Modeling chromatin dynamics using Gaussian processes and polymer physics

Guilherme Monteiro Oliveira

UNIVERSITÉ DE STRASBOURG

École Doctorale des Sciences de la Vie et de la Santé STRASBOURG

École doctorale des Sciences de la Vie et de la Santé

IGBMC – CNRS UMR 7104 – Inserm U 964

THÈSE présentée par:

Guilherme MONTEIRO OLIVEIRA

soutenue le: **23 mars 2021**

pour obtenir le grade de: **Docteur de l'université de Strasbourg**

Spécialité: Biophysique computationnelle

Modeling chromatin dynamics using Gaussian processes and polymer physics

THÈSE dirigée par:

M. Nacho MOLINA	CR, IGBMC, Illkirch-Graffenstaden, France

RAPPORTEURS:

M. Davide MARENDUZZO	Professeur, The University of Edinburgh, Edinburgh, Écosse
M. Chistof GEBHARDT	Professeur, Ulm University, Ulm, Allemagne

AUTRES MEMBRES DU JURY:

M. Daniel RIVELINE	DR, IGBMC, Illkirch-Graffenstaden, France
Mme. Angela TADDEI	DR, Institut Curie, Paris, France
Mme. Annick LESNE	DR, IGMM, Montpellier, France
M. Julien MOZZICONACCI	PR PUPH, Muséum d'histoire naturelle, Paris, France

Acknowledgment

There are several people without whom this thesis would not have happened. First I would like to thank my supervisor Nacho Molina for the opportunity to perform my PhD research in his lab and for all the nice discussions we had over the past few years. On the same note, I would like to thank all the lab members that were (or not) involved in my projects. I would like to mention specially Attila Oravecz for all the help with some of the experimental data I present in this thesis. I would also like to acknowledge the generosity of Luca Giorgetti (FMI, Basel) and his team, as they generate and shared the TetO cell line used in this thesis. I would like to thank Olivier Tassy, one of the most kind people I have ever met. He thought me a great deal about French culture and how to deal with "all the paperwork". I would also like to thank Sara Jimenez, she was always caring for my well-being and sharing the best coffees.

A central figure to my PhD was Thomas Sexton. Without him my thesis would not have happened. He was the driving force for most of the results presented in this thesis. His patience and help with complicated biological concepts was, perhaps, the only reason why my polymer model and GP-Tool came to be. Thank you, Tom. Furthermore, I would like to thank all the current (and past) members of his lab. Most importantly Dominique Kobi, who generated the ANCHOR cell lines and performed related experiments presented in the thesis. I would like to also mention Angeliki Platania, the "guinea pig" for most of my programs and pipelines.

I'm also grateful to all the members of my big thesis jury: Davide Marenduzzo, Chistof Gebhardt, Daniel Riveline, Angela Taddei, Annick Lesne and Julien Mozziconacci. In fact, I would like to present extra gratitude to Gebhardt (he was also part of my CST along with Sexton) and Riveline, the person who brought me to France, an unknown country to me all those years ago, with his Cell Physics Master program.

There are several people that are not related to my PhD work, but that are very important to me. Antony Bazir, my brother from a different country, that, even though younger than me, was the voice of wisdom several times. Kenny "Sonnenschein" Schumacher and Vincent "Poopin" Hisler, my buds for movies, lunch, beer, hiking and more.

Now, I would like to mention people that are really close to my heart. First I would like to thank my Brazilian family. Without the support and love of my parents and sister in the past almost 30 years, I would not be here today. Love you

all. Then, I would like to thank my Austrian family. Even though we speak different languages (most of the time), I feel very comfortable around them.

Finally, I would like to thank my wife-to-be Veronique Fischer. She showed me a whole new side of life I had never realized before we had met and, because of her, I feel like I am a much better person than I once was. Love you a lot!

Modélisation de la dynamique de la chromatine à l'aide de processus gaussiens et de la physique des polymères

Première partie

Introduction

Le corps principal de cette thèse sera divisé en deux parties. La première est de nature expérimentale, où l'on s'interroge sur les effets de la condensation de la chromatine sur ses propriétés diffusives. Pour répondre à cette question, notre laboratoire a utilisé une lignée cellulaire partagée par l'équipe de Giorgetti (FMI Bâle) dans laquelle un système PiggyBac est utilisé pour marquer au hasard des *loci* de chromatine dans des cellules souches embryonnaires (SE) de souris. La dynamique de déplacement de ces loci est enregistrée par microscopie fluorescente pendant une courte période de temps en interphase et en mitose. Cette expérience a été utilisée pour étudier les différences générales (ou les similitudes) entre les dynamiques de la chromatine à des stades de condensation très différents.

Pour approfondir l'étude précédente, nous voulions également déterminer si les coefficients de diffusion et d'anomalie pouvaient différer selon le *locus*. Le laboratoire de Sexton (IGBMC) a développé des lignées cellulaires dans lesquelles 3 loci spécifiques du domaine HoxA dans les cellules SE sont marqués à l'aide de sondes fluorescentes ANCHOR. De plus, lors de la culture avec de l'acide rétinoïque, ces cellules sont induites en différenciation vers des cellules précurseurs de neurones (PN). Ainsi, nous avons déterminé quels sont les effets de la différenciation sur la dynamique de ces sondes. Ceci est possible parce que les gènes HoxA sont réprimés à l'état SE, mais actifs une fois que les cellules sont différenciées.

Contrairement à ce que l'on pourrait imaginer, l'analyse de ces données n'est pas (du tout) simple, c'est pourquoi plusieurs méthodes ont été développées à cette fin. Dans la deuxième partie de cette thèse, je commence par introduire quelques concepts de la théorie des probabilités et des statistiques. Nous discutons des principales différences entre les approches statistiques bayésiennes et fréquentistes. Nous allons également discuter du théorème de la limite centrale (TCL), un concept très important en statistique. En utilisant ces concepts de base, nous allons introduire le mouvement brownien fractionnaire comme modèle pour la matrice de covariance des distributions gaussiennes multivariées. Grâce à cette méthode, nous avons pu prendre en compte au maximum les trajectoires des particules, ce qui nous a permis d'obtenir des mesures plus précises.

Dans le cadre bayésien, nous avons également développé des modèles pour améliorer la localisation des particules dans les films de microscopie. Dans le même ordre d'idées, une méthode a été mise au point pour corriger les défauts d'alignement entre les canaux du microscope, dus à des problèmes de caméra et d'aberration chromatique. Enfin, nous introduisons un nouveau modèle pour corriger les mesures de diffusion dans les situations où le substrat est en mouvement. À la différence de nombreuses approches dans la littérature, aucune installation expérimentale supplémentaire ou post-traitement des données n'est nécessaire.

Afin de mieux interpréter les résultats expérimentaux, nous introduirons un modèle basé sur la physique pour la chromatine dans la partie III de cette thèse. Je vais tout d'abord décrire la nature de la diffusion et la façon dont nous pouvons la traiter mathématiquement. Je vais également présenter des méthodes avec lesquelles nous pouvons simuler ce phénomène sur ordinateur. Ensuite, je vais introduire la chaîne de Rouse comme première approximation pour la chromatine. Sur la base de ce modèle, j'ai considérer la conformation moyenne de la chromatine dans la population, telle que visualisée par les cartes Hi-C, pour reconstruire des polymères synthétiques ayant une conformation similaire au domaine HoxA. Le premier objectif ici était de déterminer si les distances mesurées entre les sondes du domaine HoxA sont récapitulées. Ensuite, j'ai inséré la dynamique dans le système et affiné les propriétés diffusives de chaque section de chromatine en utilisant les données ChIP-seq pour évaluer le contexte dans lequel chaque section de notre polymère se trouve.

Dans les sections suivantes, je résume certains des principaux résultats présentés dans cette thèse. Pour plus de détails et d'informations, veuillez vous référer au texte principal.

Deuxième partie

Analyse des données : Mesure de la dynamique de la chromatine

1 Modélisation de la dynamique avec les processus gaussiens

Certains des modèles les plus populaires utilisés pour déduire les propriétés de diffusion sont basés sur l'analyse du déplacements des particules dans le temps, ce qui tend à ignorer les corrélations implicites entre les points temporels mesurés. De ce fait, la précision globale de l'inférence est réduite. Pour surmonter ce problème, nous avons développé une méthode utilisant le processus gaussien (PG), qui nous permet de modéliser ces corrélations temporelles et ainsi d'utiliser toutes les informations disponibles dans les trajectoires mesurées.

Bien que le PG soit un processus stochastique continu dans le temps, un sous-ensemble $\{x_t; t \in T\}$ est décrit par une distribution gaussienne multivariée donnée par

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{N/2} \ |\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \ \ \Sigma^{-1} \ (\boldsymbol{x}-\boldsymbol{\mu})\right\} \tag{1.1}$$

où $\boldsymbol{x}$ contient N variables et $\boldsymbol{\mu}$ est un autre vecteur contenant la moyenne de chaque variable x_t. La variance est maintenant représentée par une matrice symétrique Σ appelée matrice de covariance, avec des valeurs hors diagonale représentant la corrélation entre 2 variables quelconques. La variance étant définie comme positive, nous devons nous assurer que toutes les valeurs propres de cette matrice sont supérieures à zéro. En d'autres termes, la matrice de covariance doit être écrite comme une composition $\Sigma = LL^T$ ou, inversement, construite à partir de la multiplication d'une matrice L avec sa transposition.

Comme modèle pour les trajectoires stochastiques, nous utilisons le mouvement brownien fractionnaire (MBF), une moyenne mobile du mouvement brownien traditionnel, où chaque pas est pondéré en fonction de $(t-s)^{\frac{\alpha-1}{2}}$. Sa matrice de covariance est définie comme suit

$$\Sigma_{D_\alpha,\alpha}(t,s) = D_\alpha \ (t^\alpha + s^\alpha - |t-s|^\alpha), \tag{1.2}$$

pour $t > 0$, $0 < \alpha < 2$ et $D_\alpha > 0$. Dans ce cas, D_α représente le coefficient de diffusion apparente et il est associé à la mobilité d'une particule donnée. Inversement, α

est appelé le coefficient d'anomalie. À titre d'exemple, nous avons dans la figure 1.1 l'effet de α sur les trajectoires stochastiques. Sur les images du bas, nous traçons la distribution de l'angle sur 2 étapes consécutives simulées. Nous remarquons que pour $\alpha < 1$ les particules ont tendance à être plus contraint. $\alpha = 1$ représente le mouvement brownien traditionnel, où les particules sont libres d'aller n'importe où de manière aléatoire, sans direction définie. Enfin, pour $\alpha > 1$ les particules ont tendance à avoir une direction préférée à suivre.

Enfin, en fixant $t = s$, on obtient $\Sigma_{D_\alpha,\alpha}(t,s) = 2D_\alpha t^\alpha$, ce qui est connu sous le nom de déplacement quadratique moyen (DQM). Cette courbe est généralement utilisée pour estimer les valeurs de D_α et de α.

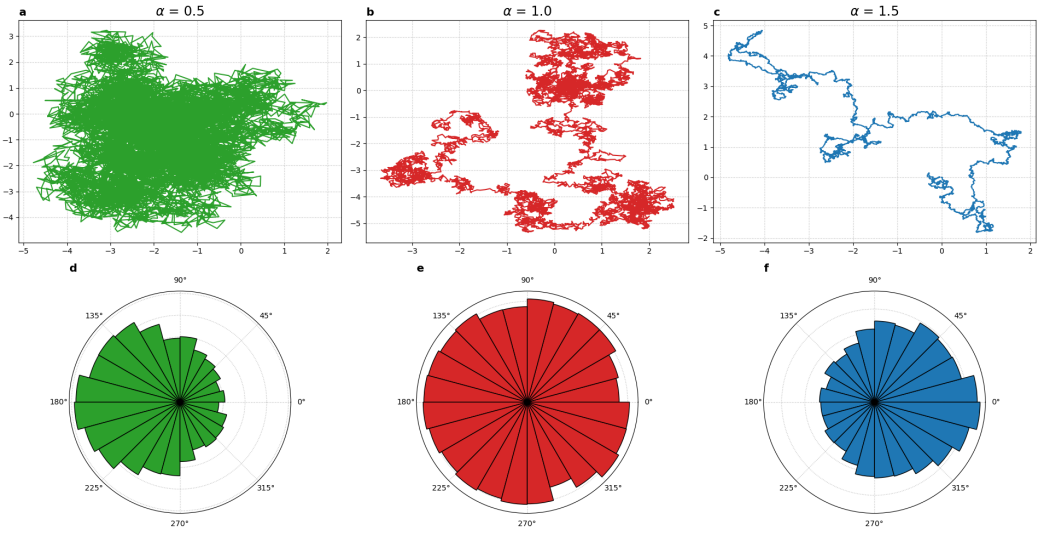

Fig. 1.1. : Nous échantillonnons une seule longue trajectoire en utilisant $D_\alpha = 1$ et $\alpha = 0{,}5$ pour (a), $\alpha = 1{,}0$ pour (b) et $\alpha = 1{,}5$ pour (c). $\alpha < 1$ confine le mouvement des particules si on le compare au mouvement brownien traditionnel $\alpha = 1$, tandis que $\alpha > 1$ le dirige.

2 Inférer les coefficients de diffusion et d'anomalie

Le PG fournit la probabilité d'observer une trajectoire $\boldsymbol{r}$ étant donné D_α et α. Ensuite, nous avons appliqué le théorème de Bayes [1] pour obtenir la distribution postérieure sur les paramètres de diffusion étant donné la trajectoire mesurée :

$$P(D_\alpha, \alpha, \boldsymbol{\mu}|\boldsymbol{r}) = \frac{P(\boldsymbol{r}|D_\alpha, \alpha, \boldsymbol{\mu})\ P(D_\alpha, \alpha, \boldsymbol{\mu})}{\int dD_\alpha\ d\alpha\ d\boldsymbol{\mu} P(\boldsymbol{r}|D_\alpha, \alpha, \boldsymbol{\mu})\ P(D_\alpha, \alpha, \boldsymbol{\mu})}, \tag{2.1}$$

où $P(D_\alpha, \alpha, \boldsymbol{\mu})$ représente la distribution préalable des paramètres du modèle. Dans

l'hypothèse d'un *flat prior* sur $\boldsymbol{\mu}$, D_α et α, le log-postérieur peut être exprimé comme

$$\log(P(D_\alpha, \alpha, \boldsymbol{\mu}|\boldsymbol{r}) \propto -\frac{1}{2}(\boldsymbol{r}-\boldsymbol{\mu})^T \Sigma_{D_\alpha,\alpha}^{-1}(\boldsymbol{r}-\boldsymbol{\mu}) - \frac{1}{2}\log|\Sigma_{D_\alpha,\alpha}| - \frac{N}{2}\log(2\pi), \quad (2.2)$$

où N représente le nombre de points mesurés et $|\cdot|$ est la fonction déterminante. Pour obtenir des estimations postérieures maximales, nous avons optimisé (2.2) en utilisant la méthode de Nelder-Mead [2]. En outre, pour calculer des intervalles crédibles pour nos estimations, nous avons utilisé la méthode de Monte Carlo par chaînes de Markov appelé Metropolis-Hastings [1] pour échantillonner à partir de la distribution de probabilité postérieure.

Pour déterminer la performance globale de cette méthode, désormais appelée GP-FBM, nous avons simulé 10000 trajectoires similaires mais avec des valeurs aléatoires de D_α et α. Nous avons fixé D_α dans la fourchette $0,01 < D_\alpha < 1,5$ et α dans la fourchette $0,01 < \alpha < 2$. À titre de comparaison, nous estimons également ces paramètres en utilisant des méthode basés sur le déplacement, tels que les DQM et l'ajustement de la distribution. Les résultats sont présentés dans la figure (2.1).

3 Méthodes développées

3.1 Améliorer la précision de localisation des particules fluorescentes

La détection et le suivi des spots pour tous les films ont été effectués avec ICY, un logiciel d'analyse d'images [3]. En supposant que les spots sont approximativement de forme gaussienne à deux dimensions, nous avons optimisé sa localisation en utilisant la méthode Nelder-Mead [2] et estimé la précision de localisation en utilisant l'algorithme de Metropolis-Hastings [1]. Pour tester cette méthode, nous générons un film synthétique avec 500 images pour une seule particule. Le spot a été généré en utilisant une forme gaussienne symétrique à deux dimensions avec un écart de 1, le signal de fond a été fixé à 100 et l'intensité du spot est égale à 200. Le bruit du signal a été généré sous la forme d'une distribution de Poisson en prenant le signal original comme moyenne. Sur la figure (3.1), nous montrons que cette méthode augmente considérablement la précision de la localisation.

3.2 Algorithme d'alignement

Certaines des expériences de microscopie réalisées ont utilisé deux caméras, c'est-à-dire une par canal. Ce système nous a permis d'enregistrer simultanément le spot dans les deux canaux, simplifiant par la suite le traitement des données. Ce-

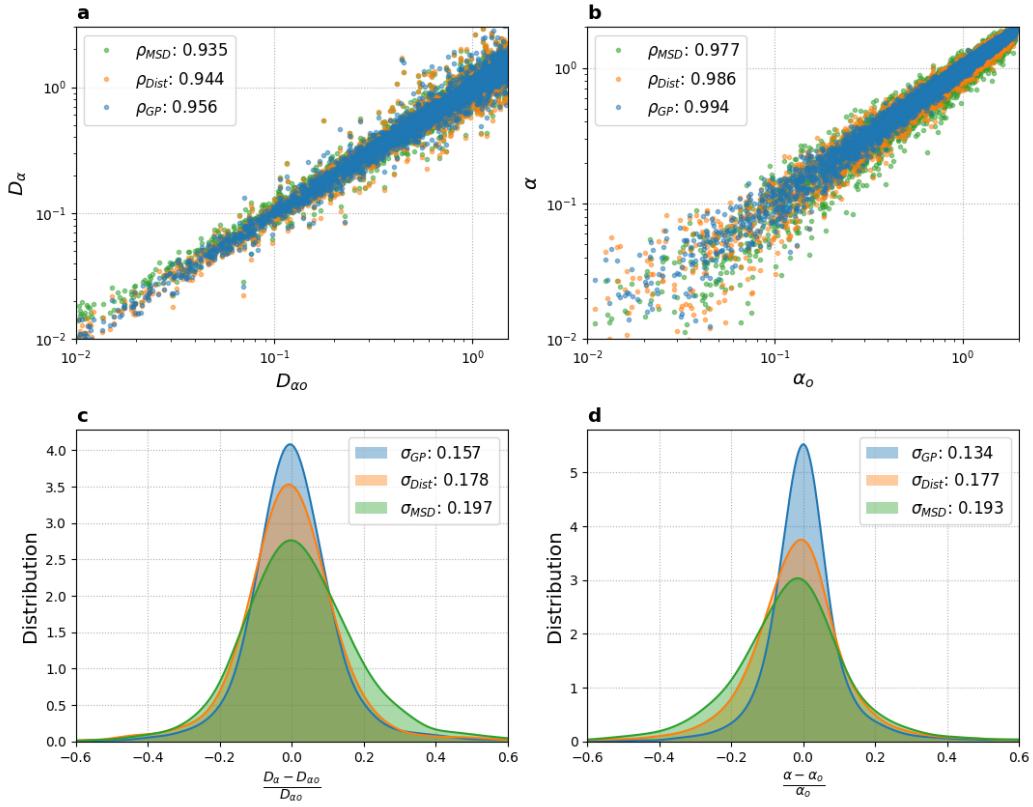

Fig. 2.1. : (a-b) Corrélation entre les paramètres fixés et estimés pour un ensemble de 2000 trajectoires simulées. (c-d) Erreur d'estimation relative pour les mêmes trajectoires. À titre de comparaison, la méthode GP-FBM est, en moyenne, plus précise que les méthodes basées sur le déplacement.

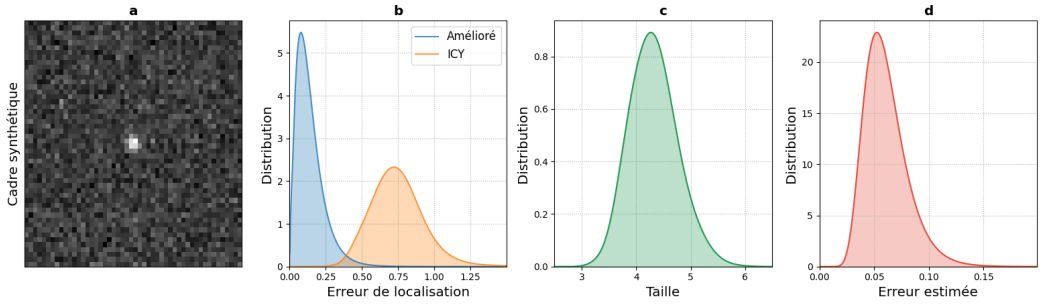

Fig. 3.1. : (a) Exemple de spot dont le signal et la taille sont similaires à ceux observés dans les films réels. (b) Notre algorithme améliore d'environ 7 fois la localisation des spots suivis. (c) Nous pouvons déterminer la taille moyenne des spots dans une trajectoire et l'utiliser pour identifier les éventuelles valeurs aberrantes. (d) Distribution des erreurs de localisation estimées pour chaque point de la trajectoire synthétique.

pendant, l'utilisation de deux caméras a introduit des écarts d'alignement non négligeables entre les canaux. Dans la figure (3.2a-c), nous avons quelques exemples.

Indépendamment de la question de la double caméra, deux longueurs d'onde différentes génèrent également des erreurs associées à l'aberration chromatinienne.

Pour corriger ces problèmes, nous avons utilisé un ensemble générique de transformations affines pour effectuer un post-alignement numérique. Le modèle est écrit comme suit

$$\Omega = \begin{pmatrix} s_x & 0 & (1-s_x)W/2 \\ 0 & s_y & (1-s_y)H/2 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & d_x + c_x \\ 0 & 1 & d_y + c_y \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos(\theta) & \sin(\theta) & 0 \\ -\sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & -c_x \\ 0 & 1 & -c_y \\ 0 & 0 & 1 \end{pmatrix}, \tag{3.1}$$

où,s_i représente la mise à l'échelle dans les directions x et y, d_i représente la translation dans les deux directions et θ est l'angle de rotation entre les deux canaux par rapport au point c_i.

Pour déduire les paramètres optimaux de la correction, nous avons utilisé 10 images de tous les films enregistrés dans la session et maximisé la probabilité suivante en utilisant la méthode de Nelder-Mead [2]

$$\ln P \propto -\frac{WH}{2} \ln \left\{ \sum_{k,l} \left[I_2(k,l|\Omega) - I_1(k,l|\mathbb{I}) \right]^2 \right\}, \tag{3.2}$$

où W et H correspondent à la largeur et à la hauteur des images et $I_r(k,l|A)$ est la valeur du pixel (k,l) dans le canal r étant donné la transformation A. Des exemples d'images désalignées et de corrections sont présentés dans la figure (3.2).

3.3 Correction des mouvements de fond

Il est assez remarquable de constater à quel point les cellules se déplacent lorsqu'on effectue une imagerie en direct. Il a été constaté que les cellules ont tendance à effectuer une sorte de mouvement brownien si on les laisse libres de se déplacer [4]. De plus, les cellules ne sont pas des corps rigides. Leur forme peut fluctuer lorsque la cellule réorganise son contenu interne. De plus, la chaleur supplémentaire introduite par le laser en microscopie fluorescente tend à rendre les cellules plus agitées, augmentant leur motilité et leurs fluctuations volumétriques. Pour résoudre ce problème, nous présentons dans la figure (3.3) un schéma représentant ce qui est observé. Les vecteurs $\boldsymbol{r}_i$ sont les positions des particules telles que mesurées dans le cadre de référence du microscope, mais ces mesures incluront le mouvement confondu $\boldsymbol{R}$. Nous nous intéressons donc à la dynamique intrinsèque décrite par les vecteurs $\boldsymbol{a}_i$. Pour simplifier, nous allons décrire un système de 2 particules, mais ce modèle peut facilement être étendu à d'autres particules.

En utilisant le modèle du processus gaussien présenté dans une section précé-

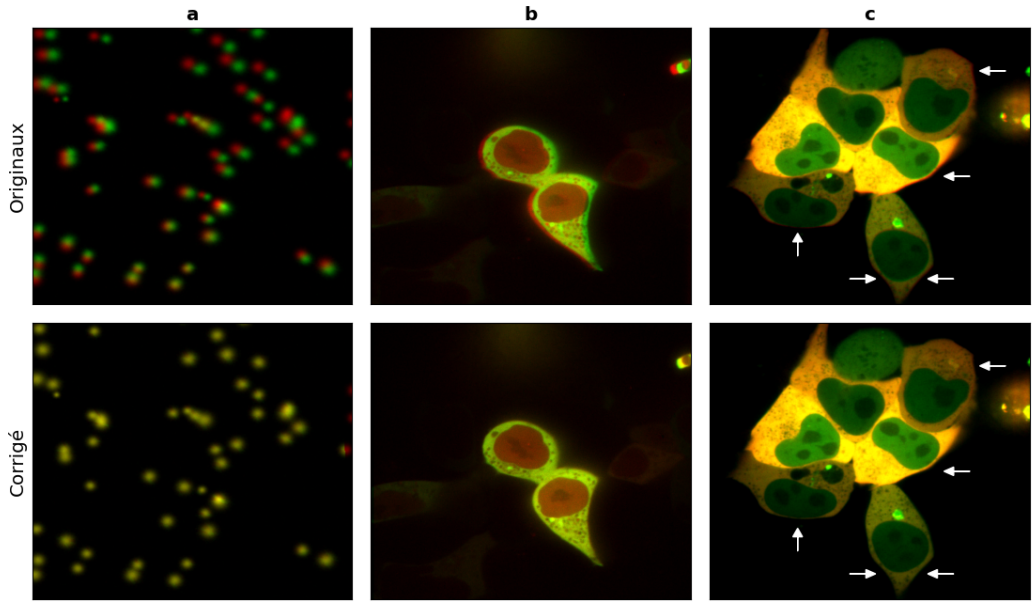

Fig. 3.2. : (a) Image synthétique générée pour illustrer les problèmes courants rencontrés dans nos expériences de microscopie. (b) Image réelle générée avec un mauvais calibrage de la caméra. (c) Image réelle générée avec un calibrage correct de la caméra. Quoi qu'il en soit, l'aberration de la chromatine est toujours perceptible.

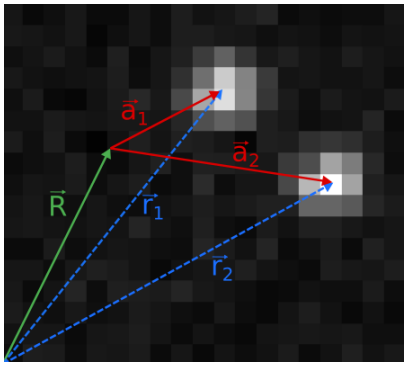

Fig. 3.3. : Schéma de correction des mouvements confondus. $\boldsymbol{r}_i$ sont des positions mesurées dans le cadre de référence du microscope. $\boldsymbol{R}$ représente le mouvement du substrat, tandis que $\boldsymbol{a}_i$ sont les positions des particules dans le cadre de référence mobile.

dente, nous pouvons exprimer ces relations comme suit

$$\rho(\boldsymbol{a}_i, \boldsymbol{R}|\alpha_i, D_i) \propto \exp\left\{-\frac{1}{2}\boldsymbol{a}_1^T\Sigma_1^{-1}\boldsymbol{a}_1 - \frac{1}{2}\boldsymbol{a}_2^T\Sigma_2^{-1}\boldsymbol{a}_2 - \frac{1}{2}\boldsymbol{R}^T\Sigma_R^{-1}\boldsymbol{R}\right\}, \tag{3.3}$$

où nous avons associé le noyau MBF Σ_i directement à $\boldsymbol{a}_i$. En raison du mouvement local de la chromatine, nous savons que les vecteurs $\boldsymbol{a}_i$ sont corrélés à travers $\boldsymbol{R}$.

En réorganisant les termes et en intégrant $\boldsymbol{R}$, nous avons obtenu le modèle final pour 2 particules

$$\rho(\boldsymbol{r}_i|\alpha_i, D_i) \propto \exp\left\{-\frac{1}{2}\begin{pmatrix}\boldsymbol{r}_1\\ \boldsymbol{r}_2\end{pmatrix}^T \begin{pmatrix}\Sigma_1+\Sigma_R & \Sigma_R\\ \Sigma_R & \Sigma_2+\Sigma_R\end{pmatrix}^{-1}\begin{pmatrix}\boldsymbol{r}_1\\ \boldsymbol{r}_2\end{pmatrix}\right\}. \tag{3.4}$$

Ce résultat nous montre que les effets du mouvement confondant introduisent des corrélations entre la trajectoire mesurée. De plus, cette source de mouvement supplémentaire augmentera la variance mesurée pour chaque point suivi. Pour cette raison, nous pouvons nous attendre à surestimer les coefficients de diffusion et d'anomalie si Σ_R n'est pas pris en compte.

Dans la figure (3.4a), nous calculons la distribution des déplacements pour 2000 particules simulées avec $D_\alpha = 0,15$ et $\alpha = 0,25$ qui se déplacent sous l'influence d'un substrat avec $D_R = 0,02$ et $\alpha_R = 1,35$. Notez que D_R est environ 10 fois plus petit que le coefficient de diffusion fixé pour la particule elle-même. En pointillés, nous montrons le déplacement moyen ainsi que les courbes de distribution en utilisant D_α et α déduites sans tenir compte du mouvement du substrat. Les lignes continues utilisent des paramètres corrigés. En chiffres (3.4b-e), nous présentons les 2000 valeurs déduites.

3.4 GP-Tool

Toutes ces méthodes développées, parmi d'autres utilitaires, font partie d'une application que j'ai développé. Ce logiciel s'appelle GP-Tool et peut être téléchargé sur ma page Github (https://github.com/guilmont). Dans la figure (3.5), nous présentons une capture d'écran de ce logiciel.

4 Mesure de la dynamique de la chromatine

4.1 Comparaison de l'interphase et de la mitose

Depuis la première fois que les cellules mitotiques ont été observées au microscope à la fin du XIXe siècle, nous avons appris que le contenu nucléaire change sauvagement d'état de condensation entre l'interphase et la mitose. Il a été mesuré par des tests volumétriques et des méthodes basées sur le FRET que la chromatine se condense 2 à 3 fois de l'interphase à la mitose [5, 6, 7], si la division cellulaire doit être accomplie dans l'espace alloué typique. En fait, la structure de la chromatine mitotique a fait l'objet d'études intensives [8, 9], mais on ne sait pas exacte-

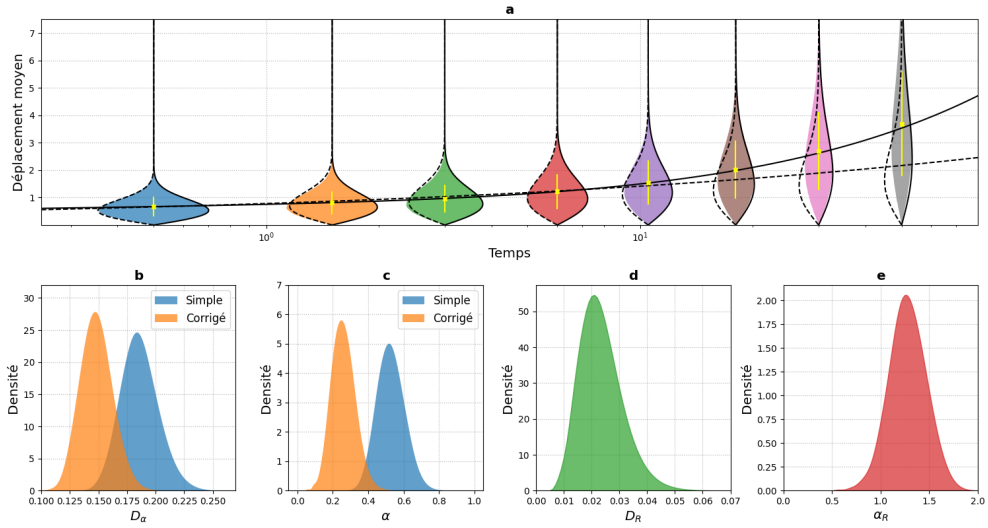

Fig. 3.4. : (a) Distribution des déplacements et moyenne dans le temps calculés pour 2000 trajectoires simulées avec $D_\alpha = 0,15$, $\alpha = 0,25$, $D_R = 0,02$ et $\alpha_R = 1,35$. Les lignes continues utilisent la valeur moyenne obtenue avec la méthode présentée au-dessus. Les lignes en pointillés négligent le mouvement du substrat. (b-c) Distribution des paramètres déduits pour les trajectoires simulées. Comme prévu, ils ont été surestimés lorsque le substrat n'a pas été pris en compte. (d-e) Distribution des paramètres pour le substrat.

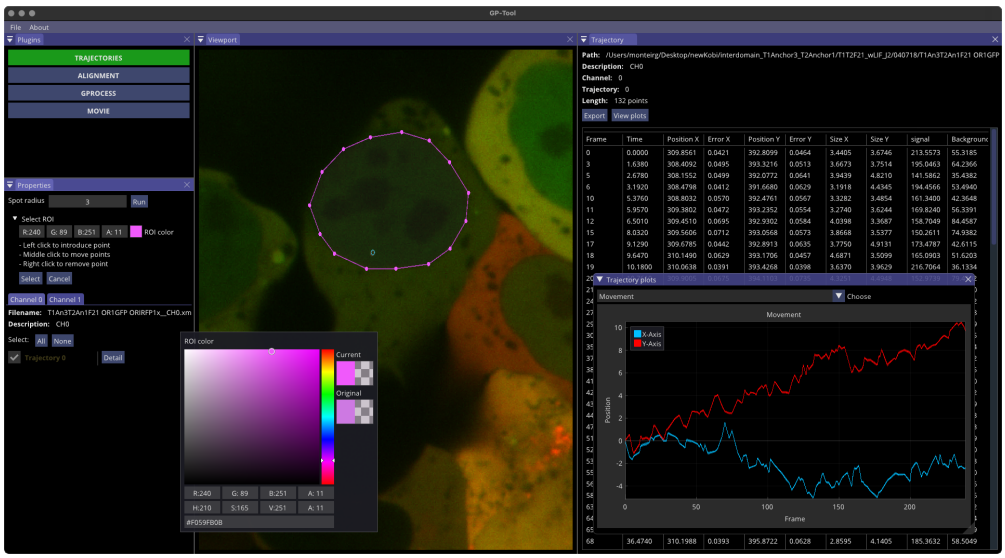

Fig. 3.5. : Exemple d'un *plugin* du GP-Tool.

ment comment les réarrangements de la chromatine influencent ses propriétés de diffusion.

Pour élucider les éventuelles différences et similitudes entre ces stades, nous

avons utilisé une lignée cellulaire SE de souris dans laquelle des réseaux TetO de 7 kb de long sont greffés sur environ 20 à 25 emplacements aléatoires du génome. Après la transfection et l'expression de GPF::TetR, ces emplacements deviennent visibles au microscope. Afin de distinguer les cellules en interphase de celles en mitose, on a eu recours à la coloration de Hoechst. Dans la figure (4.1a), nous avons un exemple de ces cellules.

Sans surprise, la probabilité de trouver des cellules mitotiques naturelles était très faible. Pour cette raison, nous avons effectué une synchronisation basée sur le nocodazole. Des masques ont été générés en attribuant des étiquettes de couleur individuelles à chaque cellule, où le canal bleu a été utilisé pour identifier les cellules en interphase, en mitose et en arrêt de nocodazole. Les loci étiquetés ont été détectés et suivis à l'aide de ICY [3]. Enfin, ces loci ont été regroupés par cellule à l'aide de nos masques et ont été intégrés dans la méthode GP-FBM.

Dans la figure (4.1b), nous montrons les résultats résumant le mouvement des loci de chromatine à l'intérieur des cellules en intrephase (bleu), en mitose (rouge) et traitées au nocodazole (vert). Nous avons analysé les valeurs moyennes de déplacement et les distributions de déplacement obtenues à partir des expériences par rapport aux expressions théoriques avec ou sans prise en compte du mouvement du substrat. Il est clair que le modèle GP-FBM étendu tenant compte du mouvement du substrat s'adapte beaucoup mieux aux données, en particulier pour les grands intervalles de temps. Notez que cette approche GP-FBM ne nécessite pas de dispositif expérimental supplémentaire ni de post-traitement des données, ce qui diffère de nombreuses approches dans la littérature.

De manière surprenante, nos résultats indiquent qu'il n'y a pas de différences significatives dans la moyenne de D_α et de α entre l'interphase et la mitose, ce qui suggère que la condensation n'affecte pas la dynamique de diffusion globale de la chromatine, comme nous pouvons le voir sur la figure (4.1c,d). Dans le cas des cellules arrêtées en mitose, nous avons observé une augmentation significative de α qui pourrait être liée à l'effet que le nocodazole a sur la formation des microtubules et donc sur la stabilité des chromosomes mitotiques. Il est intéressant de noter que nous avons obtenu un large éventail de coefficients D_α et α estimés indiquant une variabilité remarquable d'un point à l'autre, même si l'on corrige le mouvement du substrat. Cette variabilité pourrait être en partie causée par des différences dans l'état cellulaire des cellules analysées (variabilité inter-cellulaire) conduisant à une dynamique globale différente de la chromatine. Alternativement, des différences dans le contexte chromatinien des loci génomiques pourraient conduire à une dynamique de diffusion spécifique (variabilité intra-cellulaire). En utilisant le théorème de la variance totale, nous avons quantifié que seul 40% de la variabilité en D_α et

α pouvaient être expliqué par les différences entre les cellules lorsque le mouvement du substrat est pris en compte. Dans l'ensemble, cela suggère que différents loci génomiques peuvent avoir des propriétés de diffusion caractéristiques en raison d'interactions spécifiques de la chromatine avec le contexte nucléaire.

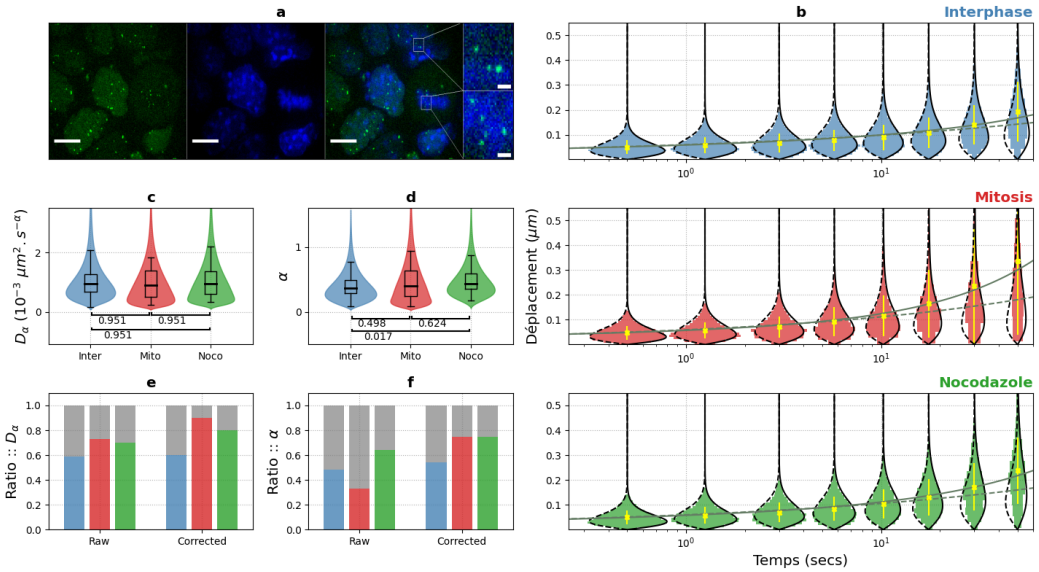

Fig. 4.1. : (a) Images de projection maximale des cellules SE contenant des taches de TetR::GFP liées à des matrices TetO et de l'ADN coloré par Hoechst (la barre d'échelle est de 10 μm). L'encadré montre la zone sélectionnée agrandie (barre d'échelle 1 μm). (b) Les lignes noires décrivent les distributions de déplacement théoriques déduites obtenues en utilisant l'approche GP-FBM avec (ligne continue) et sans (lignes pointillées) en tenant compte du mouvement du substrat. En couleurs, les distributions de déplacement calculées à partir des trajectoires mesurées. Les croix jaunes indiquent le déplacement moyen. En revanche, les lignes grises représentent les courbes de déplacement moyen théorique en utilisant les paramètres bruts (pointillés) et corrigés (continus). (c-d) Distributions de D_α et α estimées dans trois conditions corrigeant le mouvement du substrat. (e-f) Répartition de la variance totale en composantes inter- et intra-cellulaire (respectivement en gris et en couleur) pour D_α et α dans trois conditions différentes.

4.2 Le domaine HoxA

L'étude sur le système TetO nous a montré une variabilité non négligeable de D_α et de α à l'intérieur des cellules. Cet effet était d'autant plus évident lorsque l'inférence de ces paramètres était corrigée en fonction du mouvement du substrat, où, dans certains cas, on a observé que jusqu'à 90% de la variabilité provenait de cellules individuelles. Ce résultat nous a incité à spéculer sur la raison d'une telle

variabilité et sur sa possible corrélation avec leur fonction. Nous trouvons dans la littérature des études de cas montrant que les régions de chromatine à proximité des centromères et des télomères ont tendance à être moins mobiles (D_α réduit) dans la levure [10]. Nous trouvons également quelques résultats liant l'activité transcriptionnelle à un confinement local accru (α réduit) [11] et/ou une mobilité accrue des gènes [12]. Sinon, on sait peu de choses sur les effets du contexte du génome sur la dynamique de la chromatine.

Afin d'explorer la variabilité des propriétés diffusives à l'intérieur des cellules, 2 lignées SE ont été générées par double marquage avec ANCHOR [13]. Les étiquettes ANCH1 et ANCH3 ont été introduites à différents endroits dans le même allèle du chromosome 6 pour les lignées inter-TAD (T1-T2) et intra-TAD (T2-T3) [1], comme le montre la figure (4.2a). Stratégiquement, T1 et T3 sont équidistants de T2 ($\sim$ 300 kb), ce qui nous a permis d'approfondir nos recherches sur les effets de la structure TAD dans les distances à trois dimensions. En utilisant les intervalles de temps enregistrés par microscopie confocale et la méthode GP-FBM, nous évaluons D et α pour les cellules SE et les cellules induites à différenciation via l'acide rétinoïque.

Comme on peut s'y attendre, nous montrons dans la figure (4.2b) que la distance moyenne entre les sondes était plus élevée pour la combinaison inter-TAD que pour la combinaison intra-TAD, mais avec une grande hétérogénéité dans les distributions de distances [14]. Il est intéressant de noter que l'induction du gène Hox n'a pas eu d'effet sur les distances inter-TAD, mais a diminué les distances intra-TAD, ce qui soutient l'idée d'un renforcement général du TAD au fur et à mesure que la différenciation cellulaire est induite [15]. En utilisant la GP-FBM sur les trois loci, nous observons dans la figure (4.2c) que toutes les régions présentent une mobilité similaire dans les cellules SES indifférenciées, mais la région T1 est significativement plus confinée que T2 et (faiblement) T3. Un examen plus approfondi des profils épigénomiques des cellules SES (et des cellules précurseurs neuronales différenciées) autour de ces régions a montré que T1 est proche (<15 kb) du gène qui code le long ARN non codant *Haunt*, dont l'expression dans les cellules SES est liée à la suppression des gènes HoxA [16]. Une activité transcriptionnelle plus élevée autour de T1, par rapport aux régions silencieuses T2 et T3, semble donc liée à un plus grand confinement de la chromatine, conformément à une étude antérieure d'un gène induit par les oestrogènes [11]. L'induction du gène Hox par l'acide rétinoïque n'a pas eu d'effet significatif sur la diffusion de T1, mais a réduit le confinement du locus (Fig. 4.2c,d). En revanche, la région T2, qui ne présente aucune caractéristique épigénomique ou régulatrice connue, a connu des augmentations significatives de D_α et de α, ce qui indique peut-être le remodelage général

1. TAD est un acronyme en anglais que signifie *Topological Associated Domains*. TAD est une région génomique auto-interagissante.

de la chromatine dû à la différenciation. Curieusement, la région T3 est devenue plus confinée lors du traitement à l'acide rétinoïque, avec une augmentation concomitante de la diffusion. Cette région contient des sites liés par la protéine CTCF, formant un barrage pour les processus d'extrusion en boucle à médiation par la cohésine [17, 18], et on peut s'attendre à ce que cela se traduise par des altérations de la dynamique locale de la chromatine. Dans l'ensemble, ces résultats montrent une corrélation entre l'activité ou la fonction du locus d'un gène et la dynamique locale de sa chromatine, une caractéristique qui a été largement négligée dans la plupart des études précédentes.

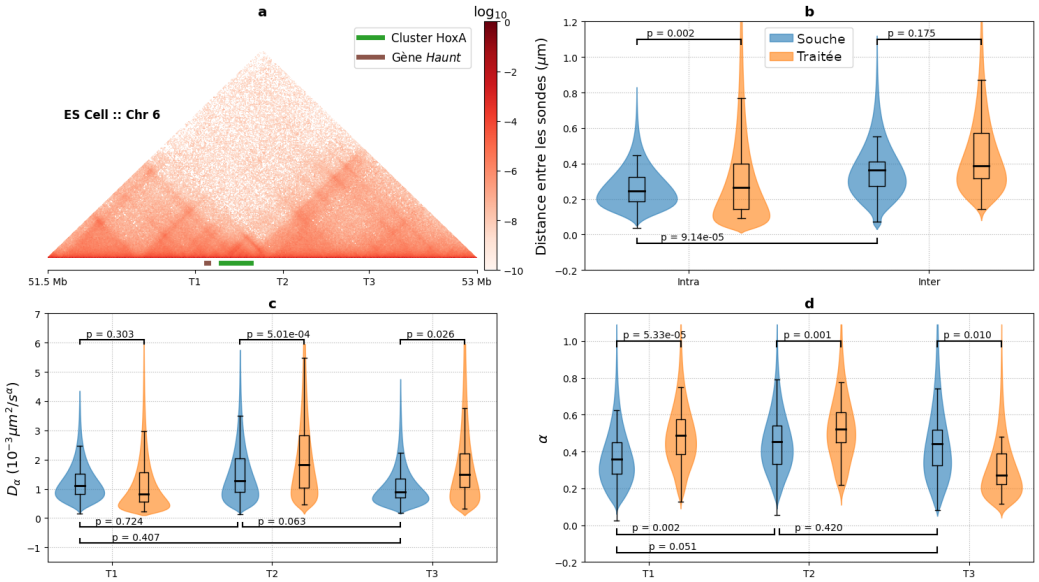

Fig. 4.2. : (a) Aperçu de la structure du locus HoxA et de la position des sondes ANCHOR. (b) Distances entre les sondes mesurées pour les cellules SE et les cellules traitées à l'acide rétinoïque. (c-d) Comparaison des coefficients de diffusion apparente et d'anomalie entre les cellules SE et les cellules traitées à l'acide rétinoïque.

Troisième partie

Biophysique : Modélisation de la chromatine

5 Reconstruire la conformation de la chromatine

Mon premier objectif est de tenter de reconstruire la conformation des polymères à l'aide de cartes de distance. À ce stade, je ne m'intéressais pas tant aux mécanismes responsables de l'existence d'une quelconque conformation préférentielle, mais simplement à la récapitulation de la position de tous les monomères dans l'espace telle que mesurée par une carte de distance. Pour s'assurer que chaque monomère va se détendre vers une position relative correcte par rapport aux autres, j'ai supposé que tous les monomères génèrent un potentiel de Lennard-Jones sur tous les autres.

