régions de plus forte densité en points sont les régions des paramètres les plus probables. La valeur retenue pour chaque paramètre est sa valeur médiane.

À la Figure 5.8 pour $\varepsilon = 0.40$, on peut observer une corrélation entre les couples de paramètres (a_1, a_2) et (a_2, a_4). Les deux couples sont anti-corrélés, cela vient du fait

que les paramètres a_2 et a_4 jouent un rôle analogue vis-à-vis de a_1. Si on s'intéresse plus particulièrement au couple (a_1, a_2), la corrélation vient du fait que, dans le régime *globule*, le couple pilote la position du maximum $t_{\max}$ de la distribution (5.11) qui, en première approximation, vérifie l'équation

$$\frac{\partial}{\partial t}\left(a_1(\varepsilon)\cdot N t_{\max} + a_2(\varepsilon)\cdot N t_{\max}^2\right) = 0$$

avec $a_1(\varepsilon) < 0$, d'où

$$t_{\max} = -\frac{1}{2}\frac{a_1(\varepsilon)}{a_2(\varepsilon)}$$

Lors de l'inférence pour maximiser la (log-)vraisemblance du modèle sur les données, il faut que le maximum $t_{\max}$ du modèle coïncide avec le maximum de la distribution échantillonnée. C'est de cette relation que découle la corrélation entre les deux paramètres.

5.4 Conclusion

Dans ce chapitre, nous venons de voir comment mettre en place les simulations nécessaires à la génération de distributions de rayon de giration pour différentes valeurs du paramètre d'attraction ε. Nous avons également vu comment inférer les paramètres $a_i(\varepsilon)$ du modèle à partir des distributions simulées.

L'inférence a permis de mettre en exergue que le modèle, en l'état à la fin du chapitre 4, ne reproduisait pas les données des simulations, notamment les médianes des distributions simulées. De ce fait, il a été nécessaire d'étendre le modèle pour qu'il puisse rendre compte des effets de tension de surface qui sont non négligeables en taille finie. Ils permettent de reproduire le *crossover*, *i.e.* le plateau observé dans les courbes, en échelle logarithmique, du rayon de giration $\langle R\rangle^{1/2}$ en fonction de la taille N.

Pour le modèle, il est important de remarquer que l'interpolation de a_1 s'annule pour $\varepsilon \approx 0.275$, marquant la transition *coil-globule*. Cette valeur est proche de l'énergie critique $\varepsilon_\Theta \approx 0.270$ obtenue pour un réseau cubique [44, 45].

L'expression de l'énergie libre, ici obtenue, permettra d'ajuster les distributions expérimentales du rayon de giration, au chapitre prochain. Elle constitue également un résultat dans le cadre de la théorie des polymères qui fera l'objet d'une publication, actuellement en cours de rédaction.

Chapitre 6

Analyse des expériences

Dans ce chapitre, on fait la synthèse de tous les résultats précédemment obtenus en vue d'analyser les données de Boettiger et al. [1], présentées dans le chapitre 1. On rappelle que les domaines épigénétiques observés sont tétraploïdes, *i.e.* le génome est constitué de quadruplets de chromosomes homologues. Chez la drosophile, il s'avère que les chromosomes homologues sont régulièrement appariés et forment ce que l'on appelle un faisceau de chromosomes.
Dans le chapitre 2, j'ai explicité la formule du rayon de giration (2.14) d'un faisceau de polymères et montré que sa distribution s'écrit comme à l'équation (2.21). Le faisceau de chromosomes se comporte comme un faisceau de polymères, *i.e.* comme un unique polymère décrit par la distribution p_N, avec une certaine épaisseur décrite par la distribution f_λ.
Dans les chapitres 3, 4 et 5, j'ai obtenu la distribution p_N par étape.
Dans le chapitre 3, j'ai obtenu l'expression de l'énergie libre (3.10) d'une marche auto-évitante.
Dans le chapitre 4, j'ai obtenu l'expression de l'énergie libre (4.13) d'une marche auto-évitante attractive.
Dans le chapitre 5, j'ai mis en place la méthode d'analyse des données qui a permis de compléter l'expression de l'énergie libre (5.11) avec un terme d'énergie de surface, ainsi que l'expression des fonctions a_i qui encodent toute la phénoménologie de la transition *coil-globule*.

6.1 Expérience

On rappelle que Boettiger et al. [1] ont imagé 46 domaines épigénétiques dans des cellules de la drosophile Kc167. Ils ont classé les domaines en trois états épigénétiques : transcriptionnellement actif, inactif et réprimé par Polycomb, sur la base de l'enrichissement des modifications sur les histones et des protéines régulatrices provenant des données ChIP-seq (immunoprécipitation chromatinienne suivie de séquençage) et DamID (identification par ADN-adénine-méthyltransférase). Les domaines actifs de chromatine ont été sélectionnés sur la base de l'enrichissement des modifications des histones H3K4me2 ou H3K79me3. Les domaines réprimés ont été sélectionnés en fonction de l'enrichissement des protéines H3K27me3 ou du groupe Polycomb. Les

domaines inactifs ont été sélectionnés en fonction de la prédominance d'histones non modifiées et de la déplétion des protéines groupe Polycomb et des activateurs transcriptionnels. Les longueurs des domaines sélectionnés vont de ~ 10 à 500 kb pour les trois états épigénétiques.

L'analyse approfondie a été menée dans le papier « Polymer coil–globule phase transition is a universal folding principle of Drosophila epigenetic domains » [61], joint avec la thèse.

6.1.1 Simple anaylse des données expérimentales

Nous avons vu au chapitre 3 que les classes de conformations d'un polymère : *stretch*, *coil* et *globule* se discriminent par la valeur de l'exposant ψ qui leur est associé, *cf.* Table 3.1. La question qui est légitime de se poser est de savoir quelles conformations sont adoptées par les domaines épigénétiques. Pour cela, on se ramène à la loi de puissance du rayon de giration (3.8), en prenant soin de remarquer que le rayon de giration et le nombre de monomères sont tous deux des nombres sans dimension.

Dans les données expérimentales, on n'a pas directement accès au nombre de segments de Kuhn N, mais à la longueur de contour L en kilobases (kb). Segments de Kuhn et longueur de contour sont reliés par l'équation (3.5) grâce à la longueur de Kuhn, également exprimée en kb, que l'on va noter K_{kb}.

Le rayon de giration est, quant à lui, exprimé en nanomètres (nm). Il existe donc une constante d'adimensionnement K_{nm}, exprimée en nm, telle que

$$\frac{\langle R^2\rangle^{1/2}}{K_{\mathrm{nm}}} \sim \left(\frac{L}{K_{\mathrm{kb}}}\right)^{\psi}.$$

C'est la longueur de Kuhn en nm. La longueur d'un segment de Kuhn peut être exprimée à la fois en paires de bases qui le constituent ou en nanomètres, sa taille physique. On peut passer de l'un à l'autre en connaissant la compaction du segment qui traduit le nombre de paires de bases contenues par nanomètres et est une caractéristique de l'architecture à petite et moyenne échelles de l'assemblage ADN-protéines qui constitue la chromatine. La compaction linéique de la chaine, en $\mathrm{kb\,nm^{-1}}$, est donnée par le rapport $K_{\mathrm{kb}}/K_{\mathrm{nm}}$.

Pour les grandeurs dimensionnées accessibles expérimentalement, la loi de puissance s'écrit

$$\langle R^2\rangle^{1/2} \sim L^{\psi} \tag{6.1}$$

où $\langle R^2\rangle^{1/2}$ est en nm et L en kb. On rappelle que Boettiger et al. [1] ont mesuré, dans plusieurs cellules d'une même lignée (Kc167), le rayon de giration de domaines épigénétiques de différentes longueurs, appartenant à trois états épigénétiques :

1. L'état *rouge*, dit actif, couvrant les régions exprimées.
2. L'état *noir*, dit inactif, caractérisés en particulier par la présence d'histones HP1.
3. L'état *bleu*, dit réprimé, caractérisés par la présence de protéines du groupe Polycomb (PcG).

Pour chaque état épigénétique et pour chaque longueur L, on a accès à un échantillon de R^2. Du fait de la nature complexe des expériences, elles ne sont pas dénuées de données aberrantes. C'est pourquoi, la médiane est un meilleur estimateur statistique de la taille à partir des échantillons mesurés, plutôt que la moyenne.

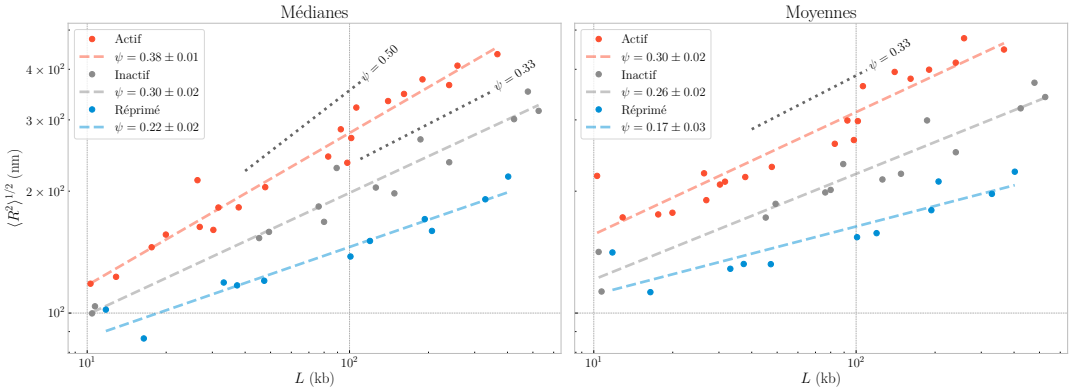

FIGURE 6.1 – Confrontation des modèles simples aux données expérimentales de BOETTIGER et al. [1].

Dans la figure 6.1, j'ai reporté en échelle logarithmique à la fois médianes et moyennes pour les ajuster avec la loi d'échelle (6.1) et les comparer aux prédictions des modèles classiques de la table 3.1. En échelle logarithmique, les lois de puissance se transforment en droites dont la pente donne la valeur de l'exposant ψ.

Qualitativement, les données suivent un comportement en loi de puissance puisqu'elles s'alignent selon des droites en échelle logarithmique. C'est le comportement attendu d'après la physique des polymères. Les exposants sont ordonnés des réprimés vers les actifs, en passant par les inactifs. Ce qui signifie que la taille des domaines est une fonction de l'état épigénétique. Plus celui-ci induit une activité transcriptionnelle, plus la taille du domaine gonfle, *i.e.* la valeur de ψ est grande.

État :	Actif	Inactif	Réprimé
Médianes	0.38 ± 0.01	0.30 ± 0.02	0.22 ± 0.02
Moyennes	0.30 ± 0.02	0.26 ± 0.02	0.17 ± 0.03

TABLE 6.1 – Récapitulatif des exposants obtenus lors d'un ajustement en loi de puissance des données de BOETTIGER et al. [1] pour les domaines épigénétiques actifs, inactifs et réprimés, à la fois en médianes et en moyennes.

Pour une analyse plus quantitative, il faut se pencher sur les valeurs des exposants, récapitulées à la Table 6.1. Étonnamment en moyennes, les domaines actifs, inactifs et réprimés ont des exposants de loi d'échelle ψ de 0.30, 0.26 et 0.17, qui sont tous trois plus petits que la valeur attendue pour des conformations globulaires de $\psi = 0.33$. Il est également intéressant de noter que les exposants apparents en médianes pour les trois types de domaines sont plus grands qu'en moyennes ($\psi = 0.38$, 0.30, 0.22 pour les actifs, inactifs et réprimés, respectivement). Or, pour les conformations des polymères

à grande valeur de N, on s'attend à ce que les lois d'échelle soient identiques pour la médiane et la moyenne. On obtient donc ici une indication assez forte d'un effet de taille finie.

6.1.2 Modèle convolué

La simple analyse des données n'est pas fructueuse, il faut tenir compte à la fois (*i*) du faisceau (2.21) et (*ii*) des effets de taille finie (5.11). On écrit (*i*) la distribution du rayon de giration d'un faisceau de polymères ($R^2_{\text{micro}} = 0$)

$$\mathcal{P}_N(r^2|\varepsilon,\sigma) = \int_0^{r^2} f_\lambda(r^2 - s^2) \cdot p_N(s^2|\varepsilon)\, ds^2$$

où

— (*i*) $f_\lambda(r^2) = \lambda e^{-\lambda r^2}$ est la distribution radiale du rayon de giration du faisceau de taille caractéristique $\sigma = \lambda^{-1/2}$.

— (*ii*) $p_N(r^2|\varepsilon) = Z_N(\varepsilon)^{-1} e^{-\beta F_N(r^2|\varepsilon)}$ est la distribution du rayon de giration d'un unique polymère attractif de taille N et de paramètre d'attraction ε.

On introduit la distribution dimensionnée $\mathcal{Q}_L$ du rayon de giration R^2, *i.e.* qui tient compte de la dimension des données où R^2 est exprimé en nm^2 et L en kb, grâce à la méthode de la transformation inverse

$$\mathcal{Q}_L(R^2|\boldsymbol{\theta}) = \left|\frac{dr^2}{dR^2}\right| \mathcal{P}_N(r^2|\varepsilon,\sigma)$$

$$\mathcal{Q}_L(R^2|\boldsymbol{\theta}) = \frac{1}{K^2_{\text{nm}}} \mathcal{P}_N(r^2|\varepsilon,\sigma)$$

où

— $\boldsymbol{\theta} = (\varepsilon, K_{\text{kb}}, K_{\text{nm}}, A_0)$ sont les paramètres du modèle.

— $r^2 = R^2/K^2_{\text{nm}}$ est le rayon de giration adimensionné.

— $\sigma = A_0/K_{\text{nm}}$ est la taille caractéristique adimensionnée du faisceau.

On construit, ensuite, l'énergie libre effective

$$\beta\mathcal{F}_L(R^2|\boldsymbol{\theta}) = -\ln \mathcal{Q}_L(R^2|\boldsymbol{\theta})$$

avec laquelle on va définir la log-vraisemblance (5.8) des données à optimiser en suivant les hypothèses de travail :

1. Le même modèle de polymère peut décrire toutes les observations quel que soit l'état épigénétique, *i.e.* la *couleur*.
2. À différentes *couleurs* correspondent différents paramètres $\boldsymbol{\theta}$.
3. L'ensemble des données d'une *couleur* donnée peut être ajusté par un unique jeu de paramètres $\boldsymbol{\theta}$, quelle que soit la taille des domaines épigénétiques, leur contexte génomique ou d'autres caractéristiques.

En toute généralité, la taille caractéristique σ du faisceau peut dépendre de la longueur N du polymère : $\sigma = \sigma(N)$. Nous avons choisi de modéliser $\sigma(N)$ comme

$$\sigma(N) = \frac{A_\infty/K_{\text{nm}}}{1 + \left(\frac{A_\infty}{A_0} - 1\right) e^{-\frac{N}{N_0}}} \tag{6.2}$$

variant d'une valeur minimale A_0 à une valeur maximale A_∞, atteinte avec une longueur caractéristique N_0. Le sens physique de cette modélisation sera discuter plus loin dans la conclusion de ce chapitre.

6.1.3 Résultats

Aux figures 6.2, 6.3 et 6.4, les histogrammes de tous les domaines actifs, inactifs et réprimés sont tracés respectivement avec les courbes théoriques obtenues pour les paramètres optimaux. La comparaison montre un accord remarquablement bon entre la distribution des données et la prédiction du modèle étant donné la taille limitée des échantillons expérimentaux.

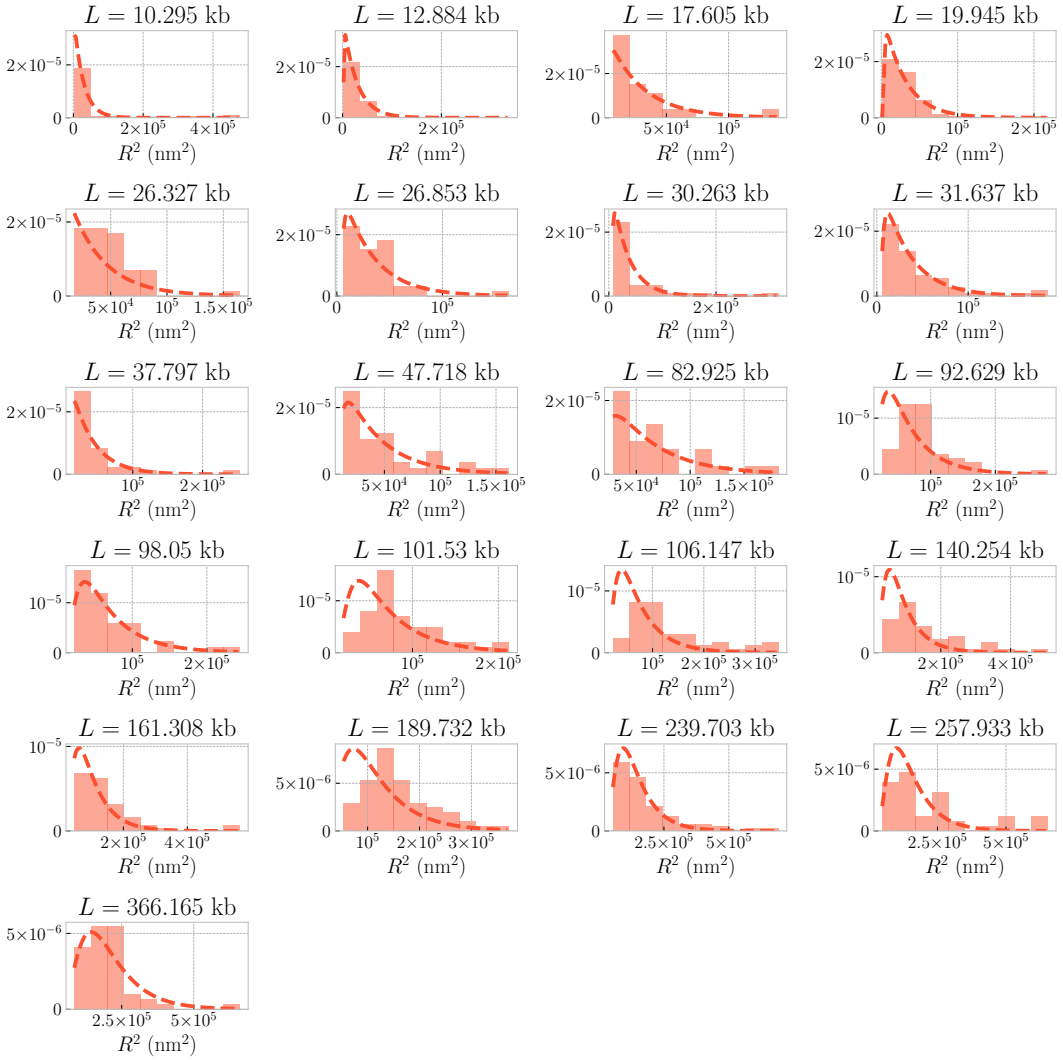

FIGURE 6.2 – Confrontation de la distribution du rayon de giration prédite par le modèle aux distributions expérimentales des domaines *actifs*.

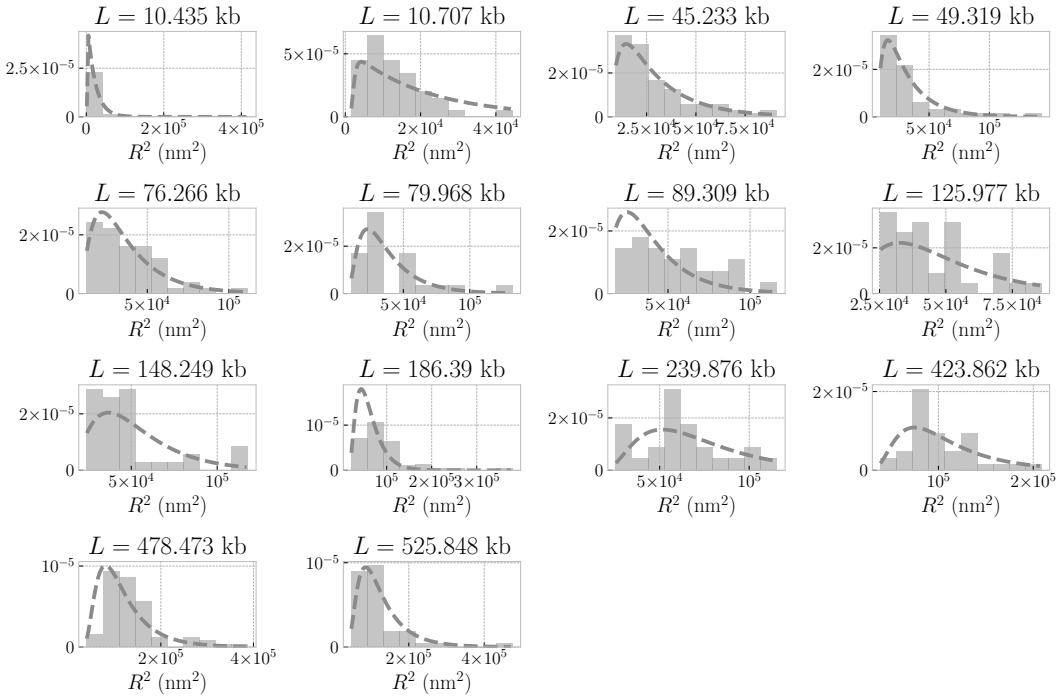

FIGURE 6.3 – Confrontation de la distribution du rayon de giration prédite par le modèle aux distributions expérimentales des domaines *inactifs*.

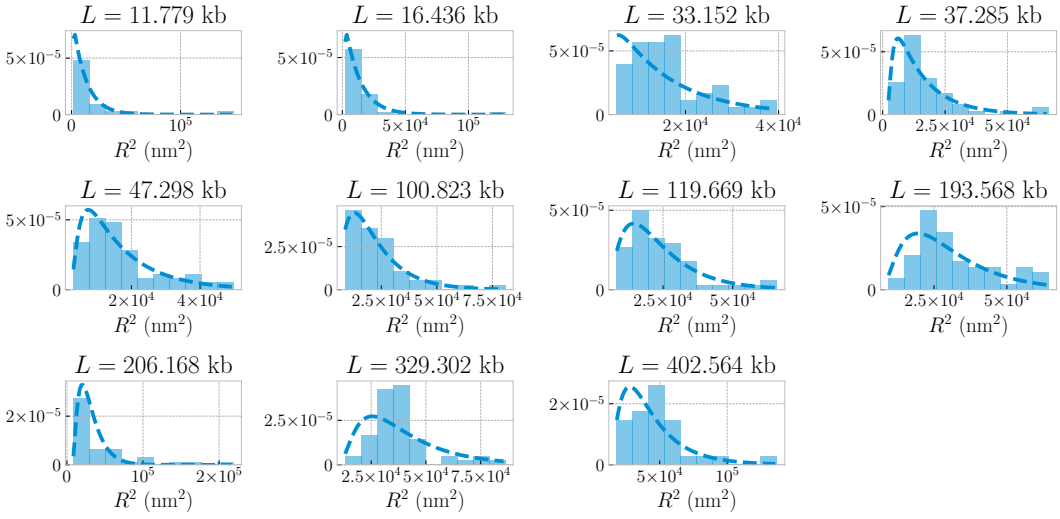

FIGURE 6.4 – Confrontation de la distribution du rayon de giration prédite par le modèle aux distributions expérimentales des domaines *réprimés*.

L'ajustement des distributions des rayons de giration permet d'utiliser les données dans leur intégralité et d'en extraire le maximum d'informations. En guise de vérifica-

tion *a posteriori* des résultats à la Figure 6.5, on compare les prédictions du modèle aux médianes et moyennes expérimentales, en fonction de la longueur L des domaines épigénétiques.

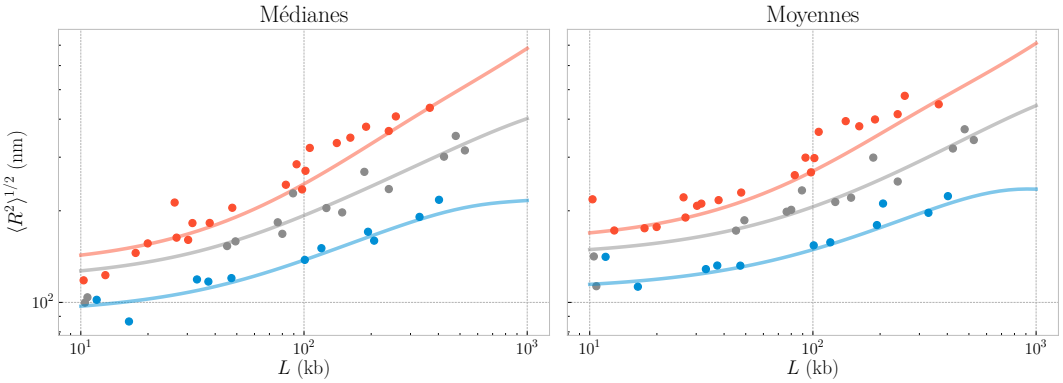

FIGURE 6.5 – Confrontation des prédictions de notre modèle aux médianes et moyennes expérimentales de BOETTIGER et al. [1] pour les trois états épigénétiques actif (en rouge), inactif (en noir) et réprimé (en bleu).

Les paramètres obtenus lors de l'ajustement pour les trois états épigénétiques sont récapitulés à la Table 6.2, avec la compaction linéaire calculée en différentes unités. Les incertitudes ont été déterminées à l'aide du 16-ième et 64-ième centiles de la distribution marginale des paramètres, correspondant à l'intervalle de confiance à 1σ pour une gaussienne.

	État	Actif	Inactif	Réprimé
Ajusté	ε (k_BT)	$0.15^{+0.03}_{-0.11}$	0.36 ± 0.03	$0.37^{+0.04}_{-0.03}$
	K_{kb} (kb)	$0.4^{+0.5}_{-0.2}$	3^{+3}_{-1}	$1.2^{+1.5}_{-0.6}$
	K_{nm} (nm)	16^{+12}_{-6}	60 ± 20	26^{+12}_{-8}
	A_0 (nm)	158 ± 6	142^{+7}_{-6}	109^{+6}_{-5}
	A_∞ (nm)	320 ± 20		
	N_0	200^{+200}_{-100}		
Calculé	c (bp nm^{-1})	25	50	46
	c_{10} (nuc/10nm)	1.4	2.6	2.4
	C (nuc/K$_{nm}$)	2.2	15.6	6.6
[62]	NRL (bp)	182.0 ± 0.5	192.6 ± 0.5	192.2 ± 0.6

TABLE 6.2 – Récapitulatif des paramètres obtenus lors de l'ajustement avec notre modèle des données de BOETTIGER et al. [1] pour les domaines épigénétiques actifs, inactifs et réprimés.

6.2 Conclusion

Dans l'ensemble, ces résultats montrent que notre méthode, ainsi que les hypothèses formulées, donnent un ajustement correct de l'ensemble des données et permettent une interprétation physiquement valable qui sera commentée dans ce qui suit.

6.2.1 L'attraction rapproche les domaines inactifs et réprimés des conditions critiques

La Figure 6.5 montre que les domaines actifs (en rouge) ont un exposant de loi d'échelle ψ très proche de l'exposant $\nu \approx 3/5$ et restent ainsi dans le régime *coil* pour toutes les longueurs observées. Ce qui est en accord avec le paramètre d'attraction $\varepsilon = 0.15$ obtenu pour les domaines actifs, bien en dessous de la valeur de transition théorique (à grand N) de $\varepsilon_\Theta \approx 0.27$.

À la différence des domaines actifs, les domaines réprimés (en bleu) sont au-dessus de l'énergie critique ε_Θ, avec $\varepsilon = 0.37$, et sont théoriquement du côté *globule* de la transition. Cependant, l'énergie critique en taille finie est plus grande que dans la limite des grands N ε_Θ : les domaines réprimés sont en fait dans la région du *crossover*. À la Figure 6.5, un plateau est en effet visible à partir des longueurs d'environ 400 kb, avec un net *crossover* entre le comportement *coil* et le comportement *globule*.

Comme on pouvait s'y attendre, les domaines inactifs (en noir) ont un comportement intermédiaire, néanmoins énergiquement proche des domaines réprimés. Avec $\varepsilon = 0.36$, ils sont au-dessus de la limite d'énergie de transition ε_Θ entre *coil* et *globule*, mais les effets de taille finie restent forts aux longueurs observées. Par conséquent, le plateau du *crossover* n'est toujours pas atteint à ces longueurs. Cependant, on observe une nette divergence par rapport au comportement *coil* aux plus petites longueurs.

Dans tous les cas, tous les domaines courts sont proches des conformations *coil* en raison des effets de taille finie qui apparaissent alors comme une caractéristique cruciale dans l'interprétation de l'imagerie super-résolution des domaines épignénétiques.

6.2.2 Géométrie du faisceau

Au cours de la procédure d'ajustement des paramètres, on a introduit la possibilité que le faisceau puisse avoir une taille caractéristique (6.2) qui dépende de la longueur du domaine. Sans quoi, on était incapable de reproduire qualitativement les données des domaines actifs. Il n'a pas été nécessaire de mettre une taille caractéristique variable pour les domaines inactifs ou réprimés. En effet, le modèle avec une taille caractéristique constante est suffisante pour reproduire quantitativement les données.

Conformément à la Table 6.2, le faisceau a une taille caractéristique minimale d'environ 100 nm pour les trois états épigénétiques. Néanmoins, on notera que le faisceau est légèrement plus grand pour les domaines actifs et inactifs. Ces valeurs sont compatibles avec les valeurs du rayon de giration observés à la Figure 6.5 pour les plus petits domaines.

Concernant les actifs, les tailles caractéristiques du faisceau s'échelonnent de $A_0 = 158$ à $A_\infty = 320$ au maximum, mais seulement pour les polymères suffisamment grands (de l'ordre de 200 monomères, soit environ 3200 nm ou encore 80 kb).

Ces résultats révèlent une corrélation entre la géométrie du faisceau et le mode de repliement du domaine. Cela semble logique puisque les deux concernent l'extension du polymère. Dans le cas extrême d'une globule dense, on ne s'attend pas à ce que la densité à l'intérieur du domaine fluctue. Elle est homogène. Par conséquent, un faisceau de plusieurs chaines est également uniformément attaché et la distance entre les chaines ne dépend pas de leur longueur. *A contrario* chez les *coil*, les chaines ne collent pas les unes aux autres et par conséquent la distance entre les chaines fluctue. Ces fluctuations augmentent avec la longueur de la chaine. Ce qui explique pourquoi les domaines actifs, plutôt décondensés, constituent un faisceau davantage large que les domaines sont longs (σ_N variable). Alors que cet élargissement est moins marqué chez les domaines plus globulaires inactifs ou réprimés (σ constant).

6.2.3 Les paramètres obtenus soutiennent la thèse des faibles longueurs de persistance

Les estimations des longueurs de Kuhn dans différents organismes restent difficiles à obtenir. L'un des principaux obstacles réside dans la nécessité de relier les longueurs de Kuhn en nanomètres et en paires de bases, au moyen de la compaction linéaire du chromosome.

Jusqu'à présent, peu de mesures *in vivo* des longueurs de Kuhn ont été effectuées. En 2004, Bystricky et al. [63] ont obtenu des longueurs de Kuhn de l'ordre de 400 nm et des densités linéaires de 6 à 9 nucléosomes par 10 nm dans la levure, grâce à des techniques d'imagerie à haute résolution.
En 2008, Dekker [64] a obtenu, de nouveau chez la levure et par 3C, des longueurs de Kuhn de l'ordre de 120 à 260 nm et des densités linéaires d'environ 1.1 à 2.2 nucléosomes par 10 nm.

Il est important de noter que la longueur de Kuhn dépend fortement de la densité linéaire. Pour les fibres d'une telle compaction, la longueur de Kuhn ne dépasse jamais 100 nm [65], qui est la limite inférieure des valeurs mesurées par Dekker. Les ADN de liaison sont plus petits chez la levure que chez la drosophile (et l'humain). On s'attend donc à ce que les estimations de Dekker soient une limite supérieure de la longueur de Kuhn chez la drosophile.

Ou et al. [66] ont réalisé des images de chromatine *in vivo* confirmant également des fibres de nucléosomes de faibles densités massiques et très flexibles.
Plus récemment chez les mammifères, les résultats de Hi-C et de probabilité de contact P(s) obtenus par Sanborn et al. [67] suggèrent une longueur de Kuhn en paires de bases d'environ 1 kb et certainement moins de 5 kb pour la chromatine. Ce qui suggère qu'à l'échelle typique du gène ($\sim$ 15 kb), la chromatine est très flexible. Cette flexibilité est également compatible avec (et indispensable pour) la formation de boucles par extrusion.
Chez la drosophile, les mêmes résultats sont obtenus par des mesures à haute résolution Hi-C.