Comme je ne m'intéressais pas au régime de non-équilibre, j'ai, par souci de simplicité, rapproché ce potentiel du quasi-équilibre, où il se comporte comme un simple potentiel harmonique quadratique. De plus, compte tenu de la viscosité dynamique élevée attendue pour le noyau de la cellule, on peux négliger les effets de l'inertie. L'équation de mouvement résultante est donnée par

$$\frac{d}{dt}\boldsymbol{r}_i = \lambda \sum_{i \neq j} \frac{r_{ij} - d_{ij}}{d_{ij}^2}\, \hat{r}_{ij}. \tag{5.1}$$

où λ module la force, tandis que d_{ij} se rapporte à la carte de distance et r_{ij} est la distance réelle entre les monomères i et j. La valeur exacte de λ n'est pas si importante, elle doit être suffisamment petite pour que le polymère puisse explorer autant de conformations que possible pendant la relaxation, mais suffisamment grande pour que les calculs ne prennent pas trop de temps. Pour nos simulations, j'ai utilisé $\lambda = 0,005 \mu m^2/s$.

Pour vérifier si l'équation (5.1) fonctionne correctement, échantillonnons une seule chaîne gaussienne et essayons de la reconstruire en nous basant sur sa carte des distances en figure (5.1a). Pour déterminer plus précisément la robustesse de cette méthode par rapport au bruit, nous introduisons un bruit lognormal avec divers σ. Un exemple avec $\sigma = 0,4$ est affiché en (b). Pour chaque niveau de bruit, je reconstruis 32 polymères pour estimer l'erreur moyenne et l'écart-type. En (c), je présente une carte de distance reconstituée pour $\sigma = 0,4$, tandis qu'en (d), les erreurs moyennes avec courbe ajustée. Une erreur moyenne d'environ 5 nanomètres

par monomère est trouvée pour $\sigma \to 0$.

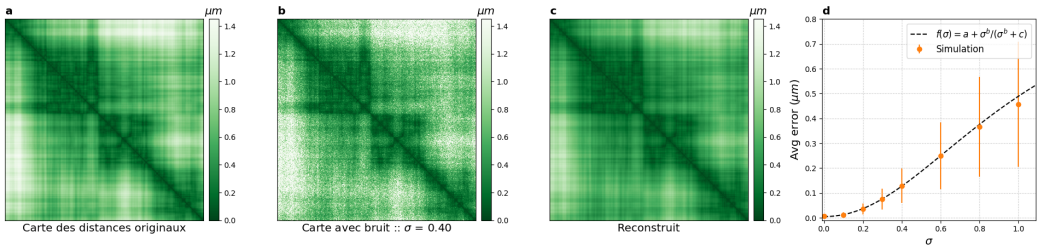

Fig. 5.1. : (a) Carte de distance calculée pour la chaîne gaussienne échantillonnée (b) En utilisant une distribution lognormale avec $\sigma = 0,4$, nous introduisons le bruit dans la distance propre. (c) Exemple de reconstruction à l'aide de la carte de bruit précédente. (d) Moyennes de l'erreur moyenne et de l'écart-type calculées en utilisant 32 polymères reconstruits pour divers σ en orange. La courbe ajustée présente $a = 0,005$, $b = 2,083$ et $c = 1,067$.

6 Modélisation de la dynamique de la chromatine

Le modèle de la chaîne de Rouse, présenté au chapitre 10, est le modèle de polymère le plus simple que l'on puisse imaginer. Il prend en compte les interactions simples de premier voisin via un ressort développant une sorte de dynamique stochastique (mais stationnaire) commandée par la température sous l'influence d'un substrat homogène. En raison des principes de symétrie, tous les monomères de ce polymère, à l'exception de ceux dans les bords, présentent un coefficient de diffusion apparente similaire avec un coefficient d'anormalie se situant à $1/2$.

A l'inverse, nos résultats expérimentaux montrent que le coefficient de diffusion varie, dans un intervalle crédible, en fonction de sa localisation relative dans la chromatine. Nous avons également déterminé que le coefficient d'anomalie est, en moyenne, inférieur au seuil théorique de $1/2$, ce qui indique que la chromatine est plus contraignante qu'un polymère libre.

Intuitivement, on se rend compte qu'il existe une relation entre le coefficient d'anomalie et le nombre d'interactions contraignantes associées à un monomère quelconque. Une seule particule se diffusant librement présente $\alpha = 1$, alors que le fait d'attacher deux ressorts raccourcit cette valeur de moitié. Par conséquent, on peut s'attendre à ce que plus un monomère a d'interactions contraignantes, plus le coefficient d'anomalie sera faible. A l'avenir, la question est de savoir quelles interactions sont importantes pour la dynamique et la conformation globale. La question la plus pertinente est peut-être celle de savoir comment ces interactions évoluent dans le temps. Malheureusement, même avec la technologie actuelle, il

est assez difficile, voire impossible, de répondre expérimentalement à cette question. Au cours de la dernière décennie, nous avons observé un grand intérêt pour les complexes en interaction, par exemple la cohésine-CTCF, qui s'est révélée être un mécanisme important pour la conformation. Néanmoins, il existe peut-être des dizaines d'autres types d'interaction qui pourraient contraindre la chromatine. Pour ne citer que quelques exemples, nous avons l'oligomérisation directe ou par médiation, les condensats, les interactions avec les repères nucléaires, entre autres [19]. En fait, la solution la plus probable serait une combinaison de ces éléments.

Quoi qu'il en soit, si différentes sections de chromatine ne sont en contact que temporellement avec une probabilité proportionnelle au nombre de lectures présentées dans une carte Hi-C, ces sections devraient également se trouver à des distances variables dans le temps. Nous ne savons pas exactement comment cette distance va évoluer dans le temps, mais nous connaissons sa moyenne. Par conséquent, en première approximation, nous utiliserons cette distance attendue comme médiateur de force pour la dynamique de chaque monomère.

Pour déterminer de manière probabiliste quelles interactions se produiront pendant une petite période de temps, j'ai utilisé une carte Hi-C. Malheureusement, elles contiennent une bonne quantité de bruit et plusieurs valeurs indéterminées en raison de problèmes expérimentaux connus, il faut donc les traiter. Afin de résoudre le problème des éléments manquants, j'ai utilisé l'interpolation. Pour réduire le bruit, j'ai convoluté cette carte avec un filtre gaussien à deux dimensions. En utilisant cette carte de contact de la population traitée, nous pouvons également calculer la carte des distances attendues en utilisant

$$P(\langle R\rangle \,|b) = \text{erf}\left\{\sqrt{\frac{4b^2}{\pi\,\langle R\rangle^2}}\right\} - \frac{4b}{\pi\,\langle R\rangle}\exp\left\{-\frac{4b^2}{\pi\,\langle R\rangle^2}\right\}, \tag{6.1}$$

qui décrit la relation entre la probabilité de contact et la distance moyenne $\langle R\rangle$ donnée à une valeur b, comme décrit au chapitre 10. Heureusement, nous avons déjà déterminé expérimentalement deux éléments de la carte de distance pour le domaine HoxA. En utilisant ces résultats, j'ai obtenu b = 56 nm pour les cellules SE et b = 93 nm pour les cellules PN.

Nous pouvons utiliser ces résultats ainsi que l'équation (5.1) pour générer des polymères à partir des interactions échantillonnées. Dans la figure (6.1a,b), nous avons des cartes Hi-C dans lesquelles les éléments manquants ont été interpolés et le bruit a été réduit avec le filtre gaussien. Au total, 2048 polymères ont été reconstruits à partir de différentes interactions échantillonnées. En (c, d), nous avons le nombre total de fois où différents monomères ont été trouvés en contact après la relaxation. En (e,f), la carte finale de la distance moyenne. Enfin, en (g,h), les

cartes de contact calculées à partir des distances moyennes. Comme nous pouvons le voir, les polymères reconstruits sont fortement corrélés aux données et peuvent expliquer aisément leur variance.

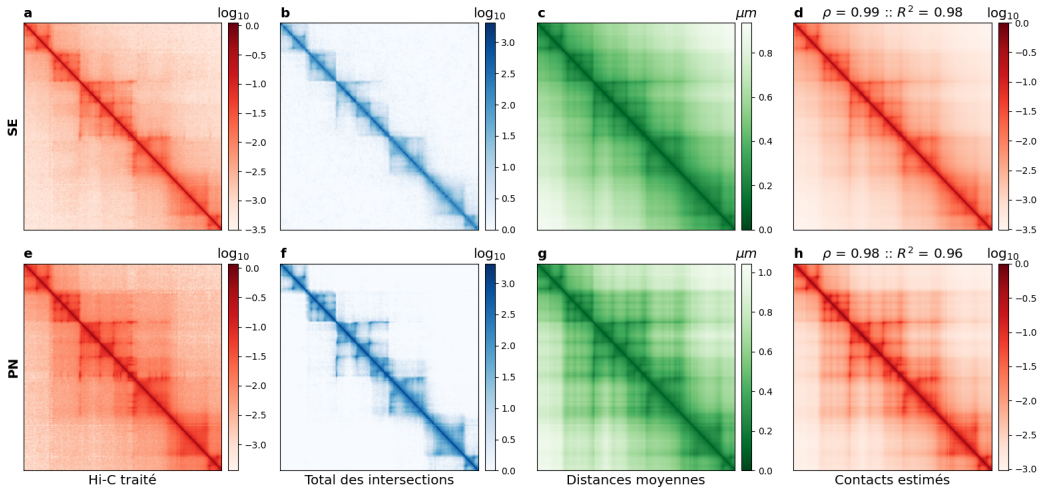

Fig. 6.1. : (a,b) Les éléments manquants des cartes Hi-C ont été interpolés et le bruit a été réduit en utilisant un filtre gaussien à deux dimensions. (c,d) Après la reconstruction de 2048, le nombre total de monomères se croisant a été compté. (e,f) Moyenne sur confirmation finale de tous les polymères. (g,h) Estimation des cartes de contact.

Comme la chromatine est fortement recouverte de protéines pour la régulation transcriptionnelle, on peut s'attendre à ce que, selon la façon dont un locus donné est régulé, les propriétés de diffusion de l'environnement différeront de celles des autres régions. Nous trouvons dans la littérature que les centromères et les télomères sont moins mobiles que la moyenne chez la levure [10], alors que les loci actifs transcriptionnels se sont avérés en corrélation avec des α plus petits et des D_α plus grands dans certains cas [11, 12].

Je modifie l'équation (5.1) pour en tenir compte et introduis une dynamique stochastique

$$d\boldsymbol{r}_i = \frac{3k_BT}{\gamma_i}dt\sum_{i\neq j}\frac{r_{ij}-d_{ij}}{d_{ij}^2}\,\hat{r}_{ij} + \sqrt{\frac{2k_BT}{\gamma_i}}d\boldsymbol{W}_t, \tag{6.2}$$

où γ_i est la viscosité dynamique associée à chaque monomère. Le bain thermique a été introduit via $d\boldsymbol{W}_t$, une force blanche aléatoire.

L'idée est de modéliser un γ_i local en utilisant les données de ChIP-seq, car il fournit des informations sur les facteurs de transcription qui se lient habituellement à la chromatine. Cela peut sembler facile au début, mais, en raison du nombre énorme de protéines nécessaires pour réguler l'ensemble du génome, ce ne serait

pas un outil durable à long terme. C'est pourquoi j'ai proposé d'utiliser des modifications d'histones à cette fin, inspirées du locus *Haunt* susmentionné, qui est enrichi en H3K4me1, H3K4me3 et H3K36me dans les cellules ES [20, 21], mais, lors de la différenciation, H3K4me3 et H3K36me3 diminuent et H3K27me3 augmente [16]. Par conséquent, nous pourrions modéliser une sorte de combinaison de signaux encapsulant des modifications majeures des histones et vérifier si nous sommes en mesure d'approximer nos résultats expérimentaux pour les coefficients de diffusion apparente et d'anomalie.

Dans la figure (6.2a), je présente les résultats pour les cellules ES utilisant H3K122ac, H3K4me1, H3K27ac et H3K64ac. Comme précédemment, 2048 polymères ont été simulés. Ils ont ensuite été divisés en 64 groupes de 32, à partir desquels nous utilisons l'DQM pour obtenir des mesures pour D_α et α ainsi qu'un intervalle de 95% de crédibilité. Ces résultats sont présentés en (b-c). Une procédure similaire a été appliquée aux cellules PN. Malheureusement, la disponibilité des données ChIP-seq pour les cellules PN est rare, c'est pourquoi seules les données H3K4me3 et H3K27ac ont été utilisées. Les résultats sont présentés dans la figure (6.3).

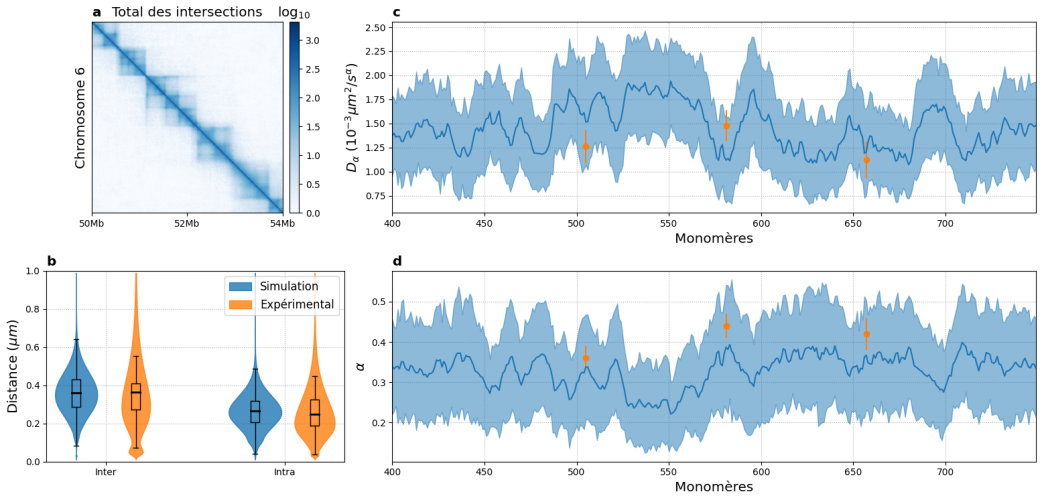

Fig. 6.2. : Comparaison des résultats obtenus à partir d'expériences et du modèle polymère pour les cellules SE. (a) Nombre de fois où des monomères ont été trouvés à proximité dans des polymères simulés. (b) Distribution des distances inter-sondes mesurées à partir de simulations et de données réelles. (c-d) En bleu, nous présentons la moyenne et un écart-type estimés pour les coefficients de diffusion apparente et de confinement à partir de 2048 simulations. En orange, intervalle de certitude de 95% pour la moyenne mesurée à partir de données réelles.

Ce travail est préliminaire et des recherches supplémentaires doivent être effectuées sur le sujet. L'une des premières choses à faire est peut-être de vérifier les

données ChIP-seq pour d'autres modifications des histones qui sont fortement associées à une activité génique spécifique. Cela devrait permettre de déterminer un meilleur système de pondération pour le signal final.

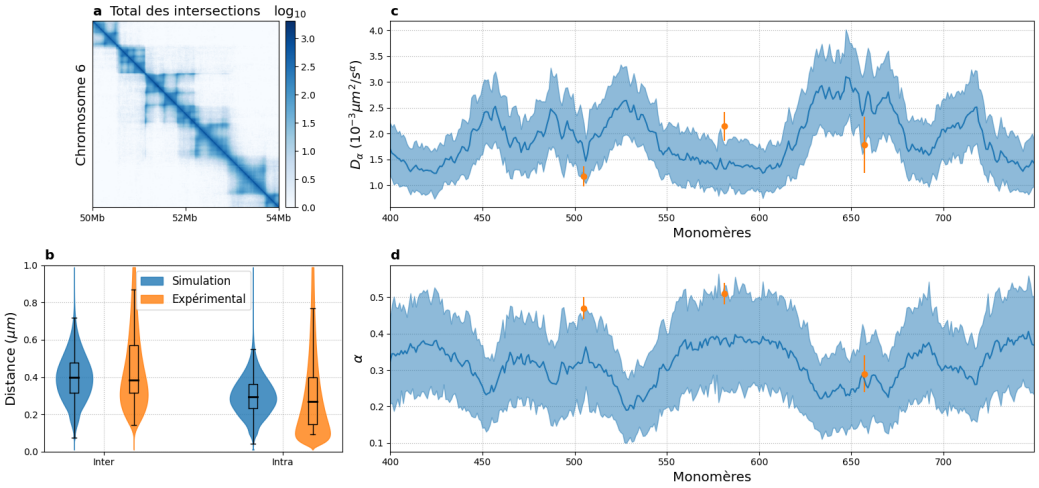

Fig. 6.3. : Comparaison des résultats obtenus à partir d'expériences et du modèle polymère pour les cellules PN. (a) Nombre de fois où des monomères ont été trouvés à proximité dans des polymères simulés. (b) Distribution des distances inter-sondes mesurées à partir de simulations et de données réelles. (c-d) En bleu, nous présentons la moyenne et un écart-type estimés pour les coefficients de diffusion apparente et de confinement à partir de 2048 simulations. En orange, intervalle de certitude de 95% pour la moyenne mesurée à partir de données réelles.

Quatrième partie

Conclusion

Nous avons présenté dans cette thèse un nouveau cadre pour analyser les coefficients de diffusion apparente et d'anomalie des taches marquées dans le noyau. Même si cette technique a été principalement utilisée pour étudier la dynamique de la chromatine, cette méthode peut être utilisée pour l'analyse de toute particule décrivant un mouvement de type brownien en supposant des déplacements gaussiens. L'utilisation de cette technique pour mesurer les coefficients de diffusion apparente et d'anomalie nous a permis de corriger nos résultats pour les effets du mouvement de fond et de clarifier les similitudes présentées dans la dynamique de la chromatine entre la mitose et l'interphase. De plus, grâce à des mesures précises, nous avons pu développer un modèle de biopolymère avec lequel nous simulons la dynamique locale de la chromatine qui récapitule les distances mesurées entre les sondes, les coefficients de diffusion apparente et d'anomalie pour les lignées cellulaires HoxA. Grâce à l'avantage d'un modèle théorique, nous avons pu établir que les propriétés de diffusion dépendent fortement du contexte chimique dans lequel les sondes sont insérées. Éventuellement, nous pouvons identifier un lien entre les propriétés dynamiques et l'activité des gènes. Quoi qu'il en soit, la méthode utilisée pour tenir compte des effets du contexte local doit être étudiée plus en profondeur.

Summary

Chromatin organization and its role in genome regulation is a fundamental concept involved in biological, pharmaceutical and health related research. There is vast literature addressing the subject from a static and chemical perspective, but the underlying associated dynamics has been largely overseen up to recent years. For that reason, this thesis encloses some results I obtained during my PhD years regarding dynamical properties of chromatin in diverse cell cycle stages and a possible connection relating gene activity to local mobility and anomalous behavior of chromatin.

To reach this goal, we developed a new computational framework based on Gaussian processes and fractional Brownian motion called GP-FBM. Using this method, I was able to infer values for apparent diffusion and anomalous coefficients more accurately, as Gaussian processes naturally account for high-order temporal correlations. For similar reason, we were also able to extend this method to correct for background movement in a natural way for systems with two or more particles, that is, no computational post-processing or extra experimental setups were necessary.

In order to extend our understanding of the experimental data provided in this thesis, I further introduce a new biopolymer model using a mean-field approach in which I use Hi-C maps to model chromatin long-range interactions and histone marks via ChIP-seq data to account for local properties of the nuclear environment. This model was able to recapitulate experimental distances and inferred values for apparent diffusion and anomalous coefficients measured via confocal microscopy for specific loci of the HoxA domain in mouse cells.

Contents

List of Figures

Part I

Introduction

1

Chromatin structure

DNA contains the basic code for the proteins necessary to maintain life, these "recipes" are called genes. It is formed as a combination of four basic units namely adenine (A), cytosine (C), guanine (G) and thymine (T) connected in pairs and organized with backbones of phosphate-deoxyribose. At the core of any mammalian cell, we find over 6 billions of these base pairs of DNA. If we organized all these units in a straight line, it would be about 2 meters long, as each of its components is 0.34 nanometers big. Now considering we have trillions of cells, how far away could we go? Farther than Voyager 1, the first human made object to leave the solar system, at the present year. We might imagine how nicely organized and packed DNA must be to fit in a rather small volume named the cellular nucleus in eukaryotes. Even more interesting is the fact that, in mammals, only about 50% or less of the genome is actually encoding for genes [22].

At first instance, DNA is wrapped about twice around a combination of 8 proteins called histones H2A, H2B, H3 and H4 (two of each), which receives the name of nucleosome [23]. Nucleosomes are the basic units of chromatin, which is later organized into functional domains such as euchromatin and heterochromatin. Euchromatin is loosely compacted and is considered to be transcriptionally active. Whereas heterochromatin is more condensed, which is believed to prevent access of transcription machinery inhibiting gene expression. Heterochromatin can be subdivided into two other subcategories [24, 25]: "Constitutive" contains genes that are permanently silenced, as in telemores and centromeres; "Facultative" presents genes that may (or not) be active in a cell type.

Euchromatin and heterochromatin can also be distinguished due to differences found in their histones. Since histone modifications were discovered [26], they

have been correlated to gene activity. For instance, their tails can protrude from one nucleosome to the next affecting inter-nucleosomal interaction. Besides that, it was also found that histone modifications might be involved in DNA repair, replication and recombination [25].

Among several recently found types of histone modifications, I would like to mention the more traditional ones: acetylation and methylation, as they will be important towards the last chapter of this thesis. Histone acetylation has been associate to nucleosome uncoiling upon action of histone acetyltransferases or HATs. As a gene must be accessible by transcriptional machinery in order to be transcribed, this modification is directly associated to gene activity. Histone methylation is not so obvious, because depending on which residue is modified it could be associated to euchromatin or heterochromatin. For example, histone H3 lysine 4 trimethylation (H3K4me3) is found at active regulatory sequences, while histone H3 lysine 27 trimethylation (H3K27me3) is found at facultative heterochromatin domains [27].

1.1 Chromosome conformation capture

As we have discussed above, chromatin is formed as a mechanism to organize long DNA molecules in the nucleus. With development of chromosome conformation capture technologies [28] in the past decade or two, we have observed that chromatin is further organize in such a way that distal regions are brought together. This clustering is known as TAD, or Topological Associated Domains, and it has been widely correlated to histone modifications and gene expression [29, 30]. There are cases in which disruption of these domains will generate malformation during development and other diseases [31]. Nonetheless, several studies have also disrupted the regular shape of TADs without greatly affecting transcriptional rates [32]. At the current day, even with the growing number of studies on the subject, it is still unclear how these structures are formed, maintained and modified through the cell cycle and differentiation.

There are several proposed mechanisms suggested in the literature for the formation of TADs and larger order chromatin organization. Some of these are associated to direct oligomerization of transcription factors and co-factors, protein clustering via condensation, histone modifications, DNA methylation, among other mechanisms [19]. Perhaps one of the most popular and successful theories presented in past years is the loop extrusion model [33, 18]. It is proposed that loop-extruding factors, such as cohesin, form ever larger loops until it encounters TAD boundary proteins like CTCF. Likewise, loops can be formed within loops as TAD

boundaries are assumed to present extrusion permissive direction. Several *de novo* simulations implementing such mechanism have displayed average TAD-like structures.

For the work developed in this thesis, we will try to determine if the above mentioned structures cause variations in the dynamics of different chromatin regions. For that purpose, we are not particularly interested to determine specific mechanisms by which all these fundamental structures are formed, but simply what are their effects on the overall dynamics. As many similar projects found in literature, we are going to utilize Hi-C data, which uses high throughput sequencing to measure chromatin conformation and generate contact maps.

First introduced by Lieberman-Aiden *et al* back in 2009 [34], this method is relatively straightforward to understand and runs as follows:

— **Cross-linking DNA:** Using formaldehyde, chromatin is cross-linked so that "sufficiently" closed regions are fixed together. A great review on the chemical aspects and how formaldehyde works is presented in [35];
— **Cut with restriction enzyme:** There are a few possible choices, popularly HindIII and NcoI are used. Importantly, these restriction enzymes should target symmetric sequences such as AAGCTT, which is important for next steps. Usually, the restriction enzyme will dictate how deep or resolved finer structures will be at the final result;
— **Fill ends and mark chromatin:** When cross-linked regions had their ends completed and marked via biotin for later purification;
— **Ligation:** When symmetric ends are re-connect;
— **Purification and shearing:** DNA is sheared and connection is purified making use of streptavidin beads;
— **Sequencing:** Purified fragments are sequenced and identified accordingly.

Upon alignment to a prior known and well defined genome, we can count the number of times long range interactions occur, that is, number of times distal sequences are found together. Naturally, this type of experiment is noisy due the stochastic dynamical nature of chromatin and technical issues. In that sense, more often than not, random sequences will be found together without any specific or functional reason. Nonetheless, upon usage of millions (maybe billions) of cells, we eventually determine some configurations that are more often observed than others. For that reason, we might expect that these configurations reflect function. Unfortunately, the interpretation of such maps are not straightforward due to all the possible experimental biases introduced in each step of its obtainment. A more detail discussion on the subject is present in [36].

There are dozens of methods and algorithms present in literature trying to overcome some of these biases. In general, different methods were developed with intention to study specific questions, hence a more complete comparison among several methods can be found [37]. Here I will present some of the most popular ones, usually correcting for the compared visibility of certain sequences by balancing the raw read maps generated via sequencing and alignment.

Used in the original work by Lieberman-Aide *et al* [34], the VC method (Vanilla-coverage) divides every element of a row by the accumulated signal of that row and, subsequently, column-wise. The KR method (Knight-Ruiz matrix balancing [38]) uses an algorithm to normalize symmetric matrices in which every row and column sums up to 1. In figure 1.1, we present an example of Hi-C matrix balanced using this method. Finally, the ICE (Iterative correction and eigenvector decomposition [39]) balances the matrix by removing experimental biases and using eigenvectors decomposition techniques to analyze different chromatin patterns.

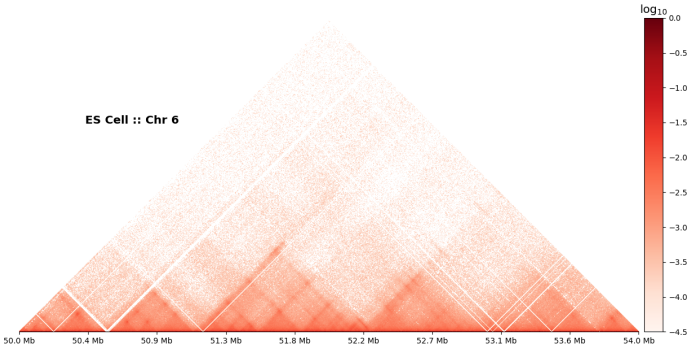

Fig. 1.1.: Hi-C map balanced using the KR method. This section of chromatin in chromosome 6 encodes for the *HoxA* gene. We are going to analyze this locus in greater detail later.

There are several models in literature where Hi-C datasets and, sometimes, FISH [1] [40] are used to reconstruct synthetic polymers. We are going to discuss the subject in greater detail later on, but, for completeness, there are two broad groups in which we can insert these models [41]. The first will try to determine chromatin consensus structure directly from these maps, disregarding the dynamic behavior of chromatin and how improbable such conformation is to occur in real life. Differently, some authors try to detect possible key interactions that will recapitulate on average the expected Hi-C map upon simulations of thousands of polymers. As we shall discuss later, I propose a method in between both approaches, that is, we are going to use Hi-C maps to determine how frequently certain interactions occur and

1. Fluorescence *in situ* hybridization (FISH) uses fluorescent probes that bind with high specificity to certain DNA sequences. Like that, we can localize these regions using fluorescent microscopy.

how far apart corresponding chromatin sections are from each others.

1.2 ChIP-seq as a measurement of heterogeneous environment

Gene expression is mediated through the action of special proteins called transcription factors (TFs) along with other co-factors. When binding to certain regulatory sequences encoded on DNA, these TFs will modify the local environment in such way that will promote or repress genes for RNA Polymerase (Pol II) transcription [42]. We find in literature two main accepted types of regulatory elements encoded by DNA: promoters or enhancers, sometimes referred as proximal and distal regulatory sequences. Promoters are found upstream, close to transcription starting site (TSS) of genes. Differently, enhancers are also found downstream or in introns of respective or unrelated genes [43, 44]. A typical human gene is usually regulated by multiple enhancers, in fact, proximal regulatory sequences are incredible outnumber (orders of magnitude) by distal ones. Which increases our believe that chromatin conformation is also correlated to transcriptional activity, by bringing enhancers and promoters close or apart in space.

But how do we know where specific TFs bind in chromatin? We can answer this question via the experimental protocol called ChIP-seq, standing for chromatin immunoprecipitation followed by sequencing, which is precise enough to determine binding sites up to about 10bp precision. Furthermore, we can indirectly measure how strongly proteins interact with those sequences depending on the final signal obtained. Nonetheless, depending on the strength of any given interaction and how specific it is, millions of cells are needed for a robust signal due the stochastic nature of diffusion dynamics described by transcription factors and other proteins. For that reason, results should be interpreted as a population average.

The protocol implemented for this method is conducted as follows

— **Cross-linking:** The first step of the protocol for a ChIP-seq experiment is cross-linking with formaldehyde. This step ensures that any protein of interest will remain attached to chromatin in the next steps of the protocol. Notice that, although possibly biased by the way proteins interact with chromatin, ChIP-seq will not provide us with any information regarding the actual underlying mechanism;
— **Fragmentation:** Once cross-linked, the material needs to be fragmented. There are a few methods for that, such as sonication or digestion using en-

zymes. Naturally, smaller fragments will increase the final resolution of the method, but the smallest size should depend on the effective interaction area of chromatin with the protein;

— **Purification:** To purify fragments of interest, one needs to use very specific antibodies for the protein of interest. Non-specificity here will increase experimental noise;

— **Reverse cross-linking:** This step can be accomplished using enzymes to digest proteins or over extensive heat incubation;

— **Sequencing and alignment:** Purified fragments are, then, sequenced and aligned to a reference genome.

For most of the purposes in this thesis, as a preliminary approach, we are not interested in the raw signal from these experiments, but to determine where interactions peak. In other words, where proteins are interacting more strongly. There are several methods for this purpose, but we are going to use MACS2 that empirically models shifts in the data to improve spatial resolution. More details are found in [45]. This method will basically give a binary result regarding where proteins of interest are bound. In figure (1.2) we bin up peaks in regions of 4 kb (same resolution of Hi-C maps used throughout the thesis) and perform Gaussian smoothing for noise removal.

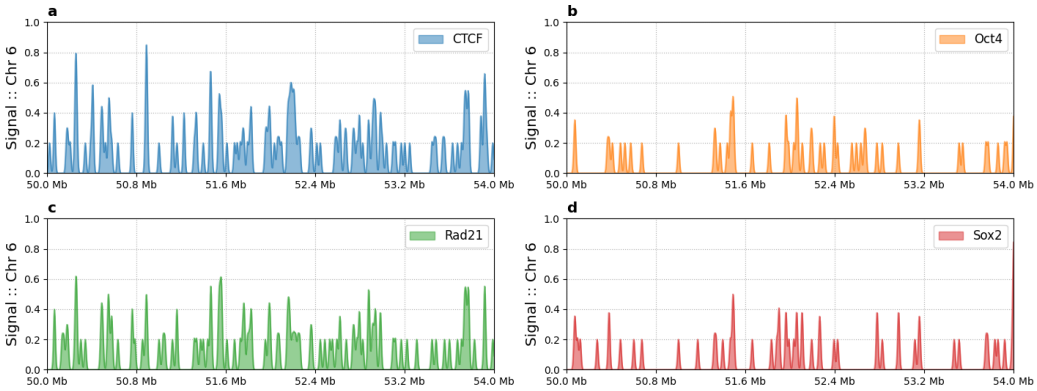

Fig. 1.2.: ChIP-seq signal for CTCF and diverse histone modifications are displayed. Raw data is binned in 4kb sections using a binary approach via a set threshold. Then, a Gaussian filter with $\sigma = 2$ was used.

This figure shows us that different sections of chromatin will interact with a wildly heterogeneous environment, where different regions will contain variable densities of molecules with diverse chemical potentials. There are several machine learning methods developed in the past few years [46, 47, 48] where authors attempt to determine probable binding sequences for certain proteins solely based on the combination of base pairs found via ChIP-seq. Notwithstanding, these meth-

ods neglect conformation and possible chemical interactions that these proteins present, thus decreasing the overall precision and predictive power of these approaches.

These measurements made via ChIP-seq will be important when we developed our biopolymer model if we want to also have increased precision in simulated dynamics. Which proteins should we consider for regions of interest? Possibly many. As a preliminary result, we shall consider an effective approach by not considering interacting proteins directly, but histone modifications correlated to chromatin states. As an example that we shall consider in more detail later, the HoxA domain is repressed by the *Haunt* gene. Hence, this genomic locus is enriched in H3K4me1, H3K4me3 and H3K36me in stem cells [20, 21]. Oppositely, upon differentiation H3K4me3 and H3K36me3 decrease, whilst H3K27me3 increases [16]. In chapter 12 we shall use these results to recapitulate measured dynamics parameters in the HoxA domain.

2

Physical models for diffusion dynamics

Molecules are subject to thermal fluctuations and stochastic interactions inside the cell. Advanced novel imaging techniques allows us to measure the movement of single particles which provides us with a better understanding of their interactions and the complex media they move in. However, to extract all the information offered by these techniques is not trivial due the degree of complexity implied in the dynamics of biological particles.

Without considering energy driven mechanisms such as molecular motors, most particles in the cell will move due diffusion. We are going to described this process in greater detail when necessary in part 3, but we shall introduce some basic ideas behind this mechanism. Generally speaking, the phenomenon of diffusion is generated when any particle of interest randomly collides with neighboring smaller particles. In the simplest model, know as Brownian diffusion, all of these collisions are elastic, hence no energy is consumed nor lost during this process, but simply distributed and balanced across the whole system of interacting particles. If we analyze the displacement of this particle over time and do some statistics on it, we can associate a parameter D to how mobile this particle is given a finite amount of time. Henceforth, we shall call this parameter as diffusion coefficient. In one of his miraculous paper, Einstein demonstrated that this coefficient depends on the particle cross-section and the fluid viscosity. Even though the overall trajectory displayed by this particle is stochastic, we can calculate an ensemble average over a sufficiently large amount of particles and determine that their mean squared displacement (MSD) is given as follows

$$\left\langle (\boldsymbol{r}(t) - \boldsymbol{r}(0))^2 \right\rangle = 2nDt, \tag{2.1}$$

where $\boldsymbol{r}(t)$ is a n-dimensional vector representing the position of particles in time. For our experiments later on, we will use 2-dimensional microscopy movies, hence n shall be considered as 2.

Unfortunately, particles in the cell do not interact with its neighborhood in an elastic fashion. There are many different types of chemical potentials with which any given particle undergo when diffusing through the cell. For instance, any section of chromatin will chemically interact via covalent bonds with its adjacent sections, hence we should expected a different MSD curve from the above. In those situations, the diffusion dynamics receives an extra term called the anomalous coefficient α, which hints us about the type of diffusion mechanics undertaken by any particle of interest. Defined in the range $0 < \alpha < 2$, the new MSD curve is given by

$$\left\langle (\boldsymbol{r}(t) - \boldsymbol{r}(0))^2 \right\rangle = 2nD_\alpha t^\alpha, \tag{2.2}$$

and accommodates 3 different types of motion: for $\alpha < 1$ we are in a sub-diffusive regime, while $\alpha > 1$ corresponds to a more directed type of motion. Notice that the diffusion coefficient now presents a dependency in α. For that reason, we shall call it apparent diffusion coefficient from now on. Nonetheless, one can restore traditional results by setting $\alpha = 1$. In figure 2.1 we can observe the expected MSD curve for these 3 regimes.

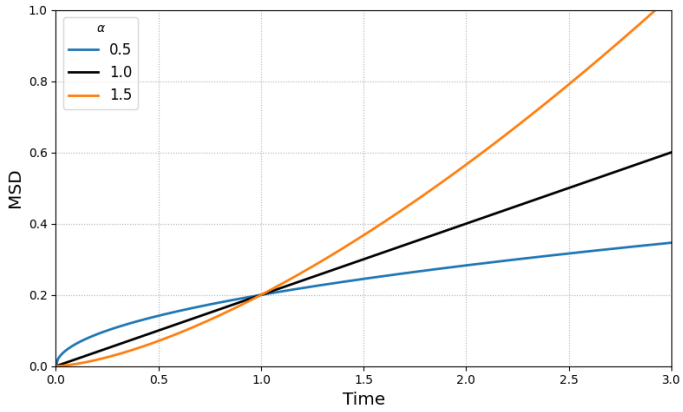

Fig. 2.1.: Comparison between values of α and expected MSD behavior.

Trying to explain such anomalous behavior, many Physics-based models have been developed since mid 20^{th} century. One of these models, namely Continuous Time Random Walks (CTRW) [49], was first presented as a generalization for the traditional diffusion process described by Einstein in 1905. In this model, the passage of time is stochastic, therefore particles might remain at their position for a random period time before next spatial step is taken. If time probability distribution is exponential with well defined mean, CTRW recapitulates the original diffusion

scheme. By modeling time distributions, one can use this model to study different instances of anomalous diffusion.

Differently, one can approach this problem by assuming environmental effects on the diffusion properties of molecules. We find in literature case studies involving effects of confined spaces on diffusive particles[50], space with randomly inserted obstacles [51] or with topology described by fractals [52]. As a common point among all these models, they attempt to recreate some features found by molecules in the cellular environment. A great review on some of these models, among others, can be found in [53]. I would also like to highlight that many of these models consider that the spatial displacement of molecules are Gaussian, which is not always the case in real life. In recent years, we find in literature increasing indications of non-Gaussian dynamics [54, 55, 56]. Curiously, the mechanisms for such non-conventional behavior is still not well understood.

As the work developed for this thesis is centered on the dynamics of chromatin, it would be preferable to develop a Physics based model in which we know how much specific features collaborate into the final measured dynamics. There are several polymer models described in literature [57, 18, 58] and a great review on modeling approaches is found in [59].

For the work presented in this thesis, we are going to consider the Langevin dynamics approach, that is, an"extension" of the Newtonian mechanics, in which stochastic forces are included. The equation behind this method is

$$m\frac{d^2\boldsymbol{r}_i}{dt^2} = \boldsymbol{F}_i - \gamma\frac{d\boldsymbol{r}_i}{dt} + \sqrt{2k_BT\gamma}\,\boldsymbol{\eta}_i(t), \tag{2.3}$$

where we split chromatin into finite sections "i" and balance all the forces acting upon them. The first term considers the effects of external forces in relation to inertial properties of each section, while $\boldsymbol{F}_i$ accumulates forces exerted by other regions of the polymer. Terms involving γ are related to effects of the solvent in which our polymer dwells including diffusion.

As an entry point to our analysis, we are going to considered the Rouse chain model, the simplest polymer model, where each section will interact solely with its first neighbors in a homogeneous solvent. This model will act as a null hypothesis for interpreting our experimental results and it will be used as base for my own model. Then, we are going to propose $\boldsymbol{F}_i$ based on contact probabilities measured via chromosome conformation capture maps and, after that, individually model environmental effects upon each section of chromatin based on ChIP-seq experimental data.

In order to calibrate our model, we are going to need experimentally obtained values for D_α and α. In the next section we will show a few methods commonly used in literature to measure such coefficients. Furthermore, we will conceptualize, in similar spirit of the work originally presented in [60, 61], the idea of using Gaussian processes to accommodate for self-correlation when dealing with single particle experiments. We are also going to use GP to interpolate particle positions and correct for background movement if more than one particle is present. All these methods will have as base assumption that the movement observed is solely due to anomalous diffusion, that is, any other mechanism of motion, if existent, will be inferred as a effective diffusive parameter further implied via anomalous behavior.

2.1 Measuring apparent diffusion and anomalous coefficients

In upcoming part of this thesis, we will want to determine an apparent diffusion and an anomalous coefficient for several chromatin loci. More than that, we will also be interested to obtain some statistics on these measurements in order to establish how much data variability we have if concerning cell-to-cell and spot-to-spot differences. For that reason, in many cases, we would like to determine measurements as precise as possible for single trajectories rather than over ensembles of particles. In the following subsections we show 2 of the popularly methods found in literature and quickly introduce a repurposed method used during my studies. As we shall see in chapter 8, chromatin tends to display Gaussian displacements for the experimental time scale we used, hence all the following methods apply.

2.1.1 Mean Squared Displacement (MSD)

As we discussed before, the traditional way to estimate D_α and α is via calculation of an ensemble average MSD curve (EA-MSD). Oppositely, if we assume the system to be ergodic, we could also estimate an MSD curve from single trajectories by oversampling its displacement. This method is know as time average MSD (TA-MSD). Unfortunately, that is not always the case, as it was found that some cell processes undergo periods where ergodicity is not held [62]. Notwithstanding, we might always expect that for short periods of time ergodicity should hold. An overview on optimal experimental setups for measurement of diffusion dynamics can be found in [63].

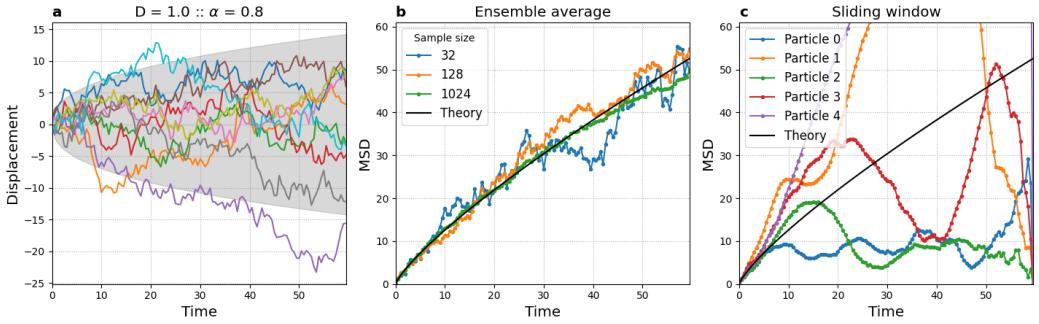

Fig. 2.2.: (a) Few examples of simulated displacements with $D_\alpha = 1$ and $\alpha = 0.8$. Shaded area correspond to range in which we expect 95% of displacements to be at any given time. (b) The mean squared displacement is calculated over a certain number of trajectories and compared to theoretical curve. (c) Single particles are over-sampled so average displacement can be calculated for different time points.

In figure 2.2, we have a comparison between MSD curves calculated via both methods on 2D trajectories simulated using $D_\alpha = 1$ and $\alpha = 0.8$ for a period of 1 minute. In (a) we present the displacement of single particles along with a shaded area corresponding to a theoretical 95% confidence interval. In (b) the MSD curve is calculated by averaging over many trajectories. As the theoretical curve accounts for an "infinity" amount of particles, we observed that simulated results asymptotically approach this limit for sets of ever larger number of particles. In (c) we estimate the average squared displacement in time from single particles. For that purpose, we divide the total amount of time recorded into ever growing intervals of time and calculate the average squared displacement for each group. As single trajectories are high correlated [63, 64], this method only works appropriately for short periods of time. Generally, only 5 or 10% of the obtained MSD curve is used to estimate diffusion and anomalous coefficients. There are attempts to solve this issue by modeling noise and auto-correlations implied in TA-MSD [65], but it is also poses new complications.

2.1.2 Displacement distribution

We can also calculate dynamic properties using theoretical displacement distributions, which is the approach used in Spot-On [66] and other methods [67, 56]. Let's considered the standard solution of the convection-diffusion equation with localization error σ for an ensemble of 2D trajectories

$$\rho(x, y|D_\alpha, \alpha, t, \sigma)\, dx\, dy = \frac{1}{2\pi(2D_\alpha\, t^\alpha + \sigma^2)} e^{-\frac{x^2+y^2}{4Dt^\alpha+2\sigma^2}} dxdy. \tag{2.4}$$

For simplicity, to deal with displacements in a more natural frame work, we convert euclidean into polar coordinates such as

$$\rho(r,\theta|D_\alpha,\alpha,t,\sigma)\ drd\theta = \frac{r}{2\pi(2D_\alpha\ t^\alpha+\sigma^2)} e^{-\frac{r^2}{4Dt^\alpha+2\sigma^2}} drd\theta. \tag{2.5}$$

In figure (2.3) we compare this result to displacement distributions measured from a single long trajectory simulated using $D_\alpha = 1$ and $\alpha = 0.8$ for 2000 time steps. As we shall see in a next chapter, this method tends to be more robust even in the single trajectory regime, but still a small percentage of time steps should be considered due biases introduced via trajectories self-correlation [66, 63].

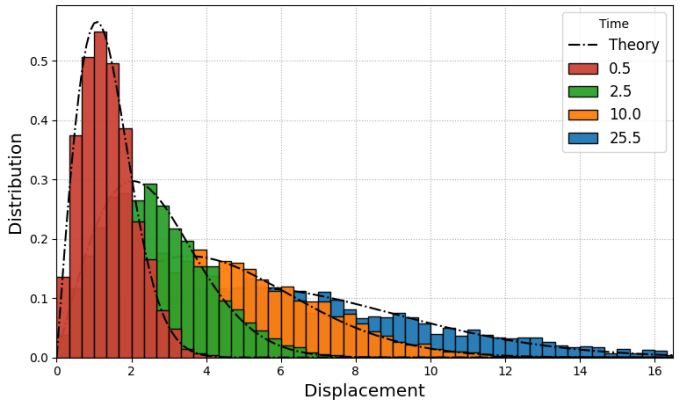

Fig. 2.3.: A single long trajectory was simulated using $D_\alpha = 1$ and $\alpha = 0.8$. Displacement distributions are calculated for several time steps and compared with theoretical curves 2.5.

Using result (2.5), we can also calculate the mean displacement curve

$$\langle r\rangle = \sqrt{\frac{\pi}{2}(2D_\alpha\ t^\alpha+\sigma^2)}. \tag{2.6}$$

and its second moment

$$\langle r^2\rangle = 4D_\alpha\ t^\alpha + 2\sigma^2, \tag{2.7}$$

which recapitulates the traditional MSD curve.

2.1.3 Gaussian process via covariance matrix

Trying to solve the problem of single particle auto-correlation in time, it has been proposed in recent years [60, 61] a different method evolving Gaussian processes (GP) [1, 68]. GP is a stochastic process encapsulating the distribution over continuous functions. For example, a function $X(t)$ can only be described as a GP if

and only if every finite set $\{X_t; t \in T\}$ is described by a multivariate normal distribution. Assuming the particle describes a fractional Brownian type of motion [50, 53], we model the GP covariance matrix of a single trajectory for time points t_1 and t_2 as

$$\mathbf{\Sigma}_{D,\alpha}(t_1, t_2) = D_\alpha \left(|t_2|^\alpha + |t_1|^\alpha - |t_2 - t_1|^\alpha \right), \tag{2.8}$$

and infer values for diffusion D_α and anomalous coefficient α via likelihood maximization of a multivariate Gaussian distribution. Further details will be discussed in chapter 6. As we shall see, this model tends to be more precise and less biased for single particles than the methods described above.

3

Aims

The work developed during the course of my PhD concerns the experimental characterization and mathematical modeling of chromatin dynamics. Thence, this thesis will be divided into 2 major parts. This organization will indubitably scatter the final goal of my PhD research, so I thought of using this chapter as a summary of what we expected to achieve and how these goals were reached (or partially).

The first part will be experimentally driven. In association with other members of our lab and the team of Thomas Sexton at IGBMC, we ask about the effects of topology, structure and locus activity on dynamic properties of chromatin. To answer this question, our lab used a cell line shared by Giorgetti's team (Basel) in which a PiggyBac system is used to randomly tag chromatin loci in mouse embryonic stem (mES) cells. The displacement dynamics of these loci are recorded via fluorescent microscopy for a short period of time (2 minutes) in interphase and mitosis. This experiment will be used to study general differences (or similarities) between chromatin dynamics under very different stages of condensation.