Pris ensemble, les résultats sont compatibles avec des densités linéaires de l'ordre de 2 nucléosomes par 10 nm et avec des longueurs de Kuhn aussi faibles que $K_{\mathrm{nm}} \sim 30\,\mathrm{nm}$ et $K_{\mathrm{kb}} \sim 1\,\mathrm{kb}$. Les valeurs relativement faibles de K_{nm} ($30-60\,\mathrm{nm}$) par rapport, en particulier, à de l'ADN seul, confirment les mesures de dynamique les plus récentes

de la grande flexibilité de la chromatine *in vivo* par Socol et al. [68]. Notons que des estimations similaires sont également utilisées dans des travaux de modélisation récents par Nuebler et al. [69] et y compris dans la partie modélisation de l'article de Boettiger et al. [1], cependant pas de manière cohérente.

6.2.4 La dépendance des paramètres aux couleurs épigénétiques indique une structure spéciale pour les domaines inactifs

Dans les études précédentes, et notamment dans les simulations d'organisation du génome 3D, on a généralement supposé une taille unique du monomère (K_{kb} ou K_{nm}) quel que soit l'état épigénétique. L'une des conclusions importantes de ce travail est que les domaines actifs, inactifs et réprimés ont tous une taille de monomères différentes. On notera, cependant, un certain recouvrement des valeurs en tenant compte des incertitudes.

Puisque les domaines réprimés de la chromatine sont dispersés dans le volume du compartiment actif [29], la similarité structurelle des fibres de nucléosomes des domaines actifs et réprimés peut faciliter les transitions d'un état à l'autre au cours de la différenciation cellulaire.

Pour faciliter la lecture, on rappelle ici les longueurs de Kuhn obtenues :

— Pour les actifs : ($K_{\mathrm{kb}} = 0.4^{+0.5}_{-0.2}, K_{\mathrm{nm}} = 16^{+12}_{-6}$).
— Pour les réprimés : ($K_{\mathrm{kb}} = 1.2^{+1.5}_{-0.6}, K_{\mathrm{nm}} = 26^{+12}_{-8}$).
— Pour les inactifs : ($K_{\mathrm{kb}} = 3^{+3}_{-1}, K_{\mathrm{nm}} = 60 \pm 20$).

La compacité et la rigidité accrues constatées dans les domaines inactifs doivent être justifiées sur une base moléculaire et nécessitent une discussion plus approfondie. Il est intéressant de noter que la chromatine noire contient près des deux tiers de tous les gènes silencieux. La plupart d'entre eux étant des gènes spécifiques des tissus, et semble inhiber activement l'expression génétique [22, 70]. La manière dont cette répression est mise en œuvre n'est pas encore claire. Les protéines qui sont maintenant connues pour marquer la chromatine noire sont, notamment, l'histone de liaison H1, qui a été précédemment liée à la répression de la transcription [70]. En réticulant les ADN d'entrée et de sortie de chaque nucléosome, la proétine H1 peut en effet aboutir à un raccourcissement efficace des ADN de liaison [71], expliquant ainsi le raidissement et le compactage de la fibre de nucléosomes. De plus, ces caractéristiques structurelles semblent raisonnablement liées à l'extinction des gènes, ce qui donne encore plus de crédibilité à l'hypothèse selon laquelle H1 serait le principal acteur de l'inactivation.

6.2.5 Recherche d'une base moléculaire pour expliquer les paramètres énergétiques inférés

Notre approche fournit la première inférence couleur-spécifique de l'énergie d'attraction ε entre les segments de Kuhn de la chromatine *in vivo*. Dans l'article de Boettiger et al. [1], une très forte attraction de $3.5\,\mathrm{k_BT}$ est utilisée pour simuler les domaines réprimés. En raison de cette énorme valeur énergétique, les conformations globulaires obtenues par les auteurs sont déjà très serrées pour $N = 400$ (Figure 4c de Boettiger et al. [1]). Les simulations ne sont là que pour reproduire le comportement en loi d'échelle mesuré expérimentalement, et ne donnent donc que des valeurs

adimensionnées pour les rayons de giration.

D'autres estimations des paramètres d'attraction de la chromatine ont été obtenues à partir de l'ajustement des données Hi-C [72, 73]. Dans une étude récente, Falk et al. [74] ont déterminé la valeur des paramètres d'énergie d'interaction dans un modèle copolymère (compartiments de la chromatine A et B). Afin de retrouver la séparation de phase expérimentale entre la chromatine A et la chromatine B, ils ont trouvé une attraction entre les monomères B de $0.55\,k_BT$ et une attraction beaucoup plus faible entre les monomères A. Ceci est compatible avec nos résultats, assimilant le compartiment A avec la chromatine active, et le compartiment B avec les réprimés.

Notons que l'hybridation FISH implique une dénaturation de l'ADN. Un effet potentiel pourrait être une décondensation partielle de la chromatine. Dans ce cas, l'énergie d'attraction effective ajustée par notre procédure serait sous-estimée par rapport aux conditions *in vivo*. Néanmoins, le protocole d'hybridation FISH, adapté par Boettiger et al. [1], assure une altération minimale de la structure de la chromatine [75]. Il serait ainsi tentant d'essayer de relier les différentes valeurs de ε obtenues, à la Table 6.2, pour les trois états épigénétiques à différents mécanismes d'interaction moléculaire. La prudence est de mise, car ε est un paramètre effectif rendant compte globalement de l'énergie d'interaction moyenne entre deux segments de Kuhn. Des simulations de fibres de nucléosomes avec un grain fin de 10 bp pour l'ADN indiquent qu'en moyenne, on ne devrait s'attendre qu'à un seul contact nucléosome-nucléosome en *trans*, *i.e.* contact de deux nucléosomes non-adjacents, par segment de Kuhn. Dans cette hypothèse, l'ordre de grandeur de ε apparait comme une estimation raisonnable pour une interaction simple en *trans*. De sorte qu'une comparaison directe entre les valeurs ajustées devient possible. Dans le cas de domaines réprimés, cette interaction est médiée par les protéines Polycomb qui sont considérées comme stabilisant les configurations de la chromatine condensée au moyen de ponts. On trouve, de manière cohérente, la plus grande énergie d'attraction $\varepsilon \approx 0.4\,k_BT$. Il n'est cependant pas clair quel mécanisme pourrait expliquer la différence d'énergie d'interaction entre les domaines actifs et inactifs. Comme on le verra dans la prochaine partie, plusieurs expériences indépendantes fournissent une explication possible, qui n'implique pas d'interactions médiées par des protéines.

6.2.6 La comparaison avec des expériences sur une solution de nucléosomes dans les noyaux révèle des caractéristiques de criticité et un rôle clé pour l'interaction nucléosome-nucléosome

L'énergie libre (5.11) exprimée en termes de densité rénormalisée $t = \rho^{1/(\nu d-1)}$ est issue d'une approche type développement du viriel. Ça consiste à supposer que les interactions sont dominées par des interactions à deux corps. Alors que celles à plusieurs corps sont plus rares, de sorte qu'un développement en basse densité est appropriée.

À la transition *coil-globule*, le coefficient $a_1(\varepsilon)$, apparenté au deuxième coefficient du viriel, s'annule et change de signe, ce qui reflète une compensation entre les interactions attractives et répulsives à deux corps. Tandis que $a_2(\varepsilon)$, apparenté au troisième coefficient du viriel, reste positif [59]. Ici, on a constaté que si les domaines actifs sont dans le régime *coil* pour toutes les longueurs observées (avec $\varepsilon = 0.15\,k_BT$), les

domaines inactifs et réprimés sont dans la région du *crossover* pour la plupart des longueurs observées, à cause des effets de taille finie (avec $\varepsilon = 0.36\,k_BT$ et $\varepsilon = 0.37\,k_BT$ respectivement). Ainsi, le second coefficient du viriel $a_1(\varepsilon)$ est proche de zéro pour les domaines non actifs, ce qui indique un grand degré de compensation entre attraction et répulsion. Une fois de plus, la question de la base moléculaire de ce comportement se pose.

Il est intéressant de décrire le système en termes de coefficients du viriel, ce qui nous permet de comparer nos résultats avec des expériences totalement indépendantes. Mangenot et al. ont caractérisé expérimentalement l'interaction entre noyaux nucléosomiques isolés à différentes concentrations de sel monovalent [76, 77]. Il est intéressant de noter que le deuxième coefficient du viriel diminue fortement jusqu'à zéro et présente une pointe dans la plage de 75 à 210 mM, *i.e.* autour des concentrations physiologiques. Ainsi, l'architecture et la biochimie des nucléosomes semblent avoir été choisies de façon à ce que la répulsion et l'attraction entre nucléosomes se compensent dans les organismes vivants.

Il est tentant de relier la transition *coil-globule* des chromosomes à l'annulation du deuxième coefficient du viriel de l'interaction nucléosome-nucléosome. C'est également conforme aux récentes mesures, réalisées par Socol et al. [68], de la dynamique des chromosomiques chez la levure. La dynamique a été modélisée comme une dynamique de Rouse ralentie par des interactions transitoires nucléosome-nucléosome d'une durée de vie de quelques secondes et caractérisés par une énergie attractive de -0.3 à $-0.5\,k_BT$.

En suivant cette ligne et en approfondissant, la chromatine inactive (noire) est très proche du point Θ, ce qui indique que les interactions nucléosome-nucléosome pourraient être dominantes dans les domaines inactifs.

Pour la chromatine active (rouge), on propose que l'attraction plus faible, par rapport aux deux autres états épigénétiques, soit liée à une attraction plus faible entre nucléosomes. Ce qui est cohérent avec l'acétylation des queues d'histones dans la chromatine transcrite [78], réduisant ainsi, sur la base de simulations, leur charge et de ce fait leur capacité à se lier aux autres nucléosomes [79]. Par ailleurs, des changements structuraux de la chromatine lors de l'acétylation des queues d'histone ont été récemment mis en évidence expérimentalement *in vitro* et *in vivo* [80–82].

Pour la chromatine réprimée (bleue), une plus grande valeur de ε indique une interaction plus forte, certainement médiée par des protéines de la famille des Polycomb, en accord avec les expériences de mutation du complexe répressif Polycomb I (PRC1) [1]. La modélisation détaillée des effets mécanistiques en jeu reste nébuleuse et souligne clairement la nécessité d'une modélisation à l'échelle moléculaire pour le complexe Polycomb.

Chapitre 7

Conclusion générale

Plusieurs indications expérimentales indiquent que la chromatine est une structure à l'organisation hautement complexe, capable de réguler de manière fine l'expression des gènes. Si des méthodes expérimentales de plus en plus sophistiquées permettent d'en explorer les caractéristiques, une modélisation de ses propriétés physiques s'avère nécessaire pour interpréter correctement les données.

Parmi les données les plus récentes chez la drosophile : des images super-résolues de domaines épigénétiques, réalisées par Boettiger et al. [1] dans la lignée Kc167, permettent d'étudier le rayon de giration de domaines fonctionnels de longueur génomique différente. Ces mesures permettent d'accéder à l'ensemble des distributions du rayon de giration à la fois pour chaque type fonctionnel et pour chaque longueur génomique étudiés. C'est une information riche qui n'avait pas été exploitée dans le papier expérimental.

Pour mener à bien cette l'analyse des données, j'ai modélisé les domaines épigénétiques comme des polymères auto-évitants dotés d'une interaction attractive entre monomères. Je me suis basé sur des propositions existantes dans la littérature pour proposer une expression pour l'énergie libre d'une marche auto-évitante attractive. Elle dépend de coefficients a_i qui sont uniquement fonction du paramètre d'attraction ε (chapitres 3 et 4).
Afin de valider cette expression et d'interpoler les coefficents, je me suis appuyé sur des simulations extensives du système à différentes longueurs et énergies d'interaction (chapitre 5). Ce travail est en cours de publication.
J'ai également proposé une modélisation du *faisceau* de chromosomes présent dans les cellules (chapitre 2).
Enfin, j'ai exprimé la distribution du rayon de giration directement en fonction des paramètres structuraux de la chromatine, notamment les longueurs de Kuhn (en nanomètres et en paires de bases) et le paramètre d'attraction. Ainsi, l'ajustement des données donne accès à ces grandeurs difficiles à mesurer directement (chapitre 6).

Un aspect important de mon travail a été de prendre en compte les contraintes techniques liées au temps de calcul. Ce qui s'est traduit par l'apprentissage en autonomie du langage de programmation Rust et par l'adaptation de bon nombre d'outils dans ce langage, pour la mise en parallèle des calculs sur le *cluster* du laboratoire. Le chapitre 5 et l'Annexe B reflètent la composante numérique de ma thèse.

7.1 Principaux résultats

Rappelons ici que l'analyse permet d'aboutir à deux principales conclusions.
La première étant que les effets de taille finie sont suffisants pour que l'on puisse estimer les longueurs de Kuhn (en nanomètres et en paires de bases) à partir des distributions de rayon de giration obtenues grâce à la microscopie super-résolue.
La seconde étant que les paramètres structuraux (longueurs de Kuhn et paramètre d'attraction) obtenus lors de l'inférence bayésienne révèlent des différences importantes entre les domaines des trois états épigénétiques. Il est intéressant de noter que les domaines actifs sont loin de la transition *coil-globule*, contrairement aux domaines inactifs ou réprimés. Cette différence qualitative pourrait expliquer les différents niveaux d'activité génétique.

7.2 Améliorations

Le travail réalisé durant ma thèse ouvre des possibilités intéressantes quant à l'analyse de données similaires. La même méthode pourrait être appliquée à des données de super-résolution équivalentes. Auquel cas, la qualité des résultats dépendrait essentiellement de la quantité de données disponibles. Pour l'améliorer, il faudrait donc avoir une meilleure statistique pour chaque domaine.
Une meilleure modélisation du faisceau pourrait cependant s'avérer nécessaire. Pour cela, il faudrait avoir accès expérimentalement à des distributions du rayon de giration pour un même domaine épigénétique avec un nombre variable de chromosomes homologues.

Sur le plan de la modélisation, il pourrait être intéressant d'étudier la fonction génératrice des cumulants K de la distribution du nombre de contacts d'une marche auto-évitante, par l'intermédiaire de simulations. Ce qui permettrait d'affiner l'écriture du modèle et ainsi obtenir des résultats toujours plus précis.

7.3 Limitations

La méthode d'analyse possède, néanmoins, quelques restrictions. Elle est fondée sur l'analyse en taille finie des différents domaines et nécessite un large éventail en taille pour fonctionner correctement. C'est-à-dire qu'on doit avoir accès à des domaines de longueurs se répartissant sur plus d'une décade.
De plus, le modèle ne rend pas compte des effets de bords. Il suppose que les domaines sont indépendants, bien qu'ils soient liés par les extrémités. Il néglige donc l'influence des domaines adjacents.
Bien que n'étant pas le mécanisme principal de formation de domaines topologiques chez la drosophile, les boucles d'extrusion restent néanmoins possibles [83, 84] et ne sont pas prises en compte par le modèle.

7.4 Perspectives à court terme

Depuis le début de ma thèse en 2016, où seules les données super-résolution de Boettiger et al. [1] étaient disponibles, de plus en plus de données sont produites [85] et peuvent être analysées. Il y a notamment des distributions de rayon bout à bout qui sont mesurées par Cattoni et al. [29], qui nécessitent une réécriture du modèle en terme de rayon bout à bout, plutôt que de rayon de giration. Au cours de la thèse, j'ai entamé ce travail de réécriture qui semble prometteur, mais n'a cependant pas encore abouti. En effet, il est nécessaire de refaire toutes les simulations en terme de rayon bout à bout à la fois pour valider l'écriture et pour inférer les paramètres. On notera, néanmoins, que le rayon de giration est une mesure plus robuste de la taille du polymère et moins sujette aux effets de bords. En effet, les domaines épigénétiques sont tous reliés par les extrémités pouvant imposer de fortes contraintes sur les bords.

7.5 Perspectives à moyen terme

Les paramètres d'attraction ε inférés pour les domaines inactifs et réprimés sont compatibles avec les récentes mesures de dynamique réalisées par Socol et al. [68] de 0.3 à 0.5 k_BT. Ce qui suggère qu'en plus de décrire la structure des domaines épigénétiques, le paramètre ε permet également de remonter à leur dynamique. Il semble alors naturel de vouloir explorer la simulation de la dynamique des chromosomes en utilisant des modèles de polymères hors-réseau.
En collaboration avec l'équipe de Thomas Gregor, il serait possible d'étudier l'évolution dynamique de la distance entre deux *loci* spécifiques (par exemple *promoter* et *enhancer*), en combinant la modélisation dynamique avec la description statistique des distances bout à bout. Ce travail fait l'objet d'un projet ANR accepté cette année.
À présent, nous avons accès aux paramètres structuraux des différents domaines épigénétiques. Il pourrait être intéressant de regarder le comportement d'une « soupe » de polymères d'énergie d'interaction différente comme représentation minimale d'un mélange de domaines épigénétiques. Cette modélisation permettrait d'étudier leurs interactions effectives et de comprendre la compartimentation de la chromatine à une échelle supérieure.