Towards the same direction, we also want to determine if the diffusion and anomalous coefficients might differ depending on the locus. Sexton's lab developed cell lines in which 3 specific loci of the HoxA domain in mES cells and tagged using ANCHOR fluorescent probes. Furthermore, upon culture with retinoic acid, these cells are induced into differentiation towards neuron precursor (NP) cells. Like so, we can determine what are the effects of differentiation, or diverse gene regulation states, over the dynamics of these probes. This is possible, because HoxA genes are repressed in ES state, but active once cells are differentiated.

Differently from what one might imagine, the analysis of these data is not (at all) straightforward, so several methods were developed for that purpose. Hoping

to reach a broader audience, I start the part II of this thesis by introducing a few concepts of probability theory and statistics. We shall discuss the main differences between Bayesian and frequency based statistical approaches, as well as the central limit theorem (CLT), a very important concept in statistics. If the reader is already comfortable with these concepts, feel free to skip these chapters.

Using these basic concepts we are going to introduce fractional Brownian motion as a model for the covariance matrix of multivariate Gaussian distributions. Using this method we will be able to consider particle trajectories with Gaussian displacements to the fullest, without discarding temporal auto-correlation, what allows us to obtained more accurate inferences for D_α and α if compared to other methods.

In the Bayesian framework, we are also going to develop models to enhanced particle localization in microscopy movies. Along this line, a method was developed to correct for misalignment between microscope channels, which occurred due to camera problems and chromatic aberration. Finally, we are going to introduce a new model to correct inferred values of D_α and α in situations where the substrate is moving. Different from many approaches in literature, no extra experimental setup or data post-processing is required.

To achieve a deeper and fundamental understanding of the experimental results in part II, we are going to develop a Physics based model for chromatin in part III. For this model, we are going to consider the population average conformation of chromatin, as visualized via Hi-C maps, to reconstruct synthetic polymers with similar conformation to the HoxA domain. Our first goal here is to determine if the distances measured between HoxA domain probes are recapitulated. Later on, we are going to insert dynamics into the system and fine tune diffusive properties of each chromatin section using ChIP-seq data as an assessment over the context in which each section of our polymer dwells.

Before that, though, at the beginning of part III, I am going to described the nature of diffusion and how we can treat it mathematically. I am also going to introduce methods with which we can simulate such phenomenon in a computer. After that, we are going to introduce the Rouse chain as a first approximation for chromatin. Based on this model, we are going to develop a new method to reconstruct chromatin and simulate dynamics. Finally, we are going to connect loci dependent context and show that experimental values for D_α and α can be recovered. Nonetheless, these last results are to be considered preliminary.

Part II

Data Analysis: Measuring chromatin dynamics

4

Probabilities

For centuries, it was believe that the universe worked in a deterministic manner. Assuming classical mechanics to be correct, if we comprehend any phenomenon well enough, we could formulate a set of equations capable to describe it with utterly precision. Very much so, Pierre-Simon Laplace created the anecdote in which an intellect, knowing the position and velocity of every particle in the universe, would have past and future to be inevitable. Nowadays we know that is not the case. With development of quantum mechanics and its many interpretations, presently we know the universe is fundamentally probabilistic. We shall not dive into more profound philosophical discussions on the subject. It is fairly out of this thesis's scope. In fact, we still do not know how all these randomness affects us in daily basis [69]. Conversely, this information is not required for our purposes.

In a general approach, probabilities should be used whenever uncertainty is present. The most popular example would be the toss of a coin or dice. Even assuming classical mechanics as utterly correct, the precision of execution necessary for an accurate/predictable toss is so great, it does not worth the effort. First we would need an object with mass as homogeneously distributed as possible. We would need a mechanical device capable to launch the object in a precise way. We would also need the whole system not to have strong vibrations and the air flow to be weak enough not to alter object's movement. On the opposite side, we can still obtain some information about such system even with so careful arrangements. There is a whole branch of Physics called Statistical Mechanics that works under this assumption. It would be close to impossible measuring the position and velocity of every particle in the atmosphere, for example. Some would say such stupendous effort is useless. Nonetheless, we can still estimate temperature, pressure and other average properties associated to our atmosphere. Essential parameters

in the development of modern engineering, architecture and biology.

If the universe is uncertain, one might ask, how can we create theories for phenomena around us? It turns out nature is great, because even random behavior presents certain intrinsic properties that are reproduced on average. Despite this randomness, we can use this expected behavior to develop models for our system and make predictions. Depending on how trust worthy our equipment is and how well we understand the subject matter, our predictions will be more or less accurate.

During the course of this thesis we shall use experimental data and stochastic simulations in order to determine and possibly explain the average behavior of dynamic properties of chromatin. The mathematical models used to describe this average behavior will be described later on. In this chapter I would like to describe some probability functions I will use to account for randomness. The first section will focus on discrete probabilities. After that I show probability functions applied over continues variables. Further along we exemplify the Central Limit Theorem and demonstrate the "law of total variance", used to discern sources of variability in our data.

4.1 Discrete probability

Discrete probability functions are used to study the randomness in systems containing only integer values. Many examples fall under this classification: coin or dice tosses, cards in a deck, distribution of birthdays in the year,etc. Perhaps more useful for our purposes is noise in microscopy images. As probability theory is better understood with aid of an example, let us use microscopy data as reference.

For the movies recorded, we had only 16 bits per pixel to store information. The amount of photons emitted from the observed object is a stochastic variable in time. Similarly, the number of photons corresponding to an single bit of information might also vary depending on the accuracy of our equipment. Gratefully, we can consider those as source of noise and treat the data via statistical tools. If we record the same image N times, we can calculated the sample average signal for each pixel as

$$\langle x \rangle = \sum_{k=0}^{N} \frac{x_k}{N}. \tag{4.1}$$

There is no reason to expect the quality of an image to be better than any other if all the images in this sample are taken under same conditions. For that reason, all images are weighted similarly, i.e. $1/N$. This is a general assumption when

sampling.

To measure how much noise is included in our measurements, we can calculate the sample variance for each pixel as follows

$$\text{var}\,[x] = \left\langle (x - \langle x \rangle)^2 \right\rangle = \sum_{k=0}^{N} \frac{(x_k - \langle x \rangle)^2}{N} = \langle x^2 \rangle - \langle x \rangle^2 . \tag{4.2}$$

In other words, the variance measures how far away from average the measurements are. Notice that we use a square exponent. This has two functions: first, to make all differences positive; second, the exponent gives higher weight to outliers. Worth noticing that, even though measured values are integers, average and sample variance are real numbers.

On a deeper level, each one of these samples is an approximation to what we call population distribution. We could count an incredibly large amount of photons and obtain a precise depiction of the object, but that could take an incredible long time and cost very expensive. In the case of live imaging that is simply impracticable, because cells tend to move or die due the extra thermal energy insert by photons. For those reasons, we can only record a few samples and treat it statistically.

Average and variance are of prime importance in any statistical analysis. For many cases, these provide us with enough information for final purposes. Nonetheless, if we aim higher and intend to generate predictions, we need to model the population data-set. There are many models described in literature for discrete population distributions, here we mention the uniform, binomial and Poisson distributions.

As a final remark, equations 4.1 and 4.2 are applied for any sample, even if events are not discrete in nature. For clarity, we should use $\langle\rangle$ and var [] as notation for sample average and variance, while E [] and Var [] for population mean and variance.

4.1.1 Uniform distribution

This is the simplest probability distribution depicted by a finite set of events where we have no reasons to believe one is favorite against others.

This distribution is expressed as

$$f(k) = \frac{1}{N}, \tag{4.3}$$

where N is the number of possible choices. In a coin, it would be head and tail, 2. In a dice, it would be 6. There are no close equations describing its mean and variance for generic sets [1], but both can be easily calculated with equations (4.1) and (4.2).

4.1.2 Binomial distribution

Assuming an event has probability p to occur, we might ask: what is the probability of measuring this event k times in a sample of size n. As an example, we might use rolling dice. What is the probability of getting 3 fives if we toss the dice 10 times? In this case $p = 1/6$ and we consider each toss to be independent of previous results. The general equation for the binomial distribution is given as

$$f(k|n,p) = \frac{n!}{k!\,(n-k)!}\,p^k\,(1-p)^{n-k}. \tag{4.4}$$

The binomial coefficient accounts for all the combinations of k successes in n trials. Considering our dice, it counts all possible outcomes with 3 fives disconsidering the order of appearance.

Notice that different values of k have different probabilities to happen. In this case, we can calculate the mean by weighting every k differently

$$\mathrm{E}\left[k|n,p\right] = \sum_{k=0}^{n} k\; f(k|n,p). \tag{4.5}$$

It is more or less straightforward to show that for the mean is given by $\mathrm{E}\left[k|n,p\right] = np$. To calculate the variance, we need to determine $\mathrm{E}\left[k^2|n,p\right]$ as

$$\mathrm{E}\left[k^2|n,p\right] = \sum_{k=0}^{n} k^2\; f(k|n,p). \tag{4.6}$$

The variance is calculated to be $\mathrm{Var}\left[k|n,p\right] = \mathrm{E}\left[k^2|n,p\right] - \mathrm{E}\left[k|n,p\right]^2 = np(1-p)$.

4.1.3 Poisson distribution

If the sample size is sufficiently big, we might consider to approximate the binomial distribution in the limit where n goes to infinity. If that is true and we expect finite mean and variance, p tends to zero. For simplicity, let's consider $\lambda \equiv np$. We

1. That is not the case to predictable sequences such as for dice and coins.

write

$$f(k|\lambda) = \lim_{n\to\infty} \frac{n!}{k!\,(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k}. \tag{4.7}$$

Moving elements independent of n out of the limit and simplifying the remaining expression, we obtain

$$f(k|\lambda) = \frac{\lambda^k}{k!} \lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^n. \tag{4.8}$$

This limit is one of the many forms of the exponential function. Concluding, the final expression describing the Poisson probability distribution is

$$f(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}. \tag{4.9}$$

Similarly to the previous subsection, we can estimate the mean and variance for the population represented. With λ greater than zero, we calculate $\mathrm{E}[k|\lambda] = \lambda$ and $\mathrm{Var}[k|\lambda] = \lambda$. As expected the variance tends to the mean as n tends to infinity and p goes to zero.

4.2 Continuous probability

In the previous section, we discussed about events that are discrete by nature. What if that is not the case? Let's consider how probable it is to meet someone that was born in a specific day of the year. Following our calendar organization, we treat days as discrete events, but time certainly is not [2]. If we can perform statistics in big enough samples, we should be able to question what is the probability of finding someone that was born within an hour of the year. We could be greedier and demand within the minute or second and so forth. Notwithstanding, we comprehend that the probability of finding a person born on March 13th between 2h30 and 2h31 is much smaller than if we had considered the whole day. Same-wise, we comprehend that the probability of finding somebody born on March 13th between 2h30 and 2h31 is much greater than between 2h30m56s and 2h30m57s. In theory, we could demand even smaller periods of times and, technically, the probability should always be greater than zero.

Does it make sense, though, to ask such a question? Does it make sense to know the probability for something to happen at a precise way? Of course, precision depends on the scale in which the study is embedded. Regardless, we could agree that it is nonsense to demand absolute precision in measurements.

2. For sanity, let's not consider Planck's time. In any fashion, 10^{-44} seconds is small enough to be considered continuous for our purpose.

Going towards that direction, continuous probability theory usually accounts for ranges. What is the probability of measuring an event "x" in the range $a < x < b$? What is the probability it rains between lunch and dinner? To answer that question we use

$$P(a \leftrightarrow b) = \int_a^b dx\ \rho(x), \tag{4.10}$$

where $\rho(x)$ is a probability density function (PDF), describing how probable event "x" is to happen in an interval $dx \to 0$. As for discrete probabilities, $\rho(x)$ is a model for the population data-set. Similarly, we can calculate mean

$$\mathrm{E}\left[x\right] = \int dx\ x\ \rho(x) \tag{4.11}$$

and population variance

$$\mathrm{Var}\left[x\right] = \int dx\ (x^2 - \mathrm{E}\left[x\right]^2)\ \rho(x), \tag{4.12}$$

with integration limits accounting for all possible events.

In the next few subsections I will present some of the popular probability density functions. More importantly, I mention the ones used for analysis and modeling later on in this thesis.

4.2.1 Uniform distribution

The uniform probability density function is described as

$$\mathcal{U}(x|a,b) = \frac{1}{b-a} \tag{4.13}$$

where every real number in $a \leq x \leq b$ presents the same probability to occur. Using equations (4.11) and (4.12) we can calculate $\mathrm{E}\left[x\right] = \frac{b+a}{2}$ and $\mathrm{Var}\left[x\right] = \frac{(b-a)^2}{12}$.

4.2.2 Beta distribution

The beta distribution can be derived from the binomial distribution. Assuming $a = k + 1$, $b = n - k + 1$ and converting the factorial numbers into the continuous gamma function, we obtain

$$\mathcal{B}(x|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\ x^{a-1}(1-x)^{b-1}, \tag{4.14}$$

for x in the range $0 < x < 1$. Using equation (4.11) we can calculate the $\mathrm{E}\left[x|a,b\right] = \frac{a}{a+b}$. Furthermore, we can show $\mathrm{Var}\left[x|a,b\right] = \frac{ab}{(a+b)^2\,(1+a+b)}$ using equation (4.12).

4.2.3 Normal distribution

The normal distribution has a special place at the heart of probability theory. This will become more apparent in the section where we talk about Central Limit Theorem (CLT). Other than that, we can prove that many PDFs can be nicely approximate by a normal distribution in specific limits. Perhaps for that reason, we find this density function so often in nature. Perhaps that is the case because it maximizes Shannon entropy, whence entropy maximization is at the core of Statistical Physics. The normal density function is described by

$$\mathcal{N}(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, \tag{4.15}$$

depending on parameters μ and σ. It is easy to prove that those are directly linked with the $\mathrm{E}\left[x|\mu,\sigma\right] = \mu$ and variance $\mathrm{Var}\left[x|\mu,\sigma\right] = \sigma^2$.

4.2.4 Log-normal distribution

The log-normal distribution can be understood as a variation of the normal distribution. In other words, the log-normal distribution represents random numbers whose logarithm are normally distributed. Hence, this distribution is defined only for positively-defined values. To prove this statement we can calculate

$$\mathcal{N}(u|\mu,\sigma)\,du = \mathcal{N}(\ln x|\mu,\sigma)\,d\left(\ln x\right), \tag{4.16}$$

using $u = \ln x$.

The log-normal distribution is defined as

$$\mathcal{L}(x|\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{\ln x-\mu}{\sigma}\right)^2\right\}, \tag{4.17}$$

from which we calculate

$$\mathrm{E}\left[x|\mu,\sigma\right] = \exp\left\{\mu+\frac{\sigma^2}{2}\right\} \tag{4.18}$$

and

$$\text{Var}\left[x|\mu,\sigma\right] = \left[\exp\left\{\sigma^2\right\} - 1\right]\ \exp\left\{2\mu + \sigma^2\right\}. \tag{4.19}$$

4.3 Central Limit Theorem (CLT)

The central limit theorem is one of the most important results in probability theory. I do not intend to demonstrate this theorem from first principles, but rather to approach its main results from a frequentist perspective and exemplify them accordingly. Aiming for that let's break down CLT in 3 statements. Assuming a sample size sufficiently large we can show that:

STATEMENT 1: *Sample averages are approximately normally distributed*

Independent of population probability distribution, even for discrete ones, we can ascertain that the distribution of samples averages should be approximately normally distributed. To test this statement we will consider two systems described by uniform discrete and Poisson distributions.

For figure (4.1a), we generate 200 samples in which we roll a dice 20 times each. In blue we have the average distribution of events for all samples. Additionally, the average was calculated for each sample and its distribution is presented in red. The black markers correspond to the theoretical population distribution. The black dashed line is the normal PDF expected from this sort of experiment in theory. A similar experiment was done using the Poisson distribution with $\text{E}\left[k\right] = \text{var}\left[k\right] = 6$. The latter is presented in figure (4.1.c).

STATEMENT 2: *Sample averages distribution has average similar to population mean*

Our dice experiment has $\text{E}\left[k\right] = 3.5$. The Poisson one has mean $\text{E}\left[k\right] = 6$ as already stated. In figures (4.1a,c) titles, we verify the average for red distributions. By inspection, we confirm that sample averages present average similar to the population mean. Moreover, in (b,d) we can see that this result is independent of sample size.

STATEMENT 3: *Sample averages variance is similar to the population distribution divided by sample size*

In other words, we can determine the variance of sample averages directly from the distribution measured or, alternatively, calculate it using the population variance as follows

$$\left\langle \left(\langle x\rangle - \langle\langle x\rangle\rangle\right)^2 \right\rangle = \frac{\text{var}\left[x\right]}{N}, \tag{4.20}$$

where N correspond to the sample size. Using our example of dice tosses and Poisson experiments with varying sample sizes, we observe in figures (4.1b,d) that the variance of averages decreases with bigger sample sizes. In simpler words, we could affirm that the confidence of our measures increase monotonically with the number of events captured in each sample.

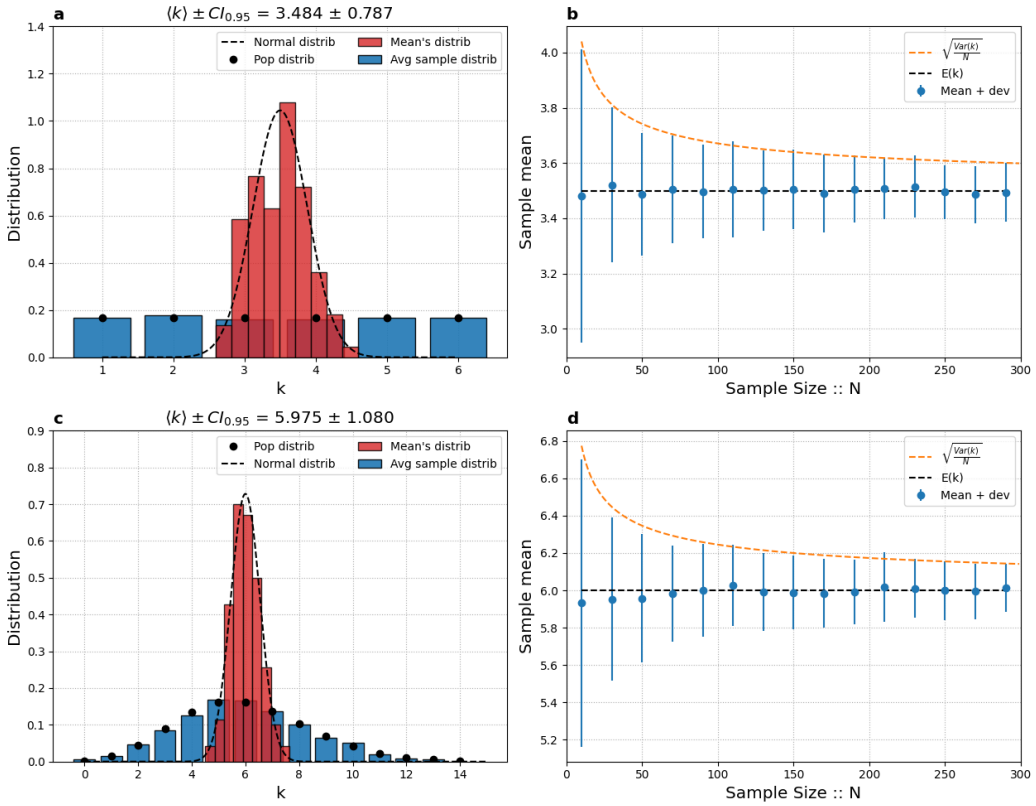

Fig. 4.1.: (a) 200 samples of 20 dices rolls are measured and average distribution in shown in blue. We have the distribution of averages calculated for each sample in red. (b) The confidence level for estimated population mean monotonically increases with larger sample sizes. (c) Similar experiment is performed using a Poisson distribution. Once again, the distribution of averages resembles a normal distribution. (d) The confidence interval also decreases with sample size for the Poisson distribution. All of these results are expected due to CLT.

4.4 Confidence intervals

As standard in many scientific areas, confidence levels for the mean are taken so that 95% of events around it are accounted. If we consider averages of several samples, CLT statement 1 tells us that their distribution is approximately normal

distributed. Using this fact associated with equation (4.10), we can estimate that the 95% confidence range lives around $\mu - 1.96\sigma < x < \mu + 1.96\sigma$.

What happens if it is impractical to measure a large number of independent samples? Or even, what if we could only obtain one single sample? In that case, we could use CLT statement 3 to help us. We could approximate the population variance directly from this single sample and estimate the confidence interval for the mean by dividing this sample variance by the number of events recorded. Regardless, we should guarantee a sample that satisfactorily represents the whole population. In figure (4.2a) we show the effect of sample size on estimation of average's variance distribution via CLT statement 3.

How big should this sample be, then? It will depend on how skewed the population distribution is, figure (4.2b). Very skewed distributions contain events with very low probability to happen, therefore a higher number of events should be recorded for a more complete representation of the population, hence better statistics.

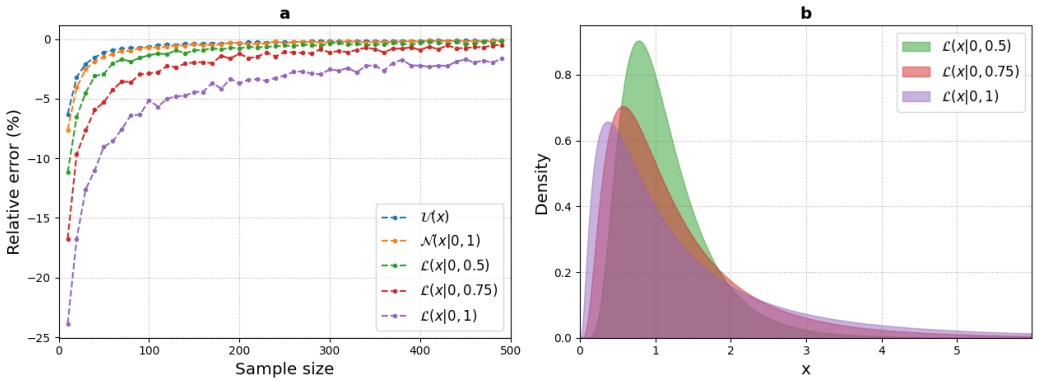

Fig. 4.2.: Effects of sample size on confidence interval calculated from single sample. (a) Average relative error in the estimation of population variance from a single sample with size N. For more skewed distributions, more data points are needed for a reliable estimation of the population variance. (b) Example of skewed distributions.

I would like to address one last remark. Given that the confidence increases with sample size, why shouldn't we merge all samples together and do our statistics with a bigger sample? It turns out there is no definitive answer for that. It depends on the question you ask. For example, in chapter 10 we will determine apparent diffusion and anomalous coefficients from a sample of 4096 independent polymers. Evidently, we could calculate mean values for D_α and α directly from this huge sample, but I decided to split it into 16 samples of equal size. Like so, I was able identify the mean along with a reasonable confidence interval.

4.5 Law of total variance

Suppose we want to isolate how much of the variance measured comes from within or across samples. For that scenario, the law of total variance can be used. In a more general approach, using the law of total expectation

$$\mathrm{E}\left[x\right] = \int dy\, \mathrm{E}\left[x|y\right] \rho(y) = \mathrm{E}\left[\mathrm{E}\left[x|y\right]\right], \tag{4.21}$$

we can easily show that

$$\mathrm{E}\left[x^2\right] = \mathrm{E}\left[\mathrm{Var}\left[x|y\right] + \mathrm{E}\left[x|y\right]^2\right], \tag{4.22}$$

holds. Subtracting $\mathrm{E}\left[\mathrm{E}\left[x|y\right]\right]^2$ from both sides we get

$$\mathrm{E}\left[x^2\right] - \mathrm{E}\left[x\right]^2 = \mathrm{E}\left[\mathrm{Var}\left[x|y\right]\right] + \mathrm{E}\left[\mathrm{E}\left[x|y\right]^2\right] - \mathrm{E}\left[\mathrm{E}\left[x|y\right]\right]^2. \tag{4.23}$$

Upon algebraic manipulation, we obtain the final result

$$\mathrm{Var}\left[x\right] = \mathrm{E}\left[\mathrm{Var}\left[x|y\right]\right] + \mathrm{Var}\left[\mathrm{E}\left[x|y\right]\right]. \tag{4.24}$$

Equation (4.24) states that the total variance in x is a combination of the var $[x]$ given sample y and $\langle x \rangle$ calculate for each y. To test this result, let's generate 20 samples with 30 points each using a Poisson distribution with $\mathrm{E}\left[k\right] = 5$. In figure (4.3a), we observe box-plots with each sample. In (4.3b), the distribution containing all the samples together. As we can see, from calculated averages and variances, the law of total variance holds.

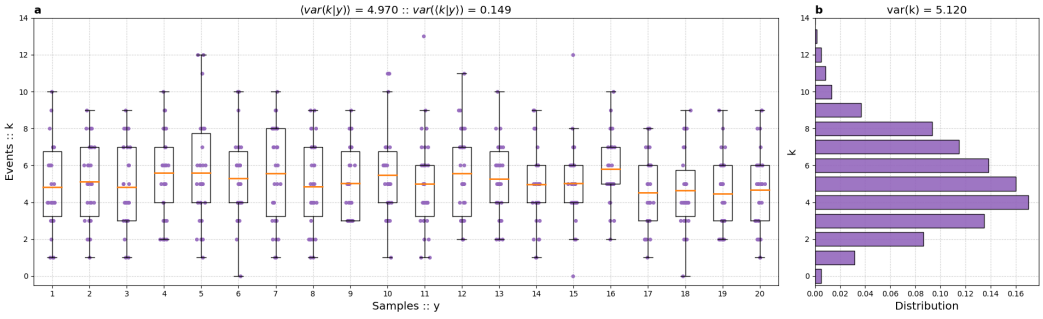

Fig. 4.3.: (a) 20 samples of 30 Poisson distributed events are recorded. We observe that means (orange line) presents some variance. Each sample also has variability. (b) Total variance containing all the data points. Using the law of total variance, we show that the variance in (b) equals the variance of averages var $[\langle k|y \rangle]$ and average sample variance $\langle \mathrm{var}\left[k|y\right] \rangle$.

5

Bayesian statistics

While describing the central limit theorem (CLT) and confidence intervals in the previous section, we were using what is called the frequentist approach to statistics. This name comes from associating probabilities and confidence intervals to the number of times given events happen. There is, however, a different approach to this problem: the Bayesian method.

Let's perform a small experiment so we can build some intuition about how the Bayesian approach works. Suppose we have a 25 cents coin and we want to determine if this coin is fair. We learn in school that this probability should be 1/2, but is that true? What if this coin is defective? At first, we have no idea if this coin favors one face over the other.

Let's first determine how likely it is to toss this coin n times and obtain k heads. As we saw in the previous chapter, this probability is given by the binomial distribution with success rate p. Let's proceed by tossing the coin 10 times and, for instance, let's assume we found 5 heads. In figure (5.1a), we show in orange our prior knowledge on the fairness of this coin, that is, none. As we have no idea which value of p is correct, all values are similarly expected. In blue (identical to green, but hidden behind), is the binomial distribution for our results, 5 heads out of 10 tosses. In green, we show the combination of our prior knowledge with the likelihood of this sample, called the posterior distribution.

The posterior distribution can be understood as an "updated knowledge", because we know more about this coin now than we did before. At this point, we have reasons to believe that the probability of obtaining heads over tails is somewhere in between 0.234 and 0.766. In Bayesian statistics, this range is called a credible interval (CI) and it might be defined in different ways. Here, I will con-

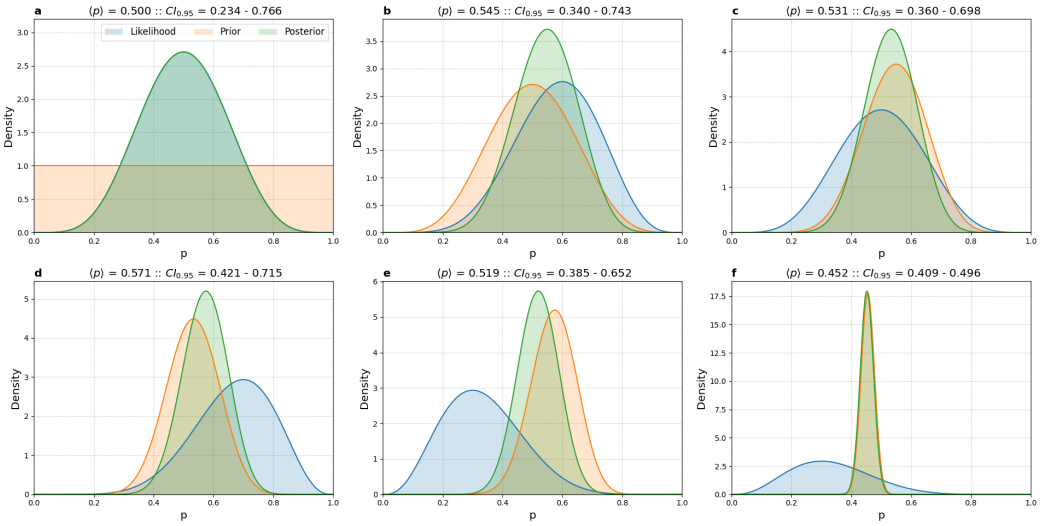

Fig. 5.1.: Evolution of posterior function with increasing number of coin tosses. Starting from no knowledge in (a), we add a sample with 10 trials each every image until (e). (f) Once 50 samples are collected (500 tosses/trials), we conclude that the coin is indeed biased.

sider as 95% of probability around the mean. Finally, we determine the average value for p is approximately 1/2.

It seems like this coin could be fair, but the CI is rather large yet. Let's toss our coin 10 times more and combine it with our current knowledge as new prior. In figure (5.1b) we observe the results. Given our new "updated knowledge", the CI is smaller now. We keep adding another sample of 10 tosses every image up to figure (5.1e). At this point we could stop and accepted that this coin could be fair. Unfortunately, probabilities can be deceiving... If we continue tossing this coin, our believe changes after a few more samples. Towards iteration 50, we determine that this coin is indeed crooked with a CI of 95% . This result is by design, of course. If we continue generating more samples, the CI would become smaller and smaller towards $\langle p \rangle = 0.45$.

The main equation behind this learning type of process is

$$P(H|E)P(E) = P(E \text{ and } H) = P(E|H)P(H), \tag{5.1}$$

for E representing measured data (or events) and H the hypothesis. Given a set with all possible events and conclusions, we can show that the probability of events given a hypothesis should be equivalent to the probability of a hypothesis given the data.

Going back to our coin example, the Bayes theorem (5.1) allowed us to write

$$P(p|\text{sample}) \propto Bin(\text{sample}|p)\ \mathcal{B}(p|a,b). \tag{5.2}$$

Due the simplicity of this model, we can normalize this result analytically and show that $P(p|\text{sample})$ is also a beta distribution. What if the system cannot be easily calculated analytically? In those situations, we will use numerical approximations. The topic of our next section.

5.1 Numerical approaches

For our coin example, the system and model associate were quite simple. So much so, we were able to solve it analytical. This is frequently not the case. When dealing with larger and more complex problems, the analytical solution might become a luxury difficult to obtain and, for those situations, we need to seek a numerical solution. There are two main approaches we can use: optimization and Monte Carlo sampling.

Put simply, optimization methods are used to search for local minima or maxima in the probability space. Most of these methods depend on numerical or analytical calculations of the first and/or second order derivatives. Some examples are "Broyden, Fletcher, Goldfarb and Shanno" (BFGS) [70], the Newton GLTR trust-region algorithm [71], among others. Diversely, if the derivatives are not an option, we can use models that do optimization by evaluating the function in multiple points and following some sort of heuristic algorithm towards a solution. Due its versatility, in this thesis we are going to use the Nelder-Mead Simplex optimizer [2]. The NMS method is very robust, but slower if compared to derivative dependent methods, as it tests several points on the way to a minimum. To compare analytical and numerical solutions, we use the 500 coin tosses from the previous experiment (fig. (5.1f)). In figure (5.2a), we show test values calculated by the NMS method towards the value of "p" that maximizes the likelihood [1]. This final result is identical to the analytical approach, but will only give us a single value, the optimal one. We might also be interested to perform some statistics on these variables and determine credible intervals.

To determine credible intervals, we can use Markov Chain Monte Carlo (MCMC) methods. I would like to focus on an algorithm called Metropolis-Hastings, used to sample a set of random numbers from a distribution that is, in principle, difficult to

1. As the NM simplex algorithm finds a minimum, we simply multiply the likelihood by minus one.

be analytically calculated.

Suppose there is a probability distribution $\pi(x)$ we want to sample. Starting from a plausible value x_0, repeat the algorithm as follows for $i = 1, ..., N$ for a large N:

— Draw a candidate $x^* \sim q(x^*|x_{i-1})$
— Calculate the ratio

$$\alpha = \frac{q(x_{i-1}|x^*)\ \pi(x^*)}{q(x^*|x_{i-1})\ \pi(x_{i-1})} \tag{5.3}$$

— If $\alpha \geq 1$, we accept proposed value and make $x_i = x^*$; else, we reject proposed value with probability $1 - \alpha$ and make $x_i = x_{i-1}$.
— Repeat while $i \leq N$

It is important to choose wisely the proposal distribution $q(x^*|x_{i-1})$. Perhaps the most popular choice is the normal distribution with $\mu = x_{i-1}$ and standard deviation σ. This choice implies that the next location depends only upon the current one, hence this kind of system is denoted a Markov Process. This one in special can also be referred as a random walk with step size associated with σ. Another good reason for choosing a $q(x^*|x_{i-1})$ normally distributed is related to its symmetry, which simplifies the detailed balance equation (5.3). Determining an appropriate sigma is, in general, a non-trivial task. Using too large values will increase the rejection rate, therefore your sample will not be representative of the whole distribution. Conversely, too small σ will portrait large acceptation rate, demanding too many steps to reach a representative sample. For random walks, it is well accepted that acceptance ratios of about $20 \sim 30\%$ are desirable.

Using normal distribution as proposal has a quirk. It does not work on bounded variables. Most optimizers, such as NMS and BFGS, are defined for variables with no boundaries indeed. Could we blindly apply these methods for bounded variables such as in our coin example where $0 < p < 1$? The answer is "it depends". Assuming that our proposal distribution presents very low probability for out of the boundary values, this algorithm should worked without problems. However, to be on the safe side, we could generate mapping functions from the bounded space into an unbounded space. Then, we convert probability calculations accordingly and run the sampler. For final results, we convert values back into the bounded space. This is going to be a trick particularly useful in the upcoming chapters of this thesis.

To complete this chapter, we have in figure (5.2b) the results of sampling the posterior distribution using the described MCMC method. As p is away from boundary conditions, we do not need to perform any mapping. Notice that the sampled

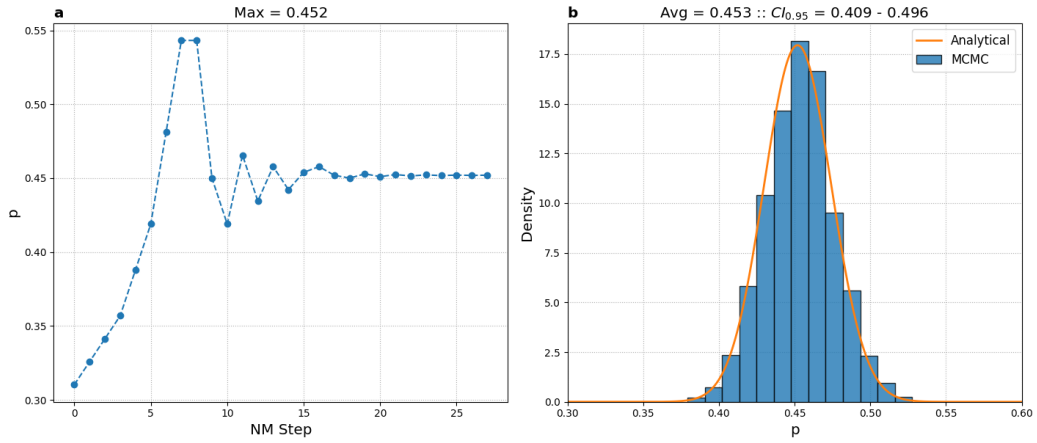

Fig. 5.2.: (a) Determining optimal value for “p” for our coin example using the Nelder-Mead Simplex numerical method. (b) We use Metropolis-Hastings algorithm to sample the probability distribution of “p”. After burn-in interval, the sampled probability distribution resembles the analytical one.

distribution is similar to the one obtained in figure (5.1f), also presenting similar credible interval.

6

Modeling diffusion dynamics with Gaussian processes

In the previous chapter we quickly verified how we can use the Bayesian theorem to infer optimal values and credible intervals for parameters of interest. We determined that given a prior knowledge on those parameters and a model that will express how likely it is to measure data given our parameters, it is possible to calculate a posterior function. This posterior function will update our knowledge given the data measured.

Now suppose we are interested to study the dynamics of Brownian-like particles diffusing in the cell. Which kind of model/likelihood should we use to estimate diffusion and anomalous coefficients, that is, D_α and α? As we saw in the introduction, some of the most popular models for that purpose are based on the analysis of particle displacements over time, but as we concluded, these models tend to ignore the implicit correlation between time points measured. Because of that, overall inference precision is reduced.

In this chapter we aim to approach this problem from a different perspective. In place of analyzing displacements over time, we are going to use Gaussian process (GP) associated with fractional Brownian motion (FBM) to model temporal correlations and, like so, determine optimal values for D_α and α as first suggested in [60, 61]. This approach, henceforth called GP-FBM, will allow us to use all the information available in measured trajectories, therefore outputting values that are more precise. To do so, we first need to get acquainted with the multivariate Gaussian distribution and its covariance matrix. Then, we present the concept of fractional Brownian motion.

6.1 Multivariate normal distribution

In the chapter 4 we discussed about the single variable normal distribution. This concept can be extended to accommodate for as many dimensions as necessary. The multivariate Gaussian distribution is defined as

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\Sigma) = \frac{1}{(2\pi)^{N/2}\ |\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\ \Sigma^{-1}\ (\boldsymbol{x}-\boldsymbol{\mu})\right\} \tag{6.1}$$

where $\boldsymbol{x}$ contains N variables and $\boldsymbol{\mu}$ is another vector with the mean for each variable x_i. The variance is now represented as a symmetric matrix Σ called the covariance matrix, with off-diagonal values representing the correlation between any two variables. As before, the variance is positive defined, which means we should ascertain that all eigenvalues of this matrix are greater than zero. In other words, the covariance matrix should be written as a composition $\Sigma = LL^T$ or, conversely, constructed from the multiplication of a matrix L with its transpose.

How do we sample from a multivariate normal distribution? Most modern programming languages allow sampling given a covariance matrix, but it is usually a good idea to know how the algorithm works. If the covariance matrix is diagonal, we have that $x_i \sim \mathcal{N}(\mu_i, \sqrt{\Sigma_{ii}})$, so we could sample each variable in $\boldsymbol{x}$ independently. Conversely, if Σ is not diagonal, we could diagonalize the covariance matrix, sample each element of $\boldsymbol{x}$ using its eigenvalues as variance and convert it back to original space with eigenvectors. It works, but it is numerically slow.

Another faster option, supposing we have a fast algorithm to decompose $\Sigma = LL^T$, is to map the cumulative distribution of our non-diagonal multivariate Gaussian into another diagonal covariance matrix, popularly the identity matrix $\mathbb{I}$ [1] using

$$\int_{-\infty}^{r_1}\cdots\int_{-\infty}^{r_N} d\boldsymbol{x}\ \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\Sigma) = \int_{-\infty}^{n_1}\cdots\int_{-\infty}^{n_N} d\boldsymbol{x}\ \mathcal{N}(\boldsymbol{x}|\boldsymbol{0},\mathbb{I}). \tag{6.2}$$

These integrals are not simple to solve [1], but we can show that

$$\boldsymbol{r} = \boldsymbol{\mu} + L\ \boldsymbol{n}, \tag{6.3}$$

where $n_i \sim \mathcal{N}(0,1)$. L can be calculated via Cholesky decomposition, an algorithm optimized for decades now, hence very fast. This method can be used even for a system with a single dimension,when L becomes the standard deviation. Clearly, the proof for this case is much simpler. In appendix B I demonstrate another conversion of vital importance in computer science: how to obtain normally distributed values

1. The identity matrix is defined to have all diagonal elements equal 1 and remaining elements 0.

from a uniform distribution. As pseudo-random generators are, in general, capable to produce only uniform distributed samples, we can use this method to amplify our spectrum of numerical samplers.

To increase our intuition over the covariance matrix and what off-diagonal terms do, let's study a simple 2D system with

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \lambda \\ \lambda & \sigma_2^2 \end{pmatrix}, \tag{6.4}$$

where each variable has its own variance, with $\sigma_i > 0$. Without loss of generality, let's consider $\lambda = \rho\sigma_1\sigma_2$ and $\boldsymbol{\mu} = \mathbf{0}$. Upon decomposition of Σ, we apply equation (6.3) to get

$$\boldsymbol{r} = \begin{pmatrix} \sigma_1 & 0 \\ \rho\sigma_2 & \sigma_2\sqrt{1-\rho^2} \end{pmatrix} \begin{pmatrix} n_1 \\ n_2 \end{pmatrix} = \begin{pmatrix} \sigma_1\, n_1 \\ \rho\sigma_2\, n_1 + n_2\, \sigma_2\sqrt{1-\rho^2} \end{pmatrix}, \tag{6.5}$$

where we notice that ρ is limited in the interval $-1 \leq \rho \leq 1$.

In figure 6.1 we show the effect of ρ on the relationship between r_1 and r_2. As ρ drops below zero, r_2 tens to oppose the behavior of r_1, while r_2 tends to behave accordingly to r_1 when $\rho > 0$. As usual, when $\rho = 0$, both variables are independent of each other. Finally, we notice that in the limit where $|\rho| = 1$, r_2 is completely determined by r_1.

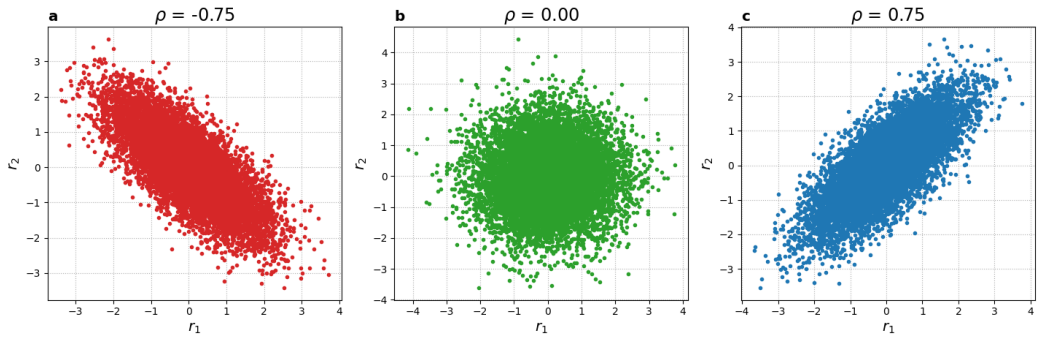

Fig. 6.1.: Effects of correlation ρ in a system $\boldsymbol{r} = \{r_1, r_2\}$. (a) In a 2D system with $\rho < 0$ we can expect that r_2 will behave oppositely to r_1. (b) For $\rho = 0$, both variables are independent of each other. (c) While for $\rho > 0$, r_2 follows the tendency observed by r_1. We chose $\sigma_1 = \sigma_2 = 1$ for these plots.

In the next subsection, we will present a covariance matrix for a N dimensional system as a model for stochastic trajectories. Notice that the degree of complexity increases exponential with the number of variables used. This happens because, *a priori*, all variables present some specific correlation to all the others.

6.2 Fractional Brownian motion (FBM)

The fraction Brownian motion is a moving average of $dB(t)$, the traditional Brownian motion, in which every past step is weighted according to $(t-s)^{\frac{\alpha-1}{2}}$. It is so defined by Mandelbrot and Ness in [72] as

$$\Gamma\left(\frac{\alpha+1}{2}\right)\{B_\alpha(t)-B_\alpha(0)\} = \int_{-\infty}^{0}\left[(t-s)^{\frac{\alpha-1}{2}}-(-s)^{\frac{\alpha-1}{2}}\right]dB(s)+\int_0^t (t-s)^{\frac{\alpha-1}{2}}\,dB(s) \tag{6.6}$$

for $t > 0$, $0 < \alpha < 2$ and $B_\alpha(0) = b_o$. Furthermore, this definition presents self-similarity $B(t)-B(s) \propto B(t-s)$ and variance $\left\langle (B(t+\tau)-B(t))^2 \right\rangle = V\tau^\alpha$. Without loss of generality, we can take constant V and calculate the covariance matrix as follows

$$\begin{aligned}
\langle B_\alpha(t)B_\alpha(s)\rangle &= \frac{1}{2}\left\langle [B_\alpha(s)-B_\alpha(s)+B_\alpha(t)]B_\alpha(s)+B_\alpha(t)[B_\alpha(t)-B_\alpha(t)+B_\alpha(s)]\right\rangle \\
&= \frac{1}{2}\left\langle B_\alpha(s)^2+B_\alpha(t-s)B_\alpha(s)-B_\alpha(t)B_\alpha(t-s)+B_\alpha(t)^2\right\rangle \\
&= \frac{1}{2}\left\langle B_\alpha(t)^2+B_\alpha(s)^2+B_\alpha(t-s)(B_\alpha(s)-B_\alpha(t))\right\rangle \\
&= \frac{1}{2}\left\langle B_\alpha(t)^2+B_\alpha(s)^2-B_\alpha(t-s)^2\right\rangle \\
&= \frac{V}{2}\left(|t|^\alpha+|s|^\alpha-|t-s|^\alpha\right).
\end{aligned} \tag{6.7}$$

In conclusion, we match the mean squared displacement of stochastic trajectories re-scaling V such that

$$\Sigma_{D,\alpha}(t,s) = \langle B_\alpha(t)B_\alpha(s)\rangle = D_\alpha\ \left(t^\alpha+s^\alpha-|t-s|^\alpha\right), \tag{6.8}$$

accomplishing $\Sigma_{D_\alpha,\alpha}(t,t) = 2D_\alpha\, t^\alpha$.

Notice that this kernel will introduce correlations between every pair of time points t and s. As an example, we sample 3 trajectories using equation 6.3 with the FBM kernel and present in figure 6.2 the effect of α over stochastic trajectories. At bottom images, we plot the distribution angles over 2 consecutive simulated steps. We notice that for $\alpha < 1$ particles tends to be more constrained to move. $\alpha = 1$ represents the traditional Brownian motion, where particles are free to randomly go anywhere, without defined direction. Finally, for $\alpha > 1$ particles tend to have a preferred direction to follow.