Annexe A

Faisceau brownien libre

A.1 Pont brownien libre

Un pont brownien libre est un mouvement brownien contraint à ses deux extrémités, partant de l'origine x_0 à l'origine des temps ($t = 0$) et attaignant sa cible x_f au bout de la durée t_f. En reprenant les notations de Majumdar et Orland [86], la probabilité de trouver une particule brownienne à l'instant t et à la position x sachant qu'elle est partie de l'origine x_0 et qu'elle atteindra le point terminal x_f au bout d'une durée t_f s'écrit :

$$\mathcal{P}(x,t) = \frac{1}{P(x_f,t_f|x_0,0)} Q(x,t)P(x,t)$$

où $P(x,t) = P(x,t|x_0,0)$ est la probabilité d'une particule brownienne se trouvant en x à l'instant t étant donné les conditions initiales $(x_0,0)$ et $Q(x,t) = P(x_f,t_f|x,t)$ est la probabilité du processus inverse, arrivant en (x_f,t_f) sachant qu'elle était précédemment en (x,t). Dans le cas d'une particule brownienne libre, $P(x,t)$ et $Q(x,t)$ s'écrivent

$$P(x,t|x',t') = \frac{1}{\sqrt{4\pi D(t-t')}} \exp\left(-\frac{(x-x')^2}{4D(t-t')}\right)$$

$$P(x,t) = \frac{1}{\sqrt{4\pi Dt}} \exp\left(-\frac{(x-x_0)^2}{4Dt}\right)$$

$$Q(x,t) = \frac{1}{\sqrt{4\pi D(t_f-t)}} \exp\left(-\frac{(x_f-x)^2}{4D(t_f-t)}\right)$$

où D est le coefficient de diffusion. On aboutit alors à l'expression gaussienne de la probabilité

$$\mathcal{P}(x,t) = \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left(-\frac{(x-\mu_t)^2}{2\sigma_t^2}\right)$$

avec

$$\mu_t = \frac{t}{t_f} x_f + (1 - \frac{t}{t_f}) x_0$$

$$\sigma_t^2 = 2Dt_f \frac{t}{t_f}(1 - \frac{t}{t_f})$$

que l'on peut également réécrire

$$\mu_s = s x_e + x_0$$

$$\sigma_s^2 = 2D\tau s(1 - s)$$

en introduisant l'abscisse curviligne $s = t/t_f$, la distance bout à bout $x_e = x_f - x_0$ et la durée $\tau = t_f$.

A.2 Faisceau dense de ponts browniens libres

Dans le cas d'un faisceau à très grand nombre de polymères ($n \gg 1$), *i.e.* dense, de longueur $m+1$, contraints à tous passer par les mêmes extrémités $\boldsymbol{r}_0$ et $\boldsymbol{r}_f$, la densité de probabilité de trouver un monomère d'indice i appartenant à n'importe lequel des poylmères s'écrit :

$$\delta_i(\boldsymbol{r}) = \mathcal{P}(x,i) \times \mathcal{P}(y,i) \times \mathcal{P}(z,i) = \frac{1}{\sqrt{2\pi\sigma_i^2}^3} \exp\left(-\frac{\boldsymbol{r} \cdot \boldsymbol{\mu}_i}{2\sigma_i^2}\right)$$

$$\boldsymbol{\mu}_i = \frac{i}{m}(\boldsymbol{r}_f - \boldsymbol{r}_0) + \boldsymbol{r}_0$$

$$\sigma_i^2 = 2Dm\frac{i}{m}(1 - \frac{i}{m})$$

Calculons à présent le rayon de giration de ce système

$$\left\langle R^2 \right\rangle = \frac{1}{m+1} \sum_{i=0}^{m} \left\langle R_i^2 \right\rangle + \frac{1}{m+1} \sum_{i=0}^{m} \left\langle (\boldsymbol{G}_i - \boldsymbol{G})^2 \right\rangle$$

avec pour centre de masse moyen à l'indice i, $\boldsymbol{G}_i = \boldsymbol{\mu}_i$ et pour centre de masse moyen $\boldsymbol{G} = \frac{1}{2}(\boldsymbol{r}_f + \boldsymbol{r}_0)$, vient le terme de droite :

$$\langle \boldsymbol{G}_i - \boldsymbol{G} \rangle = (\frac{i}{m} - \frac{1}{2})(\boldsymbol{r}_f - \boldsymbol{r}_0)$$

$$\left\langle \frac{1}{m+1} \sum_{i=0}^{m} (\boldsymbol{G}_i - \boldsymbol{G})^2 \right\rangle = \frac{m+2}{12m} (\boldsymbol{r}_f - \boldsymbol{r}_0)^2.$$

Puis, avec le rayon de giration moyen à l'indice i, $R_i^2 = \sigma_i^2$, vient le terme de gauche :

$$\frac{1}{m+1} \sum_{i=0}^{m} \left\langle R_i^2 \right\rangle = \frac{D}{3}(m-1).$$

Ainsi s'écrit le rayon de giration d'un faisceau dense de ponts browniens libres :

$$\left\langle R^2 \right\rangle = \frac{D}{3}(m-1) + \frac{m+2}{12m}(\boldsymbol{r}_f - \boldsymbol{r}_0)^2.$$

Pour une marche aléatoire sur un réseau cubique de $m+1$ pas, le coefficient de diffusion et le rayon bout à bout s'écrivent $D = L_K^2/6$ et $(\boldsymbol{r}_f - \boldsymbol{r}_0)^2 = (m+1)L_K^2$, donnant :

$$\left\langle R^2 \right\rangle = \frac{m-1}{18}L_K^2 + \frac{(m+2)(m+1)}{12m}L_K^2$$

$$\left\langle R^2 \right\rangle = \frac{5m^2+7m+6}{36m}L_K^2$$

où L_K désigne la longueur de Kuhn. Finalement, le rayon de giration peut s'exprimer dans la limite $m \gg 1$ comme :

$$R^2 = \frac{1}{18}mL_K^2 + \underbrace{\frac{1}{12}mL_K^2}_{\text{tige rigide}}$$

$$R^2 = \frac{5}{36}mL_K^2.$$

A.3 Algorithme de simulation d'un pont brownien sur réseau

Nous cherchons à mettre en place un algorithme de Monte-Carlo statique pour échantillonner des ponts browniens libres sur réseau. Reprenons l'équation obtenue par Majumdar et Orland [86]

$$\frac{dx}{dt} = \frac{x_f - x}{t_f - t} + \eta(t)\,,$$

avec un bruit blanc gaussien $\eta(t)$ tel que $\langle \eta(t) \rangle = 0$ et $\langle \eta(t)\eta(t') \rangle = 2D\delta(t-t')$. Discrétisons l'équation en temps

$$\begin{cases} \delta x(t) & = \dfrac{x_f - x(t)}{t_f - t}\delta t + \eta(t)\delta t \\ x(t+\delta t) & = x(t) + \delta x(t) \end{cases},$$

et passons en notation polymère

$$\begin{cases} \delta_i & = \dfrac{x_m - x_i}{m-i} + \eta_i \\ x_{i+1} & = x_i + \delta_i \end{cases}.$$

Le problème étant que l'amplitude des sauts δ_i est continue. On souhaite discrétiser l'espace et contraindre l'amplitude des sauts à être normée, $|\delta_i| = 1$. Pour ce faire, je propose de binariser la distribution des sauts

$$p(\delta_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\delta_i - \mu_i)^2}{2\sigma^2} \right)$$

s'écrivant avec $\mu_i = (x_m - x_i)/(m-i)$ et $\sigma^2 = 2D$. Ce qui donne les probabilités de saut et à droite $(+)$ à gauche $(-)$ du i-ème monomère :

$$P_{i+} = \int_0^{+\infty} p(\delta)\, d\delta$$

$$P_{i-} = 1 - P_{i+}.$$

À l'aide de la fonction d'erreur

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt\,,$$

on peut réécrire la probabilité de saut à droite

$$P_{i+} = \frac{1}{\sqrt{\pi}} \int_{-\frac{\mu_i}{\sqrt{2\sigma^2}}}^{+\infty} e^{-t^2}\, dt$$

$$P_{i+} = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\mu_i}{\sqrt{2\sigma^2}}\right)\right]$$

en posant $t = (\delta - \mu_i)/(\sqrt{2\sigma^2})$.

En définitive, on a les probabilités de saut à droite P_{i+} et à gauche P_{i-}, en étant à l'indice i, sachant que l'on souhaite atteindre la position finale x_m :

$$P_{i\pm} = \frac{1}{2}\left[1 \pm \mathrm{erf}\left(\frac{\mu_i}{\sqrt{2\sigma^2}}\right)\right]$$

$$P_{i\pm} = \frac{1}{2}\left[1 \pm \mathrm{erf}\left(\frac{1}{2\sqrt{D}}\frac{x_m - x_i}{m-i}\right)\right].$$

Dans le cas d'une marche aléatoire classique $\mu_i = 0$, on retrouve bien une probabilité de saut isotrope $P_{i\pm} = 1/2$.

Maintenant, si on se place dans le cas d'un réseau cubique en dimension d, les probabilités s'écrivent :

$$P_{i\pm} = \frac{1}{2}\left[1 \pm \mathrm{erf}\left(\sqrt{d/2}\frac{x_m - x_i}{m-i}\right)\right]$$

L'algorithme de construction du pont brownien libre sur réseau cubique est :

$$\begin{cases} \delta_i = \pm 1 & \text{avec la probabilité } P_{i\pm} \\ x_{i+1} = x_i + \delta_i \end{cases}.$$

Il faut noter que le processus n'atteint pas son objectif de façon déterministe.

Annexe B

Quadrature Tanh-Sinh

Une quadrature désigne génériquement, en mathématique, le calcul d'une aire. Le nom vient du latin *quadratura* qui signifie « carré, quadrature ». Dans le contexte du calcul numérique d'une intégrale, la quadrature consiste en une approximation de la valeur de celle-ci, par une somme pondérée

$$\int_a^b f(x)\,dx \approx \sum_k w_k f(x_k)$$

où les w_k désignent les poids et les x_k la subdivision.

Il existe différentes quadratures qui se distinguent par leur façon de subdiviser le domaine d'intégration en x_k et de choisir les poids associés w_k. Il y a, notamment, une méthode simple dite des trapèzes qui, comme son nom l'indique, consiste à approximer l'aire sous la fonction avec des trapèzes

$$\int_a^b f(x)\,dx \approx h\left(\frac{f(b)+f(a)}{2} + \sum_{k=1}^{n-1} f(a+kh)\right)$$

avec un pas constant h. Cette méthode est en général peu efficace face à des méthodes plus sophistiquées telles que la quadrature gaussienne dont la subdivision et les poids sont choisis pour donner un résultat exact sur des polynômes de degré inférieur ou égal à $2n-1$, où n est le nombre de points x_k. Le choix de la subdivision et des poids ne dépend que de la nature du domaine d'intégration (borné, demi-droite, droite). Ils peuvent donc être déterminés *a priori*.

Au cours de ma thèse, j'ai eu besoin de calculer numériquement des intégrales impropres de la forme

$$\int_0^{+\infty} e^{-ax^\alpha - bx^{-\beta}}\,dx \tag{B.1}$$

avec a, α, b, β des réels strictement positifs. Bien que semblant adaptée au problème, la quadrature gaussienne peut avoir des difficultés notables à converger dans une certaine gamme de paramètres, voire pouvant complètement échouer à intégrer la fonction quand l'essentiel de l'aire de celle-ci se trouve loin de l'origine. Ce qui oblige à découper l'intégrale, de manière *ad hoc*, en plus petites intégrales autour du maximum de la

fonction. Cette démarche n'est pas optimale en terme de calculs et surtout ne garantit pas la convergence bien que l'améliorant.

Une solution est d'utiliser la récente quadrature Tanh-Sinh, qui tire son nom du changement de variable $x = \phi(t) = \tanh \circ \frac{\pi}{2} \sinh t$. Cette quadrature repose sur le constat que la simple méthode des trapèzes est efficiente lorsqu'elle est appliquée à des fonctions ayant une décroissance dite « rapide » [B.1], et optimale lorsque la décroissance est dite « en exponentielle double » [B.2]. Le jeu étant de trouver le changement de variable adéquat [B.3] pour conférer une décroissance en exponentielle double à l'intégrande.

Bien que très efficace, cette méthode est peu répandue et dispose uniquement d'une implémentation en C++ dans la bibliothèque *Boost*, dont les variantes proposées (Tanh-Sinh, Exp-Sinh, Sinh-Sinh) ne recouvrent pas le cas de l'intégrale (B.1).
Dans premier temps, je propose [B.4] un algorithme formel permettant de calculer numérique l'intégrale d'une fonction à décroissance rapide. Il sera le socle commun à toutes les variantes de la quadrature.
Dans un second temps, à l'instar de l'adaptation de la variante Exp-Sinh proposé par Mori [87] que l'on retrouve dans la Table B.1, je propose [B.5] d'étendre le catalogue de changements de variable à des fonctions ayant à la fois une décroissance exponentielle en $+\infty$ et soit une décroissance en $\exp(-1/x)$ en 0, soit une décroissance exponentielle en $-\infty$.

B.1 Méthode des trapèzes

Avant toute chose, on dit d'une fonction g qu'elle est à décroissance rapide s'il existe α et β strictement positifs tels que

$$g(t) = \underset{\pm\infty}{\mathcal{O}} \left(\exp\left(-\alpha |t|^{\beta} \right) \right).$$

La notation $\mathcal{O}$ (« grand O ») désigne le caractère dominée d'une fonction par une autre au voisinage de a. Dire que g est dominée par f au voisinage de a, que l'on note $g = \underset{a}{\mathcal{O}}(f)$, signifie qu'il existe C strictement positif tel que $|g(t)| \leqslant C\,|f(t)|$ pour $t \to a$.

Une propriété remarquable de la méthode des trapèzes

$$\int_{-\infty}^{+\infty} g(t)\,dt \approx h \sum_{k=-\infty}^{+\infty} g(kh) \tag{B.2}$$

est son efficience pour estimer l'intégrale, sur l'axe réel $\mathbb{R}$, d'une fonction g régulière à décroissance rapide [88], où h désigne le pas de la subdivision.

B.2 Décroissance en exponentielle double

Il a été remarqué par Mori [87] que la méthode des trapèzes convergeait plus rapidement pour des fonctions ayant une *décroissance en exponentielle double*. Une fonction g est dite à décroissance en exponentielle double s'il existe α, β et γ strictement positifs tels que

$$g(t) = \underset{\pm\infty}{\mathcal{O}} \left(\exp\left(-\alpha \exp\left(\beta |t|^{\gamma} \right) \right) \right). \tag{B.3}$$

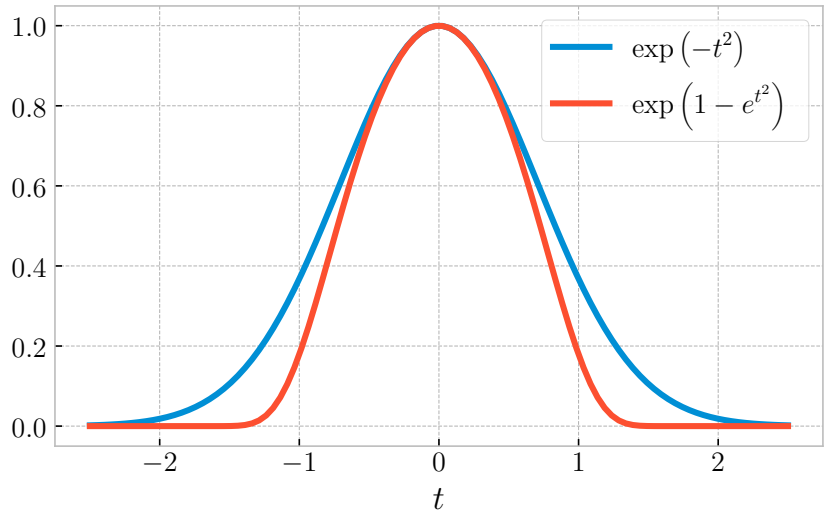

FIGURE B.1 – Exemple d'une fonction à décroissance rapide et d'une à décroissance en exponentielle double.