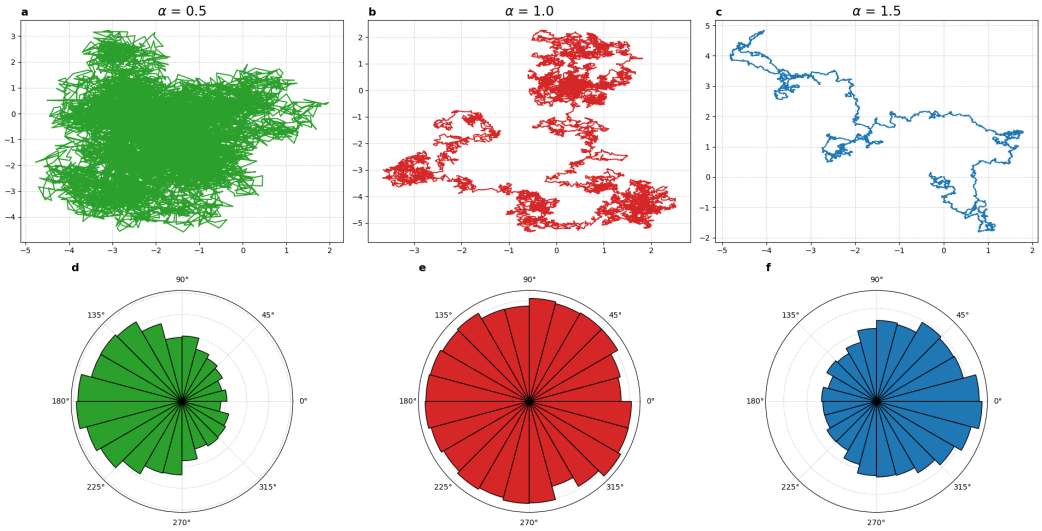

Fig. 6.2.: We sample a single long trajectory using equation 6.3 and the FBM kernel with $D_\alpha = 1$ and $\alpha = 0.5$ for (a), $\alpha = 1.0$ for (b) and $\alpha = 1.5$ for (c). $\alpha < 1$ constrains the movement of particles if compared to traditional Brownian motion $\alpha = 1$, while $\alpha > 1$ directs it. (d-f) Distribution of angles in degree calculate between two consecutive steps.

6.2.1 Gaussianity

Fractional Brownian motion is a time continuous self-similar Gaussian process, but what does that mean? Put simply, it means that the distribution of displacements are described by a Gaussian distribution. Furthermore, we can re-scale distributions for different time steps Δt and anomalous coefficients α by their standard deviation, that is, $\sqrt{2D_\alpha\,\Delta t^\alpha}$. To demonstrate what this means in a graphic fashion, in figure (6.3a,b) we calculate the re-scaled displacement distribution for 1024 simulated trajectories with 512 points, $D_\alpha = 0.1$, $\alpha = 0.44$ and $\delta t = 1.0$. Notice that all the distributions are Gaussian and self-similar, independently of correspondent time step.

6.2.2 Velocity autocorrelation function

Along with the just presented Gaussianity and self-similarity tests, one of easiest quantities to be accessed in experimental data is the velocity autocorrelation function

$$C_v^\epsilon(t) = \frac{1}{\epsilon^2}\left\langle [x(t+\epsilon) - x(t)]\,[x(\epsilon) - x(0)]\right\rangle, \tag{6.9}$$

as proposed in [73], where $v = \epsilon^{-1}\,[x(\epsilon) - x(0)]$). Having the covariance matrix for the FBM model presented in equation (6.8), we can easily show that the FBM

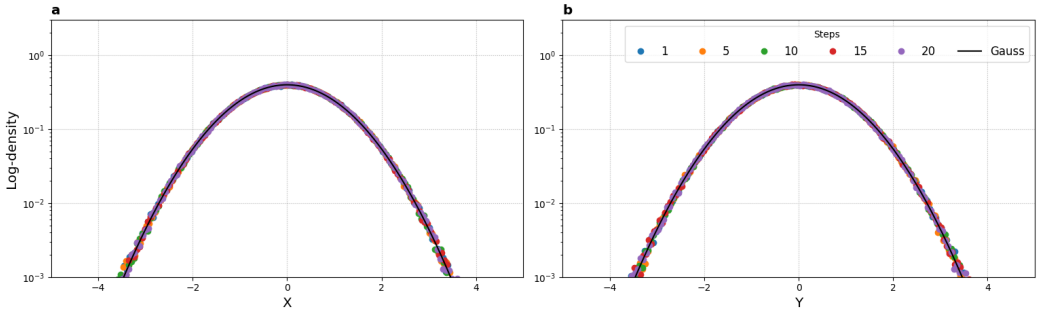

Fig. 6.3.: To demonstrate that FBM is a time continuous and self-similar Gaussian process, we simulate 1024 trajectories 512 points long using $D_\alpha = 0.1$, $\alpha = 0.44$ and $\delta t = 1.0$. As FBM is self-similar, we calculate displacement distribution for several time steps and normalize with $\sqrt{2D_\alpha\ \Delta t^\alpha}$, where $\Delta t = n\ \delta t$. As we can see, all distributions are consistent with a standard Gaussian/normal distribution.

model presents the following velocity autocorrelation curve

$$\frac{C_v^\epsilon(t)}{C_v^\epsilon(0)} = \frac{(t+\epsilon)^\alpha - 2t^\alpha + |t-\epsilon|^\alpha}{2\epsilon^\alpha}. \tag{6.10}$$

To show the theoretical shape of this curve and to demonstrate it using synthetic data for the range in which $\alpha < 1$, we simulate 4096 trajectories with 512 points each. For these simulations, we used $D_\alpha = 0.1$, $\alpha = 0.5$ and $\delta t = 0.01$. These results are present in figure (6.4). Due to the α regime under analysis, we can observe that $C_v^\epsilon(t)$ will initially "overshoot" below zero, displaying anti-correlation for a finite amount of time.

Similar can be done in the case where $\alpha = 1$ and $\alpha > 1$. The behavior for these regimes are shown in figure (6.5). In the case $\alpha = 1$, the autocorrelation plunges linearly to zero and there remains for any time points $t > \epsilon$. Differently, positive autocorrelation are always present in the situation where $\alpha > 1$.

6.3 Inferring diffusion and anomalous coefficients

In the previous section we described a model for Brownian-like trajectories of particles. This model allows us to determine how likely it is to measure a certain stochastic trajectory given apparent diffusion and anomalous coefficients, that is, D_α and α. Notwithstanding, we would like to determine the posterior probability of D_α and α given our measured data. For that purpose, we can use the Bayes theorem 5.1 as referred in the previous chapter. The next question would be: which priors

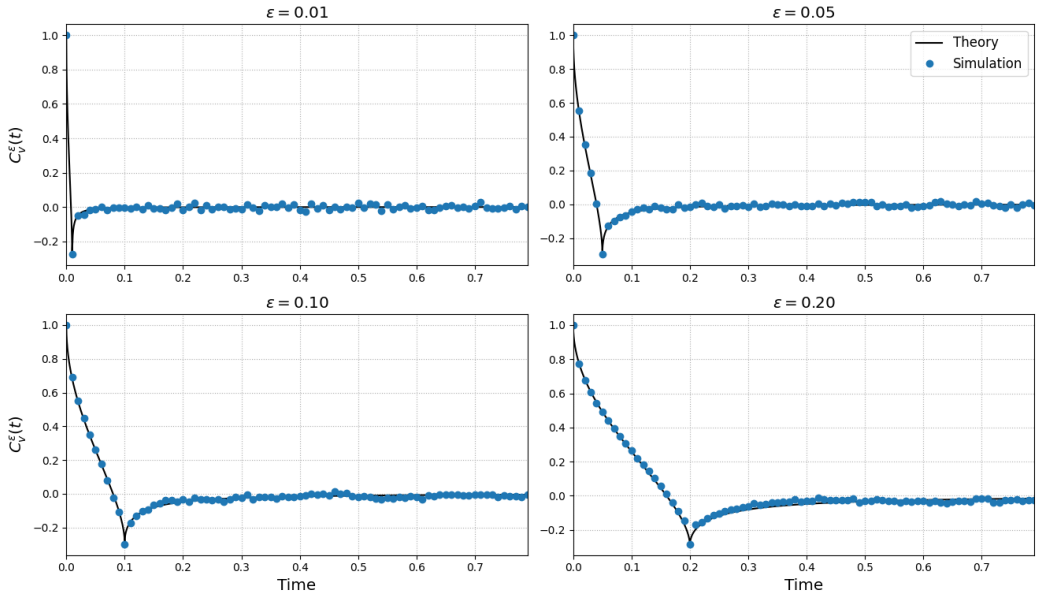

Fig. 6.4.: Comparison between theoretical and simulated velocity autocorrelation functions for the FBM model. A total of 4096 synthetics trajectories were create using $D_\alpha = 0.1$, $\alpha = 0.5$ and $\delta t = 0.01$.

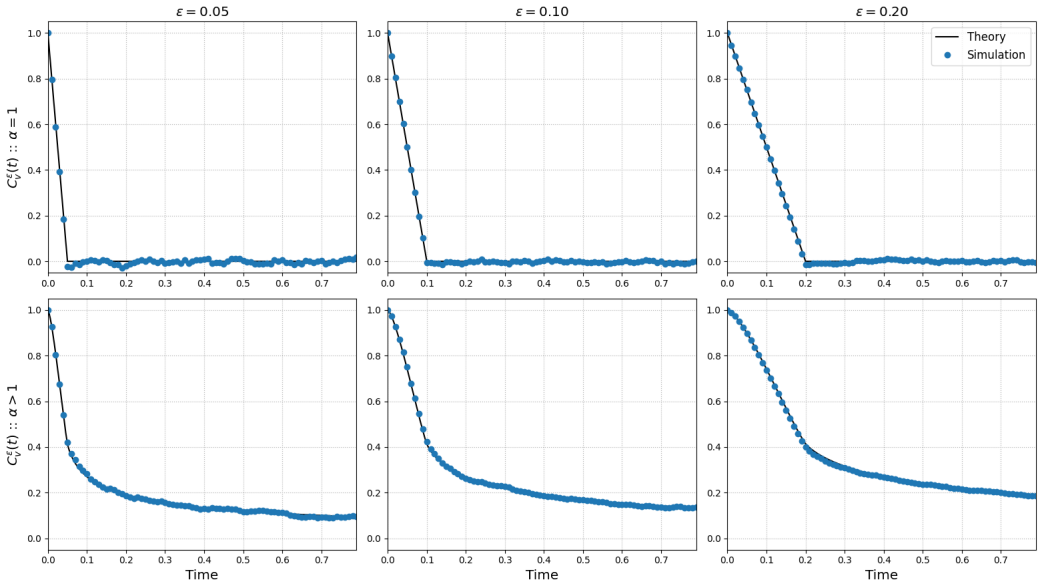

Fig. 6.5.: Comparison between theoretical and simulated velocity autocorrelation functions for the FBM model when α equals or is greater than one. A total of 4096 synthetics trajectories were create using $D_\alpha = 0.1$ and $\delta t = 0.01$. Top images were simulated using $\alpha = 1$ and bottom ones with $\alpha = 1.5$.

should we use? Values of D_α and α will vary depending on the type of particles we analyze. Different systems will present implicit physical laws that will dictated the dynamical behavior of such system. Hence, it is hard to tell what we should expect.

Mathematically, that uncertainty can be translated as a uniform prior, where every option presents similar probability. Thus, the posterior is simply given by

$$P\left(D_\alpha, \alpha | \boldsymbol{r}\right) \propto \left|\Sigma_{D_\alpha,\alpha}\right|^{-1/2} \exp\left\{-\frac{1}{2}\left(\boldsymbol{r}-\boldsymbol{\mu}\right)^T \Sigma_{D_\alpha,\alpha}^{-1} \left(\boldsymbol{r}-\boldsymbol{\mu}\right)\right\}, \tag{6.11}$$

with $\Sigma_{D_\alpha,\alpha}$ as presented in equation 6.8. $\boldsymbol{\mu}$ will be treated as a constant, *vis-à-vis*, particle position at t = 0. If necessary due occlusions, it can also be inferred as a parameter using another flat prior.

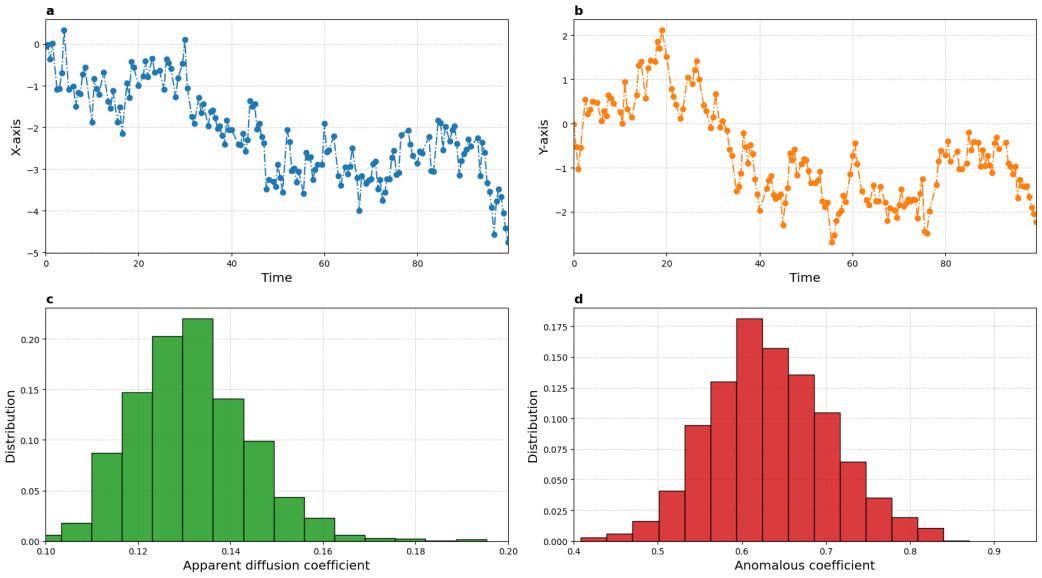

Fig. 6.6.: (a-b) Simulated 2D Brownian-like trajectory with $D_\alpha = 0.15$ and $\alpha = 0.62$. Average localization noise of 0.1 and occlusion of 10% were used. (c-d) Sampled probability distributions for D_α and α given trajectories presented. Optimal values are $D_\alpha = 0.131$ and $\alpha = 0.638$.

In figure (6.6a-b), we sample a 250 points long 2D trajectory using $D_\alpha = 0.15$ and $\alpha = 0.62$. We further use the methods presented in chapter 5 to estimate these values back. For these simulations, we are including localization noise with different variance for each point. The average variance added to the diagonal of Σ is about 0.1. On top of that, we also remove about 20% of the spots sampled as a representation of occlusions, so commonly found in microscopy.

The optimal results obtained were $D_\alpha = 0.131$ and $\alpha = 0.638$. We also sample the probability density function to estimate a credible interval. For simplicity, we are going to consider an approximate 95% probability around the mean by accumulating the area of each histogram bin. We have $CI_{\sim 0.95}(D_\alpha) = (0.11, 0.16)$ and $CI_{\sim 0.95}(\alpha) = (0.49, 0.76)$.

To determine the overall performance of this method, we simulate 10000 tra-

jectories with uniform random values of D_α and α. We set $0.01 < D_\alpha < 1.5$, $0.01 < \alpha < 2$, occlusion of 20% , localization error $\sigma = 0.1$ and $\delta t = 0.5$. For comparison, we also estimate these parameters using displacement based method via MSD curve and distribution fitting.

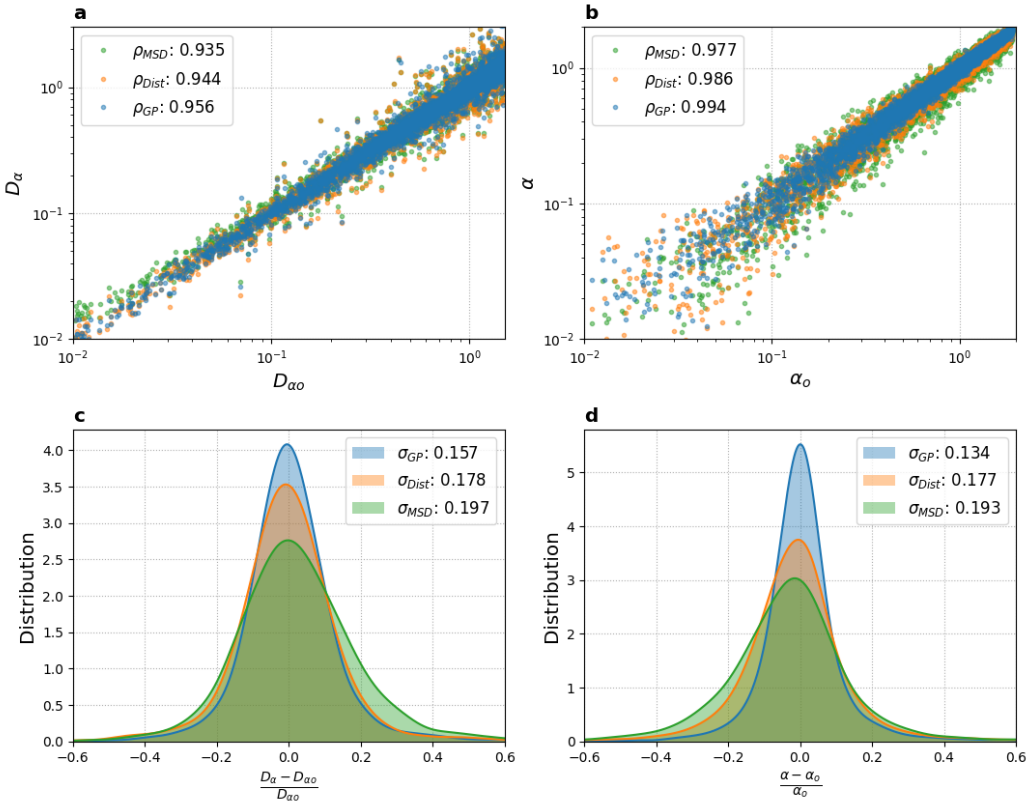

Fig. 6.7.: (a-b) Correlation between set and estimated parameters for a set of 10000 simulated trajectories. (c-d) Relative estimation error for same trajectories. As a comparison, GP-FBM is, on average, more precise than displacement based methods.

In figure (6.7a-b), we show the correlation between estimated and simulation set values. As we demonstrate previously, the closer to one, the more related estimated values are to set ones. As a second perspective, we show in (6.7c-d) the relative estimation errors for D_α and α. With this analysis, we conclude that GP-FBM is more precise as an estimation method for diffusion related parameters if compared to displacement based methods in the context of single particle trajectories.

6.4 Interpolation

I would like to present one last important feature of using Gaussian processes to analyze dynamics of stochastic particles. In the first section we show how to sample trajectories given the FBM covariance matrix. In this section we shall see how to use measured positions and inferred values of D_α and α to generate sets of probable trajectories described by the particle. With same method, we can determine which path is the most probable and how credible it is.

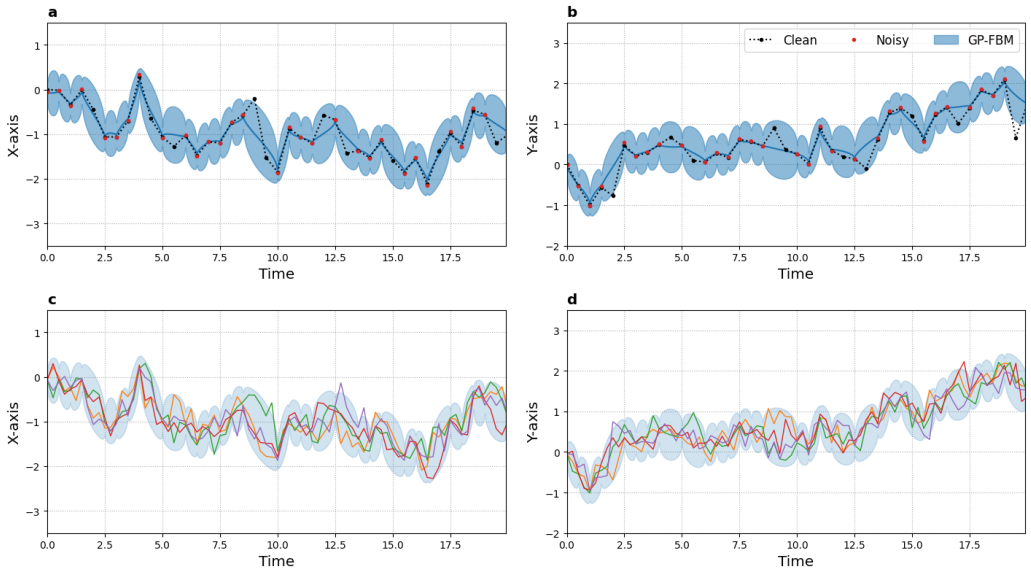

Fig. 6.8.: First time steps of simulation presented in figure (6.6). (a-b) Black line is the original trajectory without localization error. This error is included for points in red. We also remove some points to simulate occlusion. Blue line is the most probable path $\boldsymbol{\mu}_{1|2}$ predicted by GP-FBM, while shaded area represents a 95% credible interval. (c-d) Sampled trajectories given measured red points. Shaded area is the same of (a-b).

Let's start by defining the multivariate Gaussian

$$\mathcal{N}\left(\boldsymbol{x}_1, \boldsymbol{x}_2 | \boldsymbol{\mu}, \Sigma\right) \propto \exp\left\{-\frac{1}{2}\begin{pmatrix}\boldsymbol{x}_1 - \mu_1 \\ \boldsymbol{x}_2 - \mu_2\end{pmatrix}^T \begin{pmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{pmatrix} \begin{pmatrix}\boldsymbol{x}_1 - \mu_1 \\ \boldsymbol{x}_2 - \mu_2\end{pmatrix}\right\}, \tag{6.12}$$

where $\boldsymbol{x}_2$ is the vector of measured positions and $\boldsymbol{x}_1$ the vector of positions we want to sample. In this circumstance, Σ_{ij} are block matrices given by the FBM kernel 6.8. Additionally, we also add the variance associated with localization error to diagonal terms of Σ_{22}. To marginalize the distribution over $\boldsymbol{x}_1$ given measured positions, we have

$$P(\boldsymbol{x}_1|\boldsymbol{x}_2) = \mathcal{N}(\boldsymbol{x}_1|\boldsymbol{\mu}_{1|2}, \Sigma_{1|2}), \tag{6.13}$$

with

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \Sigma_{12}\,\Sigma_{22}^{-1}\,(\boldsymbol{x}_2 - \boldsymbol{\mu}_2) \qquad \text{and} \qquad \Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\,\Sigma_{22}^{-1}\,\Sigma_{21}. \quad (6.14)$$

In figure (6.6a-b) we present a dashed black line for simulated trajectory without noise, while red dots include localization error. We further remove about 10% of points to simulate data occlusion. $\boldsymbol{\mu}_{1|2}$ is presented as a continuous line with shaded area representing 95% credible interval calculated using diagonal terms from $\Sigma_{1|2}$. Finally, in figure (6.8c-d) we show a few sampled trajectories considering measured red points using equation 6.3 with $\boldsymbol{\mu}_{1|2}$ and $\Sigma_{1|2}$.

7

Methods developed

Using experimental data is never straightforward. Working in live cell imaging can be tough if we seek for accurate measurements. In general, standard deviations are fairly big if compared to average value and, in most cases, it is hard to tell if this variance is due to meaningful biology or simply experimental error. In the next few sections, I am going to present a few methods I developed with which I try to reduce data variance due to measurement inaccuracy. These methods will enhance localization precision of fluorescent spots, calibrate multi-channel imaging for improved distance measurements and improve dynamics inference by accounting for confound substrate movement.

7.1 Enhancing localization accuracy of fluorescent particles

The first step in the analysis of any movie recorded is to track fluorescent tagged chromatin spots. Fiji-ImageJ [74] and ICY [3] provide a few plugins for that end. In my experience, ICY tends to work a little better, therefore that was the software used for tracking in all the movies in this thesis. Unfortunately, ICY detector is not perfect. By inspection alone, we can determine that its localization is off-centered by approximately one pixel. Its linker algorithm [1] is very good, but some false positives tend to be included in the final result. Trying to fix, or at least improve on those issues, we apply a filter in 2 stages. In the first stage we try to enhanced the localization precision and generate confidence intervals for this position. In the

1. This plugin connects spots in different frames to form a trajectory

second stage we use those results to filter out possible outliers.

To enhance localization precision and estimate its credibility, we model each spot using a 2D Gaussian function like so

$$S_{x,y} = I_o \exp\left\{ -\frac{1}{2} \begin{pmatrix} x-\mu_x \\ y-\mu_y \end{pmatrix}^T \begin{bmatrix} L_x^2 & \theta L_x L_y \\ \theta L_x L_y & L_y^2 \end{bmatrix}^{-1} \begin{pmatrix} x-\mu_x \\ y-\mu_y \end{pmatrix} \right\} + B_G. \tag{7.1}$$

The reference signal level is given by B_G. On top of that, the spot itself has maximum signal I_o and center of mass represented as μ_x and μ_y for each axis. We use parameters L_x and L_y to fit the spot size in each direction, but allowing for rotation with parameter θ (not an angle).

Using ICY detected position as initial guess, we generate a region of interest (ROI) of appropriate size and use equation (7.1) as average signal $S_{x,y}$ for each pixel (x, y) of that ROI. Assuming the image presents Poisson-like noise, we can write the likelihood

$$\rho(g|I_o, B_G, \boldsymbol{\mu}, ...) = \prod_{x,y} \frac{(S_{x,y})^{g_{x,y}}}{g_{x,y}!} \exp\left\{-S_{x,y}\right\}. \tag{7.2}$$

where $g_{x,y}$ represents pixel values in the image.

Calculating $\rho(g|I_o, B_G, \boldsymbol{\mu}, ...)$ numerically is complicated. The factorial function $k!$ tends to produce huge numbers, making numerical simulations unstable. Fortunately, maximizing this likelihood or its natural logarithm provides identical results. So we calculate

$$\ln \rho(g|I_o, B_G, \boldsymbol{\mu}, ...) = \sum_{x,y} \left\{ g_{x,y} \ \ln S_{x,y} - S_{x,y} - \ln\left(g_{x,y}!\right) \right\}. \tag{7.3}$$

We can perform one last optimization. Notice that $\ln\left(g_{x,y}!\right)$ is a constant, that is, it won't influence in the optimization of our desired parameters. For that reason, we can neglect it for optimization purposes.

Now we should consider the boundary implied in each parameter we optimize. The position vector $\boldsymbol{\mu}$ should remain inside the ROI. L_i, I_o and B_G are positively defined. θ is limited in the range $-1 < \theta < 1$. As the NMS algorithm proposed in the previous chapter only accepts boundless variables, we need to perform a mapping

$$\mu_i = R \frac{\exp\left\{m_i\right\}}{1 + \exp\left\{m_i\right\}} \tag{7.4}$$

with R representing the ROI side. Next, all the positively defined variables should be transform according to a exponential function. Finally, θ follows a similar trans-

formation to μ_i, but limited in between -1 and 1

$$\theta = 1 - 2\frac{\exp\{t\}}{1+\exp\{t\}}. \tag{7.5}$$

Once all these parameters are optimized, we estimate the error on $\boldsymbol{\mu}$ by keeping all the other parameters fixed at optimal values and sampling $\boldsymbol{\mu}$'s distribution with Metropolis-Hasting algorithm, as presented in the previous chapter. We consider the standard deviation of sampled distribution as a measure of localization error.

To test this algorithm, we generate a synthetic movie in which a spot is randomly positioned in 500 frames. Similar signal and size to the ones observed in real movies are used. In figure (7.1a) we have an example. A comparison between the results obtained using the ICY software and the output of our optimization problem is presented in (7.1b). Notice that this method is about 7 times more precise than positions measured with ICY. A secondary result is presented in (7.1c), a distribution of sizes calculated for tracked trajectory. Finally, in (7.1d) we display a distribution of estimated error for all spots.

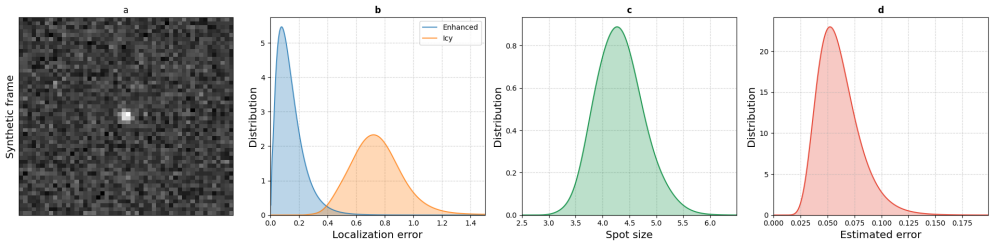

Fig. 7.1.: Testing enhancement algorithm. (a) Example of spot with similar signal and size to the ones observed in real movies. (b) Our algorithm improves about 7 times the localization of tracked spots. (c) We can determine the average size of spots in a trajectory and use it to identify possible outliers. (d) Distribution of localization errors estimated with all spots in the synthetic trajectory.

Using signal intensity, sizes and localization error measured for all points in a trajectory, we can determine possible outliers. To do that, we calculate the inter-quartile range (IQR) for these 3 parameters and neglect every spot in which any of these is higher than one IQR above and below the median.

7.2 Alignment algorithm

Some of the microscopy experiments performed used two cameras, that is, one per channel. This system allowed us to record simultaneously the spot in both channels, later on simplifying the correction for substrate movement. On the underside, using two cameras inserted non-negligible alignment discrepancies between channels. In figure (7.2a-c) we have a few examples. Akin to the dual camera issue, two different wavelengths introduce errors associated with chromatic aberration. This secondary problem would be observed even in a single camera setup.

The differences are not of vital importance when inferring dynamics of tagged spots. Stochastic trajectories are invariant upon global translations and rotations. The scaling factor associated with chromatic aberrations could be neglected, as it becomes only relevant for long trajectories further away from image center. Even like so, the effect is minor. Nonetheless, dealing with such discrepancies are of vital importance when we are also interested in the average distance between spots in different channels.

To correct such problem, we use a generic set of affine transformations to perform a digital post-alignment in 2 steps. The first step handles more grotesque errors associated with the dual camera setup. The model is written as

$$\Omega_{RT} = \begin{pmatrix} 1 & 0 & d_x + c_x \\ 0 & 1 & d_y + c_y \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos(\theta) & \sin(\theta) & 0 \\ -\sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & -c_x \\ 0 & 1 & -c_y \\ 0 & 0 & 1 \end{pmatrix}. \tag{7.6}$$

From left to right, d_x and d_y consider translations in x and y; c_x and c_y are reference points for rotation and θ is the angle to be rotated. It worth mentioning that these transformations, as it is, will generate artifacts on the resulting image. To guarantee that every pixel of the new image is properly mapped to a pixel of the original, we loop over the final image and use Ω_{RT}^{-1} to determine its original position. It would still be possible that some pixels are mapped to regions not shown by original image. In those cases, the signal is left zero.

To model the likelihood both channels observe the same object, we use the normal distribution

$$\rho(\chi|\Omega_{RT}, \sigma) = \prod_{k,l} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma^2} \left[I_2(k,l|\Omega_{RT}) - I_1(k,l|\mathbb{I}) \right]^2 \right\}, \tag{7.7}$$

where we calculate the probability for every pixel $I_2(k,l|\Omega_{RT})$ of channel 2, upon correction Ω_{RT}, represents the same object as in pixel $I_1(k,l|\mathbb{I})$ from channel 1. $\mathbb{I}$

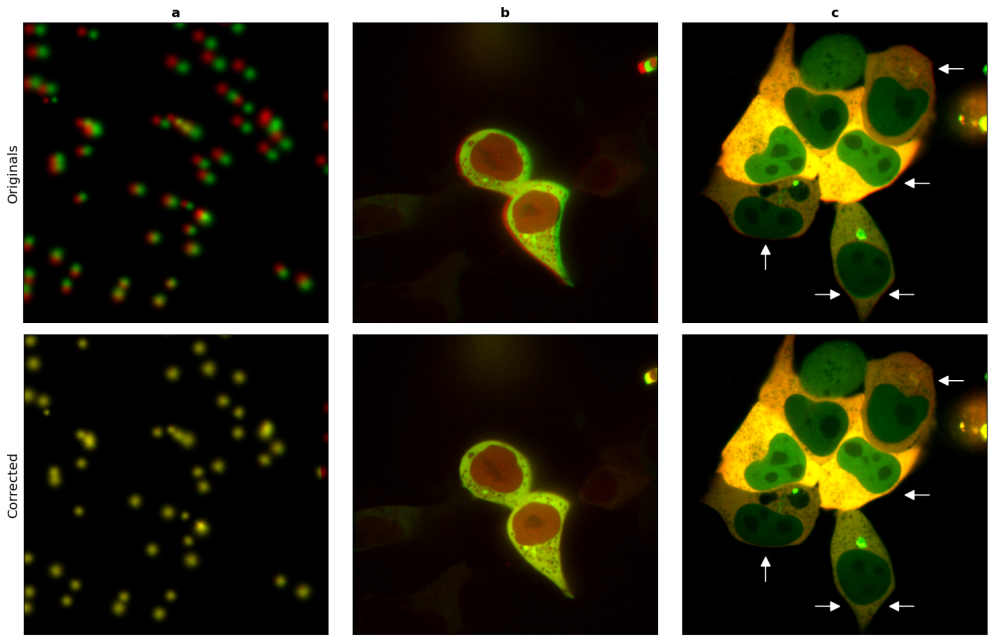

Fig. 7.2.: (a) Synthetic image generated to exemplify common problems found in our microscopy experiments. (b) Real image generated with poor camera calibration. (c) Real image generate with proper camera calibration. Regardless, chromatin aberration is still noticeable. (d-f) Corrections provided by the algorithm demonstrated in this section.

denotes a simple identity transformation.

The standard deviation σ is not known. It could be fit alongside Ω_{RT}, but we are not interested in its value. For that reason, we are going to use a non-informative Jeffrey's prior [1] for normal distribution $\rho(\sigma) \propto 1/\sigma$. Accordingly

$$\rho(\chi|\Omega_{RT}) \propto \int_0^\infty d\sigma \; \frac{1}{\sigma^{WH+1}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{k,l} \left[I_2(k,l|\Omega_{RT}) - I_1(k,l|\mathbb{I}) \right]^2 \right\}, \quad (7.8)$$

where W and H correspond, respectively, to width and height of the image. After solving this integral, we reach the final result used for optimization

$$\ln \rho(\chi|\Omega_{RT}) \propto -\frac{WH}{2} \ln \left\{ \sum_{k,l} \left[I_2(k,l|\Omega_{RT}) - I_1(k,l|\mathbb{I}) \right]^2 \right\}. \quad (7.9)$$

The absolute likelihood value is not important. For that reason, as in the previous section, we converted the final result into log space and neglected a few constants. These approximations will not alter our results, but will speed-up the numerical calculations.

For simplicity, we consider that all the movies from a given day are subjected to similar alignment discrepancies. In other words, either camera will be touched during that day. This simplification allows us to rescue movies containing weird asymmetrical agglomerations of super luminous material. Thus, we used 5 frames per channel from each movie and maximize the following likelihood using the Nelder-Mead simplex model [2].

Once all images are corrected for camera shifts and rotations, we consider chromatic aberration. The second set of affine transformations is merged in a single matrix given by

$$\Omega_S = \begin{pmatrix} s_x & 0 & (1-s_x)W/2 \\ 0 & s_y & (1-s_y)H/2 \\ 0 & 0 & 1 \end{pmatrix}, \tag{7.10}$$

where, s_i accounts for scaling in directions x and y. Finally, we optimize equation (7.9) using Ω_S.

In figure (7.2a) we present a synthetic image for easier visualization of traditional calibration problems observed in microscopy sessions. Conversely, figure (7.2b) presents a real image in which cameras were badly or not calibrated. In figure (7.2c), we can observed that, even with proper camera alignment, we can still observed a few problems. In this situation, we can associate most of these discrepancies to chromatic aberration. Figures (7.2d-f) were digitally corrected using the proposed algorithm.

One final subject remains to be addressed. More frequent than not, movies present cells in very uneven signal regimes, that is, a few cells are very bright while others are comparatively dark. To mitigate this problem, we perform an adaptive equalization of all the images via an algorithm called CLAHE (contrast limited histogram equalization) [75]. Of less importance, a median filter of size 3x3 is also applied to reduce the noise present in each image.

7.3 Correcting for background movement

It is quite noticeable how much cells move when performing live imaging. It was found that cells tend to perform some sort of Brownian motion if left free to move [4]. Not only that, cells are not rigid bodies. Their shape can fluctuate as the cell re-arranges its internal content. On top of that, the extra heat introduce by the laser in fluorescent microscopy tends to make cells more agitated, increasing their motility and volumetric fluctuations. This extra source of heat also introduces

thermal drift in the microscopy setup. For instance, we were able to observe that even calibration beads, chemically attached to the slab, present global movement for longer periods of time. Now suppose you are interested to study the intrinsic dynamics of specific chromatin regions. Which kind of movement are we observing? Do our measurements represent the values we want to study or an accumulation of all those effects? Certainly, the second.

There are a few ways that are popularly used to correct for these issues. Perhaps one of the most popular is a set of algorithms know as optical flow [76, 77, 78]. Simply put, optical flow attempts to reconstruct the next image in time by determining a velocity vector field for each pixel in the current image. This type of method has become very popular in the previous years with the increase of computation power and better graphical cards. It is widely applied in game engines, special effects, self-driving cars and more. Unfortunately, most of the algorithms developed focus on daily problems, that is, rather rigid bodies with simplified shapes. Cells, on the other hand, can be classified as soft active matter [79, 80] with dynamic shape. That is perhaps the reason why most of the tests I perform with optical flow returned fairly imprecise results. So much so, we could verify by visual inspection.

Another popular option used to calibrate experiments of this type relies on tracking a supposed immobile structure proximal to region of interest [63]. This method tends to present better results, but it is experimentally more demanding with a set of extra parameters to be optimally calibrated. We tried tracking the nuclear membrane for a couple of movies. Unfortunately, the algorithm for this detection was flawed, demanding manual correction in many cases.

Probably the easier of all "solutions" is to recorded trajectories for a short period of time, where we can neglect most of external motion. Conversely, it should be long enough to allow a good statistics for inference of dynamics parameters. Unfortunately, the equipment available for our experiments could not generate detectable spots for shorter periods than 0.5 seconds without overheating the sample.

After those 3 failed approaches, we needed to invent a different method to handle our problem. In figure (7.3) we have a scheme representing what is observed. Vectors $\boldsymbol{r}_i$ are particle positions as measure in the microscope reference frame, but these measurements will include confound movement $\boldsymbol{R}$. Hence, we are interested in the intrinsic dynamics described by vectors $\boldsymbol{a}_i$. For simplicity, we are going to describe a system of 2 particles, but this model can be easily extend for more particles.

Using the Gaussian process model presented in the previous chapter, we can

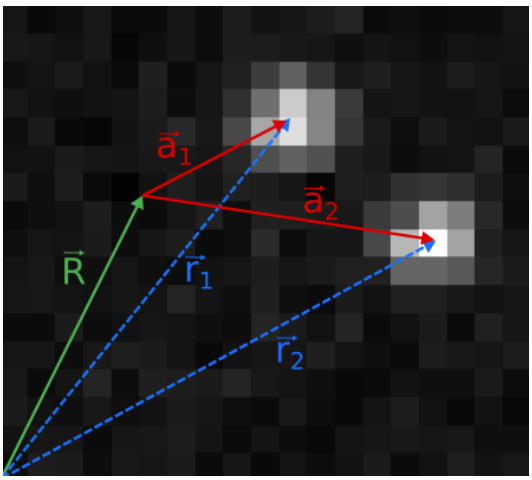

Fig. 7.3.: Scheme for confound movement correction. $\boldsymbol{r}_i$ are measured positions in the microscope reference frame. $\boldsymbol{R}$ represents the substrate movement, while $\boldsymbol{a}_i$ are the particles position in the moving reference frame.

express these relations as follows

$$\rho(\boldsymbol{a}_i, \boldsymbol{R}|\alpha_i, D_i) \propto \exp\left\{-\frac{1}{2}\boldsymbol{a}_1^T\Sigma_1^{-1}\boldsymbol{a}_1 - \frac{1}{2}\boldsymbol{a}_2^T\Sigma_2^{-1}\boldsymbol{a}_2 - \frac{1}{2}\boldsymbol{R}^T\Sigma_R^{-1}\boldsymbol{R}\right\}, \tag{7.11}$$

where we associated the FBM kernel Σ_i directly to $\boldsymbol{a}_i$. Due to local chromatin movement we know $\boldsymbol{a}_i$ are correlated through $\boldsymbol{R}$. We can write this expression in terms of measured positions $\boldsymbol{r}_i$ as follows

$$\rho(\boldsymbol{r}_i, \boldsymbol{R}|\alpha_i, D_i) \propto \exp\left\{-\frac{1}{2}\begin{pmatrix}\boldsymbol{r}_1\\ \boldsymbol{r}_2\\ \boldsymbol{R}\end{pmatrix}^T \begin{pmatrix}\Sigma_1^{-1} & 0 & -\Sigma_1^{-1}\\ 0 & \Sigma_2^{-1} & -\Sigma_2^{-1}\\ -\Sigma_1^{-1} & -\Sigma_2^{-1} & \Sigma_1^{-1}+\Sigma_2^{-1}+\Sigma_R^{-1}\end{pmatrix}\begin{pmatrix}\boldsymbol{r}_1\\ \boldsymbol{r}_2\\ \boldsymbol{R}\end{pmatrix}\right\}. \tag{7.12}$$

Presently, we are not interest in $\boldsymbol{R}$, therefore we can simply integrate this variable out and keep only $\boldsymbol{r}_i$. For that reason, we need to calculate the inverse of this central matrix in equation 7.12. To do so, we apply the results present in [81] on inverting 2x2 block matrix such as

$$\Lambda = \begin{pmatrix}A & B\\ C & D\end{pmatrix}. \tag{7.13}$$

In this publication, they show that

$$\Lambda^{-1} = \begin{pmatrix}A^{-1}+A^{-1}B(D-CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D-CA^{-1}B)^{-1}\\ -(D-CA^{-1}B)^{-1}CA^{-1} & (D-CA^{-1}B)^{-1}\end{pmatrix}. \tag{7.14}$$

Considering $A = \left[\left(\Sigma_1^{-1},\ 0\right);\ \left(0,\ \Sigma_2^{-1}\right)\right]$, $B = -\left[\Sigma_1^{-1};\ \Sigma_2^{-1}\right]$, $C = B^T$ and $D = \Sigma_1^{-1} + \Sigma_2^{-1} + \Sigma_R^{-1}$, we can show that

$$\left(\Lambda^{-1}\right)_{0,0} = \begin{bmatrix} \Sigma_1 + \Sigma_R & \Sigma_R \\ \Sigma_R & \Sigma_2 + \Sigma_R \end{bmatrix}, \tag{7.15}$$

allowing us to write the final equation of out model as

$$\rho(\boldsymbol{r}_i|\alpha_i, D_i) \propto \exp\left\{-\frac{1}{2}\begin{pmatrix}\boldsymbol{r}_1 \\ \boldsymbol{r}_2\end{pmatrix}^T \begin{pmatrix} \Sigma_1 + \Sigma_R & \Sigma_R \\ \Sigma_R & \Sigma_2 + \Sigma_R \end{pmatrix}^{-1} \begin{pmatrix}\boldsymbol{r}_1 \\ \boldsymbol{r}_2\end{pmatrix}\right\}. \tag{7.16}$$

This result shows us that the effects of confound movement will introduce correlations between the trajectory measured. Furthermore, this extra source of movement will increase the variance measured for each tracked spot. For that reason, we can expect to over-estimate the diffusion and anomalous coefficients if Σ_R is not considered.

Let's analyze an example to better understand the effects of this external source of movement over tracked spots. We can generate synthetic trajectories with uniform random values of D_α and α to infer these parameters back for comparison. Expecting more realistic simulations, we introduce localization noise of similar magnitude to those observed in real movies. We show in figure (7.4) the distribution of results over a total of 2000 simulated trajectories. As expected, D_α and α are largely over-estimated when we disregard external movement.

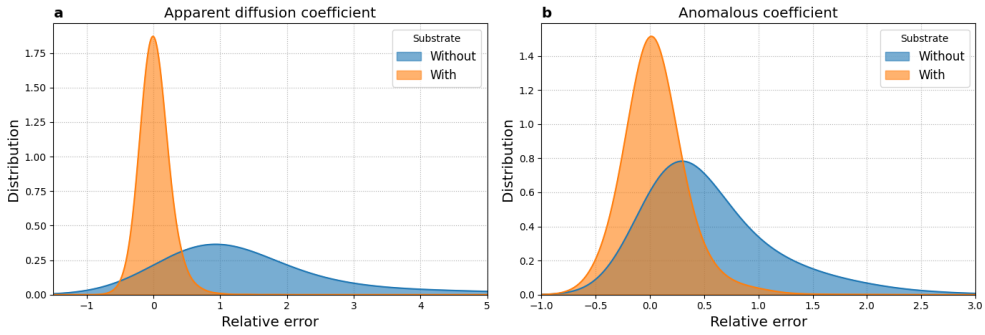

Fig. 7.4.: Comparison of apparent diffusion and anomalous coefficients inference quality on 2000 trajectories subjected to substrate movement. Extraneous movement introduces extra variance on trajectories, what cases parameters to be over-estimated if not properly considered.

Alongside optimization of apparent diffusion and anomalous coefficient for tagged particles, we infer these parameters for the background movement. In figure (7.5) we show the relative error in the inference of background movement if compared

to simulation set values.

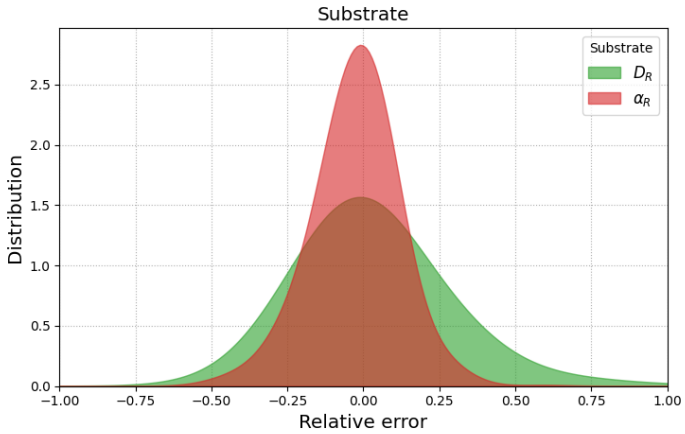

Fig. 7.5.: Relative error in the inference of susbtrate movement properties.

7.3.1 Effects on displacement distribution

Let's considered the displacement distribution of particles subjected to substrate movement[2]. In other words, how far away particles are from their original position on average when its measured trajectory includes some source of external movement. In figure (7.6a) we calculate the distribution of displacements for 2000 simulated particles with $D_\alpha = 0.15$ and $\alpha = 0.25$ that move under influence of a substrate with $D_R = 0.02$ and $\alpha_R = 1.35$. Notice that D_R is about 10 times smaller than the diffusion set for the particle itself. In dashed lines we show the displacement mean and distribution curves using values for D_α and α inferred without consideration of substrate. Continuous lines use corrected parameters. In figures (7.6b-e) we present distributions of all 2000 values inferred. These results demonstrate that substrate movement, if left unattended, will introduce significant errors on the estimation of diffusion parameters. On the same line, we show that the method present in this section is capable to resolve this issue to great accuracy.

7.3.2 Model performance over trajectories with static substrate

It is difficult to remove all possible sources of confound movement from experimental setup. As we saw, even tiny substrate movement generates non-negligible

2. The concept of displacement distribution was covered in the introduction.

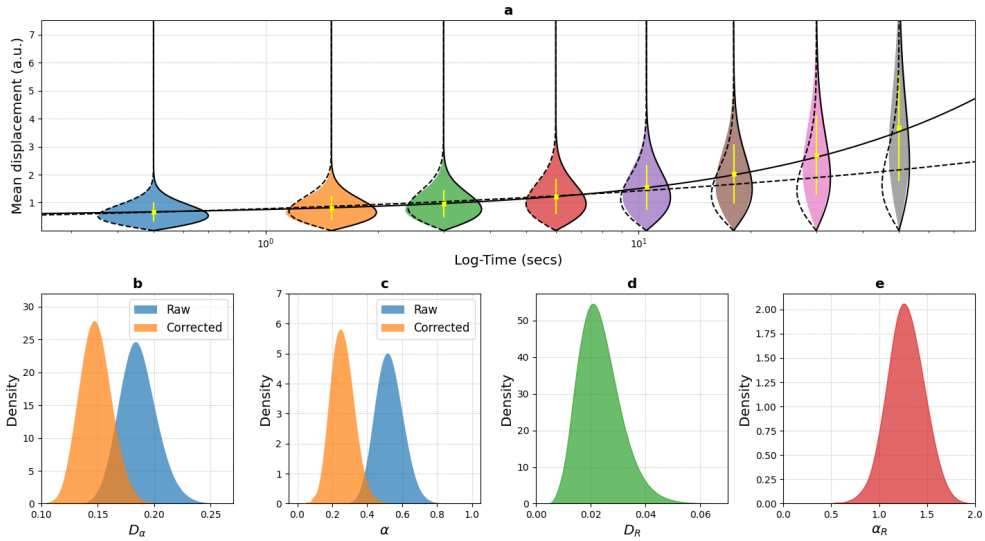

Fig. 7.6.: (a) Displacement distribution and mean over time calculated over 2000 simulated trajectories with $D_\alpha = 0.15$, $\alpha = 0.25$, $D_R = 0.02$ and $\alpha_R = 1.35$. Continuous lines use the average value obtained with method presented in this chapter. Dashed lines neglect substrate movement. (b-e) Distribution of parameters inferred for simulated trajectories.

effects for longer tracks. However, let's suppose we are 100% sure there is no substrate movement. In that case, how well does this model performs with a limited amount of data? Could we use this model as a general purpose model when 2 or more particles are present?

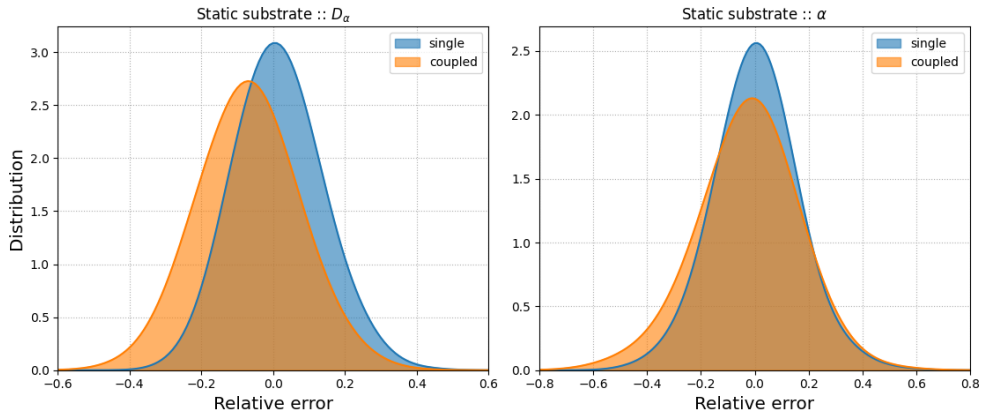

Fig. 7.7.: A sample of 2000 sets of two independent anomalous trajectories are simulate assuming a static substrate. The model presented can recapitulate results from a system where the substrate is static, however with lower precision.

To test that, we simulated another set of 2000 trajectories, but this time we do not introduce any source of external movement. In figure (7.7) we show that the

static model in the previous chapter performs better, with average relative error close to 0. Conversely, the model presented in this chapter presents an average relative error of about 10% . This result shows that when we are certain no extraneous source of movement is present, we are better off without assuming its existence. On the other side, when can we assume that? Hard to tell. For the experimental results presented in next chapter we shall always consider present background movement.

7.3.3 Estimating background trajectory

In most part of the time, we are only interested in inferring the diffusion and anomalous coefficients. What if we would also like to analyze the displacement of these particles disregarding background movement? For that purpose, let's go back to equation (7.12) and, this time, calculate the most probable path for $\boldsymbol{R}$ given trajectories $\boldsymbol{r}_i$. We have

$$\mathrm{E}\left[\boldsymbol{R}\right] = -\left(\Sigma_1^{-1} + \Sigma_2^{-1} + \Sigma_R^{-1}\right)^{-1} \begin{pmatrix} \Sigma_1^{-1} & \Sigma_2^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{r_1} - \boldsymbol{\mu}_1 \\ \boldsymbol{r_2} - \boldsymbol{\mu}_2 \end{pmatrix}. \tag{7.17}$$

The variance can be estimated using

$$\Sigma_{\boldsymbol{R}} = \left(\Sigma_1^{-1} + \Sigma_2^{-1} + \Sigma_R^{-1}\right)^{-1}. \tag{7.18}$$

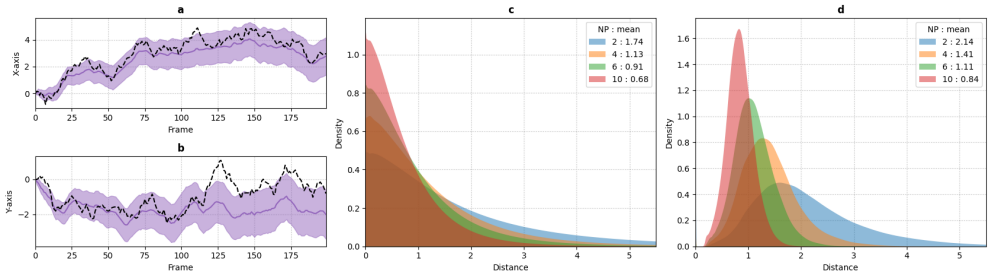

Fig. 7.8.: (a) Comparison between estimated (purple) and simulated (black) trajectories described by substrate in a system with two particles. In (b) distribution of errors when comparing inferred and simulated trajectories for systems with 2, 4, 6 and 10 particles. In (c), the distribution of standard deviation estimated for background trajectories. Overall, the greater the number of particles, the more precise is the estimation.

There is no easy analytical solution for these equations. Regardless, we can solve them numerically. In figure (7.8a) we estimate background movement with one standard deviation for an example set of 2 trajectories. In black we show the simulated substrate movement. In (7.8b-c) we display overall accuracy when

working with 2 or more particles. These results are not incredibly accurate, but good enough for many applications.

7.4 GP-Tool

All these methods developed in this chapter, among other utilities, are part of a small application I developed. This software is called GP-Tool and can be downloaded, along with extensive documentation, from my Github page (https://github.com/guilmont/GP-Tool). At the current stage, the program provides 4 plugins: movie, alignment, trajectories and g-process. The movie plugin will allow the user to open TIFF files, display basic ImageJ and OME metadata, define color-maps for each channel and manually correct for contrast. The alignment plugin will perform the algorithm described in this chapter for digitally correct possible camera calibration and chromatic aberration. Alternatively, the user can manually modify each of the parameter at will. In figure (7.9) we present examples for these plugins.

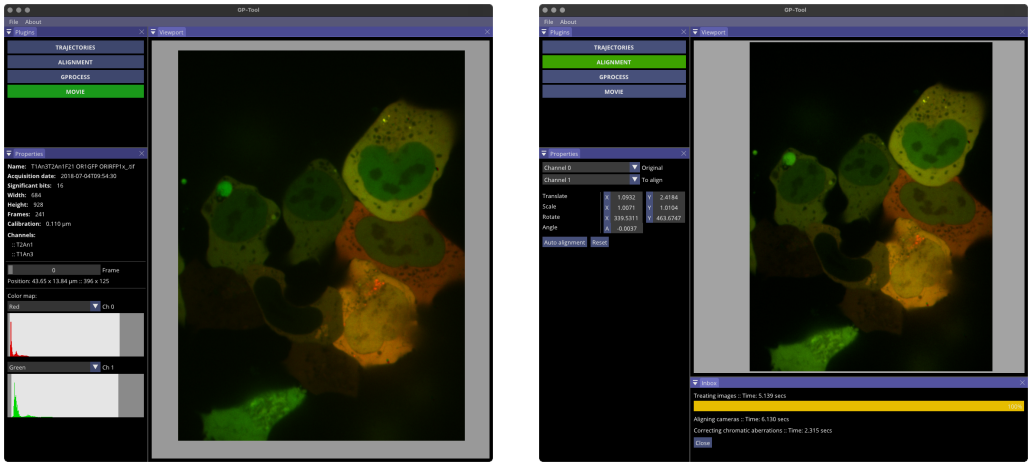

Fig. 7.9.: On left, we display a few basic metadata for the movie loaded and general utilities for visualization. On the right, a view of the alignment plugin. The user can manually setup alignment parameters or perform auto-alignment as described in this chapter.

There are many third party applications that do a great job detecting and linking spots into coherent trajectories. For this thesis we used ICY [3], for example. Due that reason, this software can import XML files exported by ICY or standard CSV files. For practical reasons, GP-TOOL will demand a different file per channel. Once loaded into the program, user will have access to all parameters used in equation (7.1). A example is presented in figure (7.10).

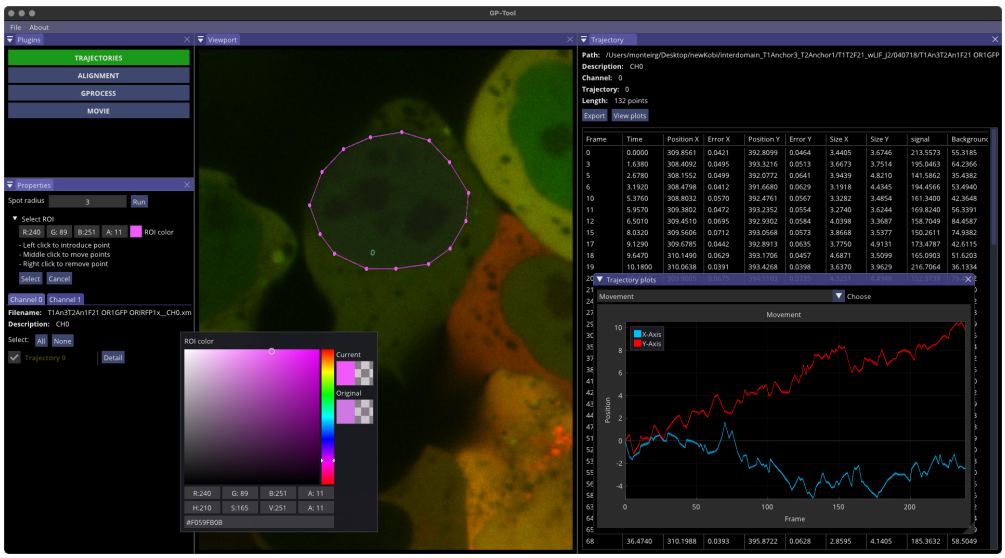

Fig. 7.10.: This plugin will display trajectories with higher localization accuracy among other parameters of interest. The user can also visualize these parameters in a graphical manner.

The program can perform the analysis of several cells in the same movie. Using the ROI utility under the trajectories plugin, the user can select spots of interest from which diffusion and anomalous coefficients are inferred and corrected for substrate movement. It is also possible to use MCMC sampler to verify the probability distribution associated with each of these parameters. Finally, we can infer substrate information for all cells with more than one particle.

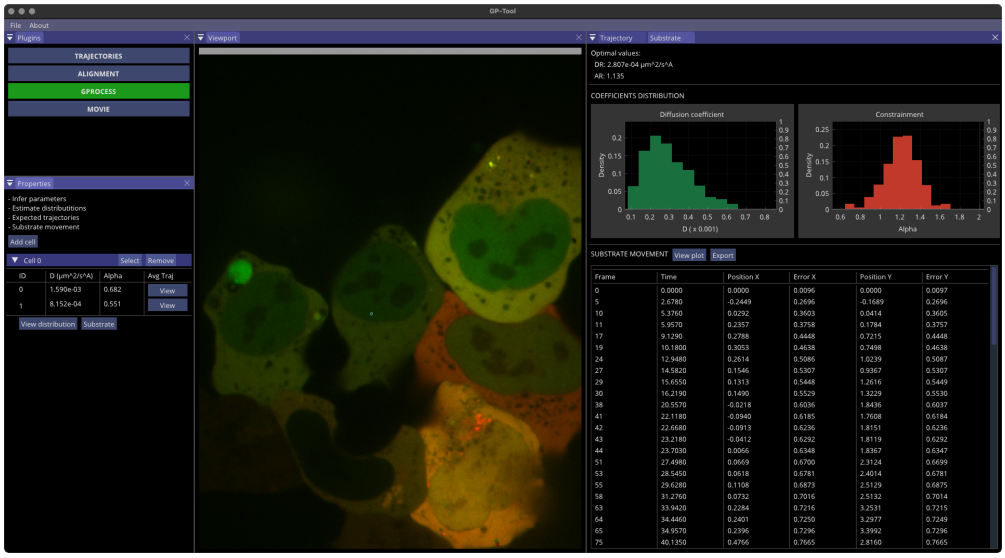

Fig. 7.11.: G-Process plugin will infer diffusion and anomalous coefficients for each spot selected. We can also infer trajectories for cell if it contains more than one particle.

Once the analysis is completed, I provide utilities so save results into 2 file formats: JSON and HDF5. I also provide export functions to save tables in CSV format. All these formats are easily parsed in all major computing languages, such as C/C++, Python and R.

8

Measuring chromatin dynamics

8.1 Comparing interphase and mitosis

Since the first time mitotic cells were observed under a microscope in the late 1800s, we learned that the nuclear content wildly changes its condensation state between interphase and mitosis. It was measured by volumetric assays and FRET-based methods that chromatin condensates 2 to 3 times from interphase to mitosis [5, 6, 7], if cellular division is to be accomplished in typical allocated space. As a matter of fact, the structure of mitotic chromatin has intensively studied [8, 9], but it is not clear how diffusion properties of chromatin are influenced by condensation.

In this chapter we want to investigate the effects of chromatin arrangement and condensation on diffusive properties. To do so, we used a mouse Embryonic Stem (mES) cell line in which TetO arrays 7kb long are piggybacked into about 20 to 25 random locations of the genome. Upon transfection and expression of GPF::TetR, these locations become visible at the microscope. In order to tell between cells in interphase and mitosis, Hoechst staining was used. In appendix A the whole protocol is described more precisely. In figure (8.1) we have an example of these cells, recorded at 4 frames/sec in a spinning disc microscope for 75 seconds.

Unsurprisingly, the probability of finding naturally occurring mitotic cells was very low. Due this reason, we perform nocodazole based synchronization. Masks were generated by assigning individual color tags for every cell, where the blue channel was used to identify cells in interphase, mitosis and nocodazole arrest. Labeled loci were detected and tracked with usage of plugins from ICY [3]. Lastly, these loci were grouped per cell using manually painted masks and batched into

the GP-FBM pipeline provided with GP-Tool.

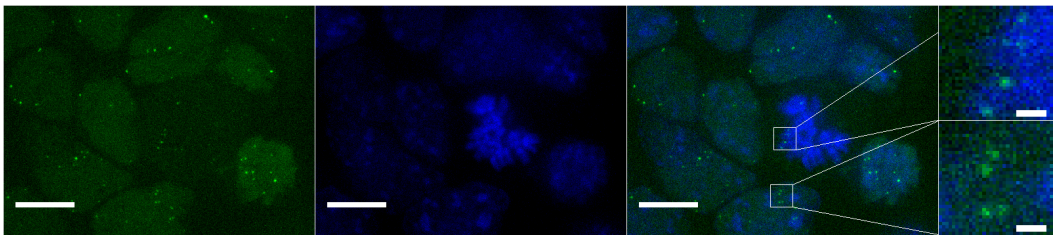

Fig. 8.1.: Approximately 20 TetO with 7kb of size are inserted in mES cells at random locations. These arrays can be visualized upon GFP::TetR expression. Hoechst staining is perform to differ interphase and mitotic cells. Scale bars on left are 1 μm.

As a first result, we calculate the displacement distribution using all the trajectories on similar cell cycle stage or treatment for several time steps and merge all of them in a single histogram. For that purpose, we normalize each displacement calculated for a given trajectory at time Δt by $\sqrt{2D_\alpha\ \Delta t^\alpha}$, where D_α and α were optimally inferred for said trajectory. In figure (8.2) we can verify that the displacement of our random insertions in chromatin are very nicely approximated by a Gaussian distribution for times similar or greater than one second. Furthermore, we also observe that their dynamics are self-similar in interphse, mitosis and nocodazole arrest.

As we have enough trajectories for a good estimation of the velocity autocorrelation for insertions in interphase, mitosis and nocodaloze arrest, we show in figure ('8.3) a comparison between the theoretical curve for the FBM model presented in equation (6.9) and results calculated from data. As we can see, the FBM model seems to be appropriate fir the analysis of chromatin.

Next, we compare the displacement curves described by each one of these cell cycle stages with their theoretical distribution curves (2.5) along with their mean displacement curve in time (2.6). For that, we shall use D and α inferred directly and with background correction. In figure (8.4a-c) we present in colors the displacement calculated from each stage in time. Dashed lines are theoretical curves without correction while continuous lines are corrected for background movement. As mentioned before, depending on how agitated cells are, we might be able to define a certain period of time in which background effects are negligible. For this case in particular, cells were moving very slowly, about 10 time slower than spots themselves, hence for intervals shorter then 10 seconds the effects of background were negligible.

Substrate movement is also noticeable when investigating the total variance of

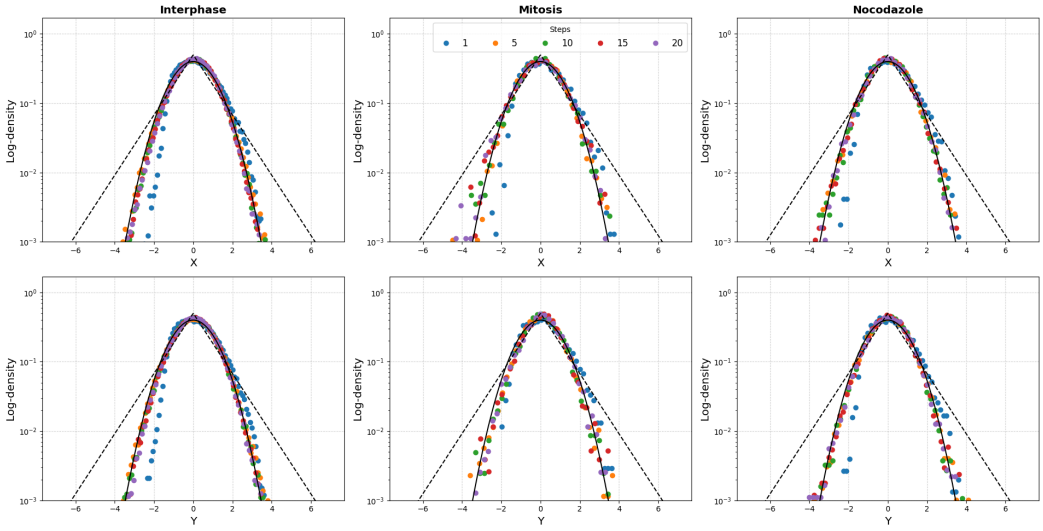

Fig. 8.2.: Normalized accumulated displacement distribution over all the trajectories under similar state, that is, interphase, mitosis and nocodazole arrest. The displacement calculated for each trajectory was normalize by $\sqrt{2D_\alpha \, \Delta t^\alpha}$, where D_α and α were optimally inferred for each trajectory and $\Delta t = n \; \delta t$, with $\delta t = 0.25$ seconds. For comparison, the continuous black line represents a standard normal distribution and the black dashed line corresponds to a Laplacian distribution. As we can see, insertions have Gaussian dynamics and are self-similar for periods similar and greater than one second.

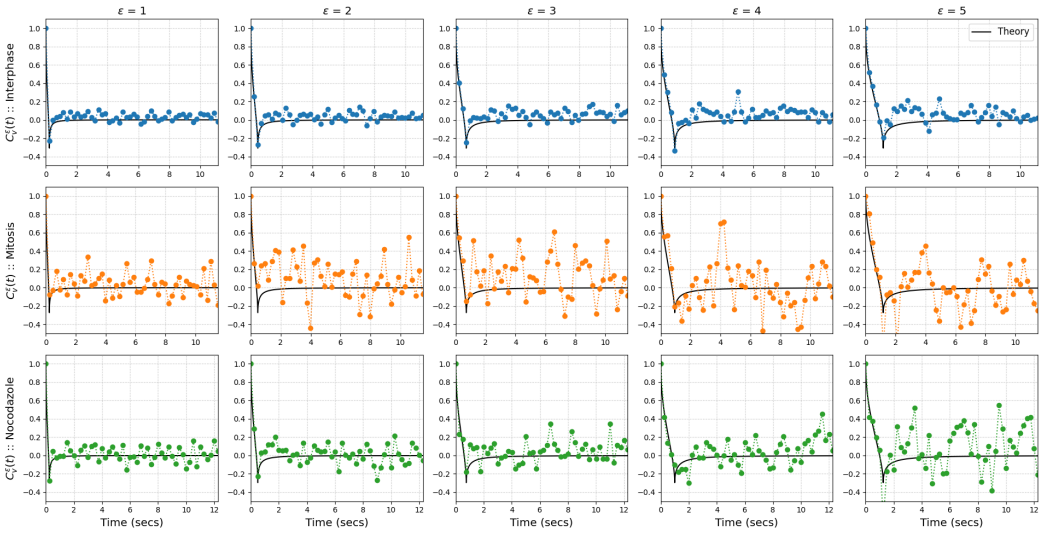

Fig. 8.3.: Average velocity autocorrelation curve calculated from all trajectories in interphase, mitosis and nodocaloze arrested cells. Theoretical curves were calculated using sample's average anomalous coefficient.

our samples. In figure (8.5a-b), we show how much of samples variance comes from spots within and among cells. In the image, gray represent among cells. Interphase

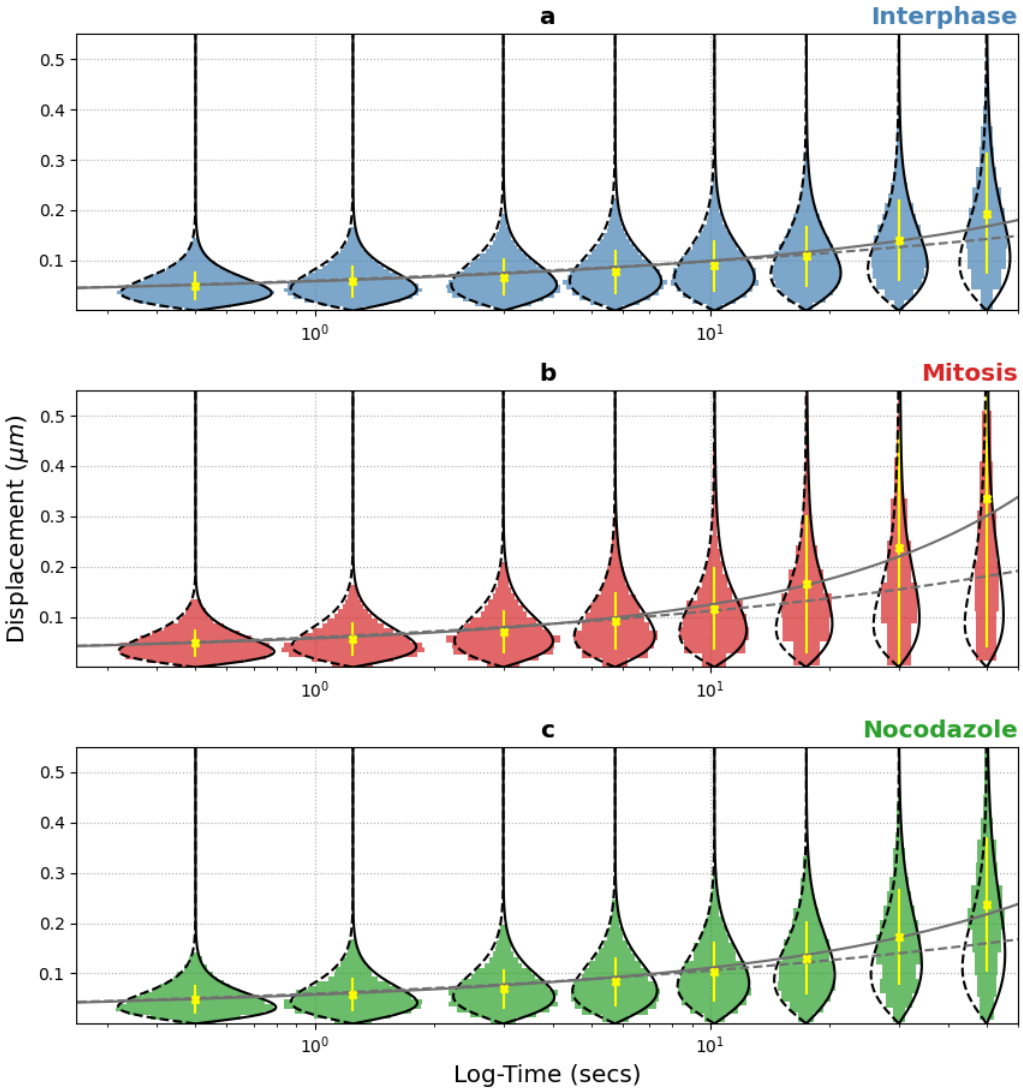

Fig. 8.4.: Evolution of mean displacement over time. In blue, red and green we have displacement distribution plots calculated from samples of interphase, mitotic and nocodazole arrested cells. Yellow points correspond to experimental distribution's mean. Dashed black lines are theoretical Gaussian distributions calculated with D_α and α inferred without substrate movement correction, while continuous black lines were calculated with corrected parameters. Gray lines correspond to the theoretical mean displacement calculated using corrected (continuous) and raw (dashed) diffusion parameters.

cells (in blue) were quite slow, so we don't observe a big difference when taking fractions before or after correction. Mitotic cells (in red) are different. As they loose adherence to the plate prior division, they become more mobile. The direct effect can be seen via the fractions presented for diffusion and, even stronger, for

the anomalous coefficients. Nocodazole cells (in green) are less mobile than mitotic cells as they have time to sediment on the plate before the movies are recorded, but cellular movement effects are still evident. In any case, we notice that most of the sample variance (more than half) come from within cells.

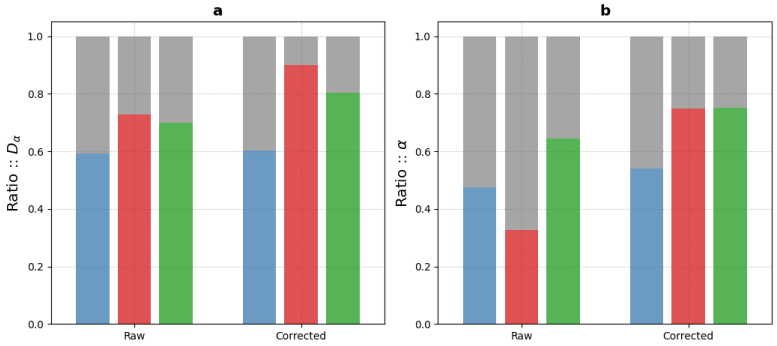

Fig. 8.5.: Determining how much of samples variance come from within cells and how much comes from different cells. As we can see, upon correction for background movement, most of the variance come from within cells. Interphase cells are in blue, mitotic cells are in red and nocodazole arrested cells in green.

The next logical question regards how great this variability is. In figure (8.6) we present the distribution of corrected apparent diffusion and anomalous coefficients. In short, D_α is approximately $1\ \mu m^2/s^\alpha \pm 50\%$ for all cycle stages. The anomalous coefficient is more variable: for interphase $\langle\alpha\rangle = 0.38 \pm 39\%$, for mitosis $\langle\alpha\rangle = 0.45 \pm 56\%$ and $\langle\alpha\rangle = 0.48 \pm 27\%$ for nocodazole arrested cells. Interestingly, the apparent diffusion coefficient have statistically similar means and distributions across samples. The same can be said about the anomalous coefficient between interphase and mitosis. Even though the nocodazole sample has similar mean and distribution to mitotic cells, they have significant distinct mean and distribution if compared to cells in interphase [1].

These results show that the average diffusion coefficient of chromatin loci are probably similar across cell cycle stages.The same applies for the anomalous coefficient with exception of arrested cells. The hypothesis to explain why the anomalous coefficient is different between interphase and arrested cells relies on how nocodazole affects micro-tubules, necessary components for chromosome alignment prior to division. It is possible that missing micro-tubules will allow chromosomes to move with increased freedom and less constraint if compared to interphase. However, such hypothesis remains to be tested.

1. Means were tested via ANOVA test, while distributions via Wilcoxon rank-sum tests. All tests were corrected for multiple sample comparison via Benjamini/Hochberg false discovery rate [82].

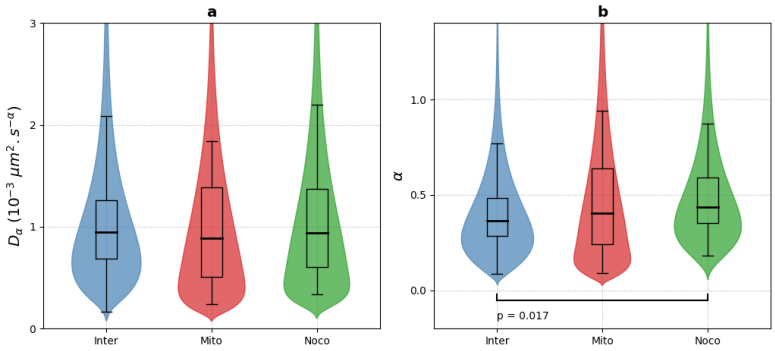

Fig. 8.6.: Distribution of apparent diffusion and anomalous coefficients measured using GP-FBM correcting for substrate movement.Diffusion coefficients are statistically similar means and distributions. The same applies for anomalous coefficient, with exception of interphase with arrested cells. Interphase in blue, mitosis in red and nocodazole arrested cells in green.

8.2 The HoxA domain

The study on TetO system showed us a non-negligible variability of D_α and α within cells. This effect was ever so more evident when inference of these parameters were corrected for substrate/background movement, where, in some cases, up to 80% of variability was observed to come from within single cells. This result hinted us to speculate over the reason for such variability and its possible correlation to function. We find in literature case studies showing that chromatin regions in proximity to centromeres and telomeres tend to be less mobile (reduced D_α) in yeast [10]. We also find some results linking transcriptional activity to increased local confinement (decrease α) [11] and/or increase gene mobility [12]. Otherwise, little is known about the effects of genome context over chromatin dynamics.

Unfortunately, the TetO system does not allow us to demand such questions, because arrays are introduced in a random fashion. Therefore, a different system in which specific loci are tagged is required. As a second point, it would also be interesting to tag a gene locus with activity profiles depending on differentiation stages. Thence, a good candidate for that purpose is the HoxA locus. Hox genes are repressed in embryonic stem (ES) cells, but active upon differentiation to neuron precursor (NP) cells, for instance. This system is also interesting from a topological perspective. In figure (8.7a) we present the Hi-C map of the HoxA domain along with insertion positions. Once cells differentiate into NP, this domain will change its topology as presented in figure (8.7b). As a general concept, it is believed that ES cells have a more open and accessible chromatin than differentiated cells. These Hi-C maps are a great example of that. The NP map presents overall greater contact

probability than if compared to ES map. This observation indicates, as we shall see, that distal loci tend to be spatially closer, thus more compacted compared to ES.

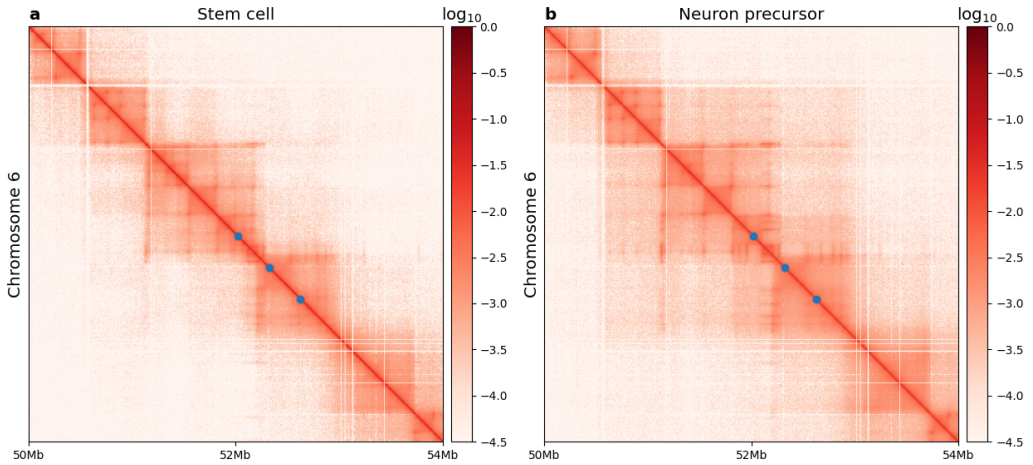

Fig. 8.7.: Hi-C map displaying the HoxA domain for a mES cell in (a) and mouse neuron precursor cells in (b). Points represent the location in which ANCHOR arrays were inserted.

In total 2 ES lines were generated by double-labeling with ANCHOR [13]. ANCH1 and ANCH3 labels were introduced into different locations within same allele of chromosome 6 for inter-TAD (T1-T2) and intra-TAD (T2-T3) lines, as specified in figure (8.7). Strategically, T1 and T3 are equidistant to T2 ($\sim$ 300 kb), which allowed us to further inquire on effects of TAD structure in 3D distances. Employing time-lapses recorded using a spinning disc microscope with one camera per channel at 2 frames/sec and the GP-FBM workflow, we assess D_α and α for ES cells and differentiation induced cells via retinoic acid. These cells are not to be considered as NP, but just enough for a differential analysis. The protocol used in the creating and treatment of these cell was provided by Tom Sexton and is presented in appendix A.

As one might expect, we found significant differences in inter-probe distances for control and differentiation induced cells, figure (8.8a-b). Interestingly, the average distance between probes remained similar post-induction [2]. Moreover, the difference in overall variance could be explained (in part) by the increase in mobility for induced cells, figure (8.9a,c). Nonetheless, Wilcoxon rank-sum test was performed to verify if distributions are different, but no statistical significance was found.

Before using GP-FBM on those three loci, we must show that these insertions present Gaussian displacement and are self-similar for both differentiation states, that is, stem and induced. Fortunately that seems to be the case. In figure (8.10),

2. Arguably, $\langle \Delta_{12} \rangle$ can be considered different with p-value = 0.051.

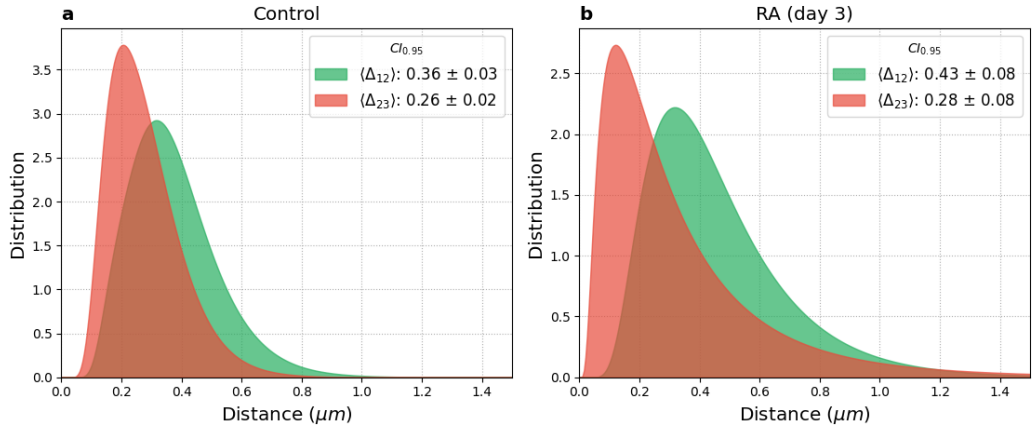

Fig. 8.8.: Distribution of inter-probe distances calculated between T1-T2 (Δ_{12}) and T2-T3 (Δ_{23}) for ES cells in (a) and induced cells in (b). Δ_{12} and Δ_{23} are statistically different against each other, but remain similar upon differentiation.

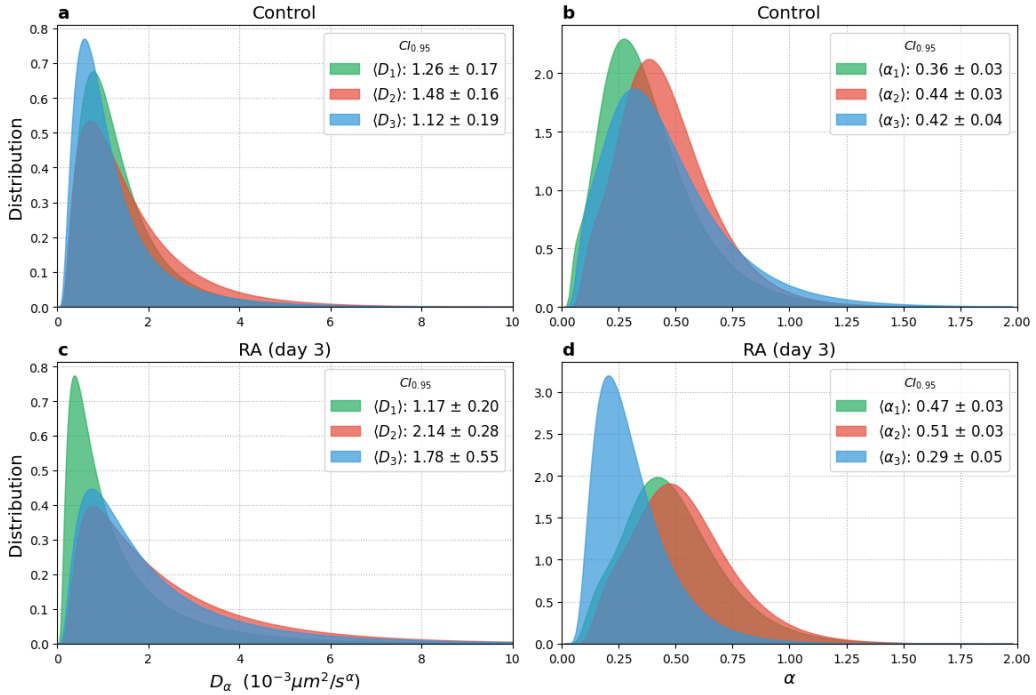

Fig. 8.9.: Distribution of apparent diffusion and anomalous coefficients measured for insertions in T1, T2 and T3 for ES and NP cells. Other than obvious significantly similar and different distributions, ES cells have $\langle D_2\rangle \neq \langle D_3\rangle$ and $\langle\alpha_1\rangle \neq \langle\alpha_2\rangle = \langle\alpha_3\rangle$. Apart from that, $\langle D_1\rangle$ is statistically similar in ES and NP cells.

we show results that support the usage of GP-FBM for the subsequent analysis. Technically speaking, we should also show that the velocity autocorrelation of these

loci is also that of a FBM, however we don't have enough data to determine semi-smooth curves, hence we omit those plots in here.

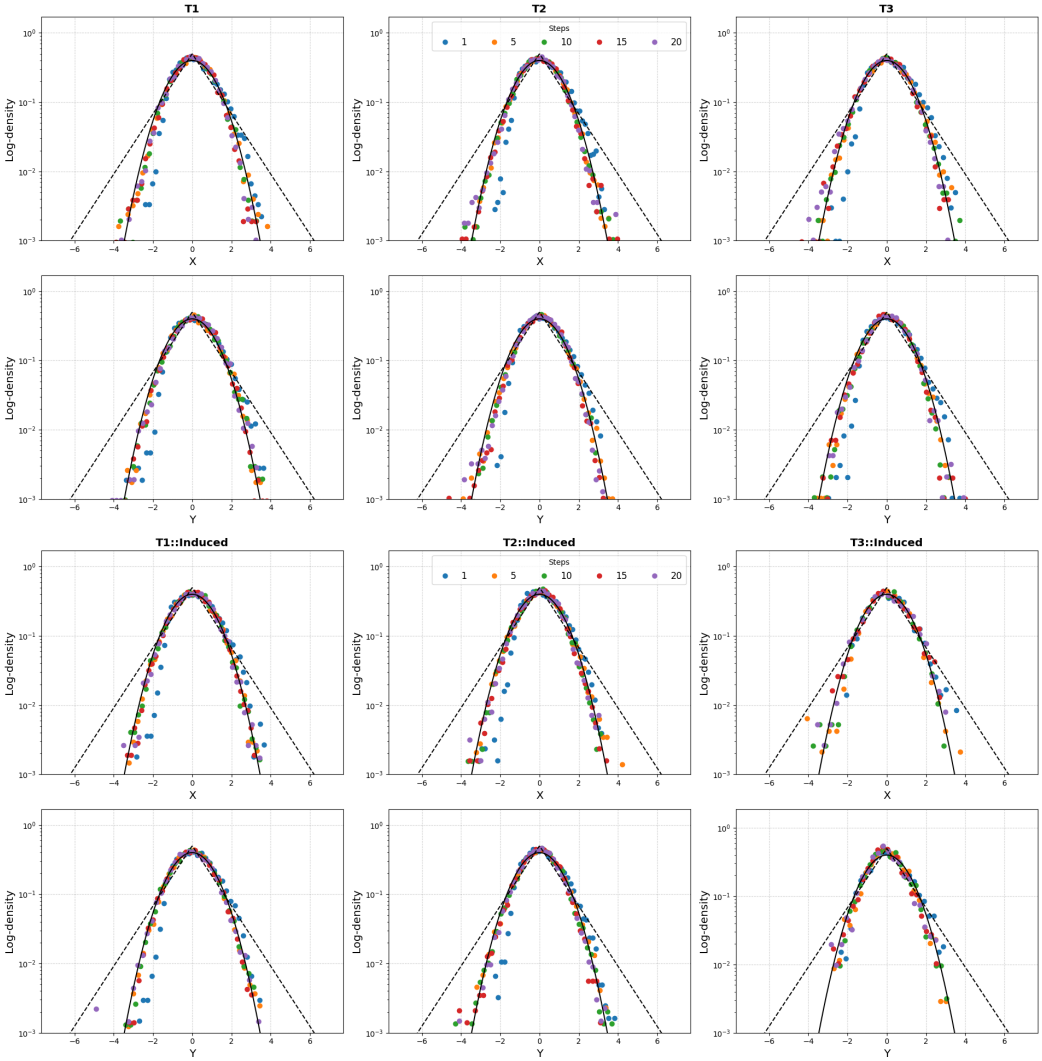

Fig. 8.10.: We group displacement distributions for all spots of the same insertion and differentiation state. For each trajectory, displacements were normalize by $\sqrt{2\ D_\alpha\ \Delta t^\alpha}$, where Δt = n/2 secs. All insertions present Gaussian self-similar displacement distributions.

We found that $\langle D_2 \rangle$ is different from $\langle D_3 \rangle$ with p-value = 0.03 in ES cells, while T1 is significantly more confined than T2 and T3 with p-values 0.001 and 0.1041. On the differential analysis, we find that $\langle \alpha_1 \rangle$ is significantly higher for induced cells than in ES cell, while we observe a significant drop in $\langle \alpha_3 \rangle$ with accompanying increase in mobility $\langle D_3 \rangle$ upon induction with retinoic acid.

Let's determine if these differences can be associate to functional reasons. We find that T1 is near a putative active enhancer of Halr1, encoding for the non-coding

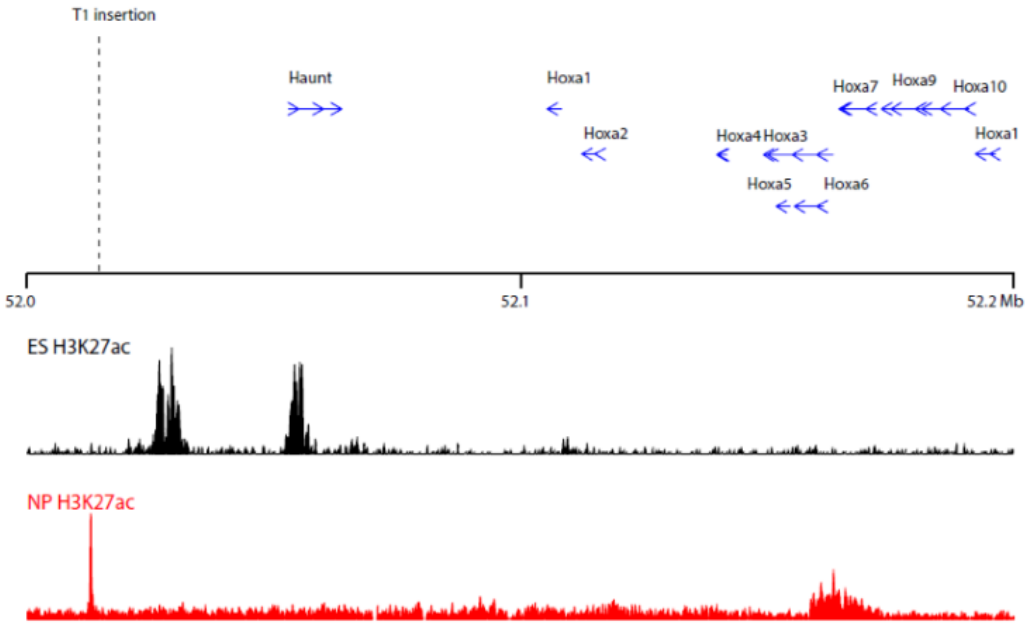

Fig. 8.11.: We identify activity chromatin marks for the HoxA repressor Haunt gene in ES cell. These marks vanished upon differentiation.

RNA Haunt, which represses Hox genes [16]. In figure 8.11 we show active histone mark H3K27ac at Haunt for ES cells, while the peak vanishes for NP cells. This result is aligned with [11]. In contrast, no significant differences was observed for the $\langle D_1 \rangle$ in relation to $\langle D_2 \rangle$ and $\langle D_3 \rangle$.

Part III

Biophysics: Modeling chromatin

9

Stochastic systems

The final aim for this part III is to build a model for chromatin that recapitulates all the measurements done for the HoxA domain in part II, including distances, apparent diffusion and anomalous coefficients. Jumping straight into the final model would occlude most of the subtleties associated with the model and all the concepts needed for a proper understanding and correct interpretation. For this part, then, we shall begin our journey by introducing stochastic processes and how to treat them numerically. In particular, we dedicate this chapter to the exploration of Langevin equations and how we can use it to study diffusion of particles. In a first instance, we are going to study the traditional Brownian motion and, later on, a small mass attached to a spring susceptible to thermal noise.

9.1 Diffusion dynamics

In principle, there is nothing stochastic about a system with many particles. Knowing their position and velocity at any given time would be sufficient to describe their motion indefinitely. Nonetheless, to analytically solve a system of equations like this is very counter productive. The solution of this problem would become impracticable even for a system of few particles due to the correlation introduced due to their interaction. Fortunately, the exact solution is not necessary, dare say undesired. Using tools of Statistical Physics, we can easily estimate the probability of any given state for the system which is more palatable for common usage. Let us consider, for simplicity, a colloid with inertial mass m bathing in a fluid of

much smaller particles. The equation of motion for this colloid is

$$m\frac{d}{dt}\boldsymbol{v}(t) = -\gamma\boldsymbol{v}(t) + \boldsymbol{\eta}(t). \tag{9.1}$$

This equation contains 3 components. On the left most, we have the inertial term, which determines how the colloid will react to forces applied over it given its inertial mass. The middle term is called drag force and can be understood as a response of the fluid to any given movement depicted by this colloid. Finally, the equation above presents a force $\boldsymbol{\eta(t)}$ accounting for the interaction of this colloid with the fluid which is, in principle, stochastic. Notice that, intuitively, this same cause of movement is also indirectly responsible by its damping [1]. In the simplest of models, we will consider the interaction between colloid and fluid to be elastic, that is, no energy is consumed, but balanced according to the inertial mass of each particle involved. Hence, we comprehend that if the colloid mass is large compared to fluid particles, the effects of each individual interaction with the fluid will be minimal on the colloid. In fact, it has been measured that the decorrelation time for each interaction is in the order of picoseconds for micron sized particles. For this reason, we can assume that the correlation between interactions will be

$$\mathrm{E}\left[\eta_i(t) \cdot \eta_j(s)\right] = g^2\ \delta_{ij}\ \delta(t-s), \tag{9.2}$$

where $\delta(t-s)$ is known as Dirac delta and tells us that interactions are to be considered instantaneous, while δ_{ij} determine that each direction is independent of the other. Concurrently, the interaction variance at any given time is g^2. We should also consider the average effect of this stochastic force. If the fluid is at equilibrium and not flowing, we should expect that $\mathrm{E}\left[\eta_i(t)\right] = 0$.