B.3 Transformation en exponentielle double

Le principe général étant le suivant, on prend une fonction f analytique sur $]a, b[$, pouvant être singulière sur les bords. Si on trouve un changement de variable $x = \phi(t)$ tel que $\phi(-\infty) = a$, $\phi(+\infty) = b$ et que l'intégrande $g = f \circ \phi \cdot \phi'$ est à décroissance en exponentielle double, alors pour calculer l'intégrale de f, il suffit d'appliquer la méthode des trapèzes (B.2) à g :

$$\int_a^b f(x)\,dx = \int_{-\infty}^{+\infty} \underbrace{f \circ \phi(t) \cdot \phi'(t)}_{g(t)}\,dt \approx h \sum_{k=-\infty}^{+\infty} g(kh).$$

On dira de g qu'elle est la transformée de f par ϕ. Cette transformation préserve la valeur de l'intégrale et permet de se placer dans les conditions où la méthode des trapèzes est optimale. Il reste à voir comment choisir la transformation adéquate en fonction du comportement de f au voisinage de ses bords. Initialement, MORI et SUGIHARA [89] propose les changements de variables ϕ suivants :

$$\text{Tanh-Sinh } \phi : \left| \begin{array}{ccc} \mathbb{R} & \to &]-1, +1[\\ t & \mapsto & \tanh \circ \frac{\pi}{2} \sinh t \end{array} \right. \tag{B.4}$$

$$\text{Exp-Sinh } \phi : \left| \begin{array}{ccc} \mathbb{R} & \to &]0, +\infty[\\ t & \mapsto & \exp \circ \frac{\pi}{2} \sinh t \end{array} \right.$$

$$\text{Sinh-Sinh } \phi : \left| \begin{array}{ccc} \mathbb{R} & \to &]-\infty, +\infty[\\ t & \mapsto & \sinh \circ \frac{\pi}{2} \sinh t \end{array} \right.$$

Cependant, la démarche à suivre pour obtenir les transformations adéquates est plus explicite chez Muhammad et Mori [90] et leurs principaux résultats sont résumés dans la Table B.1.

	$\int$	$f(x)$	$x = \phi(t)$	$\phi'(t)$
Tanh-Sinh	$\int_{-1}^{+1} f(x)\,dx$	$\begin{cases} \underset{+1}{\mathcal{O}}\left((1-x)^{-1-\alpha_+}\right) \\ \underset{-1}{\mathcal{O}}\left((1+x)^{-1-\alpha_-}\right) \end{cases}$	$\tanh \circ \frac{\pi}{2}\sinh t$	$\frac{\pi}{2}\cosh t / \cosh^2 \circ \frac{\pi}{2}\sinh t$
Exp-Sinh	$\int_{0}^{+\infty} f(x)\,dx$	$\begin{cases} \underset{+\infty}{\mathcal{O}}\left(x^{-1-\alpha_+}\right) \\ \underset{0}{\mathcal{O}}\left(x^{-1+\alpha_-}\right) \end{cases}$	$\exp \circ \frac{\pi}{2}\sinh t$	$\frac{\pi}{2}\cosh t \cdot \exp \circ \frac{\pi}{2}\sinh t$
	$\int_{0}^{+\infty} f_1(x)e^{-\alpha_+ x}\,dx$	$\begin{cases} \underset{+\infty}{\mathcal{O}}\left(e^{-(\alpha_+ - \varepsilon)x}\right) \\ \underset{0}{\mathcal{O}}\left(x^{-1+\alpha_-}\right) \end{cases}$	$\exp\left(t - e^{-t}\right)$	$\left(1 + e^{-t}\right) \cdot \exp\left(t - e^{-t}\right)$
Sinh-Sinh	$\int_{-\infty}^{+\infty} f(x)\,dx$	$\underset{\pm\infty}{\mathcal{O}}\left(\lvert x\rvert^{-1-\alpha_\pm}\right)$	$\sinh \circ \frac{\pi}{2}\sinh t$	$\frac{\pi}{2}\cosh t \cdot \cosh \circ \frac{\pi}{2}\sinh t$

Table B.1 – Récapitulatif des différents changements de variable traités par Muhammad et Mori [90], avec $\alpha_\pm > 0$.

On peut facilement étendre le cas Tanh-Sinh de la Table B.1 à un intervalle fini $]a, b[$ en le *mappant* sur l'intervalle $]-1, +1[$, via le changement de variable intermédiaire

$$x = \chi(u) = \frac{b-a}{2}u + \frac{b+a}{2}.$$

En le mettant en cascade avec $u = \phi(t)$,

$$\int_a^b f(x)\,dx = \int_{-1}^{+1} f \circ \chi(u) \cdot \chi'(u)\,du = \int_{-\infty}^{+\infty} \underbrace{f \circ \Phi(t) \cdot \Phi'(t)}_{g(t)}\,dt\,, \tag{B.5}$$

on obtient alors un changement de variable effectif : $x = \Phi(t) = \chi \circ \phi(t)$.

Il en va de même pour le cas Exp-Sinh de la Table B.1 que l'on peut étendre à un intervalle $]a, +\infty[$, où a est fini, en le *mappant* sur l'intervalle $]0, +\infty[$, via le changement de variable intermédiaire

$$x = \chi(u) = u + a.$$

B.4 Algorithme pour la méthode des trapèzes

Sur la base des travaux de Bailey et al. [91], je propose un algorithme, plus général, de quadrature d'une fonction g à décroissance rapide sur $\mathbb{R}$. Mais gardons à l'esprit que la fonction g qui nous intéresse est la transformée de f par Φ, où f est la fonction que l'on souhaite intégrer.

B.4.1 Troncature

D'abord, on tronque la somme (B.2) lorsque le k-ième terme atteint un critère arbitraire de petitesse $\sqrt{\varepsilon}$:

$$|g(kh)| \leqslant \sqrt{\varepsilon} \sim 10^{-8}. \tag{B.6}$$

Ici, ε désigne l'« epsilon machine », *i.e.* l'erreur relative maximale due à l'arrondi numérique. Dans le meilleur des cas, si g n'est pas singulière sur ses bords, un critère de troncature de $\sqrt{\varepsilon}$ est suffisant pour calculer l'intégrale à ε près, la précision de machine. Sinon, une précision de $\sqrt{\varepsilon}$ est atteinte. En double précision, l'epsilon machine vaut $\varepsilon \sim 10^{-16}$, d'où la valeur $\sqrt{\varepsilon} \sim 10^{-8}$.

On appelle N^D et N^G la plus petite valeur de $|k|$ pour laquelle (B.6) est *vraie*, avec respectivement $k > 1$ et $k < -1$. Ce sont les nombres de subdivisions à gauche et à droite. On obtient alors la troncature suivante :

$$\begin{aligned} h \sum_{k=-\infty}^{+\infty} g(kh) &\approx h \sum_{-N^G < k < N^D} g(kh) \\ &\approx h \Bigg(\underbrace{\sum_{k=1}^{N^G-1} g(-kh)}_{=G} + g(0) + \underbrace{\sum_{k=1}^{N^D-1} g(kh)}_{=D} \Bigg). \end{aligned} \tag{B.7}$$

On remarque la grande similitude entre les sommes de gauche G et de droite D. En définissant la fonction

$$\sigma_g(N, h) = \sum_{k=1}^{N} g(kh)\,, \tag{B.8}$$

on peut réécrire ces sommes comme $D = \sigma_g(N^D - 1, h)$ et $G = \sigma_g(N^G - 1, -h)$.

Par souci de simplification et pour alléger les notations, on posera pour la suite $N = \max(N^G, N^D)$. Cependant, dans une éventuelle implémetation, il est souhaitable de différencier N^G et N^D.

B.4.2 Récurrence

A priori, il est difficile de prescrire une valeur du pas de subdivision $h > 0$ qui convienne à tous les cas de figure. Bailey et al. [91] propose de calculer itérativement la somme à partir d'une valeur de $h_0 \leqslant 1$ donnée et la réduire de moitié à chaque niveau m de subdivision, *i.e.* de l'itération. Ce qui donne $h_m = h_0 \cdot 2^{-m}$. On choisira par la suite d'utiliser la valeur initiale $h_0 = 1$.

Pour un niveau m de subdivision donné, la somme s'écrit :

$$S_m = h_m \sum_{-N_m < k < N_m} g(kh_m).$$

De plus, l'extrémité $\pm N_m h_m$ est invariante au cours des subdivisions successives. Il vient alors $N_m h_m = N_0 h_0 = N$, que l'on traduit par $N_m = 2^m N$, avec $h_m = 2^{-m}$. Ce

qui donne lieu aux relations de récurrences :

$$\begin{cases} h_{m+1} = \dfrac{1}{2} h_m & \text{(récurrence)} \\ h_0 = 1 & \text{(initialisation)} \end{cases}, \tag{B.9}$$

$$\begin{cases} N_{m+1} = 2N_m & \text{(récurrence)} \\ N_0 = N & \text{(initialisation)} \end{cases}. \tag{B.10}$$

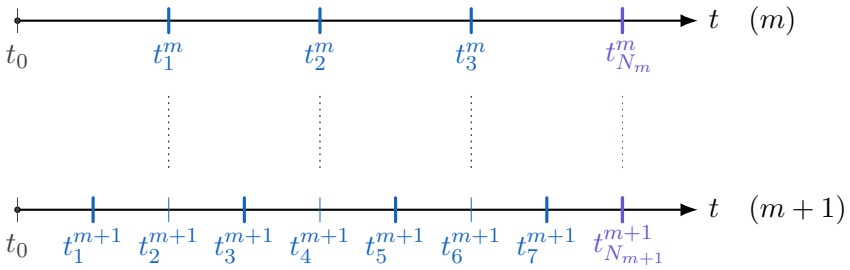

FIGURE B.2 – Subdivision successive de l'intervalle d'intégration. L'axe des t représente le domaine d'intégration que l'on subdivise en deux lors du passage d'une itération m à la $m+1$ suivante. En toute généralité, t_0 est constant d'une itération à l'autre et $t_k^m = t_0 + kh_m$. Jusqu'à maintenant, on a considéré le cas $t_0 = 0$. Néanmoins, il peut être fixé à une valeur plus adaptée *a priori*.

Conformément au schéma de la Figure B.2, lors de la subdivision, on constate que la totalité des points de l'itération m est présente dans l'itération $m+1$, $t_k^m = t_{2k}^{m+1}$. Ce qui permet d'exhiber la relation de récurrence :

$$\begin{cases} S_{m+1} = \dfrac{1}{2} S_m + h_{m+1} \displaystyle\sum_{-N_{m+1}<k<N_{m+1},\, k \text{ impair}} g(t_k^{m+1}) & \text{(récurrence)} \\ S_0 = \displaystyle\sum_{k=1-N}^{N-1} g(t_k^0) & \text{(initialisation)} \end{cases}, \tag{B.11}$$

en se contentant d'évaluer la fonction g pour les valeurs de $t_k^m = t_0 + kh_m$ où k est impair, sauf lors de l'initialisation.

B.4.3 Cas pathologique du dépassement de capacité

Définition mathématique

Mathématiquement, le changement de variable Tanh-Sinh (B.4) est à valeur dans $]-1,+1[$. Ce qui ne pose aucun problème si jamais la fonction que l'on cherche à intégrer est singulière en ± 1. Prenons l'exemple de la dérivée de arcsin $/\pi$:

$$f : \left| \begin{array}{ccc}]-1,+1[& \to & \mathbb{R}_+ \\ x & \mapsto & \left(\pi\sqrt{1-x^2}\right)^{-1} \end{array} \right. .$$

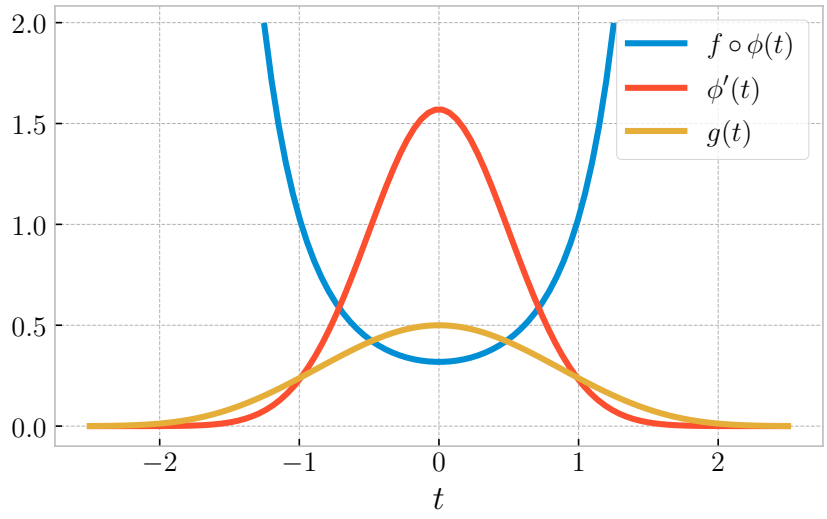

FIGURE B.3 – Exemple d'une fonction, $f(x) = \left(\pi\sqrt{1-x^2}\right)^{-1}$, analytique sur $]-1,+1[$ possédant deux singularités aux bords. En effectuant le changement de variable Tanh-Sinh (B.4), la fonction ϕ' porte la décroissance en exponentielle double. La composée $g = f \circ \phi \cdot \phi'$ est alors une fonction à décroissance en exponentielle double et de ce fait un bon candidat pour la méthode des trapèzes.

La composée $f \circ \phi$ demeure toujours à valeur dans $\mathbb{R}_+$ et la croissance en exponentielle double est portée par $\phi'(t) = \frac{\pi}{2}\cosh t/\cosh^2 \circ \frac{\pi}{2}\sinh t$. Ce qui permet à la fonction $g = f \circ \phi \cdot \phi'$ d'être intégrable par la méthode des trapèzes.

Définition numérique

Cependant ! Numériquement, les nombres réels sont représentés par un nombre fini de bits en mémoire. On appelle cette représentation *nombre à virgule flottante* ou plus simplement *flottant*, dont leur arithmétique est décrite par « IEEE Standard for Floating-Point Arithmetic » [92]. Les flottants sont généralement encodés sur 32 ou 64-bits. Dans ces deux formats, les plus grands nombres représentables (en valeur absolue) sont, respectivement, $\sim 10^{38}$ et $\sim 10^{308}$. Tout nombre à virgule flottante résultant d'une opération qui excéderait ce maximum se voit attribuer la valeur $\pm\infty$. On parle alors de dépassement de capacité ou *overflow* en anglais.

On peut se représenter un nombre à virgule flottante comme étant un réel appartenant à la droite réelle achevée que l'on note $\overline{\mathbb{R}}$, où $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\} = [-\infty, +\infty]$. Numériquement, le changement de variable Tanh-Sinh

$$\phi : \left| \begin{array}{ccc} \overline{\mathbb{R}} & \to & [-1,+1] \\ t & \mapsto & \tanh \circ \frac{\pi}{2}\sinh t \end{array} \right.$$

est à valeur dans $[-1,+1]$. Ce qui signifie qu'en cas de dépassement de capacité, la

fonction

$$f : \left| \begin{array}{ccc} [-1,+1] & \to & \overline{\mathbb{R}}_+ \\ x & \mapsto & \left(\pi\sqrt{1-x^2}\right)^{-1} \end{array} \right.$$

serait évaluée en ± 1 qui résulterait dans notre cas à $f(\pm 1) = +\infty$. Il suffit d'un seul dépassement de capacité lors de l'évaluation de f pour que s'en suive une cascade de dépassement de capacité tout le long de la récurrence (B.11), aboutissant à l'erroné résultat que l'intégrale de f serait infinie.

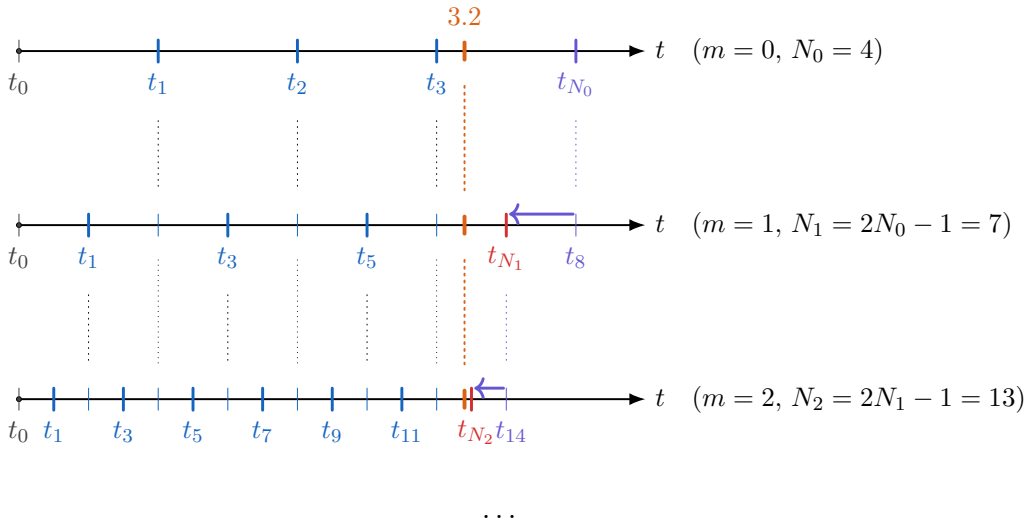

FIGURE B.4 – Exemple de subdivision *restrictive* successive de l'intervalle d'intégration avec dépassement de capacité en $t \approx 3.2$.