Notice we had used the mean symbol $E[]$ twice, but how exactly is it calculated? As the decorrelation time for interactions between colloid and fluid are approximately instantaneous, we can naturally perform our calculations over a long period of time. Notwithstanding, ensemble averages should wield similar results of those calculated over time. In a simple view, this possibility to exchange between ensemble and time averages defines ergodicity. In other words, given enough time the system will "visit" all allowed states, which is equivalent to measuring several systems at random time points.

However, what is the distribution of this stochastic force $\eta_i(t)$? To answer this question, we can recall the central limit theorem stating that the added effect of measured forces in a picosecond, or the average affect of those forces, should be

1. For a macro life example, take for instance biking on a calm sunny day with no wind. The faster you go, the stronger will be the "wind" against you. This "wind" is called a drag force, generate by your pedaling.

normally distributed. Hence, the stochastic force in our equation of motion is to be considered as white noise. Later on, we shall connect this effect to thermal energy.

As we are not particularly interested in the non-equilibrium regime of this system, we should consider only the dynamics for large periods of time, that is, when the system undergoes stationary dynamics. Assuming this condition, it is easier to solve this equation (9.1) in the Fourier space. We have

$$v_i(\omega) = \frac{\eta_i(\omega)}{\gamma + i\omega m} \tag{9.3}$$

for a certain frequency ω, where i as the complex factor.

In order to calculate the auto-correlation function for the velocity, we use the Wiener–Khinchin theorem [83] and derive the power spectrum to be

$$|v_i(\omega)|^2 = \frac{|\eta_i(\omega)|^2}{\gamma^2 + \omega^2 m^2}, \tag{9.4}$$

The power spectrum of our stochastic force is constant for all frequencies, a property of white noise. This constant is simply given by the variance g^2. Thence, we calculate auto-correlation as the inverse transform

$$\mathrm{E}\left[v_i(t+\tau)\; v_i(t)\right] = \frac{g^2}{2\pi m^2}\int_{-\infty}^{\infty} d\omega\; \frac{\exp\left\{-i\omega\tau\right\}}{(\gamma/m)^2 + \omega^2} = \frac{g^2}{2\gamma m}\exp\left\{-\frac{\gamma}{m}\tau\right\}, \tag{9.5}$$

for $\tau \geq 0$. Notice that for any arbitrary time t, the correlation involved depends solely on the interval τ, which indicates one aspect of stationary dynamics. Also, notice that interactions with the fluid are considered to be instantaneous, but its overall effect on the colloid is propagated into much larger periods of time, exponentially distributed.

Equation (9.5) can be used to determine the value of g if we ponder over the energy implied for this system. Using results from Maxwell and others, we know that the energy disposed by the fluid in which our colloid dwells is thermal. Concomitantly, the kinetic energy of this colloid should emerge from fluid temperature. This results in the following relation

$$g = \sqrt{2\gamma k_B T}, \tag{9.6}$$

where k_B is the Boltzmann constant and T is temperature. This verification is commonly called fluctuation-dissipation theorem, because we associate the cause of movement to be directly associated with is resistance γ. Hence, we write velocity's

auto-correlation function in its final format

$$\mathrm{E}\left[v_i(t+\tau)\ v_i(t)\right] = \frac{k_B T}{m} \exp\left\{-\frac{\gamma}{m}\tau\right\}. \tag{9.7}$$

We can use this result to determine the mean squared displacement(MSD) as follows

$$\begin{aligned}
\mathrm{E}\left[\Delta r_i^2(t)\right] &= \int_0^t d\tau \int_0^t ds\ \mathrm{E}\left[v(\tau)\ v(s)\right] \\
&= 2\int_0^t d\tau\ (t-\tau)\ \mathrm{E}\left[v(0)\ v(\tau)\right] \\
&= 2\frac{k_B T}{\gamma}\ t - 2\frac{m k_B T}{\gamma^2}\left[1 - \exp\left\{-\frac{\gamma}{m}t\right\}\right].
\end{aligned} \tag{9.8}$$

In the limit of t much larger than m/γ, we can simplify the MSD equation to

$$\mathrm{E}\left[\Delta r_i^2(t)\right] = 2\frac{k_B T}{\gamma}\ t = 2D\ t, \tag{9.9}$$

where we defined $D = k_B T/\gamma$. Notice that the diffusion coefficient depends not only on the temperature, but on how the object/colloid interacts with this fluid. The feature will be largely used later on, when we model chromatin. There, the loci will interact differently with the substrate depending on its context.

For completeness, we present the probability density function (PDF) for velocities and displacements for our colloid. Using the auto-correlation function for velocity and CLT, we have

$$\rho(v_i|T,m) = \sqrt{\frac{m}{2\pi\ k_B T}}\ \exp\left\{-\frac{m v_i^2}{2 k_B T}\right\}, \tag{9.10}$$

the Maxwell-Boltzmann PDF for velocities. This equation represents how kinetic energy is distributed or partitioned in the system. For that reason, this result is also called the partition function of the system. The displacement distribution is given by

$$\rho(r_i|D,t) = \frac{1}{\sqrt{4\pi D t}}\ \exp\left\{-\frac{r_i^2}{4Dt}\right\}, \tag{9.11}$$

where we assume the colloid started in $\boldsymbol{r}(0) = \mathbf{0}$. For t in limit of zero, this equation becomes a delta Dirac for position zero, but the variance increases with t. In the limit for t going to infinity, we could expect a homogeneous probability of find the particle in any point in space.

9.2 Wiener process

The previous section gave us a reasonable intuition over traditional Brownian motion. However, we cannot directly use those equations to generate stochastic trajectories. Needless to say, we could use the GP-FBM approach presented in the chapter 6 to generate these trajectories, but due the simplicity presented by $\alpha = 1$, we do not need to go through the trouble. Let us suppose we are in diffusion time scale [2], that is, effects of the random force are approximately instantaneous, so we can neglect inertial term in equation (9.1) and write

$$dx = \frac{\eta(t)}{\gamma}\, dt \longrightarrow dx = \sqrt{2D} dW_t, \tag{9.12}$$

where dW_t is defined in the Itô sense [84] and is called the Wiener process. It has a few interesting properties such as $\mathrm{E}[W_t] = 0$ and $\mathrm{Var}[W_t] = t$. On can also show it is a stationary process $W_t - W_s \sim \mathcal{N}(0, \sqrt{|t-s|})$ with temporal correlation $\mathrm{E}[W_t W_s] = \min(t, s)$, which is a special case of the fractional Brownian motion kernel (6.8) for $\alpha = 1$.

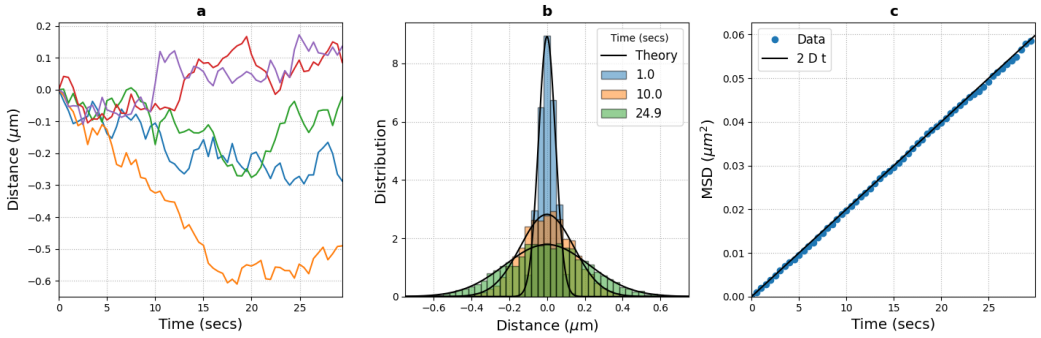

Fig. 9.1.: (a) Few examples of 1D Wiener process using $D = 0.001\mu m^2/s$ and $dt = 0.5$ seconds. (b) Displacement distribution in time. (c) Mean squared displacement curve calculated over 4096 synthetic trajectories and analytically.

In figure (9.1), we generate 4096 one dimensional trajectories using the Wiener process with $D = 0.001\mu m^2/s$ and $dt = 0.5$ seconds. In (a) we have a few examples of simulated trajectories. We calculated the displacement distribution for different times and compare with the theoretical result of equation (9.11) in (b). Finally, we calculate the average squared displacement among all particles and compare with the theoretical result (9.9). It worth mentioning that a large number of particles are necessary for smooth plots that mimic the analytical results. In the next section we will further discuss how to solve stochastic differential equations numerically.

2. Conversely, we could assume that the fluid is very viscous.

9.3 Numerical integration of stochastic equations

We will solve many stochastic/Langevin equations in the next chapters. For that, we are going use a Verlet-like integration method modified to take into consideration the non-continuum behavior of stochastic Langevin forces. The complete derivation and analytical examples can be found in [85]. For simplicity, I transcribe equations (21) and (22) of that paper here

$$r_{n+1} = r_n + b\ dt\ v_n + \frac{bdt^2}{2m} f_n + \frac{b\ dt}{2m}\ \eta_{n+1}, \tag{9.13}$$

$$v_{n+1} = av_n + \frac{dt}{2m}\left(af_n + f_{n+1}\right) + \frac{b}{m}\ \eta_{n+1}, \tag{9.14}$$

with

$$a = \frac{2m - \gamma dt}{2m + \gamma dt} \tag{9.15}$$

$$b = \frac{2m}{2m + \gamma dt}. \tag{9.16}$$

For more integration methods and comparisons [86, 87, 88].

To exemplify this method, let's study the traditional 1D harmonic oscillator in a thermal fluid. This system is a good choice due its simplicity and facility to obtain an analytical solution. The Langevin equation describing its behavior is

$$\begin{aligned} m\frac{d}{dt}v(t) &= -\gamma v(t) - kx(t) + \eta(t), \\ \frac{d}{dt}x(t) &= v(t), \end{aligned} \tag{9.17}$$

where m corresponds to the mass of our object, γ the dynamic friction constant, k the elastic coefficient and $\eta(t)$ the thermal forces due the fluid. As usual, $\langle \eta(t) \rangle = 0$ and $\langle \eta(t)\eta(s) \rangle = 2k_BT\gamma\ \delta(t-s)$. For simplicity, let us suppose that all parameters are given in terms of k_BT.

Before jumping straight into solving these equations numerically, we shall determine which kind of behavior we can expect. As Langevin equations are used to study systems embedded in a thermal bath, we can expect that total equilibrium energy should not depend on initial conditions, but solely on the thermal energy associated to the environment. Furthermore, there are 2 types of energy associated to this system, elastic potential given by $U(x) = kx^2/2$ and kinetic $K(v) = mv^2/2$. In such notice, we use the Boltzmann distribution

$$\rho(x, v|k, m, k_BT) \propto \exp\left\{-\frac{kx^2 + mv^2}{2k_BT}\right\} dx\ dv \tag{9.18}$$

to infer some basic statistics such as its most probable position and velocity. We can easily demonstrate that $\langle x \rangle = 0$, $\langle x^2 \rangle = k_B T/k$, $\langle v \rangle = 0$ and $\langle v^2 \rangle = k_B T/m$, which shows that the total mean energy in equilibrium is $\langle E \rangle = k_B T$.

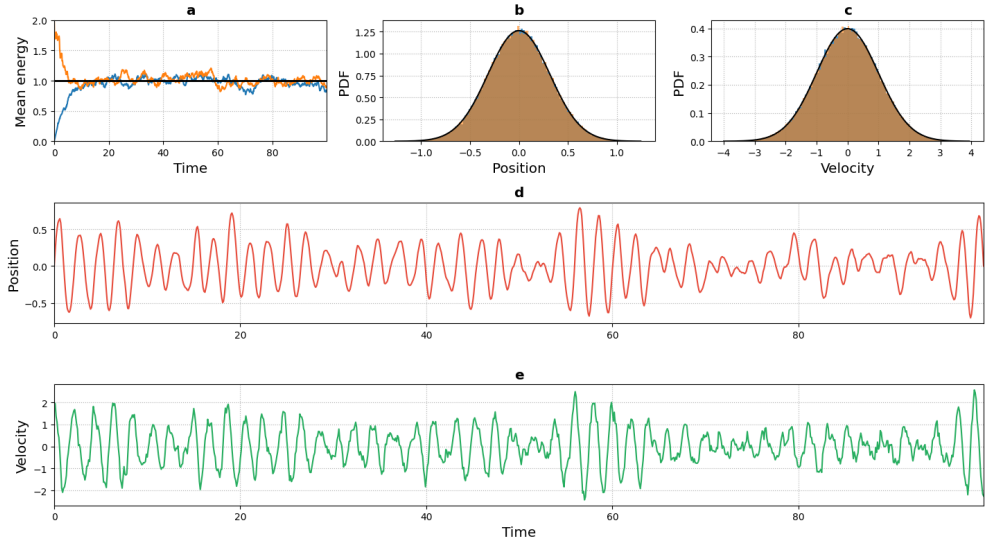

Fig. 9.2.: Numerical solution of a harmonic oscillator in a thermal bath. (a) The oscillator will absorb or release energy into the thermal bath, reaching equilibrium. (b-c) Upon relaxation, displacement and velocity are Boltzmann distributed. (d-e)) Example of position and velocity in time for a single oscillator starting without potential energy. Parameters used are $k_B T = 1$, $k = 10$, $\gamma = 0.25$ and $m = 1$.

To have quick simulations with large amount of repetitions to perform ensemble averages, I wrote a CUDA algorithm that runs 5120 of this oscillator in parallel for a period of 100 time units. The parameter values used were $k_B T = 1$, $k = 10$, $\gamma = 0.25$ and $m = 1$. Two sets of oscillators were simulated in total, both depart from equilibrium position, but the first starts off with no kinetic energy while the second begins with 2 $K_B T$. In figure (9.2a) we observe the evolution of system's total energy (potential + kinetic). As expected from Canonical ensemble, both initial conditions relaxed to total equilibrium energy proportional to bath temperature. In (9.2b-c) we plot the distribution of positions and velocities, respectively, for both sets of oscillator once in equilibrium, displaying that both situations are describe by the Boltzmann distribution in black. In (d-e) we display the evolution of position and velocity of a single realization departing with non-zero kinetic energy. Differently from the perfect sinusoidal curves expected from a traditional spring system, the amplitude of motion is modulated by the stochastic effects of the bath.

Before concluding this section, I would like to adventure into the non-equilibrium regime, that is the period of time in which the system relaxes to steady dynamics.

I'm mostly interested in the role of mass (m) and dynamic friction (γ) in the extension of such period. In the limit where $m \gg \gamma$, the fluid in which the body is emerged might present negligible effect if compared to other forces acting on this object, hence the relaxation time would be large. Oppositely, we have the situation where $m \ll \gamma$ which is usually the case for cytoplasm and nucleoplasm. Let's consider eGFP for a moment. It consists of a cylindrical-like molecule with cross-section $2.4 \times 4.2\ nm$ and molecular mass of $27\ kDa$ (about $10^{-23}\ kg$). It is used to tag proteins in the cellular content where the viscosity is about $2\ cPa$ or $10^{-9} \frac{kg}{\mu m\ s^2}$ [89]. In this situation, the relaxation time in the order of picoseconds or faster. Given the magnitude of this number and the usual time scale we work in microscopy, for instance, we can approximate equation (9.13) as

$$r_{n+1} = r_n + \frac{dt}{\gamma} f_n + \frac{1}{\gamma}\ \eta_{n+1}, \tag{9.19}$$

which relates to the more traditional Euler-Maruyama method [87]. This equation is at the heart of subsequent models we present in this thesis. For that end, we model all forces f acting on a section of polymer and add thermal noise η to simulate diffusion.

10

Rouse chain

The Rouse chain is the base for our chromatin model and should give us some basic understanding on what to expect from a polymer submerged in a homogeneous substrate without long range interactions. In the next few chapter we shall study more in depth the effects of long range interactions and a heterogeneous substrate. Without further ado, let us consider a segment of chromatin which is partitioned in N sections with a given number of base pairs in each. For simplicity, let us call these sections monomers $\boldsymbol{r}_i$ with size given by a probability density function (PDF) $\Psi(\boldsymbol{r_i})$ with mean zero and variance b^2, that is, individual segments have no preferred direction, but an average size proportional to b. With this assumption we have the partition function

$$\Omega(\boldsymbol{R}, N) = \mathrm{E}\left[\delta\left(\boldsymbol{R} - \sum_i \boldsymbol{r}_i\right)\right], \tag{10.1}$$

where $\boldsymbol{R}$ is the end-to-end vector. It is clear that the movement of proximal chromatin will be correlated, but we might imagine that for larger number of base pairs per monomer, the correlation between each end will decrease. Making use of the central limit theorem, we can show that adding up independent monomers will produce

$$\Omega(\boldsymbol{R}, N) = \left(\frac{3}{2\pi N B^2}\right)^{\frac{3}{2}} \exp\left\{-\frac{3R^2}{2Nb^2}\right\}, \tag{10.2}$$

where each degree of freedom collaborates with a variance $Nb^2/3$. This is called the ideal chain model or Gaussian chain model.

Now, let us estimate a PDF for each monomer. To do so, we shall convert equa-

tion (10.1) space into the Fourier space such that

$$\begin{aligned}\Omega(\boldsymbol{R},N) &= \frac{1}{(2\pi)^3}\int_{-\infty}^{\infty} d\boldsymbol{k}\ \mathrm{E}\left[\exp\left\{i\boldsymbol{k}\cdot\left(\boldsymbol{R}-\sum_{i=0}^{N}\boldsymbol{r}_i\right)\right\}\right] \\ &= \frac{1}{(2\pi)^3}\int_{-\infty}^{\infty} d\boldsymbol{k}\ e^{i\boldsymbol{k}\cdot\boldsymbol{R}}\ \mathrm{E}\left[\exp\left\{-i\sum_{i=0}^{N}\boldsymbol{k}\cdot\boldsymbol{r}_i\right\}\right],\end{aligned} \tag{10.3}$$

as each segment is independent of others, we can write

$$\Omega(\boldsymbol{R},N) = \frac{1}{(2\pi)^3}\int_{-\infty}^{\infty} d\boldsymbol{k}\ e^{i\boldsymbol{k}\cdot\boldsymbol{R}}\left[\int_0^{\infty} d\boldsymbol{r}\ e^{i\boldsymbol{k}\cdot\boldsymbol{r}}\Psi(\boldsymbol{r})\right]^N. \tag{10.4}$$

The PDF $\Psi(\boldsymbol{r})$ has average $\mathbf{0}$ and a well defined second moment, consequently its maximum is at $\boldsymbol{k}=\mathbf{0}$ and tends to zero for large $\boldsymbol{k}$. Associating this fact with a large number N of monomers, we have

$$\begin{aligned}\Omega(\boldsymbol{R},N) &= \frac{1}{(2\pi)^3}\int_{-\infty}^{\infty} d\boldsymbol{k}\ e^{i\boldsymbol{k}\cdot\boldsymbol{R}}\left[\int_0^{\infty} d\boldsymbol{r}\ \left(1-i\boldsymbol{k}\cdot\boldsymbol{r}-\frac{1}{2}|\boldsymbol{k}\cdot\boldsymbol{r}|^2+\cdots\right)\Psi(\boldsymbol{r})\right]^N \\ &\approx \frac{1}{(2\pi)^3}\int_{-\infty}^{\infty} d\boldsymbol{k}\ e^{i\boldsymbol{k}\cdot\boldsymbol{R}}\left[1-\frac{N}{2}\left\langle|\boldsymbol{k}\cdot\boldsymbol{r}|^2\right\rangle+\cdots\right] \\ &\approx \frac{1}{(2\pi)^3}\int_{-\infty}^{\infty} d\boldsymbol{k}\ e^{i\boldsymbol{k}\cdot\boldsymbol{R}}\ e^{-\frac{N}{6}k^2b^2}.\end{aligned} \tag{10.5}$$

Converting this result back into normal space we obtain equation (10.2). Hence, in order to satisfy this relation given that our monomer are independent from each other

$$\Omega(\boldsymbol{R},N) = \prod_i \Psi(\boldsymbol{r_i}), \tag{10.6}$$

we must have

$$\Psi(\boldsymbol{r_i}) = \left(\frac{3}{2\pi b^2}\right)^{\frac{3}{2}}\exp\left\{-\frac{3}{2b^2}r_i^2\right\}, \tag{10.7}$$

that is, each monomer is also described by a Gaussian distribution.

We are interested to determine the dynamics of this chain, hence we can use this probability density function to calculate the total entropy

$$S(R,N) = k_B\ln\Omega(\boldsymbol{R},N) = S_0 - \frac{3k_BR^2}{2Nb^2} \tag{10.8}$$

and, consequently, the free energy

$$dF = -S\ dT + p\ dV = S_0\ dT - \frac{3k_B R^2}{2Nb^2}\ dT \tag{10.9}$$

resulting in

$$F = F_0 + \frac{3k_B T}{2Nb^2}\ R^2. \tag{10.10}$$

Based on this result, we observe that this polymer can be approximated by an entropic spring with elastic constant

$$K_c = \frac{3k_B T}{Nb^2}. \tag{10.11}$$

Interestingly, this equation represents a set of N harmonic oscillators in series with elastic constant as follows

$$K_c = \left(\sum_{i=0}^{N} \frac{1}{k_i}\right)^{-1} = \left(\sum_{i=0}^{N} \frac{b^2}{3k_B T}\right)^{-1} = \left(\frac{Nb^2}{3k_B T}\right)^{-1} = \frac{3k_B T}{Nb^2}. \tag{10.12}$$

which is the initial hint for the Rouse chain model.

The Rouse chain uses these results described so far and adds dynamics into the theory. This extra component is described by hydrodynamic properties of the solvent in which the polymer is immersed (nuclear content, for chromatin). We shall consider a Brownian dynamic approach following equations

$$m\ \partial_t \boldsymbol{v}_0 = -\gamma \boldsymbol{v}_0 - \frac{3k_B T}{b^2}\ (\boldsymbol{r}_0 - \boldsymbol{r}_1) + \boldsymbol{\eta}_0 \tag{10.13}$$

$$m\ \partial_t \boldsymbol{v}_i = -\gamma \boldsymbol{v}_i - \frac{3k_B T}{b^2}\ (2\boldsymbol{r}_i - \boldsymbol{r}_{i+1} - \boldsymbol{r}_{i-1}) + \boldsymbol{\eta}_i \tag{10.14}$$

$$m\ \partial_t \boldsymbol{v}_N = -\gamma \boldsymbol{v}_N - \frac{3k_B T}{b^2}\ (\boldsymbol{r}_N - \boldsymbol{r}_{N-1}) + \boldsymbol{\eta}_N, \tag{10.15}$$

where $\boldsymbol{\eta_i}$ is a stochastic thermal force with $\mathrm{E}\left[\boldsymbol{\eta}_i\right] = 0$ and $\mathrm{E}\left[\boldsymbol{\eta_i} \cdot \boldsymbol{\eta_k}\right] = 6k_B T\gamma\ \delta_{ik}\ \delta(t_1 - t_2)$.

As the equations display, each particle reacts to forces due their interaction with first neighbors and the substrate. As first approximation, we consider that the context does not reverberate far from the location where interactions occur, which is a good approximation for the regime of polymer melt, i.e., in the case of crowded environments. The Zimm chain model [90], developed by Bruno Zimm in 1956, takes hydrodynamic effects into account. Nonetheless, this model is an over-complication for our current purposes.

Another approximation made by Rouse regards the limit of low inertia. As ex-

plained before, we assume that the effect of inertial mass is negligible in comparison to thermal noise and friction. Using this approximation we obtain the following set of equations

$$\partial_t \boldsymbol{r}_0 = -\frac{3k_BT}{\gamma b^2}(\boldsymbol{r}_0 - \boldsymbol{r}_1) + \frac{\boldsymbol{\eta}_0}{\gamma} \tag{10.16}$$

$$\partial_t \boldsymbol{r}_i = -\frac{3k_BT}{\gamma b^2}(2\boldsymbol{r}_i - \boldsymbol{r}_{i+1} - \boldsymbol{r}_{i-1}) + \frac{\boldsymbol{\eta}_i}{\gamma} \tag{10.17}$$

$$\partial_t \boldsymbol{r}_N = -\frac{3k_BT}{\gamma b^2}(\boldsymbol{r}_N - \boldsymbol{r}_{N-1}) + \frac{\boldsymbol{\eta}_N}{\gamma}. \tag{10.18}$$

To solve this model analytically, we can assume periodic boundary conditions. As we are not interested in the end monomers and if the chain is long enough to a point where the bulk is not affected, this assumption will help in the math without loss of information. Let us proceed with the Fourier relation

$$\boldsymbol{r}_i(t) = \frac{1}{N+1}\sum_{k=0}^{N} e^{-iq_k n}\ \boldsymbol{\rho}_k(t), \tag{10.19}$$

with relations

$$q_k = \frac{2\pi k}{N+1} \tag{10.20}$$

$$q_{n+N+1} = \frac{2\pi(n+N+1)}{N+1} = \frac{2\pi n}{N+1} + 2\pi = q_n. \tag{10.21}$$

Incidentally, transforming into the Fourier space can be written as

$$\boldsymbol{\rho}_{\boldsymbol{k}}(t) = \sum_{n=0}^{N} e^{iq_k n}\ \boldsymbol{r}_n(t). \tag{10.22}$$

Using the boundary conditions and the Fourier transform of equations (10.16), we have

$$\frac{d}{dt}\boldsymbol{\rho}_{\boldsymbol{k}}(t) = -\frac{6k_BT}{\gamma b^2}\left\{1 - \cos(q_k)\right\}\ \boldsymbol{\rho}_{\boldsymbol{k}}(t) + \frac{\boldsymbol{\xi}_{\boldsymbol{k}}(t)}{\gamma}. \tag{10.23}$$

Notice that we were able to decouple the dynamics of all monomers, making the solution of this model much simpler. Assuming first a homogeneous solution for the differential equation, further correcting for the extra term and transforming back into normal space, we determine that

$$\boldsymbol{r}_{\boldsymbol{n}}(t) = \frac{1}{N+1}\sum_{k=0}^{N} e^{-iq_k n}\ \int_0^t ds\ \frac{\boldsymbol{\xi}_{\boldsymbol{k}}(s)}{\gamma}\ \exp\left\{-\frac{6k_BT}{\gamma b^2}\left[1 - \cos(q_k)\right](t-s)\right\}. \tag{10.24}$$

.

As a possible first step to estimate the shape of this equation, we might determine the solution for $k = 0$, that is, for long wavelengths,

$$\boldsymbol{r_n}^0(t) = \frac{1}{N+1}\int_0^t ds\ \frac{\boldsymbol{\xi_k}(s)}{\gamma} \tag{10.25}$$

and calculate the mean squared displacement as follows

$$\left\langle |\boldsymbol{r_n}^0(t)|^2 \right\rangle = \frac{1}{(N+1)^2\gamma^2}\int_0^t ds\ \int_0^t da\ \left\langle \boldsymbol{\xi_k}(a)\cdot\boldsymbol{\xi_k}(s)\right\rangle. \tag{10.26}$$

Assuming the transform

$$\left\langle \boldsymbol{\xi_0}(t)\cdot\boldsymbol{\xi_0}(s)\right\rangle = 6k_BT\gamma(N+1)\ \delta(t-s), \tag{10.27}$$

we determine that

$$\left\langle |\boldsymbol{r_n}^0(t)|^2 \right\rangle = \frac{6k_BT}{(N+1)\gamma}t = 6D_Gt, \tag{10.28}$$

which is a interesting result. Being the longest wavelength in the system, we expected that it acts upon all the monomers with same strength, therefore we can interpret this result as the mean squared displacement of the polymer as a whole. Expectantly, this result should be different from dilute solutions. Zimm demonstrated [90] $\left\langle |\boldsymbol{r_n}^0(t)|^2 \right\rangle$ is actually proportional to $\frac{1}{(N+1)^\nu}$, which tends to fit better experimental results in dilute solutions.

Now, let us determine the mean squared displacement for each monomer. We shall multiply the result in equation (10.24) by itself and solve to

$$\begin{aligned}\left\langle |\boldsymbol{r_n}(t)|^2 \right\rangle &= \frac{6k_BT}{(N+1)\gamma}\sum_{k=0}^{N}\int_0^t ds\ \exp\left\{-\frac{12k_bT}{\gamma b^2}\left[1-\cos(q_k)(t-s)\right]\right\} \\ &= \frac{b^2}{2(N+1)}\sum_{k=0}^{N}\frac{1-\exp\left\{-\frac{12k_bT}{\gamma b^2}\left[1-\cos(q_k)\right]\right\}}{1-\cos(q_k)}.\end{aligned} \tag{10.29}$$

Solving this summation can be difficult, demanding us for some sort of trick. In this situation, we can suppose our polymer chain is very long and transform this sum into the integral

$$\left\langle |\boldsymbol{r_n}(t)|^2 \right\rangle = \frac{b^2}{2\pi}\int_0^\infty dq\ \frac{1-\exp\left\{-\frac{12k_bT}{\gamma b^2}\left[1-\cos(q)\right]t\right\}}{1-\cos(q)}. \tag{10.30}$$

We can further approximate this solution in order to simplify the math using a low

pass filter, that is, approximate for shorter frequencies, yielding

$$\left\langle |\boldsymbol{r_n}(t)|^2 \right\rangle = \frac{b^2}{\pi} \int_0^\infty \frac{dq}{q^2} \left(1 - \exp\left\{ -\frac{6k_bT}{\gamma b^2} q\, t \right\} \right). \tag{10.31}$$

This integral can be easily solved by deriving and integrating the time variable as a trick to remove q^2 from denominator. The final solution for our problem is

$$\left\langle |\boldsymbol{r_n}(t)|^2 \right\rangle = \sqrt{\frac{12k_BTb^2}{\gamma\pi}} t \tag{10.32}$$

$$= 6\sqrt{\frac{k_BTb^2}{3\gamma\pi}}\, t^{1/2} \tag{10.33}$$

$$= D_{loc}\, t^{1/2} \tag{10.34}$$

Differently from traditional diffusion, we observe that the anomalous coefficient is smaller than one, that is $\alpha = 1/2$. In fact, this result is expected due the spring-like interaction with first neighbors. In the next section we shall solve this system numerically to verify all results are correct.

10.1 Numerical solution

Many approximations were made to obtain solution (10.32). In order to validate our results, let us solve this model numerically and compare the analytical results obtained. For that end, I wrote a CUDA algorithm for a Rouse chain with 512 monomers using $b = 50$ nm. In total we run 4096 chains starting off random conformations as given by a Gaussian chain. The Langevin method was used to simulate these chains for 60 seconds and positions were measured every half a second (polymer time). We further use $\frac{k_BT}{\gamma} = 0.00377 \mu m^2/s$, expecting to measure an apparent diffusion coefficient $D_\alpha = 0.001 \mu m^2/s^\alpha$, approximate value we measured with the TetO system. In order to estimate D_α and α with ranges of credibility, we split out 4096 chains into 16 groups and we use ensemble average mean squared displacement (EA-MSD) in each group.

Our main objective thus far is to verify that the MSD curve is well described by the analytical results, hence all monomers (excluding some towards both ends) should present identical diffusion and anomalous coefficients. Moreover, we also want to verify that the center of mass is properly described by a slow Brownian motion. Finally, we want to verify that the movement of nearby neighbors is correlated due the presence of springs. I am also interested to know how far away this

correlation is relevant.

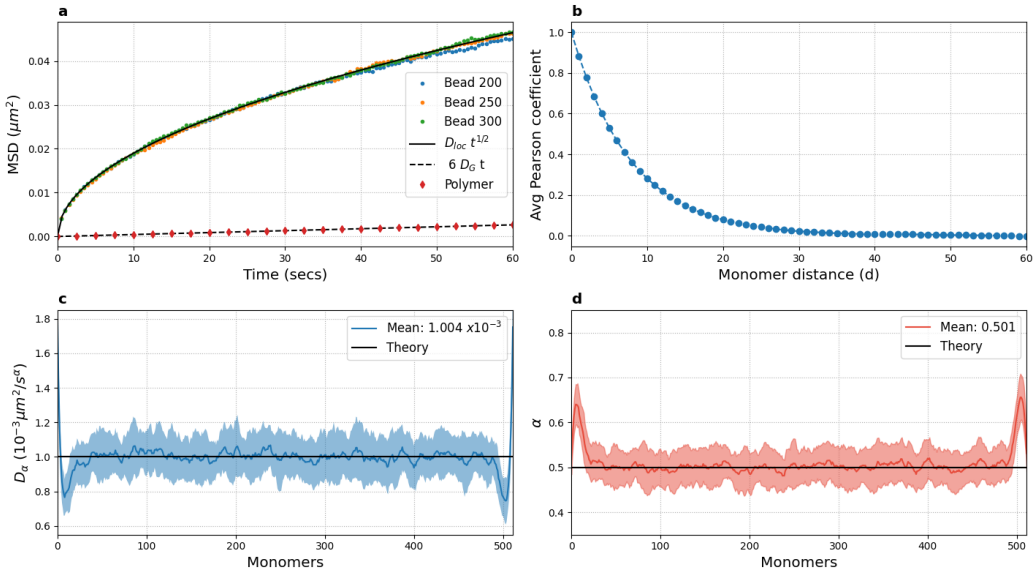

Fig. 10.1.: Results of 4096 simulated polymers with b=50 nm and $\frac{k_B T}{\gamma} = 0.00377 \mu m^2/s$. Even with all the approximations done, the analytical results match quite nicely the numerical ones with exception of boundary monomers. In (a) we compared the MSD curves of central monomers and the center of mass. In (b) we present movement correlation along nearby monomers. In (c-d), we compare analytical apparent diffusion and anomalous coefficients with the ones obtained numerical via ensemble average MSD fitting. Shades are one standard deviation.

In figure (10.1), we verify that our analytical results are correct. In (a), we have the MSD curve for single monomers being well described by the analytical result. The red diamonds were calculated over the center of mass for all simulated polymers. In (b), we observe how far away the dynamics of a monomer is propagated into its neighbors via movement correlation. This result is important so we know how big any simulated polymer should be so we can neglect boundary effects. Finally, we have the average D_α and α calculated over the MSD curve described by individual monomers at the bottom with 95% credible interval. Theoretical values are recapitulated with good accuracy. As expected, the boundary values are different due to lack of symmetry.

10.1.1 EA-MSD and GP-FBM

For figure 10.1, we calculate apparent diffusion and anomalous coefficients using samples of EA-MSD. I would like to compare those results with measurements

performed using the GP-FBM method presented in chapter 6. As the amount of time needed for optimization is comparable to a second per bead, we will analyze only 128 polymers. In figure 10.2 we present results. As we can see, the values for D_α and α are, undeniably constant at the bulk, but optimal values obtained using GP-FBM are biased by 2 to 3 percent.

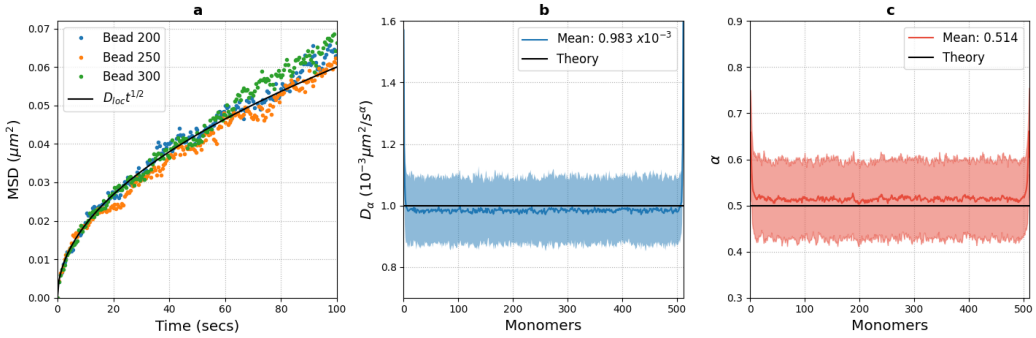

Fig. 10.2.: Results obtained using only 128 polymers via GP-FBM method. Even though MSD curves are very noisy in (a), the distribution of apparent diffusion and anomalous coefficients are fairly constant in (b-c).

10.2 Contact maps and distance measurements

Using techniques such as DNA Fish, we can determine positions and distances of well determined sequences of chromatin in fixed cells. More recently, a new super resolution, single cell tracing method based on DNA Fish [91] was developed to generate distance maps with resolution reaching the low mega bases, precise enough to observe TAD like structures in single cells. Nonetheless, as we shall see in figure (10.5), even a Gaussian chain presents such structures when analyzed in a single polymer fashion. Perhaps a more interesting approach would be to consider the average behavior over an ensemble of cells. Gladly, this technology exists and is called Chromosome Conformation Capture. These methods use a large number of cells to determine an average chromatin conformation on the form of a contact map, that is, each term in this map will tell us about the probability of finding any two regions in proximity. Of course, due to experimental unknowns, we cannot determine quite precisely what "proximity" means other than speculate on the reach of cross-linking upon cell fixation. By no means, a perfect and complete technique, but very informative. For more practical purposes, we are going to assume that contact is defined as monomers sharing a sufficiently small region in space so that, upon stochastic dynamics, different monomers present higher chance to be cross-liked. In this section we aim to determine, or estimate, the relationship between

average 3D distances and contact probability.

At the beginning of this chapter, we determined that the end-to-end distance probability distribution is given by

$$\Omega(\boldsymbol{R}, N) = \left(\frac{3}{2\pi N b^2}\right)^{\frac{3}{2}} \exp\left\{-\frac{3R^2}{2Nb^2}\right\}. \tag{10.35}$$

To calculate the average distance between any two monomers separated by N monomers, we can simply do

$$\begin{aligned}\langle R(N)\rangle &= \int_0^{2\pi} d\phi \int_0^{\pi} d\theta \int_0^{\infty} dR\, R^3 \sin(\theta)\Omega(\boldsymbol{R}, N) \\ &= \sqrt{\frac{8b^2}{3\pi}N}.\end{aligned} \tag{10.36}$$

Calculating the contact probability between these monomers is a little more subtle, but just as easy

$$\begin{aligned}P(R \leq b|N) &= \int_0^{2\pi} d\phi \int_0^{\pi} d\theta \int_0^{b} dR\, R^2 \sin(\theta)\Omega(\boldsymbol{R}, N) \\ &= \text{erf}\left(\sqrt{\frac{3}{2N}}\right) - \sqrt{\frac{6}{N\pi}}\exp\left\{-\frac{3}{2N}\right\},\end{aligned} \tag{10.37}$$

where we consider that two monomer are in contact if their center of mass are in distance inferior to their average diameter, that is, b. Using literature nomenclature, N would be called the linear distance between chromatin regions, while b is commonly called Kuhn length. The next logical step would be to write $P(R \leq b|N)$ as a function of the average 3D distance, hence

$$P(R \leq b|N) = \text{erf}\left(\sqrt{\frac{4b^2}{\pi \langle R\rangle^2}}\right) - \frac{4b}{\pi \langle R\rangle}\exp\left\{-\frac{4b^2}{\pi \langle R\rangle^2}\right\}. \tag{10.38}$$

Admittedly, this equation is more complicated than the ones we observe in literature [41]. We can approximate to a simpler shape in the limit where $N \gg 1$. We have

$$P(R \leq b) \approx \frac{16b^3}{\pi^2} \langle R\rangle^{-3} \tag{10.39}$$

which is frequently found in literature as proposed for mammalian cells. Further more, we usually find in Bayesian approaches to chromatin reconstruction the exponent of $\langle R\rangle$ to be -2.5. We have a comparison between both coefficients in figure (10.3).

To test if our results represent correctly our simulations, in figure 10.4 we have in

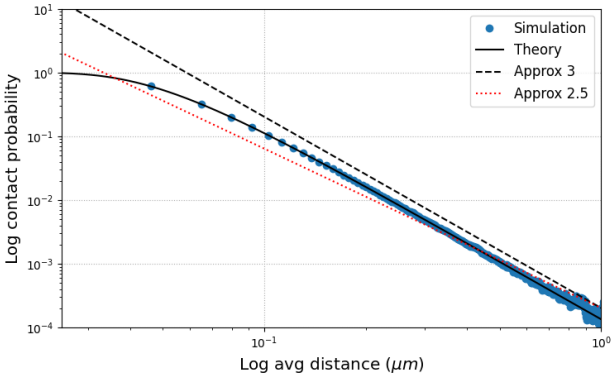

Fig. 10.3.: Comparison between our results and common approximations found in literature. The blue dots were obtained from those 8192 Rouse chains simulated in the previous section.

blue the average values calculated from the simulation of 4096 polymers presented in the previous sections. As we can see our analytical results fit the Rouse chain simulations perfectly.

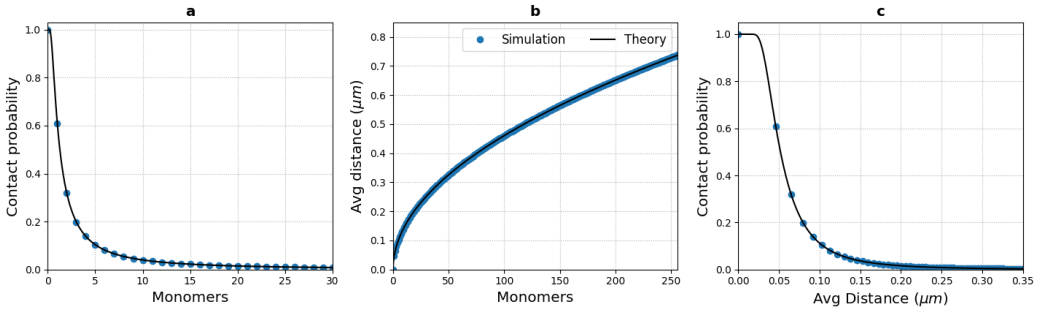

Fig. 10.4.: We have a comparison between contact probabilities and average distances as obtained by simulations and the curves we just calculated.

Finally, I would like to address the differences between an average contact and distance maps, that is, calculated over an ensemble of many polymers (or many cells in an experimental situation), to a single polymer map. In figure (10.5), we have such comparison.

Notice that even though the average results in (a-b) present a monotonic exponential decay with increasing linear distances, the single polymer distance maps in (c) present certain structures that could be interpreted as TAD-like formations in a misleading fashion. These structures are dynamic and expected from the stochastic type of motion performed by monomers.

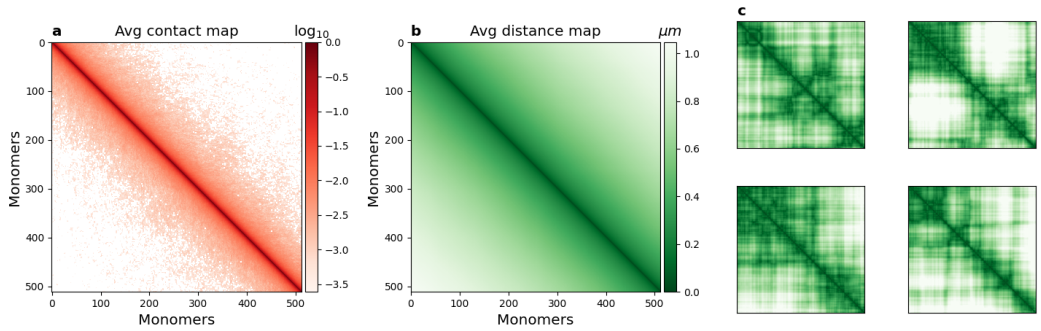

Fig. 10.5.: (a-b) Average contact and distance maps calculated over an ensemble of 8192 polymers. No interesting structures are found. (c) Distance maps calculated at the single polymer level. TAD-like structure are dynamics and not to be taken as causal or functional importance.

10.3 Comparing to real data

As expected, real life chromatin behaves differently from a Gaussian chain. In fact, this result is already known for decades. In [92, 93] is said that linear chromatin scales with 3D space following $\langle R \rangle \propto N^{1/3}$ due to its fractal properties. Moreover, a single Hi-C map element represents thousands of base pairs, which could correspond to several of the monomers we developed for the Rouse chain model. Hence, I would like to use experimental data to confirm the value of this fractal exponent to be $1/3$ and determine how many monomers are enclosed per HI-C map element. Evidently, these results shall be interpreted with care, as we will appreciate averages for a whole chromosome.

Before we start, it might be worth remembering that we currently have $1/2$ as the exponent for a polymer in random conformation (equation 10.36), hence we should adapt the Gaussian chain distribution accordingly and recalculate the relationships between linear distance (N), average 3D distances ($\langle R \rangle$) and contact probability.

Let us assume the relationship $N = L^{\beta}$ with $L = \epsilon n$, where n corresponds to single monomers. We expect that ϵ will count the number of monomers in a Hi-C map element, while β corrects the fractional exponent. Consequently, we re-write equation 10.2 as

$$\Omega(\boldsymbol{R}|\beta, L) = \left(\frac{3}{2\pi L^{\beta} b^2}\right)^{3/2} \exp\left\{-\frac{3R^2}{2L^{\beta} b^2}\right\}, \tag{10.40}$$

notice that $\epsilon = \beta = 1$ for our original Rouse chain results. The average distance

between n monomers can be estimated as usual

$$\begin{aligned}\langle R\rangle &= \int_0^{2\pi} d\phi \int_0^{\pi} d\theta \int_0^{\infty} dR\, R^3 \sin(\theta)\, \Omega(\boldsymbol{R}|\beta, L) \\ &= \sqrt{\frac{8b^2}{3\pi}} L^{\beta/2}, \end{aligned} \tag{10.41}$$

so we might expect that $\beta = 2/3$. To calculate the function connecting contact probability and the re-scaled linear distance L we do

$$\begin{aligned} P(R \leq b|\beta, L) &= \int_0^{2\pi} d\phi \int_0^{\pi} d\theta \int_0^{b} dR\, R^2 \sin(\theta)\, \Omega(\boldsymbol{R}|\beta, L) \\ &= \text{erf}\left\{\sqrt{\frac{3}{2L^\beta}}\right\} - \sqrt{\frac{6}{\pi L^\beta}}\, e^{-\frac{3}{2L^\beta}}. \end{aligned} \tag{10.42}$$

This result seems to be a bit more complex than the traditional power law results found in literature. In [34] is suggested that $P(L) \propto L^{-1}$ and present some experimental data supporting this result. Nonetheless, this coefficient should be expected for great linear distances. Let us Taylor expand our result to $L \gg 1$

$$\begin{aligned} P(R \leq b|\beta, L \gg 1) &\approx \frac{2}{\sqrt{\pi}} \sqrt{\frac{3}{2L^\beta}} - \sqrt{\frac{6}{\pi L^\beta}} \left(1 - \frac{3}{2L^\beta}\right) \\ &= \sqrt{\frac{27}{2\pi}} L^{-3\beta/2}. \end{aligned} \tag{10.43}$$

As we determined before, $\beta = 2/3$ should recover results presented in [34]. Finally, we can recalculate the function mapping contact probability to average 3D distance, which is identical to equation (10.38).

Inferring genome-to-model scale ϵ and fractional exponent β

For the purpose of this inference, we use 5kb resolution Hi-C map obtained for the chromosome 20 of IMR90 human fibroblast cells downloaded from the paper [94]. Associated with this Hi-C map, we also had access to data corresponding to genome wide expected number of reads per linear genomic distance.

Let us assume the normalized maps present Gaussian distributed errors and assign the probability for our measured values $\boldsymbol{\chi}$ given relation (10.42) is

$$P(\boldsymbol{\chi}|\epsilon, \beta, \sigma) = \prod_{n=1}^{N} \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left\{-\frac{[\chi_n - P(R \leq b|\beta, L_{\epsilon,n})]^2}{2\sigma^2}\right\}. \tag{10.44}$$

Using the Bayesian approach discussed in chapter 5, we can rewrite this equation as the probability of having ϵ, β and σ given our data assigned

$$P(\epsilon, \beta, \sigma|\boldsymbol{\chi})P(\boldsymbol{\chi}) = P(\boldsymbol{\chi}|\epsilon, \beta, \sigma)P(\epsilon, \beta, \sigma). \tag{10.45}$$

Our task is simple, to find the values of these parameters that maximize this function. In practice, however, the effects of σ is not the most relevant for our goal, so we will simply integrate this parameter out assuming $P(\sigma) = 1/\sigma$. We have

$$P(\epsilon, \beta|\boldsymbol{\chi}) \propto \left\{ \sum_{n=1}^{N} \left[\chi_n - P(n, \epsilon, \beta)\right]^2 \right\}^{-N/2} P(\epsilon, \beta). \tag{10.46}$$

To maximize this probability we shall use the Nelder-Mead optimization algorithm [2]. To map these parameters into boundless space, we used the function $y = e^x$. The final equation used will be converted into log-space for numerical stability, yielding

$$\ln P(\epsilon, \beta|\boldsymbol{\chi}) = -\frac{N}{2} \ln \left\{ \sum_{n=1}^{N} \left[\chi_n - P(n, \epsilon, \beta)\right]^2 \right\} + \ln \epsilon + \ln \beta, \tag{10.47}$$

where the extra terms were added to account for a flat prior given mappings done.

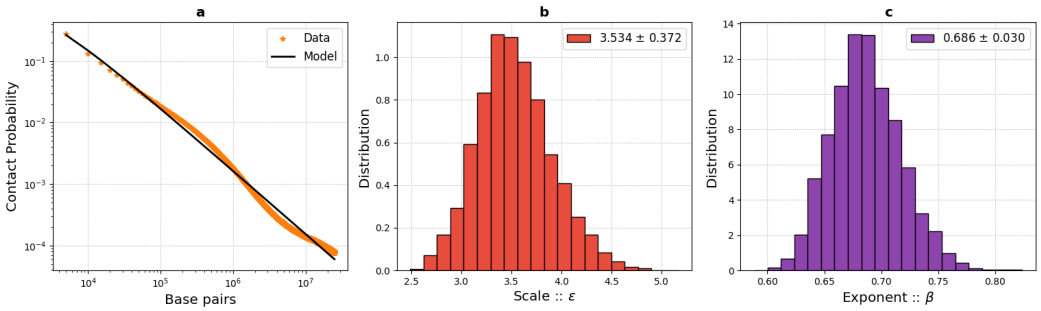

Fig. 10.6.: Inferring chromatin-to-model genomic scale ϵ and fractional exponent β using expected reads for 1-dimensional genomic distance present in [94]. (a) Fortunately, our model offers a good first approximation for expected contact probability. (b-c) Posterior distributions for parameters.

In figure (10.6) we have the results inferred for parameters ϵ and β. As already expected, we find that β is approximately $2/3$, corroborating the hypothesis chromatin presents fractional properties due long range interactions. On second instance, we also conclude that our model presents a good first approximation for the expected contact probability. Interestingly, most of the differences are found for linear distances in which average chromatin structure differs from the model, that is, where we normally observe the presence of topological associated domains

(TADs). Finally, we conclude that each element of this 5kb resolution map should contain about 3.5 of our monomers or, in other words, each monomer contains about 1.5 kb.

Inferring Kuhn length

The final parameter we are interested to determine is an average Kuhn length. Once again, we are going to use Hi-C maps for the chromosome 20 of IMR90 human fibroblast cells downloaded from the paper [94]. Differently, we also need experimental data containing relative distances for this chromosome. We shall borrow the results presented in [95], where a smart 3D DNA Fish protocol was used to map the position of specific chromatin regions and measure relative distances among them using 3D microscopy. The data we are interested the most is presented in figure (10.7).

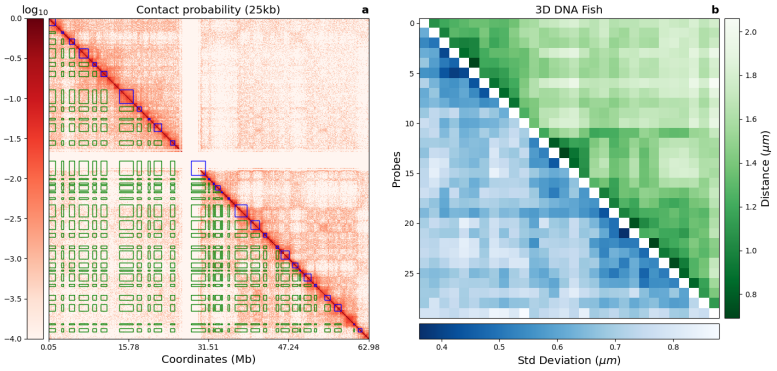

Fig. 10.7.: Summary of results in [95]. Blue rectangles in (a) mark different probed locations, while green rectangles represent contact probabilities between regions. Using the average contact probability among probed locations and actual 3D distance measured via 3D microscopy in (b), we can infer Kuhn length values for chromatin using different Hi-C map resolutions.

With this data and equation (10.38) as a model, we can fit the Kuhn length for normalized Hi-C maps with different resolution and, with that, clarify how this value changes with chromatin scale. Hence, we can use the average probability calculate for green rectangles in figure (10.7a) and find a b such that the respective distances in figure (10.7b) are optimized. We shall use a similar method to previous subsection, that is, maximizing a likelihood given by equation 10.44. Fortunately, we have measured standard error this time and they are shown in blue for relative positions in figure (10.7a).

The final results are disclosed in figure 10.8. As we can observe, depending on

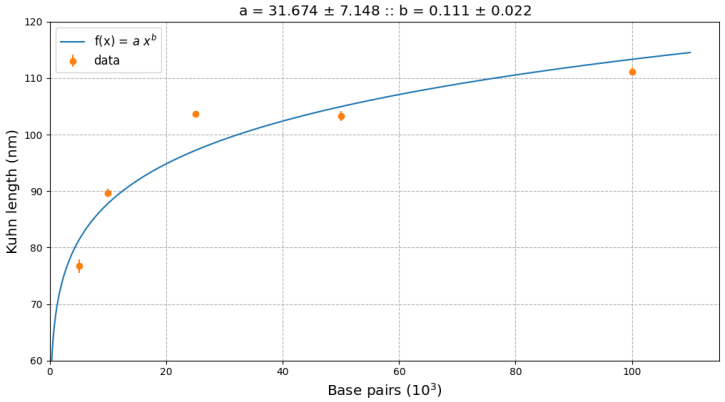

Fig. 10.8.: Different chromatin scales present difference values for Kuhn length. By extrapolation, we find a 95% expected Kuhn length range for one of our monomer (1.5 kb) to be 71.189 ± 38.564 nm.

the scale analyzed, different values for Kuhn length were inferred. Intuitively, we expect this happens due the long range interactions. In the next chapters we will discuss further about this subject. Nonetheless, we can approximate these results with a curve $f(x) = ax^b$ to determine an approximate diameter for our monomers. Upon extrapolation towards 1.5 kb, we find $b = 71.189 \pm 38.564$ nanometers as a 95% credible range. We shall consider these results during our simulations later on.

11

Reconstructing chromatin conformation

The biggest purpose for all the results deduced in the last chapter is to attempt reconstructing polymer conformation using contact maps or, as more traditionally known in biology, Hi-C maps. At his point, I am not so interested about the dynamic mechanisms responsible for the existence of any predetermined conformation, but simply recapitulating the position of all monomers in space as measured by the contact map or, concurrently, a distance map. Thus, by knowing how far away all the monomers are from each other, we can determine by cross-validation when we have the correct shape. For that purpose, we are going to take as ground truth the conformation obtained for a sampled Gaussian chain to test our efforts.

Given the known physics involved in our toy model, one possible approach would be to simply simulate a dynamic polymer for a great period of time until its distance map matches the input one. Statistically speaking via ergodicity arguments, it is guaranteed that, if we wait long enough, our toy model will eventually reach a conformation that resembles the population average for a short period of time. This is most definitely a very inefficient approach. To redeem our expected map, we may take a different approach.

There are some established methods in literature to reconstruct 3D conformation based on contact maps. A great review on models can be found in [41], but I would like to recall some in here. One of the most traditional models is associated with the minimization of a error function $\sum_{ij}(d_{ij} - \delta_{ij})^2$, where d_{ij} corresponds to a set of attempt positions and δ_{ij} are estimated somehow from Hi-C maps and/or measured with use of DNA FISH experiments. Some examples can be found in [96, 97]. In a similar fashion, some studies conclude that this weight function is biased towards mostly longer range distances, therefore they should be corrected by re-scaling it

as $\sum_{ij}(1 - d_{ij}/\delta_{ij})^2$.

A more statistical approach is presented in [98], where they assume a conversion between contact map and distances of the form $c_{ij} = \beta d_{ij}^{\alpha}$ [1]. In the previous chapter, we determined that this is a valid approximation for polymers well described by a Gaussian chain model [99]. Furthermore, they suggest that the Hi-C map presents Poisson noise [2]. Hence, their approach consists of optimizing positions, α and β that maximize the likelihood

$$\mathcal{L}(d_{ij}, \alpha, \beta) = \prod_{ij} \frac{(\beta\ d_{ij}^{\alpha})^{c_{ij}}}{c_{ij}!} \exp\left\{-\beta\ d_{ij}^{\alpha}\right\}. \tag{11.1}$$

This approach is particularly interesting due the attempt to remedy the huge amount of noise associated with this type of experiment. Another Bayesian approach to chromatin reconstruction from experimental data can be found in [100], but now offering more physical argumentation on the nature of polymers.

Finally, I would like to mention the work developed in [101]. To reconstruct chromatin conformation, they apply a set of spring-like forces to model short range interactions between neighboring monomers of a chain as well as long range interactions based on Hi-C map and volume exclusion assuming soft monomers. Another valid point of their paper is related to reconstruction based on single cell data. In general, hundreds of thousands of cells are used to generate a single Hi-C map, therefore its experimental results should be interpreted as a population average. We shall discuss further into the matter when we reach models to understand the dynamic behavior of chromatin. For now, let's focus on reconstructing conformation from the contact/distance map of a single toy polymer generated using Gaussian chain model, hence there is no need to address noise just yet.

At first instance, let's assume that all monomers should generate some force over all the others and that this force should be modulated by the distance map. As traditional for such simulations, our forces are described by the Lennard-Jones potential

$$U_{ij} = \epsilon \left(\frac{d_i j}{r_{ij}}\right)^n \left[\left(\frac{d_i j}{r_{ij}}\right)^n - 2\right], \tag{11.2}$$

where $r_{ij}^2 = (x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2$ is the distance between monomers i and j, d_{ij} is determined by the distance map generated for our toy polymer, ϵ controls the strength of all those forces while n its reach.

Knowing the energy potential a monomer applies on another, we can calculate

1. This α is not to be confused with anomalous coefficient
2. In that sense, a Hi-C map measures the number of times chromatin sections where found in contact given a large number of cells

the force associated using $\boldsymbol{F_{ij}} = -\boldsymbol{\nabla} U_{ij}$. Converting to a spherical coordinates system, we have that

$$\boldsymbol{F}_{ij} = \frac{2n\epsilon}{r_{ij^2}} \left(\frac{d_i j}{r_{ij}}\right)^n \left[\left(\frac{d_i j}{r_{ij}}\right)^n - 1\right] \boldsymbol{r}_{ij}, \tag{11.3}$$

and summing up the total force a monomer feels given the interaction with all the others is represented as

$$\boldsymbol{F}_{ij} = 2n\epsilon \sum_{i \neq j} \frac{\boldsymbol{r}_{ij}}{r_{ij^2}} \left(\frac{d_i j}{r_{ij}}\right)^n \left[\left(\frac{d_i j}{r_{ij}}\right)^n - 1\right], \tag{11.4}$$

where we discard self-interaction.

Now we know all the forces associated to our expected conformation, we could initialize all monomers to a Gaussian chain and let these forces handle the dynamics from there. However, one key element is missing: the annealing factor. This term should remove energy from monomers allowing them to relax towards correspondent equilibrium positions. Let us then work with the following set of equations

$$m\frac{d}{dt}\boldsymbol{v}_i = -\xi \boldsymbol{v}_i + 2n\epsilon \sum_{i \neq j} \frac{\boldsymbol{r}_{ij}}{r_{ij^2}} \left(\frac{d_i j}{r_{ij}}\right)^n \left[\left(\frac{d_i j}{r_{ij}}\right)^n - 1\right], \tag{11.5}$$

$$\frac{d}{dt}\boldsymbol{r}_i = \boldsymbol{v}_i, \tag{11.6}$$

where m is the mass of each monomer and ξ is a dynamics friction coefficient.

Let's pause our deductions for awhile and, before we continue, ponder about all those parameters:

Mass: We do not have an experimental value for this parameter, but we can make a rough estimation based on the resolution of H-C maps we will usually use (4 kb). Part of this mass comes from the approximate 8000 nucleotides (forward and reverse strands) with 325 Dalton each. We also have all histones necessary for their organization into nucleosomes. With average of 200 base pairs (core and linker), we get 20 nucleosomes per 4 kb, each with 2 H2A (2×12843 Dalton), 2 H2B (2×13936 Dalton), 2 H3 (2×15388 Dalton) and 2 H4 (2×11367 Dalton) histones. All these values were taken from *bio-protocols* and *uniprot.org* for mouse cells. Adding all those numbers together we have 4,741,360 Dalton. For better handling later, we shall convert this value to kilograms, thence $m = 7.87 \times 10^{-21}$ kg, without considering interaction energies [3].

Lennard-Jones strength (ϵ): This parameter should be fit somehow. One way

3. As effective interaction energy is negative, this mass should be actually a little smaller.

of dealing with it is to associate its value to the spring constant from the Rouse model. We can do that due to the fact that our potential energy behaves like a harmonic oscillator for small displacements

$$\begin{aligned} U(r_{ij}) &= U(d_{ij}) + \frac{d}{dt}U(d_{ij})\ (r_{ij} - d_{ij}) + \frac{1}{2}\frac{d^2}{dt^2}U(d_{ij})\ (r_{ij} - d_{ij})^2 + \cdots, \\ \frac{d}{dt}U(d_{ij}) &= 0, \\ \frac{d^2}{dt^2}U(d_{ij}) &= \frac{2\epsilon n^2}{d_{ij}^2}. \end{aligned} \tag{11.7}$$

Therefore, when d_{ij} is the distance between adjacent monomers, we could have

$$\frac{2\epsilon n^2}{d_{i\ i\pm1}^2} = \frac{3k_BT}{b^2}, \tag{11.8}$$

resulting in

$$\epsilon = \frac{3k_BT}{2n^2}. \tag{11.9}$$

Before calling it done, we might do another approximation. At the current state we are mostly interested to determine a equilibrium conformation for our polymer, thus we are not aiming to identity precise values for interaction forces when distances are much akin to equilibrium. In fact, even when dealing with dynamics and stochastic displacements, our simulations will occur around the equilibrium conformation, forthwith we shall approximate our Lennard-Jones forces 11.4 to near equilibrium forces, resulting in

$$\boldsymbol{F}(r_{ij}) = \frac{3k_BT}{d_{ij}^2}(r_{ij} - d_{ij})\ \hat{r}_{ij}, \tag{11.10}$$

with $\hat{r}_{ij}$ representing an unitary directional vector from i to j. As expected, the near equilibrium force is calculated to be a harmonic oscillator force with elastic constant depending on the equilibrium distance between monomers.

Friction: The friction parameter is somehow arbitrary as we do not know for sure the hydrodynamic effects of solvent (nucleoplasm) over our monomer abstraction. In fact, if we are simply interested on the equilibrium conformation, but not the effective dynamic behavior of the polymer, we can choose it to be weak enough so the polymer can explore many conformations before relaxing to equilibrium, but high enough so the relaxation algorithm runs quickly.

In reason of the large difference between the order of magnitude found for friction and other forces associated with this system, we find that the effects of inertia

(related to mass) can be neglected due to that monomers would react to interaction forces almost instantaneously. For that reason we set

$$m\frac{d}{dt}\boldsymbol{v}_i \approx \mathbf{0} \tag{11.11}$$

and re-write the differential equations (11.6) as

$$\frac{d}{dt}\boldsymbol{r}_i = \frac{3k_BT}{\xi}\sum_{i\neq j}\frac{r_{ij}-d_{ij}}{d_{ij}^2}\,\hat{r}_{ij}. \tag{11.12}$$

To verify if equation (11.12) works appropriately, let's sample a single Gaussian chain and try to reconstruct it based on its map of distances. In figure (11.1a), we have a distance map calculated for this sampled polymer. To further determine how robust this method is regarding noise, we introduce lognormal noise with parameter[4] σ for each term of the distance map. An example with $\sigma = 0.4$ is displayed in (b). For different noise levels, we reconstruct 32 polymers to estimate average error and standard deviation. In (c), we present a reconstructed distance map for $\sigma = 0.4$. Finally, we show in (d) the average of averages with fitted curve. An expected error of about 5 nanometers per monomer is found for $\sigma \to 0$.

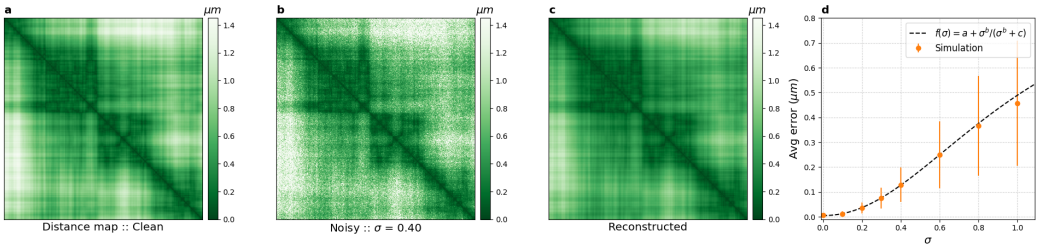

Fig. 11.1.: Testing polymer reconstruction using equation 11.12. (a) Distance map calculated for sampled Gaussian chain (b) Using a lognormal distribution with $\sigma = 0.4$, we introduce noise into clean distance. (c) Example of reconstruction using previous noisy map. (d) Averages of average error and standard deviation calculated using 32 reconstructed polymer for diverse σ in orange. Fitted curve presents $a = 0.005$, $b = 2.083$ and $c = 1.067$.

4. View equation (4.17) for reference.

11.1 Reconstructing conformation from population maps

So far we have studied a method to reconstruct conformations based on a distance map of a single polymer. What if we do not have such map, but only a contact map? As a first step, let us first determine how a single polymer contact map looks like. For that purpose, we shall sample another Gaussian chain of length 512 monomers with b = 50 nm. Then, we can calculate the contact map by determining which monomers touch or intersect others.

In figure (11.2a-b), we have contact maps for 2 sampled chains. Due to the method with which we generate contact maps, we know that all the blue elements of those maps are within 50 nm from each other. Oppositely, these maps provide no further information concerning non-touching monomers other than there are not intersecting. What we expect, however, is that if we sample many polymers and add up the overall contacts, we shall re-obtain the average behavior predicted by theory in chapter 10. To verify this situation, we sample 256 polymers and present the results in (c). For a closer inspection, in (d) we compare the actual theoretical curve (10.37) to a middle row of the matrix presented in (c). By symmetry arguments, we can conclude that all monomers toward the center of this chain should present similar theoretical curves.

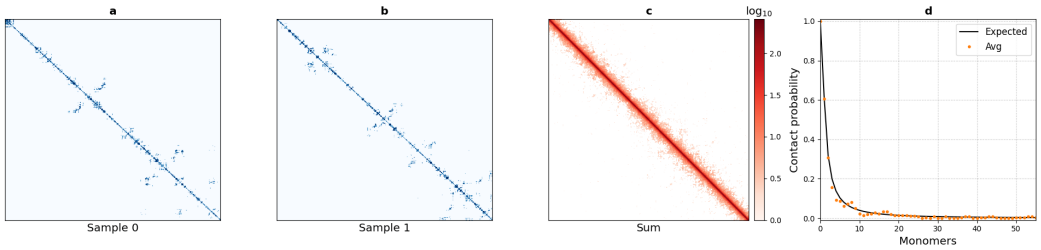

Fig. 11.2.: (a-b) Examples of contact maps measured for 2 Gaussian chains with 512 monomers and Kuhn length of 50 nm. (c) Summing enough maps such as in (a-b), we recapitulate theoretical ensemble averaged results. (d) Direct comparison between theory and a row of the matrix in (b).

Let's explore a little further the concept of using a single Gaussian chain contact map to reconstruct chromatin. In figure (11.3a), we show the distance map calculated for a chain, while in (b) we present all touching monomers. To reconstruct this monomer assuming only these contacts is simple, we just need to run the equation 11.12 over touching monomers and set d_{ij} equal to the Kuhn length. Due to lack of information, there is an infinite amount of possible conformations this polymer could take and still satisfy these measured contacts. For that reason, we

reconstruct and calculate the average distance map for 256 polymers as presented in (c). From this average map we can estimate an contact map in (d), where just 124 out of 130816 terms where wrong. By inspection, most of the major structures are still represented, but we see in (e) that the reconstructed contacts are not exactly similar to the original map in the single reconstruction bases. Nonetheless, this structure is not representative of the average or population behavior. For that, we should do similar procedure over a large number of other possible conformations and, like so, recover the expected population average.

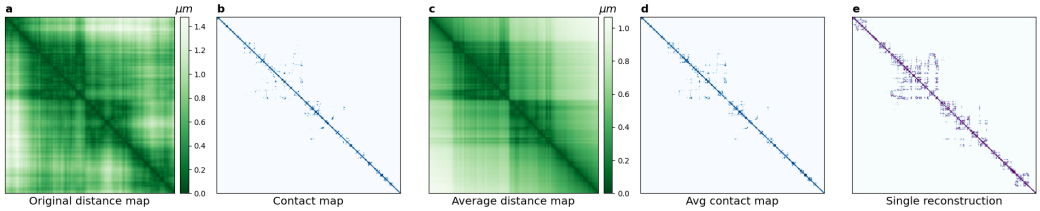

Fig. 11.3.: Recosntructing polymer from a single chain contact map. (a) Starting of from a distance map calculated for a sampled Gaussian chain with b=50 nm, we determine all monomers that are intersecting and build a contact map in (b). (c) Average distance map calculated over 256 polymers reconstructed using measured contacts in (b). (d) Contact map generate from average distance map. Most of the main structures are nicely recovered. (e) Contact map generate from a single reconstructed polymer. As we can see, it doesn't quite resemble the original contact map.

Notice that, up to now, we are studying average properties of these polymers by generating independent samples. Otherwise, we could summon ergodicity and show that similar results would be obtained by simulating a single Rouse chain for a large period of time and capturing its conformation every so often. For obvious reasons, ergodicity is not always maintained in cellular context [62], but we assume as an approximation for shorter periods of time. This result hints us towards a rather obvious idea, but important one, that the population behavior is determined upon specific features associated with polymer dynamics. The Rouse chain presents unseasoned properties, hence its population behavior is just as bland. Consequently, we could implement specific mechanisms and/or interesting interactions that will bring to light different types of average behaviors on the long run. We are going to take a different approach and find a way to sample polymers that, on average, recover the overall population behavior without considering specific biophysical mechanisms.

Sampling polymers is no easy task due to the implicit correlation existent among nearby monomers. In the sense that it might be more or less probable for any given monomer i to contact j if a neighbor of i is already interacting with monomer j.

Even for a Gaussian chain, upon quick inspection of figure (11.2a-b), we recognize that off-diagonal marks present correlation. Evidently, there are many methods to sample structured noise, such as modeling the covariance function of a Multivariate Gaussian distribution. In other words, we could create a model to sample these polymers using a Gaussian process, for example. However, we are going to take a much simpler approach and sample contacts based on their probability to occur given an expected distance. This method will generate polymers that always resemble the population average, but, at the same time, allows monomers to explore the space with increased liberty. We are going to discuss this choice in more detail when we start to talk about dynamics.

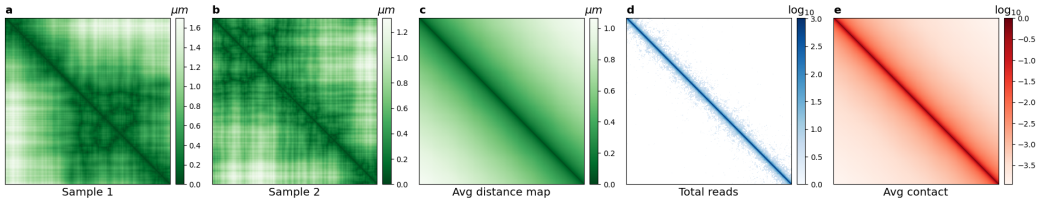

Fig. 11.4.: Testing reconstruction method from average conformation of the Gaussian chain with b = 50 nm. (a-b) Results for 2 sampled interaction matrices. (c) Average distance map calculated for 1024 sampled interactions. (d) We add up all the contacts measured from reconstructed distance maps. (e) We use the average distance map to estimate expected contact map.

To better appreciate this method, let's employ one last time the Gaussian chain model with Kuhn length $b = 50$ nm. Using a random uniform distribution, we sample interactions $i - j$ from equation 10.37. These terms will be accounted for during the reconstruction procedure via equation 11.12. As usual, self-interactions are neglected and, due to polymer constrains, we always consider first neighbors. Distances d_{ij} are taken as the expected distance in equation 10.36. Figure (11.4a-b) show the reconstruction results for 2 out of 1024 sampled interaction matrices. In (c), we show the average distance map calculated over all reconstructions. Further along, we determine intersecting monomers for each reconstructed polymer and add them up as a analog to total number of reads in a Hi-C maps, for example. The results are displayed in (d). Finally, we can use equation 10.38 to calculate the final contact map in (e) from the average distance map. Needless to say, the final contact map is nicely described by theoretical values.

11.2 HoxA domain

In the previous section we determined a method to sample polymers from the theoretical contact map expected for the Gaussian chain model. In this section, however, we apply this methodology for the HoxA domain in ESC via already balanced and properly normalized Hi-C maps, so that each row resembles equation 10.40. For completeness, we are also going to present results for NPC, even though our RA induced cells cannot really be consider as such.

Unfortunately, as we can see in figure (8.7), these maps contain a fair amount of noise and undetermined values due known experimental issues. For those reason, prior to applying our sampling method further treatment needs to be done. In order to solve the problem of missing interaction probabilities, we are going to use the results displayed in figure (10.6), that is, the fact that our re-scaled Gaussian model is a fairly good approximation for real contact probabilities. In that sense, we are going to use the re-scaled model to interpolate missing contact probabilities. Finally, to help blend in this approximation and results overall noise, we convolve this map with a Gaussian filter with standard deviation one. The final result can be seen in figure (11.5a). From treated map, we can also calculate interaction matrices. In (b) we show how this matrix looks like for a single polymer. Upon accumulation and normalization of interaction matrices for many polymers, these results will resemble more and more to the original contact map.

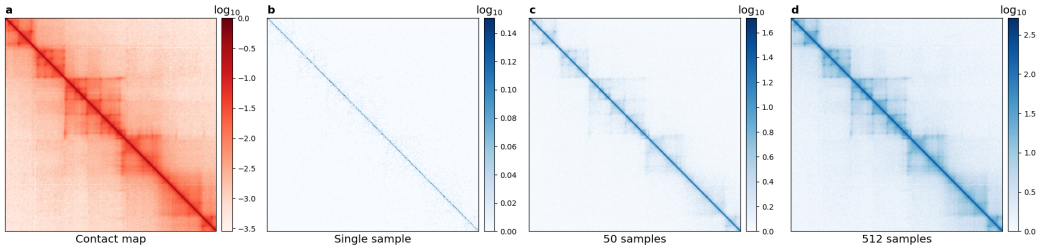

Fig. 11.5.: (a) Treated contact map for HoxA domain in ES cells. For reference, original map is presented in figure (8.7.a). (b) Interaction matrix sampled for (a). (c-d) The more samples we use, the more the cumulative interaction matrix resembles original contact map.

Using this treated population contact map and assuming once more the re-scaled Gaussian chain model as a good approximation for chromatin, we can also calculate the expected distance map using equation (10.38) given a value for Kuhn length. Fortunately, we have already experimentally determined the values of two elements in this population distance map in chapter 8, more specifically in figure (8.8), hence we can fit the Kuhn length. We obtain $b = 56$ nm for ESC and $b = 93$ nm for NPC.

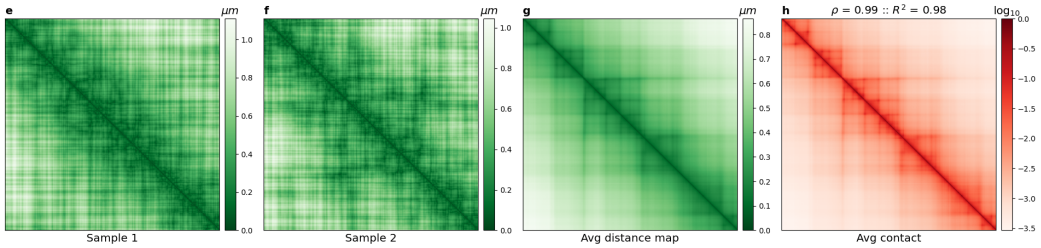

Fig. 11.6.: Reconstruction results for 2048 polymers with independently sampled interaction matrices for ESC. (a-b) Example of reconstructed distance maps for 2 polymers. (c) With average of maps calculated for all polymers, we estimate the population distance map. (d) Using these distances, we estimate population contact probabilities.

In figure (11.6) we have the results obtained upon reconstruction of 2048 polymers for the ESC. In (a-b), the distance map calculated for 2 sampled interaction matrices are display. The average distance map over these 2048 polymers is presented in (c), from which we calculate the expected contact map in (d). As measured of goodness, we show that this reconstructed expected contact map is correlated with $\rho = 0.99$ with original map. We also calculated the explained variance, where we have that our reconstructed map accounts for 98% of the variability in treated map. Out of curiosity, we display in figure (11.7) the evolution of ρ and R^2 with the number or polymers used.

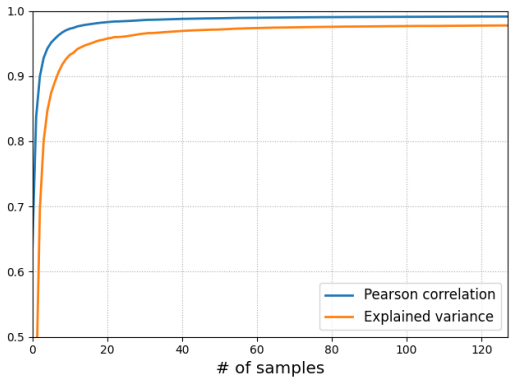

Fig. 11.7.: Relationship between number of reconstructed polymers to goodness of fit parameters, that is, Pearson correlation and explained variance.

Similar procedure is applied for the Hi-C maps of neuron precursor cells. In figure (11.8a), we show the treated population Hi-C map, from which we sample interaction matrices in (b-d). Once more, we reconstruct 512 polymers to determine an average distance map in (g) followed by the calculated average contact map that correlated 98% with experimental map and explains 96% of its variability.

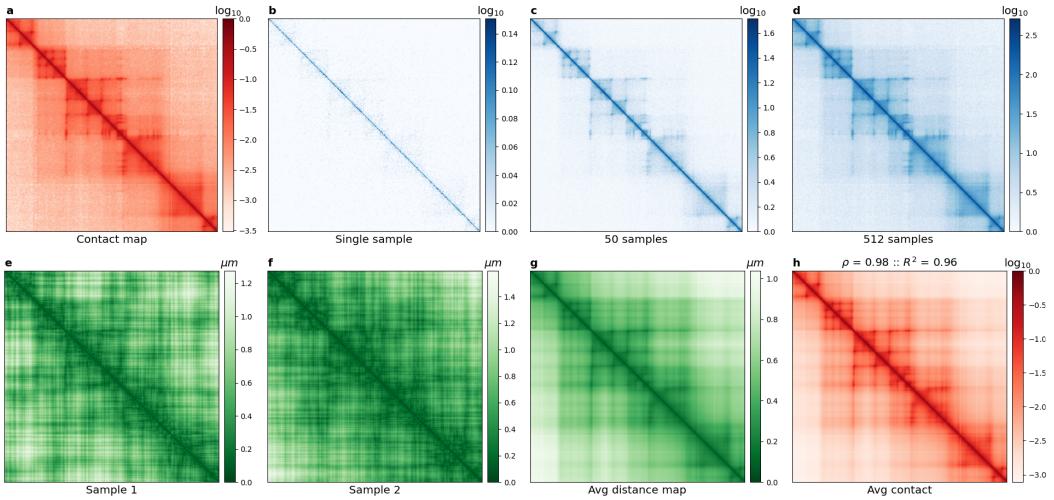

Fig. 11.8.: Reconstruction results for 2048 polymers with independently sampled interaction matrices for NPC. (a) Treated Hi-C map with estimated missing components and reduced noise. (b-d) Accumulation of sampled interaction matrices. (e-f) Example of 2 distance maps calculated from reconstructed polymers. (g) Estimated the population distance map. (h) Estimate contact probabilities. ρ and R^2 are calculated in comparison to (a).

Before closing this chapter, let's determine the distribution of distances measured in between our probes. Results can be appreciated in figure (11.9). The values approximated quite nicely our past measurements. Nonetheless, we must consider that all these polymers are static, that is, no dynamics apply. Hence, we might expect that these values will be slightly different, possibly a little more apart. On that note, we are going to introduce dynamics into our model in the next chapter.

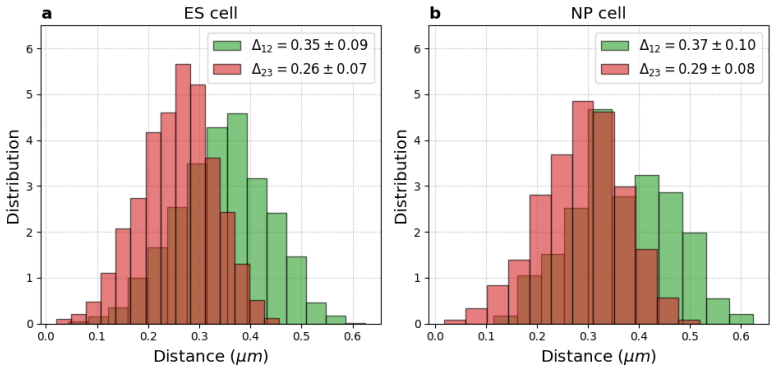

Fig. 11.9.: Reconstructed distances between probes for ES and NP cells. These results are very good approximations for values measured experimentally in figure 8.8, even though our RA induced cells are far from NP state.

12

Modeling chromatin dynamics

When we studied the Rouse chain model in chapter 10, we determined what one can expect from the most simple polymer model one can image. For that model, we account for simple first neighbor interactions via a spring developing some sort of temperature driven stochastic (but stationary) dynamics under influence of homogeneous substrate. Due to symmetry principles, all monomers of that polymer, with exception of boundary ones, present similar diffusion coefficient with anomalous behavior dwelling in 1/2. We also show that this model, with a few tweaks, can be taken as a first approximation to chromatin, consequently, we could propose more promising interpretation of our experimental results in chapter 8.

In that chapter, we determined that the apparent diffusion coefficient varies, within a credible interval, depending on its relative chromatin location. We also determined that the anomalous coefficient is, on average, bellow the theoretical 1/2 threshold, indicating that chromatin is more constraint than a free polymer. Among these divergences between data and Rouse model, perhaps the easiest one to explain is related to the anomalous coefficient. We can actually find a few interesting publications in which polymers can be simulated to present anomalous coefficients to ones will. In *Weber et al* [102], they achieve this by supposing that the polymer interacts differently to its surrounding substrate if compared to the Rouse chain. In *Amitai and Holcman* [103], they present a polymer model with long-range interactions that tend to decrease the Rouse anomalous coefficient. Notwithstanding, these models present homogeneous values for α, which is not what we have seem experimentally. For that reason, we might expect that the observed dynamics is, but a combination of both mechanisms, that is, substrate and long-range interactions.

Intuitively, we realize a connection between anomalous coefficient and the num-

ber of constraining interactions associated to any monomer. A single particle freely diffusing presents $\alpha = 1$, while attaching springs on either side shortened this value by half. Hence, we might expect that the more constraining interactions one monomer has, the lower shall be the anomalous coefficient. Forthcoming, the question relates to which interactions are important for dynamics and overall conformation. Perhaps, the more appropriate question is how these interactions change in time. Unfortunately, even with current technology, it is pretty hard or impossible to experimentally answer this question. In the previous decade, we have observed large interest in interacting complexes cohesin-CTCF, for example, which has been shown as important regulator of chromatin conformation. Nevertheless, there are possibly dozens of other interaction types that could constrain chromatin. Just to cite a few examples we have direct or mediate oligomerization, condensates, interactions with nuclear landmarks, among others [19]. In fact, the most probable solution would be a combination of several agents.

Following this reasoning, I reached the scheme of interactions proposed in the reconstruction chapter. If different sections of chromatin are only temporally in contact with probability proportional to the number of reads presented in a Hi-C map, these sections should also be at varying distances over time. We don't know exactly how this distance will change over time, but we do know its average. Hence, as a first approximation, I chose to use this expected distance as a force mediator for the dynamics of each monomer.

Upcoming next, we are going to extend our reconstruction model to account for stationary Langevin dynamics and determine, numerically, how it is affected by our sampled constraining interactions. To conclude this chapter, we will address the effects of a heterogeneous environment over dynamics as well.

12.1 Stationary dynamics with sampled long range interactions

We shall modify our reconstruction equation (11.12) to add stochastic white noise onto it as follows

$$d\boldsymbol{r}_i = \frac{3k_BT}{\xi}dt\sum_{i\neq j}\frac{r_{ij}-d_{ij}}{d_{ij}^2}\,\hat{r}_{ij} + \sqrt{2\frac{k_BT}{\xi}}d\boldsymbol{W}_t. \tag{12.1}$$

As explained, the summation will happen based on sampled interactions. We should choose the value of ξ more carefully than we did for reconstruction, but let us

suppose that the solvent or substrate affects all monomers exactly the same way. For simplicity, we won't create any fancy procedure to find its optimal value, so we shall fit it by inspection. In figure (12.1) we use those 2048 reconstructed polymers for ES cells from last chapter to analyze dynamics. For this case, a value of $\frac{k_B T}{\xi} = 2 \times 10^{-3} \mu m^2/s$ was chosen. To assess apparent diffusion and anomalous coefficients, we split the simulated polymers into 64 groups of 32 and use EA-MSD. The mean with 95% interval are presented in images (e) and (f) alongside a 95% CI of the experimentally measured mean. Even though our RA induced cells are no quite NP, similar procedure was performed for NP cells, these presented in figure (12.2).

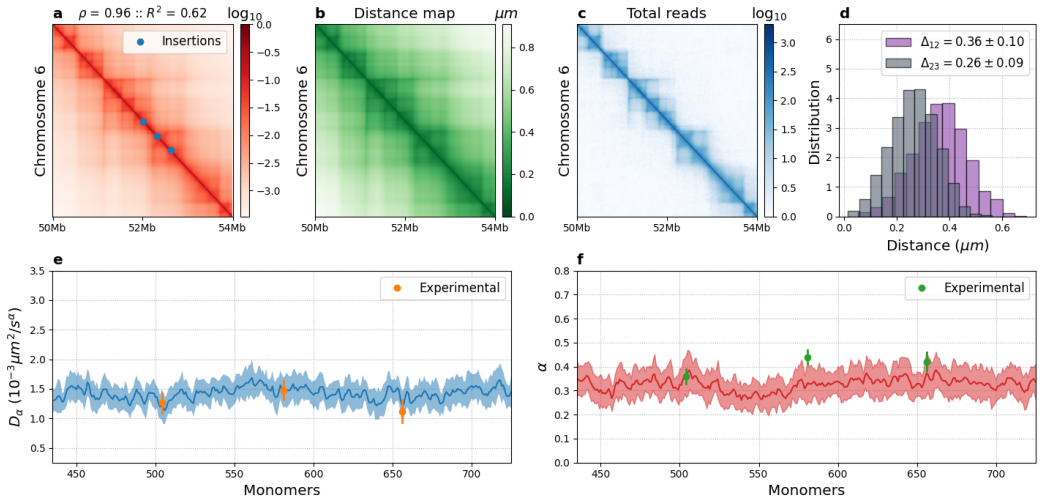

Fig. 12.1.: Simulating the HoxA domain of ES cells assuming homogeneous environment. (a-b) Average contact and distance maps obtained for our 2048 simulated polymers. (c) We accumulate contact over all polymers at the last time point simulated, that is, 60 seconds. (d) Inter-probe distances measured for last simulated time point in all polymers. (e) Shaded blue represents 95% credible interval for simulated apparent diffusion coefficients, while points are experimentally assessed with 95% credible interval to the mean. (f) Likewise, but for the anomalous coefficient.

As we could have expected, the effect of extra long range interactions is quite apparent and it had the average anomalous coefficient to drop by around 40% . Regardless though, to affirm that D_α and α are different for each monomer is still questionable. The correlation ρ to original Hi-C maps is still significant, but its explained variance R^2 has dropped non-negligibly in both cell types. Regardless, the distribution of distances are still is great agreement with experimental values.

This model poses a good first step towards explaining experimental results, but it fails to recapitulate some values measured for D_α and α. Nonetheless, there is

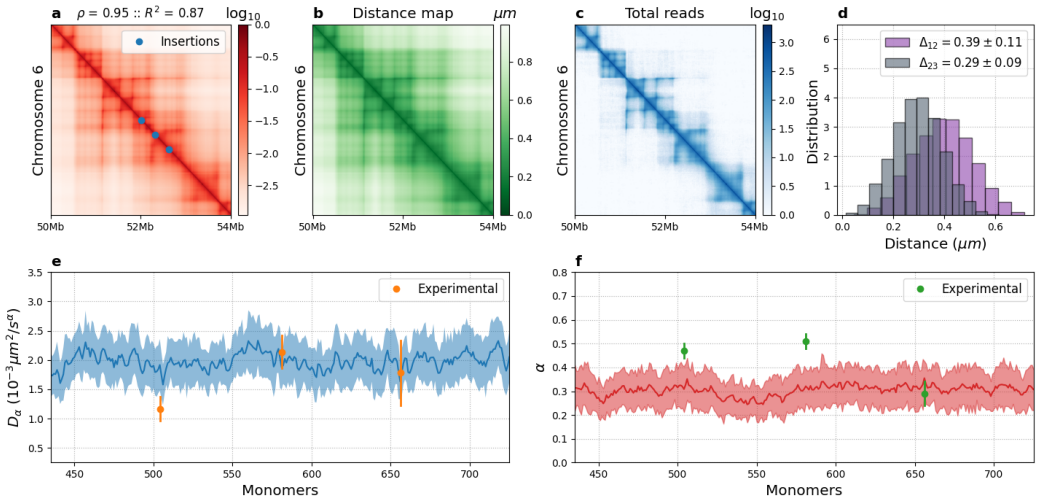

Fig. 12.2.: Simulating the HoxA domain of NP cells assuming homogeneous environment. (a-b) Average contact and distance maps calculated from 2048 reconstructed for NP cells. (c) Accumulated contacts over all polymers after 60 seconds of temporal evolution. (d) Distribution of inter-probe distances measured for all polymers. (e-f) Shaded colors accounts for 95% of the measured values for apparent diffusion and anomalous coefficients, respectively. Points represent 95% credible interval for experimentally measured mean for retinoic acid induced cells.

another parameter we have not utilized or, rather, has been kept constant up to now, the dynamics viscosity ξ. In the next section are going to explore its effects over dynamics and its practical interpretation.

12.2 Effects of chromatin context on dynamics

Up to now we have ignored one very important component in nuclear diffusion processes, the substrate in which particles undergo dynamics. For sake of simplicity, we have consider that our polymers are diffusing in a homogeneous substrate, even though we discussed during the introduction that it is not true. As chromatin is heavily coated by proteins for the purpose of transcriptional regulation and DNA replication, we might expect that, depending on how a given locus is regulated, its environmental diffusive properties will differ from others regions. As already discussed, we find in literature content showing that centromeres and telomeres are less mobile than average in yeast [10], while transcriptional active loci were found to correlate with smaller α and greater D_α in some instances [11, 12]. For

those reasons, I propose a slight modification in our polymer dynamics equation

$$d\boldsymbol{r}_i = 3\lambda_i dt \sum_{i \neq j} \frac{r_{ij} - d_{ij}}{d_{ij}^2}\, \hat{r}_{ij} + \sqrt{2\lambda_i} dW_t, \tag{12.2}$$

where we consider $\lambda_i \equiv k_B T/\xi_i$ as given in the monomer basis, with a different values depending on ChIP-seq signal accumulated over the number of base pairs enclosed in each monomer.

The idea of modeling a local λ_i based on ChIP-seq data of transcription factors might sound easy at first, but, due the huge number of proteins necessary to regulate the whole genome, it would not be a sustainable tool for modeling. Hence, I would like to propose using histone modifications for that purpose. As a example, we saw that the *Haunt* locus is enriched in H3K4me1, H3K4me3 and H3K36me in ES cells [20, 21], but, upon differentiation, H3K4me3 and H3K36me3 decrease and H3K27me3 increases [16]. Therefore, we could model some sort of signal combination encapsulating major histone modifications and verify if we are able to approximate values measured in chapter 8 for diffusion and anomalous coefficients. As described in the introduction, ChIP-seq signal will be balanced via MACS [45] and peaks where selected in a binary fashion.

Using the results from previous section, we will say that λ_i should have an average similar to $2 \times 10^{-3} \mu m^2/s$. To modulate this value, we simply bin peaks within the sequence encapsulated by each monomer and combine histograms obtained for each modification assuming equal weights. Final result goes through a Gaussian filter for smoothing.

In figure (12.3a) I present the results for ES cells using H3K122ac, H3K4me1, H3K27ac and H3K64ac. As before, 2048 polymers where simulated as in the previous sections, but now using λ_i. These were then divided into 64 groups of 32, from which we use EA-MSD to obtained measurements for D_α and α along with a 95% range. These results are presented in (b-c). Similar idea and approximations were applied for NP cells. Unfortunately, availability of ChIP-seq data for NP cells is scarce, thence only H3K4me3 and H3K27ac were utilized. Results are presented in figure (12.4).

This work is preliminary and more research needs to be done on the subject. Perhaps one of the first things to be done is to check ChIP-seq data for other histone modifications that are strongly associated with specific gene activity. This should help into determining better weighting system for final signal. Another possible improvement would be to use the raw ChIP-seq data, where non-specific interactions would also be considered.

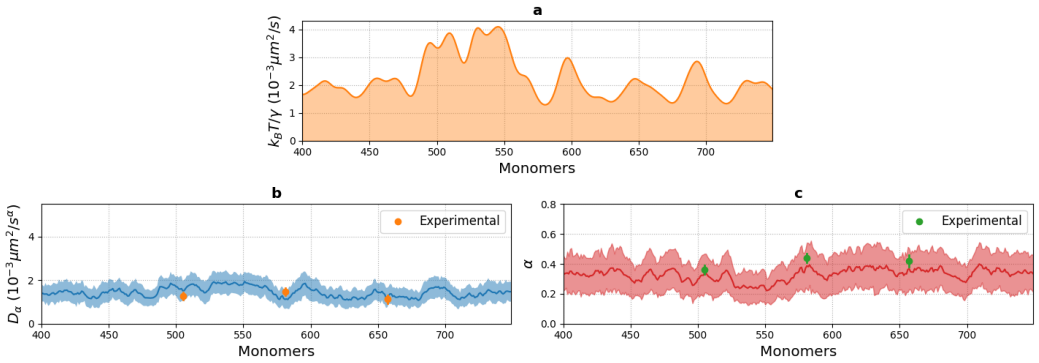

Fig. 12.3.: Dynamics of the HoxA domain in ES cells. (a) Empirical values for k_BT/ξ calculated via ChIP-seq data. (b-c) Comparison between experimental results and values of D_α and α calculated via EA-MSD for simulated polymers. Points and bars represent 95% of the experimental means. Shaded areas correspond to 95% of simulated values.