Palliatif

En double précision, *i.e.* flottant 64-bits, le dépassement de capacité a lieu vers $t \approx 3.2$. Il convient alors de mettre en place une stratégie de subdivision *restrictive*, en deux points, visant à restreindre le domaine d'intégration aux valeurs pour lesquelles la fonction est évaluable sans aboutir à un dépassement de capacité.

Premièrement, il faut étendre le critère de troncature (B.6) pour évincer les termes divergeants,

$$|g(kh_0)| \leqslant \sqrt{\varepsilon} \sim 10^{-8} \quad \text{ou} \quad |g(kh_0)| = +\infty.$$

Deuxièmement, il faut vérifier, lors de la récurrence, que le dernier terme des sommes de gauche et de droite de (B.11) soient finis pour les comptabiliser, sinon il faut décrémenter N_m en conséquence.

$$\text{si} \quad |g(t_0 \pm (N_m - 1)h_m)| = +\infty \quad \text{alors} \quad N_m \leftarrow N_m - 1 \tag{B.12}$$

D'un point de vue algorithmique, il apparaît indispensable de différencier N_m^G et N_m^D.

B.4.4 Critère de convergence

Il reste à apporter un critère de convergence (B.13) pour s'assurer que l'on ait choisi un N suffisamment grand ; et (B.14) pour interrompre les subdivisions successives de l'intervalle d'intégration.

En dépit du fait que le critère de petitesse soit vérifié, il apparaît nécessaire pour certaines fonctions de s'assurer qu'elles décroissent bien en valeur absolue. En définitive, on étend le critère de troncature comme suit :

$$\Big(|g(kh_0)| \leqslant \sqrt{\varepsilon} \sim 10^{-8} \text{ et } |g(kh_0)| \leqslant |g((k-1)h_0)|\Big) \text{ ou } |g(kh_0)| = +\infty. \quad \text{(B.13)}$$

Enfin, il semble raisonnable d'arrêter la récurrence, soit lorsque le progrès relatif de deux itérations consécutives atteint une précision de $\sqrt{\varepsilon}$, soit au bout d'un nombre prédéfini d'itérations,

$$\frac{S_m - S_{m-1}}{A_m} \leqslant \sqrt{\varepsilon} \sim 10^{-8} \quad \text{ou} \quad m = 6. \quad \text{(B.14)}$$

A_m désigne l'approximation itérative de la norme 1 de g, notée $\|g\|_1$,

$$A_m = h_m \sum_{-N_m<k<N_m} |g(kh_m)|\,,$$

et suit la même récurrence qu'en (B.11)

$$\begin{cases} A_{m+1} = \dfrac{1}{2}A_m + h_{m+1} \displaystyle\sum_{-N_{m+1}<k<N_{m+1},\, k \text{ impair}} |g(t_k^{m+1})| & \text{(récurrence)} \\ A_0 = \displaystyle\sum_{k=1-N}^{N-1} |g(t_k^0)| & \text{(initialisation)} \end{cases}. \quad \text{(B.15)}$$

En pratique sur un ensemble de fonctions test, le niveau maximal de subdivisions atteint est $m = 4$. C'est pourquoi on propose, ici, de fixer un garde-fou à $m = 6$.

B.4.5 Heuristique de t_0

Pour des fonctions f à décroissance en exponentielle *simple*, avec leur changement de variables ϕ adéquat, il s'avère qu'au voisinage de zéro, on se trouve dans l'une des queues de la fonction $g = f \circ \phi \cdot \phi'$, où celle-ci est nulle après arrondi numérique.

Le schéma de la subdivision du domaine d'intégration de la Figure B.2 est général et t_0 est la valeur où l'on scinde la somme,

$$h \sum_{k=-\infty}^{+\infty} g(kh) \approx h \left(\sum_{k=1}^{N^G-1} g(t_0 - kh) + g(t_0) + \sum_{k=1}^{N^D-1} g(t_0 + kh) \right). \quad \text{(B.16)}$$

L'abscisse t_0 est constant d'une itération à l'autre. Jusqu'à maintenant, on a considéré le cas $t_0 = 0$.

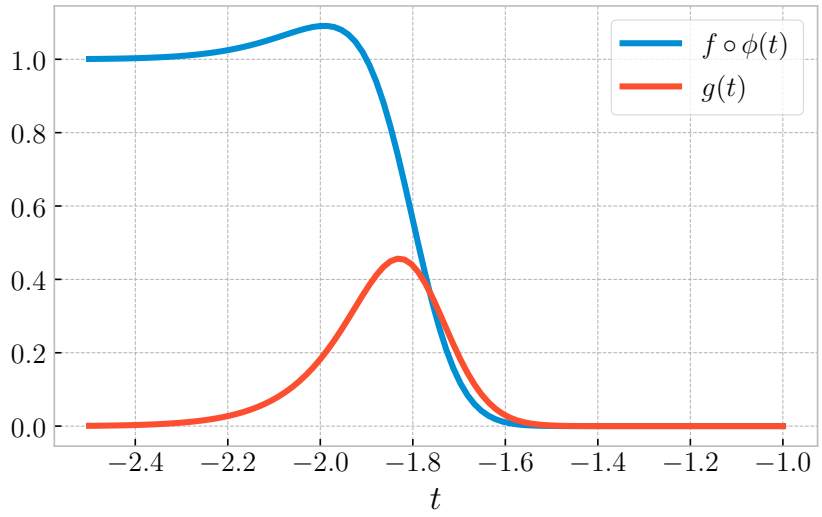

FIGURE B.5 – Exemple de fonction $f(x) = (1+x)^2/(1+x^2)^2 \exp(-x)$ à décroissance en exponentielle simple en $x \to +\infty$. La transformée $g = f \circ \phi \cdot \phi'$ est nulle au voisinage de 0, avec notamment $g(0) = g(\pm 1) = 0$. On a ici : $\phi(t) = b(1 - \exp(-e^t))$, $b = 2000$ et $h_0 = 0$.

Or, dans le cas où $g(0) = g(\pm 1) = 0$, la condition (B.13) est *vraie* et l'algorithme retient $N = 1$, conduisant à une intégrale nulle (B.7). Ce qui est *faux* !

Néanmoins, elle peut être fixée à une valeur plus adaptée *a priori*, soit par une heuristique ; soit par une valeur connue où la fonction ne s'annule pas, *i.e.* le maximum de celle-ci, $t_0 = \underset{t}{\operatorname{argmax}}\, g$. Dans le cas $g = f \circ \phi \cdot \phi'$, on peut écrire $t_0 \approx \phi^{-1}(x_0)$ avec $x_0 = \underset{x}{\operatorname{argmax}}\, f$, si le changement de variable ϕ est inversible.

L'heuristique que je propose, consiste à attribuer alternativement à t_0 les valeurs de la suite $(-np)_{n\geqslant 0} = (0, -p, -2p, \dots)$ et de la suite $(np)_{n\geqslant 1} = (p, 2p, \dots)$ [1] jusqu'à satisfaire la condition,

$$g(t_0) \neq 0 \quad \text{ou} \quad n = 2\,, \tag{B.17}$$

où $p \geqslant h_0$ désigne l'amplitude des pas lors de l'exploration du domaine d'intégration. On utilisera par la suite la valeur $p = 3$. La profondeur d'exploration $n = 2$ permet d'imposer une limite arbitraire au cas où la condition $g(t_0) \neq 0$ n'est jamais atteinte.

B.4.6 Synthèse

En résumé, l'algorithme se fonde sur :

1. l'heuristique de t_0 (B.17)
2. la réécriture de la somme tronquée en trois termes (B.16) avec notamment les sommes de gauche G et de droite D (B.7), grâce à la fonction σ (B.8)
3. les relations de récurrence :

1. Il en résulte la suite $(0, p, -p, 2p, -2p, \dots)$.

(a) du pas de la subdivision h (B.9)

(b) du nombre de subdivisions à gauche N^G et à droite N^D (B.10)

(c) de l'intégrale de g (B.11), ainsi que de sa norme 1 (B.15), en prenant en compte, lors du calcul, la correction sur le nombre de subdivisions (B.12)

4. les critères de convergence :

(a) de la troncature (B.13)

(b) de la subdivision successive de l'intervalle d'intégration (B.14)

B.5 Extension du catalogue de changements de variable

Jusqu'à présent, on s'est focalisé sur l'établissement d'un algorithme [B.4] générique à l'aide des changements de variables [B.3] proposés par Mori. Dans cette section, on se propose d'introduire des nouveaux changements de variables Exp-Exp [B.5.1], Exp [B.5.2] et Sinh [B.5.3] pour des fonctions f possédant déjà une décroissance exponentielle en $\pm\infty$ ou en 0. Dans toute la suite, on considère $\alpha_\pm > 0$.

B.5.1 Exp-Exp

Bornes canoniques : $]0, +\infty[$

On se propose d'étendre le cas 4.b de Muhammad et Mori [90], aux fonctions f sur l'intervalle $]0, +\infty[$, possédant déjà une décroissance exponentielle en $+\infty$. On considère le changement de variable qui *mappe* l'intervalle $]0, +\infty[$ sur $]-\infty, +\infty[$.

$$\int_0^{+\infty} f_1(x) e^{-\alpha x}\, dx\,, \quad f(x) = \begin{cases} \underset{+\infty}{\mathcal{O}}\left(e^{-(\alpha_+ - \varepsilon)x}\right) \\ \underset{0}{\mathcal{O}}\left(x^{-1+\alpha_-}\right) \end{cases} \tag{B.18}$$

L'auteur prescrit le changement de variable,

$$x = \phi(t) = \ln\left(1 + \exp \circ \frac{\pi}{2} \sinh t\right)\,,$$

car inversible analytiquement. On prendra le soin de reconnaître la fonction $\mathrm{softplus}(x) = \ln(1 + \exp(x))$, couramment utilisée comme fonction d'activation dans les réseaux de neurones. Dans cette écriture, la fonction softplus est sujette aux dépassements de capacité à cause du calcul numérique de l'exponentielle. C'est pourquoi, je préconise de l'utiliser sous sa forme

$$\mathrm{softplus}(x) = \ln\left(1 + e^{-|x|}\right) + \max(x, 0)\,,$$

où le terme contenant l'exponentielle $\ln\left(1 + e^{-|x|}\right)$ est majoré par $\ln 2$.

Cependant ! Par souci d'optimisation, je vais utiliser le changement de variable

$$\begin{aligned} x &= \phi(t) = \exp\left(t - e^{-t}\right) \\ t &= \phi^{-1}(x) = W\left(\frac{1}{x}\right) + \ln x\,, \end{aligned} \tag{B.19}$$

dont le temps d'exécution est ~ 2 fois plus court. W désigne la fonction W de Lambert, réciproque de la fonction $t \mapsto t \cdot e^t$, qui, quant à elle, n'est pas représentable simplement [93]. On aura besoin de cette fonction, ponctuellement, lorsque l'on voudra estimer la valeur de t_0. C'est pourquoi, je propose l'approximation par morceaux pour $x > 0$

$$W(\frac{1}{x}) + \ln x = \begin{cases} \phi_1^{-1}(x) = \dfrac{\sum_{i=1}^{4} a_i^1 (x - x_1)}{1 + \sum_{i=1}^{3} b_i^1 (x - x_1)} & x < e^{-1} \\ \phi_2^{-1}(x) = \dfrac{\sum_{i=0}^{4} a_i^2 (x - x_2)}{1 + \sum_{i=1}^{3} b_i^2 (x - x_2)} & x < 5 \\ \phi_3^{-1}(x) = \dfrac{1}{x \exp(1/x)} + \ln x & \text{sinon} \end{cases},$$

où ϕ_1^{-1} et ϕ_2^{-1} sont des approximants de Padé dont les coefficients sont résumés dans la Table B.2.

i	0	1	2	3	4
a_i^1		1.35914	4.56488	3.52982	0.24866
b_i^1		4.37801	5.44349	1.70222	
a_i^2	0.567143	1.55702	1.21562	0.275018	0.00515443
b_i^2		1.62026	0.753076	0.0911261	

TABLE B.2 – Coefficients des approximants de Padé ϕ_1^{-1} et ϕ_2^{-1}.

En combinant les propriétés de décroissance de la fonction (B.18) avec le changement de variable (B.19) et de sa dérivée

$$\phi'(t) = \left(1 + e^{-t}\right) \cdot \exp\left(t - e^{-t}\right),$$

on obtient bien la propriété (B.3) de décroissance en exponentielle double de l'intégrande :

$$f \circ \phi(t) \cdot \phi'(t) = \underset{\pm\infty}{\mathcal{O}} \left(\exp\left(-(\alpha_\pm - \varepsilon) e^{|t|}\right)\right).$$

Bornes quelconques : $-\infty < a < b < +\infty$

On souhaite, à présent, généraliser aux cas de bornes quelconques.

$$\int_a^b f_1(x) e^{-\alpha x}\, dx, \quad f(x) = \begin{cases} \underset{+\infty}{\mathcal{O}} \left(e^{-(\alpha_+ - \varepsilon)x}\right) \\ \underset{a}{\mathcal{O}} \left((x-a)^{-1+\alpha_-}\right) \end{cases}$$

Là où il est facile d'effectuer une translation, en posant $x = \chi(u) = u + a$, pour se ramener au cas d'une borne supérieure quelconque $b - a > 0$,

$$\int_0^{b-a} f_1 \circ \chi(u) e^{-\alpha \chi(u)}\, du, \quad f \circ \chi(u) = \begin{cases} \underset{+\infty}{\mathcal{O}} \left(e^{-(\alpha_+ - \varepsilon)u}\right) \\ \underset{0}{\mathcal{O}} \left(u^{-1+\alpha_-}\right) \end{cases},$$

il est moins aisé de calculer l'intégrale pour une borne supérieure quelconque. Muhammad et Mori [90] utilise la formule proposé par Haber [94] qui repose sur la méthode Sinc. Cette méthode consiste à développer une fonction g à l'aide de la translatée du sinus cardinal

$$g(t) = \sum_{k=-\infty}^{+\infty} g(kh)\,\mathrm{sinc}\left(\frac{t}{h} - k\right), \tag{B.20}$$

où $h > 0$ est la largeur du pas de subdivision. Le sinus cardinal étant, ici, définit comme :

$$\mathrm{sinc}(t) = \frac{\sin(\pi t)}{\pi t}.$$

En injectant (B.20) sous le signe intégral de la méthode des trapèzes (B.2), on obtient,

$$\int_{-\infty}^{\tau} g(t)\,dt \approx h \sum_{k=-\infty}^{+\infty} g(kh)\left(\frac{1}{2} + \frac{1}{\pi}\,\mathrm{Si}\left(\pi\frac{\tau}{h} - \pi k\right)\right),$$

où Si désigne le sinus intégral

$$\mathrm{Si}(x) = \int_0^x \frac{\sin t}{t}\,dt\,.$$

Muhammad et Mori [90] précise que $\left(\frac{1}{2} + \frac{1}{\pi}\,\mathrm{Si}\left(\pi\frac{\tau}{h} - \pi k\right)\right)$ joue le rôle de fonction de coupure qui va évincer les termes supérieures à τ. On peut la voir comme l'analogue régulière de la fonction de Heaviside. Dans notre cas, on a $g = f \circ \phi$ et $\tau = \phi^{-1}(b)$. Ce qui nécessite que le changement de variable ϕ soit inversible.