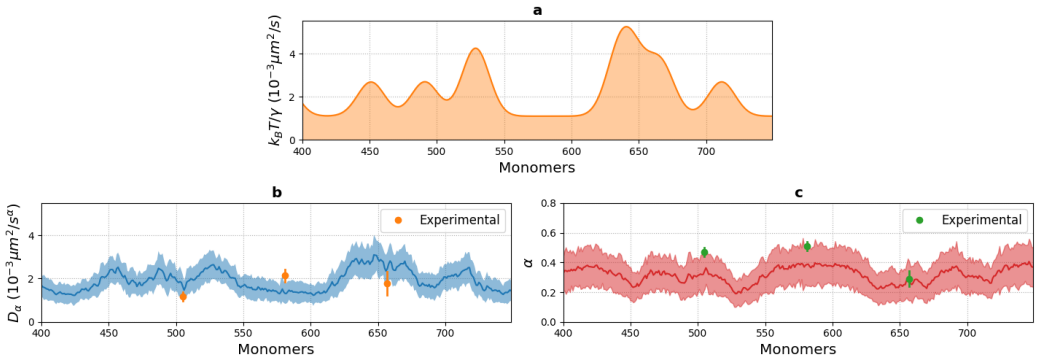

Fig. 12.4.: Dynamics of the HoxA domain in NP cells. (a) Empirical values for k_BT/ξ calculated via Chip-Seq data. (b-c) Comparison between experimental results for RA induced cells and values of D_α and α calculated via EA-MSD for simulated polymers. Points and bars represent 95% of the experimental means. Shaded areas correspond to 95% of simulated values.

Part IV

Conclusion

13

Discussion

On the course of my PhD, I was mostly interested on the dynamics of chromatin, but I was not simply concerned with measuring diffusion coefficient and determining how confined the movement of chromatin is from an experimental perspective. My curiosity drove me into exploring possible reasons to explain measured values, to question the very nature of the experimental data in many instances and, somehow, tell apart meaningful results from noise. These questions were the sole inspiration for most of the methods presented in part II.

From the beginning, we knew that measuring loci dependent diffusion properties of chromatin in such a heterogeneous environment (the nucleus) would demand greater precision and consideration than popular methods, such as MSD, could offer. Perhaps the most evident experimental bias that needed attending was background movement. Even without the knowledge proposed by the Rouse chain model, it was quite discernible that tagged chromatin loci were not supposed to be near a freely diffusive regime. Sooner than later, we verified that in many of those instances the whole region (or cell) was moving due to diverse reasons, from membrane fluctuations to heat induced stress. Several approaches were tested to resolve this issue, from Optical Flow to machine learning techniques, but all delivered unsatisfactory results due to simple imprecision and/or amount of work demanded. In the end, we approach the problem in a novel manner, we model background induced correlation among spots under similar conditions using Gaussian processes and fractional Brownian motion (GP-FBM). This approach allowed us to "save" most of the data recorded over the years without extra experimental needs and demanding computational pipelines. Nonetheless, we never cross-validated this methodology using properly calibrated experiments, that is, where we could estimate cellular movement experimentally. All the tests were conducted

over synthetic data. Despite that, computationally generated data was sufficient to inform us about its limitations and scaled of variance to be expected.

Interestingly, our experimental data was still much noisier than what we would expect based on *in silico* experiments. Unquestionably, the low signal-to-noise ratio presented by our tracked chromatin spots was an important component of great standard deviations due to imprecise localization and false positive detection. After several iterations, we reached to our localization enhancing method, the one presented in chapter 7. Of course, using two (or three) dimensional bell curves to fit spots in microscopy images is not new (by all means). Regardless, the pipeline allowed to statistically reduce falsely detected spots, improved localization several times and allowed us to estimate positional errors, vital for GP-FBM.

All this preparation was key for the analysis in which we compared chromatin dynamics between interphase and mitosis in chapter 8. We showed that cell average diffusion and anomalous coefficients are similar in both stages of the cell cycle, even though chromatin is two to three times denser during mitosis. Similarly, we were also able to show that great part of the variability found for such diffusion properties was within cells. Unsurprisingly, the variation could be intuitively deduced to rise from the difference between constitutive heterochromatin, known to be denser and more constraint, and euchromatin. But, is our intuition correct? Hard to say and more research (experimental or literature) should be performed. Unfortunately, our results for the HoxA domain (chapter 8), suggested that these differences are more complicated than a simple matter of difference in chromatin organization. We verified that local effects due to gene expression are also relevant.

The problem at this point was that we knew very little about which values of D_α and α we could expect from a theoretical perspective. We knew that free particles present anomalous coefficient equal to one (that is, not anomalous at all) and we expected that α should have been smaller than that for polymers. In fact, upon consideration of the Rouse chain model, α should be about 1/2. Otherwise, we measured in chapter 8 an average α to be near 1/3. Evidently, that is to be expected if Hi-C maps are to be considered. Obviously chromatin is biased towards some non-random conformation, indicating the presence of mechanisms driving long range interactions. Consequently or collaterally, these constraining forces reduce α bellow 1/2.

Possible molecular components involved in such mechanisms are studied for over a decade now. Several have been identified and tested via computational modeling. Nonetheless, I was not really interested in accounting for all those different types of tools used for genome regulation. My approach was constructed

empirically and does not explicitly consider any particular mechanism underlying the formation and maintenance of the 3D structure. I use a mean-field approach where all of these mechanisms are considered on average. This model consists in sampling long range interactions from Hi-C maps in such a way that the average dynamic distances are maintained. Interestingly, this model was able to recapitulate the average behavior we measured using the TetO system in chapter 8. This simple model was also able to recover inter-probe distances measured at the HoxA domain for stem cells and satisfactorily approximate values for neuron precursor cells. Despite of these nice results, upon comparison of diffusion and anomalous coefficients measured for our synthetic polymers and the values obtained via experimental data, we noticed non-negligible differences. Evidently, a polymer that behaves like average chromatin would not precisely represent local behavior. Discrepancies are anticipated to be even higher if we are to consider regulatory sequences, a minority if compared to entirety of chromatin.

To mitigate such differences, we needed to consider more carefully what makes regulatory sequences not akin to other regions of chromatin. Depending on the way such question is formulated, the answer might be fairly complicated to approach. My first idea, and still to be considered preliminary, was to adapt the equation to local surrounding environment. Transcription factors (and other molecules) interact non-specifically with the whole chromatin, but it is biased towards regulatory sequences due to specific DNA-protein interactions. Because of that, we might expect that the density of molecules and the strength of chemical forces are higher in surrounding regions. Not even to mention that usually dozens of different proteins are biased towards these regions. In short, it becomes quite unassailable that our assumption in which chromatin dwells in a homogeneous environment will fail locally. To fix these problem, we could use ChIP-seq data for all transcriptional factors associated to each of these regulatory sequences. Unfortunately, that is no simple task. Even with modern technology, to identify which proteins are involved in the regulation of different genes is far from being satisfactorily solved. Despite of that, there are plenty of material in literature correlating gene expression with histone modifications. For that reason, I decided to take ChIP-seq data for such marks as indications of heterogeneity in chromatin surroundings. If this approach is the correct way of dealing with the problem remains to be confirmed. Independent of that, we were able to improve our fittings and recover (or almost) experimental measurements utilizing the simple ideas introduced.

14

Perspectives

In part II we introduced the usage of Gaussian processes for inference of diffusion and anomalous coefficients using the fractional Brownian motion covariance matrix. This combination proved to be quite efficient in doing its job, but we have unleashed very little of its potential. We have indeed used this approach to filter out substrate movement from diffusing particle, however we always considered that each trajectory enclosed a single type of motion. Let us consider transcription factors (TFs), for example. It has been shown with single molecule tracking (SMT) that TFs alternate between two or more types of diffusion, where they may diffuse freely or bound to chromatin. Assuming that switching time between modes is slow compared with the recording time, GP-FBM could be adapted to infer diffusion parameters from single trajectories. Like that, we could infer binding rates and other kinetic parameters. In that sense, GP-Tool could eventually be modified to allow the user to specify a model for analysis.

Regarding the experiments presented in chapter 8. More specifically concerning the HoxA domain. The original idea for the experiment was to detect if TADs would play an important role in diffusion properties. As our results suggest, that is not the case, as differences in D_α and α seem to be more striking near regulatory sequences. For that reason, it would be nice to analyze data of probes inserted into proximal and distal regulatory sequences. If differences between these values and the average chromatin model are more relevant, it would indicate further that gene regulation is more important for diffusion dynamics than chromatin organization.

It would be nice to further investigate if histone modifications could be used to model locus based context for chromatin dynamics. As the work develop was based on simplified/binary ChIP-seq signal of a small set of marks, it would be better to

develop a model that used raw data for a larger number of histone modifications. Eventually, we could also use ChIP-seq for TFs.

Finally, it would be interesting to extend the polymer model developed to study enhancer-promoter dynamics and its relationship to transcriptional activity. Perhaps a good experimental approach would be to use the MS2 tagging system to monitor transcriptional output of the *Sox2* gene in ES cells and other differentiation states. Like so, we could analyze local structural re-arrangements that are relevant for transcription initiation.

Bibliography

[1] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[2] John A Nelder and Roger Mead. « A Simplex Method for Function Minimization ». In: *The Computer Journal* 7.4 (Jan. 1965), pp. 308–313. ISSN: 0010-4620. DOI: 10.1093/comjnl/7.4.308. URL: https://academic.oup.com/comjnl/article-lookup/doi/10.1093/comjnl/7.4.308.

[3] Fabrice de Chaumont et al. « Icy: an open bioimage informatics platform for extended reproducible research ». In: *Nature Methods* 9.7 (July 2012), pp. 690–696. ISSN: 1548-7091. DOI: 10.1038/nmeth.2075. URL: http://www.nature.com/articles/nmeth.2075.

[4] D. Selmeczi et al. « Cell motility as random motion: A review ». In: *European Physical Journal: Special Topics*. 2008. DOI: 10.1140/epjst/e2008-00626-x.

[5] David Llères et al. « Quantitative analysis of chromatin compaction in living cells using FLIM-FRET ». In: *Journal of Cell Biology* (2009). ISSN: 00219525. DOI: 10.1083/jcb.200907029.

[6] Felipe Mora-Bermúdez, Daniel Gerlich, and Jan Ellenberg. « Maximal chromosome compaction occurs by axial shortening in anaphase and depends on Aurora kinase ». In: *Nature Cell Biology* (2007). ISSN: 14657392. DOI: 10.1038/ncb1606.

[7] Robert M. Martin and M. Cristina Cardoso. « Chromatin condensation modulates access and binding of nuclear proteins ». In: *The FASEB Journal* (2010). ISSN: 0892-6638. DOI: 10.1096/fj.08-128959.

[8] Marc Kschonsak and Christian H. Haering. « Shaping mitotic chromosomes: From classical concepts to molecular mechanisms ». In: *BioEssays* (2015). ISSN: 15211878. DOI: 10.1002/bies.201500020.

[9] Ewa Piskadlo and Raquel A. Oliveira. « Novel insights into mitotic chromosome condensation ». In: *F1000Research* (2016). ISSN: 2046-1402. DOI: 10.12688/f1000research.8727.1.

[10] P. Heun et al. « Chromosome dynamics in the yeast interphase nucleus ». In: *Science* (2001). ISSN: 00368075. DOI: 10.1126/science.1065366.

[11] Thomas Germier et al. « Real-Time Imaging of a Single Gene Reveals Transcription-Initiated Local Confinement ». In: *Biophysical Journal* (2017). ISSN: 15420086. DOI: 10.1016/j.bpj.2017.08.014.

[12] Bo Gu et al. « Transcription-coupled changes in nuclear mobility of mammalian cis-regulatory elements ». In: *Science* (2018). ISSN: 10959203. DOI: 10.1126/science.aao3136.

[13] Bernard Mariamé et al. « Real-Time Visualization and Quantification of Human Cytomegalovirus Replication in Living Cells Using the ANCHOR DNA Labeling Technology ». In: *Journal of Virology* (2018). ISSN: 0022-538X. DOI: 10.1128/jvi.00571-18.

[14] Elizabeth H. Finn et al. « Extensive Heterogeneity and Intrinsic Variation in Spatial Genome Organization ». In: *Cell* (2019). ISSN: 10974172. DOI: 10.1016/j.cell.2019.01.020.

[15] Boyan Bonev et al. « Multiscale 3D Genome Rewiring during Mouse Neural Development ». In: *Cell* (2017). ISSN: 10974172. DOI: 10.1016/j.cell.2017.09.043.

[16] Yafei Yin et al. « Opposing roles for the lncRNA haunt and its genomic locus in regulating HOXA gene activation during embryonic stem cell differentiation ». In: *Cell Stem Cell* (2015). ISSN: 18759777. DOI: 10.1016/j.stem.2015.03.007.

[17] Adrian L. Sanborn et al. « Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes ». In: *Proceedings of the National Academy of Sciences of the United States of America* (2015). ISSN: 10916490. DOI: 10.1073/pnas.1518552112.

[18] Geoffrey Fudenberg et al. « Formation of Chromosomal Domains by Loop Extrusion ». In: *Cell Reports* (2016). ISSN: 22111247. DOI: 10.1016/j.celrep.2016.04.085.

[19] Seungsoo Kim and Jay Shendure. « Mechanisms of Interplay between Transcription Factors and the 3D Genome ». In: *Molecular Cell* 76.2 (2019), pp. 306–319. ISSN: 10974164. DOI: 10.1016/j.molcel.2019.08.010. URL: https://doi.org/10.1016/j.molcel.2019.08.010.

[20] Chloe M. Rivera and Bing Ren. *Mapping human epigenomes*. 2013. DOI: 10.1016/j.cell.2013.09.011.

[21] Warren A. Whyte et al. « Master transcription factors and mediator establish super-enhancers at key cell identity genes ». In: *Cell* (2013). ISSN: 00928674. DOI: 10.1016/j.cell.2013.03.035.

[22] Feng Yue et al. « A comparative encyclopedia of DNA elements in the mouse genome ». In: *Nature* 515.7527 (2014), pp. 355–364. ISSN: 14764687. DOI: 10.1038/nature13992.

[23] Anthony T. Annunziato. « DNA Packaging: Nucleosomes and Chromatin ». In: *Nature Education*. 2008.

[24] Aniek Janssen, Serafin U. Colmenares, and Gary H. Karpen. *Heterochromatin: Guardian of the Genome*. 2018. DOI: 10.1146/annurev-cellbio-100617-062653.

[25] Andrew J. Bannister and Tony Kouzarides. *Regulation of chromatin by histone modifications*. 2011. DOI: 10.1038/cr.2011.22.

[26] V G ALLFREY, R FAULKNER, and A E MIRSKY. « Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. » In: *Proc. Natl. Acad. Sci. USA* (1964). ISSN: 0027-8424.

[27] Bing Li, Michael Carey, and Jerry L. Workman. *The Role of Chromatin during Transcription*. 2007. DOI: 10.1016/j.cell.2007.01.015.

[28] Elzo de Wit and Wouter de Laat. « A decade of 3C technologies: Insights into nuclear organization ». In: *Genes and Development* (2012). ISSN: 08909369. DOI: 10.1101/gad.179804.111.

[29] Tom Sexton et al. *Gene regulation through nuclear organization*. 2007. DOI: 10.1038/nsmb1324.

[30] Tom Sexton and Giacomo Cavalli. *The role of chromosome domains in shaping the functional genome*. 2015. DOI: 10.1016/j.cell.2015.02.040.

[31] Peter Hugo Lodewijk Krijger and Wouter De Laat. *Regulation of disease-associated gene expression in the 3D genome*. 2016. DOI: 10.1038/nrm.2016.138.

[32] Natalia Sikorska and Tom Sexton. *Defining Functionally Relevant Spatial Chromatin Domains: It is a TAD Complicated*. 2020. DOI: 10.1016/j.jmb.2019.12.006.

[33] Geoffrey Fudenberg et al. « Emerging Evidence of Chromosome Folding by Loop Extrusion ». In: *Cold Spring Harbor symposia on quantitative biology* 82 (2017), pp. 45–55. ISSN: 19434456. DOI: 10.1101/sqb.2017.82.034710.

[34] Erez Lieberman-Aiden et al. « Comprehensive mapping of long-range interactions reveals folding principles of the human genome ». In: *Science* 326.5950 (2009), pp. 289–293. ISSN: 00368075. DOI: 10.1126/science.1181369.

[35] Elizabeth A. Hoffman et al. « Formaldehyde crosslinking: A tool for the study of chromatin complexes ». In: *Journal of Biological Chemistry* 290.44 (2015), pp. 26404–26411. ISSN: 1083351X. DOI: 10.1074/jbc.R115.651679.

[36] Bryan R. Lajoie, Job Dekker, and Noam Kaplan. « The Hitchhiker's guide to Hi-C analysis: Practical guidelines ». In: *Methods* (2015). ISSN: 10959130. DOI: 10.1016/j.ymeth.2014.10.031.

[37] Mattia Forcato et al. « Comparison of computational methods for Hi-C data analysis ». In: *Nature Methods* (2017). ISSN: 15487105. DOI: 10.1038/nmeth.4325.

[38] Philip A Knight and Daniel Ruiz. « A fast algorithm for matrix balancing ». In: *IMA Journal of Numerical Analysis* 33.3 (2013), pp. 1029–1047.

[39] Maxim Imakaev et al. « Iterative correction of Hi-C data reveals hallmarks of chromosome organization ». In: *Nature Methods* (2012). ISSN: 15487091. DOI: 10.1038/nmeth.2148.

[40] Daynna J. Wolff. « Fluorescence in situ hybridization(FISH) ». In: *The Principles of Clinical Cytogenetics, Third Edition*. 2013. ISBN: 9781441916884. DOI: 10.1007/978-1-4419-1688-4_17.

[41] Oluwatosin Oluwadare, Max Highsmith, and Jianlin Cheng. « An Overview of Methods for Reconstructing 3-D Chromosome and Genome Structures from Hi-C Data ». In: *Biological Procedures Online* 21.1 (2019), pp. 1–20. ISSN: 14809222. DOI: 10.1186/s12575-019-0094-0.

[42] Michael Levine, Claudia Cattoglio, and Robert Tjian. *Looping back to leap forward: Transcription enters a new era*. 2014. DOI: 10.1016/j.cell.2014.02.009.

[43] Eileen E.M. Furlong and Michael Levine. *Developmental enhancers and chromosome topology*. 2018. DOI: 10.1126/science.aau0320.

[44] Stefan Schoenfelder and Peter Fraser. *Long-range enhancer–promoter contacts in gene expression control*. 2019. DOI: 10.1038/s41576-019-0128-0.

[45] Yong Zhang et al. « Model-based analysis of ChIP-Seq (MACS) ». In: *Genome Biology* (2008). ISSN: 14747596. DOI: 10.1186/gb-2008-9-9-r137.

[46] Zhen Shen, Wenzheng Bao, and De Shuang Huang. « Recurrent Neural Network for Predicting Transcription Factor Binding Sites ». In: *Scientific Reports* (2018). ISSN: 20452322. DOI: 10.1038/s41598-018-33321-1.

[47] Sungjoon Park et al. « Enhancing the interpretability of transcription factor binding site prediction using attention mechanism ». In: *Scientific Reports* (2020). ISSN: 20452322. DOI: 10.1038/s41598-020-70218-4.

[48] Peter K. Koo and Matt Ploenzke. *Deep learning for inferring transcription factor binding sites*. 2020. DOI: 10.1016/j.coisb.2020.04.001.

[49] Elliott W. Montroll and Harvey Scher. « Random walks on lattices. IV. Continuous-time walks and influence of absorbing boundaries ». In: *Journal of Statistical Physics* 9.2 (1973), pp. 101–135. ISSN: 00224715. DOI: 10.1007/BF01016843.

[50] Jae-Hyung Jeon and Ralf Metzler. « Fractional Brownian motion and motion governed by the fractional Langevin equation in confined geometries ». In: *Physical Review E* 81.2 (Feb. 2010), p. 021103. ISSN: 1539-3755. DOI: 10.1103/PhysRevE.81.021103. URL: https://link.aps.org/doi/10.1103/PhysRevE.81.021103.

[51] M. J. Saxton. « Anomalous diffusion due to obstacles: a Monte Carlo study ». In: *Biophysical Journal* (1994). ISSN: 00063495. DOI: 10.1016/S0006-3495(94)80789-1.

[52] Shlomo Havlin and Daniel Ben-Avraham. « Diffusion in disordered media ». In: *Advances in Physics* 51.1 (2002), pp. 187–292. ISSN: 00018732. DOI: 10.1080/00018730110116353.

[53] Felix Höfling and Thomas Franosch. « Anomalous transport in the crowded world of biological cells ». In: *Reports on Progress in Physics* 76.4 (2013). ISSN: 00344885. DOI: 10.1088/0034-4885/76/4/046602. arXiv: 1301.6990.

[54] Ralf Metzler. « Gaussianity fair: the riddle of anomalous yet non-Gaussian diffusion ». In: *Biophysical journal* 112.3 (2017), p. 413.

[55] R Metzler, J-H Jeon, and AG Cherstvy. « Non-Brownian diffusion in lipid membranes: Experiments and simulations ». In: *Biochimica et Biophysica Acta (BBA)-Biomembranes* 1858.10 (2016), pp. 2451–2467.

[56] Thomas J Lampo et al. « Cytoplasmic RNA-protein particles exhibit non-Gaussian subdiffusive behavior ». In: *Biophysical journal* 112.3 (2017), pp. 532–542.

[57] Luca Giorgetti et al. « Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription ». In: *Cell* (2014). ISSN: 10974172. DOI: 10.1016/j.cell.2014.03.025.

[58] Peter R. Cook and Davide Marenduzzo. « Transcription-driven genome organization: A model for chromosome structure and the regulation of gene expression tested through simulations ». In: *Nucleic Acids Research* (2018). ISSN: 13624962. DOI: 10.1093/nar/gky763. arXiv: 2010.00551.

[59] Chris A. Brackey, Davide Marenduzzo, and Nick Gilbert. « Mechanistic modeling of chromatin folding to understand function ». In: *Nature Methods* 17.8 (2020), pp. 767–775. ISSN: 15487105. DOI: 10.1038/s41592-020-0852-6. URL: http://dx.doi.org/10.1038/s41592-020-0852-6.

[60] Jens Krog et al. « Bayesian model selection with fractional Brownian motion ». In: *Journal of Statistical Mechanics: Theory and Experiment* 2018.9 (2018), p. 093501.

[61] Samudrajit Thapa et al. « Bayesian analysis of single-particle tracking data using the nested-sampling algorithm: maximum-likelihood model selection applied to stochastic-diffusivity data ». English. In: *Physical chemistry chemical physics : PCCP* 20.46 (Sept. 2018), pp. 29018–29037. ISSN: 1463-9084. DOI: 10.1039/C8CP04043E.

[62] Aubrey V. Weigel et al. « Ergodic and nonergodic processes coexist in the plasma membrane as observed by single-molecule tracking ». In: *Proceedings of the National Academy of Sciences of the United States of America* (2011). ISSN: 00278424. DOI: 10.1073/pnas.1016325108.

[63] Christian L. Vestergaard. « Optimizing experimental parameters for tracking of diffusing particles ». In: *Physical Review E* 94.2 (2016), pp. 1–17. ISSN: 24700053. DOI: 10.1103/PhysRevE.94.022401.

[64] Nilah Monnier et al. « Bayesian approach to MSD-based analysis of particle motion in live cells ». In: *Biophysical Journal* 103.3 (Aug. 2012), pp. 616–626. ISSN: 00063495. DOI: 10.1016/j.bpj.2012.06.029.

[65] Christian L Vestergaard, Paul C Blainey, and Henrik Flyvbjerg. « Optimal estimation of diffusion coefficients from single-particle trajectories ». In: *Physical Review E* 89.2 (Feb. 2014), p. 022726. ISSN: 1539-3755. DOI: 10.1103/PhysRevE.89.022726. URL: https://link.aps.org/doi/10.1103/PhysRevE.89.022726.

[66] Anders S Hansen et al. « Robust model-based analysis of single-particle tracking experiments with Spot-On ». In: *eLife* 7 (Jan. 2018), e33125. ISSN: 2050-084X. DOI: 10.7554/eLife.33125. URL: https://elifesciences.org/articles/33125.

[67] Indrani Chakraborty and Yael Roichman. « Disorder-induced Fickian, yet non-Gaussian diffusion in heterogeneous media ». In: *Physical Review Research* 2.2 (2020), p. 022020.

[68] Carl Edward Rasmussen. « Gaussian Processes in machine learning ». In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (2004). ISSN: 16113349. DOI: 10.1007/978-3-540-28650-9_4.

[69] Adam Becker. *What is real? The unfinished quest for the meaning of quantum physics*. 2018. ISBN: 978-0-465-09606-0.

[70] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. 2006. ISBN: 9780387303031.

[71] Nicholas I.M. Gould et al. « Solving the trust-region subproblem using the Lanczos method ». In: *SIAM Journal on Optimization* (1999). ISSN: 10526234. DOI: 10.1137/S1052623497322735.

[72] Benoit B. Mandelbrot and John W. Van Ness. « Fractional Brownian Motions, Fractional Noises and Applications ». In: *SIAM Review* (1968). ISSN: 0036-1445. DOI: 10.1137/1010093.

[73] Stas Burov et al. « Single particle tracking in systems showing anomalous diffusion: the role of weak ergodicity breaking ». In: *Physical Chemistry Chemical Physics* 13.5 (2011), pp. 1800–1812.

[74] Johannes Schindelin et al. *Fiji: An open-source platform for biological-image analysis*. 2012. DOI: 10.1038/nmeth.2019.

[75] Stephen M. Pizer et al. « Contrast-limited adaptive histogram equalization: Speed and effectiveness ». In: *Proceedings of the First Conference on Visualization in Biomedical Computing*. 1990. ISBN: 0818620390. DOI: 10.1109/vbc.1990.109340.

[76] Ajay K. Prasad. « Particle image velocimetry ». In: *Current Science* (2000). ISSN: 00113891. DOI: 10.1201/b19031-55.

[77] Andreas Wedel et al. « An Improved Algorithm for TV-L 1 Optical Flow ». In: 2009, pp. 23–45. DOI: 10.1007/978-3-642-03061-1_2. URL: http://link.springer.com/10.1007/978-3-642-03061-1_2.

[78] Anshuman Agarwal, Shivam Gupta, and Dushyant Kumar Singh. « Review of optical flow technique for moving object detection ». In: *Proceedings of the 2016 2nd International Conference on Contemporary Computing and Informatics, IC3I 2016*. 2016. ISBN: 9781509052554. DOI: 10.1109/IC3I.2016.7917999.

[79] M. C. Marchetti et al. « Hydrodynamics of soft active matter ». In: *Reviews of Modern Physics* (2013). ISSN: 00346861. DOI: 10.1103/RevModPhys.85.1143.

[80] F. Jülicher et al. *Active behavior of the Cytoskeleton*. 2007. DOI: 10.1016/j.physrep.2007.02.018.

[81] Tzon-Tzer Lu and Sheng-Hua Shiou. « Inverses of 2 x 2 block matrices ». In: *Computers and Mathematics with Applications* 43.1-2 (Jan. 2002), pp. 119–129. ISSN: 08981221. DOI: 10.1016/S0898-1221(01)00278-4. URL: https://linkinghub.elsevier.com/retrieve/pii/S0898122101002784.

[82] Yoav Benjamini and Yosef Hochberg. « Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing ». In: *Journal of the Royal Statistical Society: Series B (Methodological)* (1995). ISSN: 2517-6161. DOI: 10.1111/j.2517-6161.1995.tb02031.x.

[83] T. K. Boehme and Ron Bracewell. « The Fourier Transform and its Applications. » In: *The American Mathematical Monthly* (1966). ISSN: 00029890. DOI: 10.2307/2314845.

[84] G. F. Newell, K. Ito, and H. P. McKean. « Diffusion Processes and their Sample Paths. » In: *The American Mathematical Monthly* (1968). ISSN: 00029890. DOI: 10.2307/2315169.

[85] Niels Grønbech-Jensen and Oded Farago. « A simple and effective Verlet-type algorithm for simulating Langevin dynamics ». In: *Molecular Physics* 111.8 (2013), pp. 983–991.

[86] Donald L Ermak and Helen Buckholz. « Numerical integration of the Langevin equation: Monte Carlo simulation ». In: *Journal of Computational Physics* 35.2 (1980), pp. 169–182.

[87] Peter E Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*. Springer, 1992. ISBN: 978-3-662-12616-5.

[88] Wei Wang and Robert D Skeel. « Analysis of a few numerical integration methods for the Langevin equation ». In: *Molecular Physics* 101.14 (2003), pp. 2149–2156.

[89] Lifang Liang et al. « Noninvasive determination of cell nucleoplasmic viscosity by fluorescence correlation spectroscopy ». In: *Journal of Biomedical Optics* 14.2 (2009), p. 24013.

[90] Bruno H Zimm. « Dynamics of polymer molecules in dilute solution: viscoelasticity, flow birefringence and dielectric loss ». In: *The journal of chemical physics* 24.2 (1956), pp. 269–278.

[91] Bogdan Bintu et al. « Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells ». In: *Science* 362.6413 (2018). ISSN: 10959203. DOI: 10.1126/science.aau1783.

[92] A Grosberg et al. « Crumpled globule model of the three-dimensional structure of DNA ». In: *EPL (Europhysics Letters)* 23.5 (1993), p. 373.

[93] A Y Grosberg. « Crumpled globule model of DNA packing in chromosomes: from predictions to open questions ». In: *Biomat 2010: International Symposium on Mathematical and Computational Biology*. World Scientific. 2011, pp. 17–28.

[94] Suhas S P Rao et al. « A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping ». In: *Cell* 159.7 (2014), pp. 1665–1680.

[95] Siyuan Wang et al. « Spatial organization of chromatin domains and compartments in single chromosomes ». In: *Science* 353.6299 (2016), pp. 598–602.

[96] Zhijun Duan et al. « A three-dimensional model of the yeast genome ». In: *Nature* 465.7296 (2010), pp. 363–367.

[97] Hideki Tanizawa et al. « Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation ». In: *Nucleic acids research* 38.22 (2010), pp. 8164–8177.

[98] Nelle Varoquaux et al. « A statistical approach for inferring the 3D structure of the genome ». In: *Bioinformatics* 30.12 (2014), pp. i26–i33.

[99] Marius Socol et al. « Rouse model with transient intramolecular contacts on a timescale of seconds recapitulates folding and fluctuation of yeast chromosomes ». In: *Nucleic acids research* 47.12 (2019), pp. 6195–6207.

[100] Siyu Wang, Jinbo Xu, and Jianyang Zeng. « Inferential modeling of 3D chromatin structure ». In: *Nucleic Acids Research* 43.8 (2015), pp. 1–12. ISSN: 13624962. DOI: 10.1093/nar/gkv100.

[101] Takashi Nagano et al. « Single-cell Hi-C reveals cell-to-cell variability in chromosome structure ». In: *Nature* 502.7469 (2013), pp. 59–64. ISSN: 00280836. DOI: 10.1038/nature12593. URL: http://dx.doi.org/10.1038/nature12593.

[102] Stephanie C Weber, Julie A Theriot, and Andrew J Spakowitz. « Subdiffusive motion of a polymer composed of subdiffusive monomers ». In: *Physical Review E* 82.1 (2010), p. 011913.

[103] Assaf Amitai and David Holcman. « Polymer model with long-range interactions: Analysis and applications to the chromatin structure ». In: *Physical Review E* 88.5 (2013), p. 052604.

[104] Hicham Saad et al. « DNA Dynamics during Early Double-Strand Break Processing Revealed by Non-Intrusive Imaging of Living Cells ». In: *PLoS Genetics* (2014). ISSN: 15537404. DOI: 10.1371/journal.pgen.1004187.

[105] Osamu Masui et al. « Live-cell chromosome dynamics and outcome of X chromosome pairing events during ES cell differentiation ». In: *Cell* (2011). ISSN: 00928674. DOI: 10.1016/j.cell.2011.03.032.

[106] Josef Redolfi et al. « DamC reveals principles of chromatin folding in vivo without crosslinking and ligation ». In: *Nature Structural and Molecular Biology* (2019). ISSN: 15459985. DOI: 10.1038/s41594-019-0231-0.

[107] Makoto Matsumoto and Takuji Nishimura. « Mersenne Twister: A 623-Dimensionally Equidistributed Uniform Pseudo-Random Number Generator ». In: *ACM Transactions on Modeling and Computer Simulation*. 1998. DOI: 10.1145/272991.272995.

[108] W. H. Payne, J. R. Rabung, and T. P. Bogyo. « Coding the Lehmer pseudo-random number generator ». In: *Communications of the ACM* (1969). ISSN: 15577317. DOI: 10.1145/362848.362860.

[109] B. W. Silverman. *Density estimation: For statistics and data analysis*. 2018. ISBN: 9781351456173. DOI: 10.1201/9781315140919.

[110] David P. Doane. « Aesthetic frequency classifications ». In: *American Statistician* (1976). ISSN: 15372731. DOI: 10.1080/00031305.1976.10479172.

Appendix

A

Cell lines

A.1 ANCHOR system

Cell lines and microscopy were performed in Thomas Sexton's lab, mostly by Dominique Kobi. The following protocol was provided by Tom.

A.1.1 ES cell culture and transgenic lines

J1 mouse ES cells were grown on gamma-irradiated mouse embryonic fibroblast cells under standard conditions (4.5 g/L glucose-DMEN, 15% FCS, 0.1 mM non-essential amino acids, 0.1 mM beta-mercaptoethanol, 1 mM glutamine, 500 U/mL LIF, gentamicin), then passaged onto feeder-free 0.2% gelatin-coated plates for at least two passages to remove feeder cells before subsequent transfections. The two ("inter-TAD" and "intra-TAD") ANCHOR transgenic lines were generated by sequential CRISPR/Cas9-mediated knock-in experiments in the following manner. First, flanking homology arms (mm9 chr6: 52,320,061-52,321,144, and chr6: 52,321,145-52,322,244) were introduced by PCR amplification and Gibson assembly into a vector containing ANCH1 sequence [104]. This vector (1 μg) was co-transfected with 3 μg of a vector containing Cas9-GFP, a puromycin resistance marker and the scaffold to transcribe the sgRNA specific to the T2 insertion site (CGGCGCGCACTTAACACCAA; vector generated by the IGBMC Molecular Biology platform) in 1 million cells with Lipofectamine-2000. Two days after transfection, the cells were cultured for 24 h with 3 $\mu g/ml$ puromycin, then 48 h with 1 $\mu g/ml$ puromycin to enrich for transfected cells, before sorting individual GFP-positive cells on to feeders to amplify individual clones. Clones with the correct sequence were screened by PCR and sequencing, then

the CRISPR knock-in process was repeated to insert the ANCH3 sequence [11] into either the T1 site ("inter-TAD" line; homology arms at chr6: 52,013,471-52,014,370 and chr6: 52,014,371-52,015,270; gRNA sequence AATCGAGCTCACGCCATTAG) or the T3 site ("intra-TAD" line; homology arms at chr6: 52,622,955-52,623,855 and chr6: 52,623,856-52,624,755; gRNA sequence TATGCTGAGGCGTGTCGCAA). Final clones were verified for maintained pluripotency by qRT-PCR to assess Oct4, Nanog and Sox2 expression. Subsequent microscopy experiments (see below) confirmed heterozygous incorporation of the ANCH sequences (detection of one specific spot per ANCH sequence per cell) within the same allele (two spots were always in close proximity).

A.1.2 OR transfection

150,000 cells are plated two days prior to imaging off feeder cells onto laminin-511-coated 35 mm glass bottom petri dishes, and transfected with 3 μg OR1-EGFP and 3 μg OR3-IRFP plasmids using Lipofectamine-2000. After two days, the medium is exchanged for imaging medium and ready for microscopy.

A.1.3 ES differentiation/Hox induction

ES cells were passaged without feeders and cultured on laminin-511 for two days without LIF, then for a subsequent three days without LIF and with the addition of 5 μM retinoic acid. One day after the addition of retinoic acid, the cells are transfected with the OR proteins as previously.

A.1.4 Image acquisition

Imaging was performed on an inverted Nikon Eclipse Ti microscope equipped with a PFS (perfect focus system), a Yokogawa CSU-X1 confocal spinning disk unit, two sCMOS Photometrics Prime 95B cameras for simultaneous dual acquisition to provide 95% quantum efficiency at 11 μm x 11 μm pixels and a Leica 100x oil objective (HC PL APO 1,4 oil immersion). We excited EGFP and IRFP with a 491 nm ($\sim$ 100mw) and a 635-nm laser (>28mW),respectively. We detected green and far red fluorescence with an emission filter using a 525/50 nm and a 708/75 nm detection window, respectively. A thermostated heater (Tokai Hit Stage Top Incubator) allowed for heating at 37 C, humidity and CO2 control (5%). Time-lapse analysis of GFP and IRFP foci was performed in 2D acquiring 241 time points at a 0.5 s time interval. The system was controlled using Metamorph 7.10 software.

A.2 TetO system

The experiments were conducted by Attila Oravecz, working in the lab at the time. The following protocol was provided by him.

A.2.1 Cell culture

The mouse ES cell line was kindly provided by Dr. Luca Giorgetti. It is derived from a X0 clone of the PGKT2 sub-clone of the feeder-independent PGK12. This mESC line was engineered by co-transfection with pBROAD3-TetR-ICP22NLS-eGFP and pcDNA3. Hygromycin selection (250 $\mu g/ml$) was used to provide stable expression of TetR-eGFP recombinant protein after random integration [105, 106]. The piggy-bac transposon system was then used to generate cells with 20-25 stable random integrations of 150 TetO binding site array as described in [106]. Cells were cultured on 0.1% gelatin-coated culture plates in DMEM (4,5g/l glucose) supplemented with GLUTAMAX-I, 15% fetal calf serum (ESC culture tested), 0.1 mM beta-mercaptoethanol, 1,500 U/ml leukemia inhibitory factor (produced in house) and 0.1 mM non-essential amino acids in 5% CO_2 at 37^o C. Mitotic arrest was performed by treating the cells for 5 h with 100 ng/ml Nocodazole (Sigma, M1404-2MG).

A.2.2 Live cell imaging

35 mm glass-bottom dishes (Ibidi 81158) were coated with 10 $\mu g/ml$ fibronectin human plasma (Sigma, F2006-1MG) in PBS for 45 minutes at room temperature. 3-5x105 cells were seeded one day before imaging, then the medium was replaced by phenol-red-free medium containing 500 ng/ml Hoechst 33342 (Invitrogen, H3570). Mitotic arrested cells were collected on the day of imaging by "shake-off", incubated with 0.25% Trypsine-1mM EDTA (Invitrogen, 25200-072) for 1 min at 37^o C and washed, and placed on fibronectin-coated glass-bottom dishes in phenol-red-free medium containing 100 ng/ml Nocodazole and 500 ng/ml Hoechst 33342. Confocal life cell imaging was performed on a Nikon Eclipse Ti-E inverted wide-field microscope (Perfect Focus System) equipped with a CSU-X1 confocal scanner unit and an Evolve back-illuminated EMCCD camera (Photometrics). Images were recorded using 100x HC Plan APO oil immersion objective (Leica, NA 1.4). Intensities were set to 10% for the 405 nm and 30% or 50% for the 491 nm lasers, with exposure times of 100 ms and 50 ms or 25 ms, respectively. 5 z-stacks with 0.5 μm distances were recorded for each channel. 301 time-laps images were recorded only in the 491 channel.

B

Box-Muller algorithm

The generation of normally distributed random numbers is required for many numerical applications. Oppositely, most of pseudo-random number generators used for computations produce uniform distributed numbers via bit operations [107, 108]. Gladly, there are methods to convert these uniformly distributed numbers into which-ever distribution we might be interested. Most modern programming languages offer these algorithms implicitly, but I thought it would be useful to describe one of these Box-Muller algorithm. The Box-Muller algorithm converts uniform sample into the normal space.

As the math for single dimension normal distribution is complicated, we shall simplify the problem by using 2 dimensions in polar coordinates. Like so, we can write the following relationship be cumulative distributions

$$\int_0^R \int_0^\lambda d\theta \; dr \frac{r}{2\pi\sigma^2} \exp\left\{-\frac{r^2}{2\sigma^2}\right\} = \int_0^U \int_0^V du \; ds, \tag{B.1}$$

where the left term corresponds to the cumulative normal distribution, while on the right we have the cumulative uniform distribution. Upon integration we have

$$\frac{\lambda}{2\pi} \left\{1 - \exp\left\{-\frac{R^2}{2\sigma^2}\right\}\right\} = UV. \tag{B.2}$$

Without loss of generality, we can split this result as

$$R = \sqrt{-2\sigma^2 \ln V'} \tag{B.3}$$

$$\lambda = 2\pi U \tag{B.4}$$

where $V' = 1 - V$, which is a uniform random number by itself. Converting back to original variables we have

$$x_1 - \mu_1 = R\cos\lambda \tag{B.5}$$

$$x_2 - \mu_2 = R\sin\lambda. \tag{B.6}$$

Using this algorithm we produce 2 normal random numbers with variance σ^2, but we can modulate the mean as preferred.

C

Kernel density estimation

In several density plots throughout this thesis, a smoothed distribution function for measured or simulated data is presented. Those plots were generate using something called a kernel density estimation. At the core of this method, we assume that there is a continuous distribution that is generated by the convolution of probability density functions for each data point. Mathematically, we can write it like so

$$f(x) \propto \sum_{n=1}^{N} K\left(x|\chi_n, h_n\right), \tag{C.1}$$

where we have N data points $\boldsymbol{\chi}$ each with error $\boldsymbol{h}$. There are several functions we can use as models for K, a few examples are the uniform, triangular, Epanechnikov, among other distributions. For this thesis, we chose the normal distribution for boundless datasets.

Regarding $\boldsymbol{h}$, we have 2 options. Naturally, estimating the error for each measurement would be the option, but that is not always possible. Hence, we could assume that all the points have similar associated error and set $\boldsymbol{h}$ by hand, which is sometimes the best option. Nonetheless, due some properties associated with normally distributed error, there is a "rule-of-thumb" we can use [109]

$$h \approx 1.06 \, \frac{\text{var}\left[\chi\right]}{n^{1/5}}. \tag{C.2}$$

In figure (C.1), we sampled 100 points with probability described by curve in black. The histogram was calculated using the Doane binning method [110]. The curve in orange was calculated using equation (C.1) with h_n given by rule-of-thumb (C.2) and further normalized.

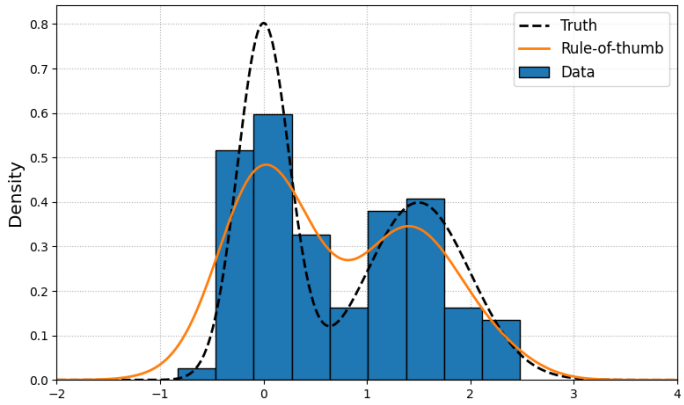

Fig. C.1.: Example were generated using 100 data points sampled from density in black. Kernel estimation in orange was calculated using the rule of thumb.

A few modifications should be done if a similar treatment is to be performed for semi-defined parameters. In that case, we assumed values to present lognormal error, hence we convert measurement into normal space as follows

$$f_h(x)dx \propto \sum_{n=1}^{N} \mathcal{N}\left(\ln x | \ln \chi_n, h_n\right) d\ln x = \sum_{n=1}^{N} \mathcal{L}\left(x | \ln \chi_n, h_n\right) dx. \tag{C.3}$$

Like so, we can still use the rule-of-thumb.

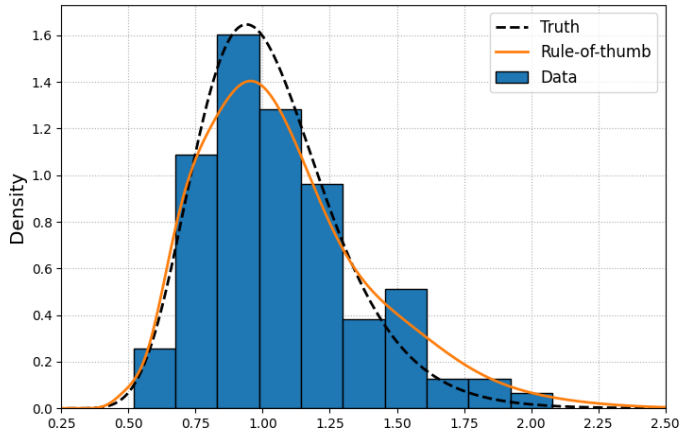

Fig. C.2.: Example were generated using 100 data points sampled from the lognormal distribution in black. Kernel density estimation in orange was calculated using the rule of thumb.

In figure (C.2), we sampled 100 lognormally distributed points from black density function. Continuous approximation was calculated using equation (C.3), where $\boldsymbol{h}$ was estimate using the rule-of-thumb over converted dataset.

Guilherme MONTEIRO OLIVEIRA

Modeling chromatin dynamics using Gaussian processes and polymer physics

Résumé

Comprendre l'organisation de la chromatine et son rôle dans la régulation des gènes sont d'une importance majeure, mais sa dynamique a été largement négligée jusqu'à ces dernières années. Je présente des résultats concernant les propriétés dynamiques de la chromatine dans diverses étapes du cycle cellulaire et une connexion possible entre l'activité des gènes et les propriétés de diffusion locale. Je développe la GP-FBM, une nouvelle méthode basée sur les processus gaussiens et le mouvement brownien fractionnel, que infère les coefficients de diffusion apparente et d'anomalie avec plus de précision que d'autres méthodes populaires et corrige pour les mouvements de fond. Je présente également un nouveau modèle de biopolymère dans laquelle les cartes Hi-C sont utilisées pour modéliser les interactions à longue portée de la chromatine. En outre, les données ChIP-seq sont utilisées pour calibrer les propriétés locales de l'environnement nucléaire. Ce modèle a permis de récapituler les résultats expérimentaux pour certains loci du domaine HoxA dans des cellules de souris.

Mots-clés : processus gaussien, mouvement brownien fractionné, GP-FBM, statistiques bayésiennes, physique des biopolymères, dynamique de la chromatine, HoxA, interphase, mitose

Summary

Understanding chromatin organization and its role in gene regulation is of major importance, however its underlying dynamics has been overseen up to recent years. Here I present results regarding dynamical properties of chromatin in diverse stages of the cell cycle and a possible connection between gene activity and local diffusion properties. I develop a new computational framework based on Gaussian processes and fractional Brownian motion called GP-FBM. This method infers apparent diffusion and anomalous coefficients more accurately than other popular methods and corrects for confound background movement. I further introduce a new biopolymer model using a mean-field approach in which Hi-C maps are used to model chromatin long-range interactions. Further, ChIP-seq data is used to calibrate local properties of the nuclear environment. This model was able to recapitulate experimental results for specific loci of the HoxA domain in mouse cells.

Keywords: Gaussian process, fractional Brownian motion, GP-FBM, Bayesian statistics, biopolymer physics, chromatin dynamics, HoxA, interphase, mitosis