Toutefois, le fait que l'argument du sinus intégral s'écrive comme $\pi(\tau - kh)/h$, empêche d'écrire $g(kh)\left(\frac{1}{2} + \frac{1}{\pi}\,\mathrm{Si}\left(\pi\frac{\tau}{h} - \pi k\right)\right)$ comme une fonction qui ne dépendrait uniquement de la variable $t = kh$. Il en résulte que l'on ne peut plus écrire la relation de récurrence $f(t_k^m) = f(t_{2k}^{m+1})$, déduite de la Figure B.2, avec $t_k^m = kh_m$. Ce qui signifie que l'on ne pourrait plus utiliser l'approche itérative développée dans la section B.4. C'est pourquoi, je vais emprunter une approche fondée sur un changement de variable intermédiaire

$$\begin{aligned} x &= \chi(u) = b - (b-a)\exp\left(-\frac{u}{b-a}\right) \\ u &= \chi^{-1}(x) = -(b-a)\ln\frac{b-x}{b-a}\,, \end{aligned} \tag{B.21}$$

pour nous ramener au cas des bornes canoniques avec un changement de variable effectif

$$\begin{aligned} x &= \Phi(t) = \chi \circ \phi(t) \\ t &= \Phi^{-1}(x) = \phi^{-1} \circ \chi^{-1}(x)\,, \end{aligned} \tag{B.22}$$

qui vérifie la propriété (B.3)

$$f \circ \Phi(t) \cdot \Phi'(t) = \underset{\pm\infty}{\mathcal{O}}\left(\exp\left(-(\alpha_\pm - \varepsilon)e^{|t|}\right)\right),$$

avec les dérivées

$$\chi'(u) = \exp\left(\frac{u}{b-a}\right)$$
$$\Phi'(t) = \chi' \circ \phi(t) \cdot \phi'(t) = \exp\left(\frac{1}{b-a}e^t + t\right).$$

En résumé, on peut écrire

$$\int_a^b f(x)\,dx = \int_0^{+\infty} f \circ \chi(u) \cdot \chi'(u)\,du = \int_{-\infty}^{+\infty} f \circ \Phi(t) \cdot \Phi'(t)\,dt\,.$$

Remarque. *Le changement de variable effectif Φ proposé ici est similaire au changement de variable Tanh-Sinh (B.5) qui aurait pu être utilisé à la place. Il est, néanmoins, plus rapide d'un facteur ~ 1.5.*

B.5.2 Exp

Bornes canoniques : $]0, +\infty[$

On s'intéresse aux fonctions f sur l'intervalle $]0, +\infty[$, possédant déjà une décroissance exponentielle à la fois en 0 et en $+\infty$.

$$\int_0^{+\infty} f_1(x)e^{-\alpha_+ x}e^{-\alpha_- \frac{1}{x}}\,dx\,, \quad f(x) = \begin{cases} \underset{+\infty}{\mathcal{O}}\left(e^{-(\alpha_+ - \varepsilon)x}\right) \\ \underset{0}{\mathcal{O}}\left(e^{-(\alpha_- - \varepsilon)\frac{1}{x}}\right) \end{cases}$$

On considère le changement de variable qui *mappe* l'intervalle $]0, +\infty[$ sur $]-\infty, +\infty[$.

$$x = \phi(t) = \exp t$$
$$t = \phi^{-1}(x) = \ln x$$

Il confère la convergence attendue en double exponentielle :

$$f \circ \phi(t) \cdot \phi'(t) = \underset{\pm\infty}{\mathcal{O}}\left(\exp\left(-(\alpha_\pm - \varepsilon)e^{|t|}\right)\right).$$

Bornes quelconques : $-\infty < a < b < +\infty$

On utilise la même démarche que précédemment pour se ramener du cas des bornes quelconques aux bornes canoniques.

$$\int_a^b f_1(x)e^{-\alpha_+ x}e^{-\alpha_- \frac{1}{x-a}}\,dx\,, \quad f(x) = \begin{cases} \underset{+\infty}{\mathcal{O}}\left(e^{-(\alpha_+ - \varepsilon)x}\right) \\ \underset{a}{\mathcal{O}}\left(e^{-(\alpha_- - \varepsilon)\frac{1}{x-a}}\right) \end{cases}$$

On réutilise le même changement de variable effectif Φ (B.22), ainsi que le même changement de variables intermédiaire χ (B.21)

$$f \circ \Phi(t) = \begin{cases} \underset{+\infty}{\mathcal{O}}(1) \\ \underset{-\infty}{\mathcal{O}}\left(\exp\left(-(\alpha_- - \varepsilon)e^{|t|}\right)\right) \end{cases} \qquad \Phi'(t) = \begin{cases} \underset{+\infty}{\mathcal{O}}\left(\exp\left(-\frac{1-\varepsilon}{b-a}e^t\right)\right) \\ \underset{-\infty}{\mathcal{O}}(1) \end{cases}$$

qui confère la convergence attendue en double exponentielle :

$$f \circ \Phi(t) \cdot \Phi'(t) = \begin{cases} \underset{+\infty}{\mathcal{O}} \left(\exp\left(-\dfrac{1-\varepsilon}{b-a} e^{t} \right) \right) \\ \underset{-\infty}{\mathcal{O}} \left(\exp\left(-(\alpha_- - \varepsilon) e^{|t|} \right) \right) \end{cases} .$$

B.5.3 Sinh

Bornes canoniques : $]-\infty, +\infty[$

On s'intéresse aux fonctions f sur l'intervalle $]-\infty, +\infty[$, possédant déjà une décroissance exponentielle à la fois en $-\infty$ et en $+\infty$.

$$\int_{-\infty}^{+\infty} f_1(x) e^{-\alpha|x|}\, dx\,, \quad f(x) = \underset{\pm\infty}{\mathcal{O}} \left(e^{-(\alpha-\varepsilon)|x|} \right)$$

On considère le changement de variable qui *mappe* l'intervalle $]-\infty, +\infty[$ sur lui-même.

$$x = \phi(t) = \frac{\pi}{2} \sinh t$$
$$t = \phi^{-1}(x) = \operatorname{asinh}\left(\frac{2}{\pi} x \right)$$

Il confère la convergence attendue en double exponentielle :

$$f \circ \phi(t) \cdot \phi'(t) = \underset{\pm\infty}{\mathcal{O}} \left(\exp\left(-(\alpha - \varepsilon) e^{|t|} \right) \right).$$

Bornes quelconques : $-\infty < a < b < +\infty$

L'extension des bornes canoniques aux bornes quelconques est simple.

$$\int_{a}^{b} f_1(x) e^{-\alpha|x|}\, dx\,, \quad f(x) = \underset{\pm\infty}{\mathcal{O}} \left(e^{-(\alpha-\varepsilon)|x|} \right)$$

Il suffit de remarquer que f est bornée au voisinage de ses bornes

$$f(x) = \begin{cases} \underset{b}{\mathcal{O}}(1) \\ \underset{a}{\mathcal{O}}(1) \end{cases} ,$$

ce qui correspond au cas Tanh-Sinh (B.5).

B.5.4 Bibliothèque *dequad*

J'ai crée une bibliothèque, dans le langage Rust, nommée *dequad* pour l'anglais *Double Exponential QUADrature*, mettant en œuvre l'algorithme [B.4.6] avec toutes les variantes résumées dans les tables B.3 et B.4. L'Extrait B.1 vise à calculer la valeur

de l'intégrale $\int_{-1}^{+1} \left(\pi\sqrt{1-x^2}\right)^{-1} dx$ et de vérifier qu'elle est précise à $\sqrt{\varepsilon} \sim 10^{-8}$ près en double précision.

```rust
use dequad::tanh_sinh::Integrator;
use std::f64::consts::FRAC_1_PI;
use std::f64::EPSILON;

let integrate = Integrator::default()
    .integrate(-1.0, 1.0, |x: f64| {
        FRAC_1_PI * (1.0 - x * x).sqrt().recip()
    });
let truth = 1.0;

assert!((integrate.integral - truth).abs() <= EPSILON.sqrt());
```

Extrait B.1: Exemple d'utilisation de la bibliothèque *dequad* pour le calcul de l'intégrale $\int_{-1}^{+1} \left(\pi\sqrt{1-x^2}\right)^{-1} dx$.

Quant à l'Extrait B.2, il illustre l'utilisation faite de la bibliothèque pour la normalisation de la distribution de probabilité du rayon de giration dans le code d'inférence bayésienne. On peut également voir comment s'aider de la position du maximum x_0 de la distribution de probabilité pour découper l'intégrale autour de la valeur $t_0 = \phi^{-1}(x_0)$, avec dans le cas présent $\phi = \exp$.

```rust
let integral = dequad::exp::Integrator::with_heuristic_start(x0)
    .integrate(0.0, std::f64::INFINITY, pdf)
    .integral;
```

Extrait B.2: Exemple d'utilisation de la bibliothèque *dequad* dans le code d'inférence bayésienne pour le calcul de la norme de la distribution de probabilité du rayon de giration.

B.6 Tables récapitulatives

	$\int$	f	ϕ	ϕ'
Tanh-Sinh	$\int_{-1}^{+1} f(x)\,dx$	$\begin{cases} \underset{+1}{\mathcal{O}}\left((1-x)^{-1-\alpha_+}\right) \\ \underset{-1}{\mathcal{O}}\left((1+x)^{-1-\alpha_-}\right) \end{cases}$	$\tanh \circ \frac{\pi}{2}\sinh t$	$\frac{\pi}{2}\cosh t / \cosh^2 \circ \frac{\pi}{2}\sinh t$
Exp-Sinh	$\int_{0}^{+\infty} f(x)\,dx$	$\begin{cases} \underset{+\infty}{\mathcal{O}}\left(x^{-1-\alpha_+}\right) \\ \underset{0}{\mathcal{O}}\left(x^{-1+\alpha_-}\right) \end{cases}$	$\exp \circ \frac{\pi}{2}\sinh t$	$\frac{\pi}{2}\cosh t \cdot \exp \circ \frac{\pi}{2}\sinh t$
Sinh-Sinh	$\int_{-\infty}^{+\infty} f(x)\,dx$	$\underset{\pm\infty}{\mathcal{O}}\left(\lvert x\rvert^{-1-\alpha_\pm}\right)$	$\sinh \circ \frac{\pi}{2}\sinh t$	$\frac{\pi}{2}\cosh t \cdot \cosh \circ \frac{\pi}{2}\sinh t$
Exp-Exp	$\int_{0}^{+\infty} f_1(x)e^{-\alpha_+ x}\,dx$	$\begin{cases} \underset{+\infty}{\mathcal{O}}\left(e^{-(\alpha_+-\varepsilon)x}\right) \\ \underset{0}{\mathcal{O}}\left(x^{-1+\alpha_-}\right) \end{cases}$	$\exp\left(t-e^{-t}\right)$	$\left(1+e^{-t}\right)\cdot\exp\left(t-e^{-t}\right)$
Exp	$\int_{0}^{+\infty} f_1(x)e^{-\alpha_+ x}e^{-\alpha_-\frac{1}{x}}\,dx$	$\begin{cases} \underset{+\infty}{\mathcal{O}}\left(e^{-(\alpha_+-\varepsilon)x}\right) \\ \underset{0}{\mathcal{O}}\left(e^{-(\alpha_--\varepsilon)\frac{1}{x}}\right) \end{cases}$	$\exp t$	$\exp t$
Sinh	$\int_{-\infty}^{+\infty} f_1(x)e^{-\alpha\lvert x\rvert}\,dx$	$\underset{\pm\infty}{\mathcal{O}}\left(e^{-(\alpha-\varepsilon)\lvert x\rvert}\right)$	$\frac{\pi}{2}\sinh t$	$\frac{\pi}{2}\cosh t$

Table B.3 – Récapitulatif des changements de variable sur leur intervalle canonique.

	$\int$	$f(x)$	$x = \chi(u)$	$u = \phi(t)$
Tanh-Sinh	$\int_a^b f(x)\,dx$	$\begin{cases} \underset{b}{\mathcal{O}}\left((b-x)^{-1-\alpha_+}\right) \\ \underset{a}{\mathcal{O}}\left((x-a)^{-1-\alpha_-}\right) \end{cases}$	$\frac{b-a}{2}u + \frac{b+a}{2}$	$\tanh \circ \frac{\pi}{2}\sinh t$
Exp-Sinh	$\int_a^{+\infty} f(x)\,dx$	$\begin{cases} \underset{+\infty}{\mathcal{O}}\left(x^{-1-\alpha_+}\right) \\ \underset{a}{\mathcal{O}}\left((x-a)^{-1+\alpha_-}\right) \end{cases}$	$u + a$	$\exp \circ \frac{\pi}{2}\sinh t$
Sinh-Sinh $\hookrightarrow$	Tanh-Sinh ou Exp-Sinh			
Exp-Exp	$\int_a^b f_1(x)e^{-\alpha_+ x}\,dx$	$\begin{cases} \underset{+\infty}{\mathcal{O}}\left(e^{-(\alpha_+-\varepsilon)x}\right) \\ \underset{a}{\mathcal{O}}\left((x-a)^{-1+\alpha_-}\right) \end{cases}$	$b-(b-a)\exp\left(-\frac{u}{b-a}\right)$	$\exp\left(t-e^{-t}\right)$
Exp	$\int_a^b f_1(x)e^{-\alpha_+ x}e^{-\alpha_- \frac{1}{x-a}}\,dx$	$\begin{cases} \underset{+\infty}{\mathcal{O}}\left(e^{-(\alpha_+-\varepsilon)x}\right) \\ \underset{a}{\mathcal{O}}\left(e^{-(\alpha_- -\varepsilon)\frac{1}{x-a}}\right) \end{cases}$	$b-(b-a)\exp\left(-\frac{u}{b-a}\right)$	$\exp t$
Sinh $\hookrightarrow$	Tanh-Sinh ou Exp-Exp			

TABLE B.4 – Récapitulatif des changements de variable sur un intervalle quelconque : $-\infty < a < b < +\infty$.

Bibliographie

[1] A. N. Boettiger et al., « Super-resolution imaging reveals distinct chromatin folding for different epigenetic states », Nature **529**, 418–422 (2016) (cf. p. 10–13, 15, 25, 31, 77–79, 83, 86–89, 91).

[2] J. D. Watson et F. H. C. Crick, « Molecular structure of nucleic acids : a structure for deoxyribose nucleic acid », Nature **171**, 737–738 (1953) (cf. p. 1).

[3] W. Commons, *File :dna structure and bases fr.svg — wikimedia commons, the free media repository*, 2017 (cf. p. 1).

[4] eBiologie.fr, *La cellule eucaryote, schéma simplifié*, https://www.ebiologie.fr/cours/s/305/schema-cellule-eucaryote (cf. p. 2).

[5] D. E. Olins et A. L. Olins, « Chromatin history : our view from the bridge », Nature Reviews Molecular Cell Biology **4**, 809–814 (2003) (cf. p. 2).

[6] P. Takizawa, *Euchromatin and heterochromatin*, http://medcell.med.yale.edu/histology/cell_lab/euchromatin_and_heterochromatin.php (cf. p. 2, 3).

[7] A. P. Wolffe et J. J. Hayes, « Chromatin disruption and modification », Nucleic Acids Research **27**, 711–720 (1999) (cf. p. 3).

[8] MBInfo, *What are nucleosomes ?*, https://www.mechanobio.info/genome-regulation/what-are-nucleosomes/ (cf. p. 4).

[9] R. Cortini et al., « The physics of epigenetics », Reviews of Modern Physics **88** (2016) 10.1103/revmodphys.88.025002 (cf. p. 4–6).

[10] S. T. Milner, « Polymer brushes », Science **251**, 905–914 (1991) (cf. p. 6).

[11] J.-M. Arbona et al., « Inferring the physical properties of yeast chromatin through bayesian analysis of whole nucleus simulations », Genome Biology **18** (2017) 10.1186/s13059-017-1199-x (cf. p. 6).

[12] T. Sexton et al., « Three-dimensional folding and functional organization principles of the drosophila genome », Cell **148**, 458–472 (2012) (cf. p. 6, 8).

[13] G. Fudenberg et al., « Formation of chromosomal domains by loop extrusion », Cell Reports **15**, 2038–2049 (2016) (cf. p. 6).

[14] G. Fudenberg et al., « Emerging evidence of chromosome folding by loop extrusion », Cold Spring Harbor Symposia on Quantitative Biology **82**, 45–55 (2017) (cf. p. 6).

[15] A. Routh et al., « Nucleosome repeat length and linker histone stoichiometry determine chromatin fiber structure », Proceedings of the National Academy of Sciences **105**, 8872–8877 (2008) (cf. p. 6).

[16] K. Maeshima et al., « Chromatin as dynamic 10-nm fibers », Chromosoma **123**, 225–237 (2014) (cf. p. 6).

[17] G. Felsenfeld et M. Groudine, « Controlling the double helix », Nature **421**, 448–453 (2003) (cf. p. 7).

[18] T. Sexton et G. Cavalli, « The role of chromosome domains in shaping the functional genome », Cell **160**, 1049–1059 (2015) (cf. p. 7).

[19]J. Dekker, « Capturing chromosome conformation », Science **295**, 1306–1311 (2002) (cf. p. 7).

[20]J. Dostie et al., « Chromosome conformation capture carbon copy (5c) : a massively parallel solution for mapping interactions between genomic elements », Genome Research **16**, 1299–1309 (2006) (cf. p. 7).

[21]N. L. van Berkum et al., « Hi-c : a method to study the three-dimensional architecture of genomes. », Journal of Visualized Experiments (2010) 10.3791/1869 (cf. p. 7).

[22]G. J. Filion et al., « Systematic protein location mapping reveals five principal chromatin types in drosophila cells », Cell **143**, 212–224 (2010) (cf. p. 8, 86).

[23]C. Flors et W. C. Earnshaw, « Super-resolution fluorescence microscopy as a tool to study the nanoscale organization of chromosomes », Current Opinion in Chemical Biology **15**, 838–844 (2011) (cf. p. 9).

[24]J. A. Thorley et al., « Super-resolution microscopy », in *Fluorescence microscopy* (Elsevier, 2014), p. 199–212 (cf. p. 10).

[25]W. Commons, *File :haploid, diploid, triploid and tetraploid.svg — wikimedia commons, the free media repository*, 2018 (cf. p. 12).

[26]B. R. Williams et al., « Disruption of topoisomerase II perturbs pairing in drosophila cell culture », Genetics **177**, 31–46 (2007) (cf. p. 12, 15).

[27]T. N. Senaratne et al., « Investigating the interplay between sister chromatid cohesion and homolog pairing in drosophila nuclei », PLOS Genetics **12**, sous la dir. de G. Bosco, e1006169 (2016) (cf. p. 12, 15).

[28]A. M. C. Gizzi et al., « Microscopy-based chromosome conformation capture enables simultaneous visualization of genome organization and transcription in intact organisms », Molecular Cell **74**, 212–222.e5 (2019) (cf. p. 23, 24).

[29]D. I. Cattoni et al., « Single-cell absolute contact probability detection reveals chromosomes are organized by multiple low-frequency yet specific interactions », Nature Communications **8** (2017) 10.1038/s41467-017-01962-x (cf. p. 24, 25, 86, 91).

[30]J. R. C. van der Maarel, « Introduction to biopolymer physics », in (WORLD SCIENTIFIC, déc. 2007) chap. 2 (cf. p. 33).

[31]K. Minagawa et al., « Direct observation of the coil-globule transition in dna molecules », Biopolymers **34**, 555–558 (1994) (cf. p. 39).

[32]K. Yoshikawa et al., « Large discrete transition in a single DNA molecule appears continuous in the ensemble », Physical Review Letters **76**, 3029–3031 (1996) (cf. p. 39).

[33]M. Ueda et K. Yoshikawa, « Phase transition and phase segregation in a single double-stranded DNA molecule », Physical Review Letters **77**, 2133–2136 (1996) (cf. p. 39).

[34]P. Flory, *Principles of polymer chemistry*, Baker lectures 1948 (Cornell University Press, 1953) (cf. p. 39, 54).

[35]P.-G. de Gennes, *Scaling concepts in polymer physics* (Cornel University Press, nov. 1979) (cf. p. 39, 42, 44, 54).

[36]P. J. Flory, « The configuration of real polymer chains », The Journal of Chemical Physics **17**, 303–310 (1949) (cf. p. 39).

[37]H. Orland, « Flory theory revisited », Journal de Physique I **4**, 101–114 (1994) (cf. p. 39).

[38]D. Lhuillier, « A simple model for polymeric fractals in a good solvent and an improved version of the flory approximation », Journal de Physique **49**, 705–710 (1988) (cf. p. 40).

[39]J.-M. Victor et al., « The number of contacts in a self-avoiding walk of variable radius of gyration in two and three dimensions », The Journal of Chemical Physics **100**, 5372–5377 (1994) (cf. p. 40).

[40]J. M. Victor et D. Lhuillier, « The gyration radius distribution of two-dimensional polymer chains in a good solvent », The Journal of Chemical Physics **92**, 1362–1364 (1990) (cf. p. 40).

[41]J. B. Imbert et J. M. Victor, « Beyond flory's theory : a computer aided phenomenology for polymers », Molecular Simulation **16**, 399–419 (1996) (cf. p. 40, 43, 48, 54).

[42]J. B. Imbert et al., « Distribution of the order parameter of the coil-globule transition », Physical Review E **56**, 5630–5647 (1997) (cf. p. 40, 43, 44, 48, 71).

[43]A. L. Owczarek et al., « New scaling form for the collapsed polymer phase », Physical Review Letters **70**, 951–953 (1993) (cf. p. 44, 71).

[44]P. Grassberger et R. Hegger, « Simulations of three-dimensional ϑ polymers », The Journal of Chemical Physics **102**, 6881–6899 (1995) (cf. p. 47, 48, 54, 76).

[45]P. Grassberger, « Pruned-enriched rosenbluth method : simulations of ϑ polymers of chain length up to 1 000 000 », Physical Review E **56**, 3682–3693 (1997) (cf. p. 47, 76).

[46]B. R. Caré et al., « Finite-size conformational transitions : a unifying concept underlying chromosome dynamics », Communications in Theoretical Physics **62**, 607–616 (2014) (cf. p. 47).

[47]M. Kardar, « Series expansions », in *Statistical physics of fields* (Cambridge University Press, 2007), p. 123–155 (cf. p. 52).

[48]P.-G. de Gennes, « Collapse of a polymer chain in poor solvents », Journal de Physique Lettres **36**, 55–57 (1975) (cf. p. 54).

[49]C. Wu et X. Wang, « Globule-to-coil transition of a single homopolymer chain in solution », Physical Review Letters **80**, 4092–4094 (1998) (cf. p. 54).

[50]W. K. Hastings, « Monte carlo sampling methods using markov chains and their applications », Biometrika **57**, 97–109 (1970) (cf. p. 57).

[51]F. T. Wall et F. Mandel, « Macromolecular dimensions obtained by an efficient monte carlo method without sample attrition », The Journal of Chemical Physics **63**, 4592–4595 (1975) (cf. p. 63).

[52]M. Socol et al., « Contraction and tumbling dynamics of DNA in shear flows under confinement induced by transverse viscoelastic forces », Macromolecules **52**, 1843–1852 (2019) (cf. p. 65).

[53]J. Goodman et J. Weare, « Ensemble samplers with affine invariance », Communications in Applied Mathematics and Computational Science **5**, 65–80 (2010) (cf. p. 68).

[54]D. Foreman-Mackey et al., « Emcee : the MCMC hammer », Publications of the Astronomical Society of the Pacific **125**, 306–312 (2013) (cf. p. 68).

[55]J. A. Nelder et R. Mead, « A simplex method for function minimization », The Computer Journal **7**, 308–313 (1965) (cf. p. 69).

[56]A. N. Rissanou et al., « Monte carlo study of the coil-to-globule transition of a model polymeric system », Journal of Polymer Science Part B: Polymer Physics **44**, 3651–3666 (2006) (cf. p. 70).

[57]I. M. Lifshitz et al., « Some problems of the statistical physics of polymer chains with volume interaction », Reviews of Modern Physics **50**, 683–713 (1978) (cf. p. 71).

[58]A. Khokhlov, « Theory of the polymer chain collapse for the d-dimensional case », Physica A: Statistical Mechanics and its Applications **105**, 357–362 (1981) (cf. p. 71, 72).

[59]A. Y. Grosberg et A. R. Khokhlov, *Statistical physics of macromolecules*, AIP series in polymers and complex materials (AIP Press, 1994) (cf. p. 71, 87).

[60]D. Foreman-Mackey, « Corner.py : scatterplot matrices in python », The Journal of Open Source Software **1**, 24 (2016) (cf. p. 74).

[61]A. Lesage et al., « Polymer coil–globule phase transition is a universal folding principle of drosophila epigenetic domains », Epigenetics & Chromatin **12** (2019) 10.1186/s13072-019-0269-6 (cf. p. 78).

[62]A. Scacchetti et al., « CHRAC/ACF contribute to the repressive ground state of chromatin », Life Science Alliance **1**, e201800024 (2018) (cf. p. 83).

[63]K. Bystricky et al., « Long-range compaction and flexibility of interphase chromatin in budding yeast analyzed by high-resolution imaging techniques », Proceedings of the National Academy of Sciences **101**, 16495–16500 (2004) (cf. p. 85).

[64]J. Dekker, « Mappingin VivoChromatin interactions in yeast suggests an extended chromatin fiber with regional variation in compaction », Journal of Biological Chemistry **283**, 34532–34540 (2008) (cf. p. 85).

[65]E. Ben-Haïm et al., « Chromatin : a tunable spring at work inside chromosomes », Physical Review E **64** (2001) 10.1103/physreve.64.051921 (cf. p. 85).

[66]H. D. Ou et al., « ChromEMT : visualizing 3d chromatin structure and compaction in interphase and mitotic cells », Science **357**, eaag0025 (2017) (cf. p. 85).

[67]A. L. Sanborn et al., « 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 (2015) (cf. p. 85).

[68]M. Socol et al., « Rouse model with transient intramolecular contacts on a timescale of seconds recapitulates folding and fluctuation of yeast chromosomes », Nucleic Acids Research **47**, 6195–6207 (2019) (cf. p. 86, 88, 91).

[69]J. Nuebler et al., « Chromatin organization by an interplay of loop extrusion and compartmental segregation », Proceedings of the National Academy of Sciences **115**, E6697–E6706 (2018) (cf. p. 86).

[70]B. van Steensel, « Chromatin : constructing the big picture », The EMBO Journal **30**, 1885–1895 (2011) (cf. p. 86).

[71]H. Schiessel, « How short-ranged electrostatics controls the chromatin structure on much larger scales », Europhysics Letters (EPL) **58**, 140–146 (2002) (cf. p. 86).

[72]N. Haddad et al., « IC-finder : inferring robustly the hierarchical organization of chromatin folding », Nucleic Acids Research, gkx036 (2017) (cf. p. 87).

[73]S. K. Ghosh et D. Jost, « How epigenome drives chromatin folding and dynamics, insights from efficient coarse-grained models of chromosomes », PLOS Computational Biology **14**, sous la dir. de S. Zhong, e1006159 (2018) (cf. p. 87).

[74]M. Falk et al., « Heterochromatin drives organization of conventional and inverted nuclei », (2018) 10.1101/244038 (cf. p. 87).

[75]Y. Markaki et al., « The potential of 3d-FISH and super-resolution structured illumination microscopy for studies of 3d nuclear architecture », BioEssays **34**, 412–426 (2012) (cf. p. 87).

[76]S. Mangenot et al., « Salt-induced conformation and interaction changes of nucleosome core particles », Biophysical Journal **82**, 345–356 (2002) (cf. p. 88).

[77]S. Mangenot et al., « Interactions between isolated nucleosome core particles : a tail-bridging effect ? », The European Physical Journal E **7**, 221–231 (2002) (cf. p. 88).

[78]C. Lavelle, « Transcription elongation through a chromatin template », Biochimie **89**, 516–527 (2007) (cf. p. 88).

[79]R. Collepardo-Guevara et al., « Chromatin unfolding by epigenetic modifications explained by dramatic impairment of internucleosome interactions : a multiscale computational study », Journal of the American Chemical Society **137**, 10205–10215 (2015) (cf. p. 88).

[80]T. Nozaki et al., « Dynamic organization of chromatin domains revealed by super-resolution live-cell imaging », Molecular Cell **67**, 282–293.e7 (2017) (cf. p. 88).

[81]A. Allahverdi et al., « The effects of histone h4 tail acetylations on cation-induced chromatin folding and self-association », Nucleic Acids Research **39**, 1680–1691 (2010) (cf. p. 88).

[82]M. Ricci et al., « Chromatin fibers are formed by heterogeneous groups of nucleosomes in vivo », Cell **160**, 1145–1158 (2015) (cf. p. 88).

[83]Y. Ogiyama et al., « Polycomb-dependent chromatin looping contributes to gene silencing during drosophila development », Molecular Cell **71**, 73–88.e5 (2018) (cf. p. 90).

[84]N. E. Matthews et R. White, « Chromatin architecture in the fly : living without CTCF/cohesin loop extrusion ? », BioEssays, 1900048 (2019) (cf. p. 90).

[85]J. Xu et Y. Liu, « A guide to visualizing the spatial epigenome with super-resolution microscopy », The FEBS Journal (2019) 10.1111/febs.14938 (cf. p. 91).

[86]S. N. Majumdar et H. Orland, « Effective langevin equations for constrained stochastic processes », Journal of Statistical Mechanics: Theory and Experiment **2015**, P06039 (2015) (cf. p. 93, 95).

[87]M. Mori, « Quadrature formulas obtained by variable transformation and the DE-rule », Journal of Computational and Applied Mathematics **12-13**, 119–130 (1985) (cf. p. 98).

[88]N. Eggert et J. Lund, « The trapezoidal rule for analytic functions of rapid decrease », Journal of Computational and Applied Mathematics **27**, 389–406 (1989) (cf. p. 98).

[89]M. Mori et M. Sugihara, « The double-exponential transformation in numerical analysis », Journal of Computational and Applied Mathematics **127**, 287–296 (2001) (cf. p. 99).

[90]M. Muhammad et M. Mori, « Double exponential formulas for numerical indefinite integration », Journal of Computational and Applied Mathematics **161**, 431–448 (2003) (cf. p. 100, 107, 109).

[91]D. H. Bailey et al., « A comparison of three high-precision quadrature schemes », Experimental Mathematics **14**, 317–329 (2005) (cf. p. 100, 101).

[92]« IEEE standard for floating-point arithmetic », 10.1109/ieeestd.2008.4610935 (cf. p. 103).

[93]Wikipedia contributors, *Lambert w function — Wikipedia, the free encyclopedia*, https://en.wikipedia.org/w/index.php?title=Lambert_W_function&oldid=895496128#Representations, 2019 (cf. p. 108).

[94]S. Haber, « Two formulas for numerical indefinite integration », Mathematics of Computation **60**, 279–279 (1993) (cf. p. 109).

Antony LESAGE **20 septembre 2019**

Sujet : Inférence bayésienne des paramètres structuraux de la chromatine de la drosophile sur des images super-résolues de domaines épigénétiques

Résumé : Récemment, de grandes avancées ont permis de mieux caractériser l'architecture et la dynamique des génomes. Au cœur de ce problème, un domaine particulier suscite aujourd'hui un intérêt croissant dans plusieurs communautés scientifiques : l'épigénétique. C'est l'étude du vaste ensemble des modifications biochimiques de l'ADN et des protéines architecturales qui dirigent le destin cellulaire sans en altérer l'information génétique.
Des domaines fonctionnels localisés dans les chromosomes ont été mis en évidence. Chez la drosophile, ces domaines fonctionnels sont biochimiquement définis par des marques épigénétiques, ce qui suggère que l'arrangement spatial peut être le « chainon manquant » entre l'épigénétique et l'activité génétique.
Depuis peu, la microscopie de super-résolution a permis d'imager la structure de ces domaines avec une résolution sans précédent. Elle a, en particulier, donné accès à la distribution des rayons de giration pour des domaines de différentes longueurs et associés à des états d'activité de transcription différents : actif, inactif ou réprimé. Les lois d'échelle observées nécessitaient le développement d'un cadre théorique pour les interpréter à la lumière de la physique des polymères.
Ce travail de thèse a consisté à contribuer à la modélisation physique de la chromatine, ainsi qu'à mettre en place une méthodologie d'analyse pour extraire les paramètres structuraux des domaines fonctionnels, à partir de ces données de super-résolution.
En corollaire, les résultats de ce travail pourraient, dans une moindre mesure, permettre de trouver la bonne modélisation à adopter pour comprendre l'organisation de la chromatine chez la drosophile.

Mots clés : Inférence bayésienne, structure, chromatine, drosophile, super-résolution, épigénétique, transition, *coil-globule*, nucléosome

Subject : Bayesian inference of structural parameters of Drosophila chromatin on super-resolved images of epigenetic domains

Abstract : Recently, major advances have made it possible to better characterize the architecture and dynamics of genomes. At the heart of this problem, a particular field that is now attracting growing interest in several scientific communities : epigenetics. It is the study of the vast set of biochemical modifications of DNA and architectural proteins that control cellular fate without altering genetic information.
Functional domains located in the chromosomes have been identified. In Drosophila, these functional domains are biochemically defined by epigenetic markers, suggesting that the spatial arrangement may be the "missing link" between epigenetics and genetic activity.
Recently, super-resolution imaging has made it possible to image the structure of these domains with unprecedented resolution. In particular, it provided access to the distribution of the radius of gyration for domains of different lengths and associated with different states of transcription activity : active, inactive or repressed. The observed scaling laws required the development of a theoretical framework to interpret them in the light of polymer physics.
This thesis work consisted in contributing to the physical modelling of chromatin, as well as setting up an analysis methodology to extract structural parameters from functional domains, based on these super-resolution data.
As a corollary, the results of this work could, to a lesser extent, make it possible to find the appropriate model to adopt to understand the organization of chromatin in Drosophila.

Keywords : Bayesian inference, chromatin, structure, Drosophila, super-resolution, epigenetics, coil-globule, transition, nucleosome