Développement de méthodes mathématiques pour l'analyse de trajectoires conformationnelles en dynamique moléculaire

Sharad Goulam Abas

Thèse de doctorat

NNT: 2020UPASN021

université PARIS-SACLAY

Développement de méthodes mathématiques pour l'analyse de trajectoires conformationnelles en dynamique moléculaire

Thèse de doctorat de l'Université Paris-Saclay

École doctorale n° 574 Mathématiques Hadamard (EDMH)
Spécialité de doctorat: Mathématiques aux interfaces
Unité de recherche: Centre Borelli (CGB UMR 9010 Centre Borelli)
Référent: : ENS Paris-Saclay

Thèse présentée et soutenue en visio-conférence, le 23 Juin 2020, par

Sharad GOULAM ABAS

Composition du jury:

Philippe Derreumaux Professeur, Univsersité de Paris	Président
Alessandra Carbone Professeure, Sorbonne Université	Rapportrice
Frédéric Cazals Directeur de recherches, INRIA Sophia-Antipolis	Rapporteur
Aurélien de la Lande Chargé de recherches HDR, Université Paris-Saclay	Examinateur
Yann Ponty Chargé de recherches HDR, Institut Polytechnique de Paris	Examinateur
Alain Trouvé Professeur, Université Paris-Saclay	Directeur
Luba Tchertanov Directrice de recherches, Université Paris-Saclay	Codirectrice

Remerciements

Je remercie très vivement Madame Alessandra Carbone et Monsieur Frédéric Cazals d'avoir accepté d'évaluer de cette thèse en tant que rapporteurs, et effectué une lecture attentive et détaillée du manuscrit, pour des retours qui ont été grandement encourageants et profitables à l'amélioration de ce travail.

Je remercie également Messieurs Philippe Derreumaux, Aurélien de la Lande, et Yann Ponty d'avoir accueilli favorablement l'invitation à faire partie du Jury, malgré leurs charges de travail respectives et le contexte actuel bien particulier.

J'en viens à remercier mon directeur de thèse Monsieur Alain Trouvé, pour l'exceptionnelle qualité de son encadrement, sa grande disponibilité, sa patience à toute épreuve et l'extrême bienveillance dont il a su faire preuve à mon égard. Si je ne peux à mon très modeste niveau, que laisser la considération de ses paires témoigner de ses qualités de mathématicien, je peux en revanche saluer sa profonde humanité, et son attachante humilité, qui du haut de son savoir reflètent bien la marque des grands esprits. Travailler aux côtés de Monsieur Trouvé pendant ces quelques années fut incontestablement une expérience des plus formatrices, enrichissantes et inspirantes.

Je remercie également ma codirectrice de thèse Madame Luba Tchertanov de m'avoir accueilli au sein de l'équipe BiMoDyM, et de m'avoir offert des conditions idéales pour réaliser ce travail dans un contexte pluridisciplinaire privilégié. Je remercie également Madame Tchertanov pour toutes nos discussions scientifiques, toujours nourries par quelque nouveau point de vue, ainsi que pour ses très nombreux et pertinents conseils prodigués tout au long de ce travail, plus particulièrement sur les thématiques en lien avec la biologie et la physique. Ce fut pour moi un véritable honneur que d'avoir pu bénéficier de son expertise pendant toute la durée de cette thèse.

Je remercie Monsieur Bernard Chalmond pour son encadrement durant mon stage de Master, ainsi que pour ses nombreuses visites au laboratoires qui ont toujours su apporter un angle rafraîchissant aux problématiques abordées.

Je me dois de remercier très chaleureusement Messieurs Nolan Chatron et Zoltan Palmai, avec qui ce fut un réel plaisir de partager mon environnement de travail pendant plus d'une année, dont je garde d'excellents souvenirs. Par ailleurs, bien au-delà de ses engagements en tant que doctorant au laboratoire, Monsieur Chatron m'a accordé beaucoup de son temps et à de nombreuses reprises, notamment au sujet de questions pratiques, que je n'aurais sans doute jamais su résoudre sans son aide précieuse.

Je remercie également Mesdemoiselles Julie Ledoux, Irène-Mauricette Mendy et Myriam Hanna ainsi que Monsieur Maksim Stolyarchuk pour l'aide qu'ils m'ont apportée. Je remercie également Messieurs Marco Pasi et Atman Kendira pour l'intérêt qu'ils ont manifesté à l'égard de mon travail.

Je remercie Madame Agnès Desolneux pour sa gentillesse, sa bienveillance, et l'oreille attentive qu'elle aura su prêter à mes tourments de doctorant. Ma reconnaissance s'étend à l'Ecole Doctorale de Mathématiques Jacques Hadamard pour m'avoir accordé un financement pendant plus de trois années, ainsi qu'à tous le personnel du CMLA pour leur sympathie.

Je remercie mon entourage le plus proche pour son infaillible soutien durant ces années, qui furent parfois difficiles.

Ma pensée la plus intime va évidemment à Monsieur François Inizan, victime innocente d'un monde malade.

Table des matières

Chapitre 1

Introduction

Cette première partie est consacrée à la formulation des questions abordées au cours de cette thèse, et de leur contextualisation vis-à-vis des enjeux contemporains en biologie structurale computationnelle ainsi qu'à une présentation des méthodes classiquement employées et de leurs contraintes. La dynamique moléculaire (DM), et notamment la simulation de trajectoires dites de DM, vise à étudier la fonction d'une protéine au regard de son comportement dynamique : il s'agit aujourd'hui de l'approche la plus couramment employée en biologie et en bio-médecine. A ce titre, de nombreux efforts ont été mis en oeuvre afin de développer des outils de simulation sophistiqués (AMBER, GROMACS, CHARMM, NAMD), lesquels s'appuient sur les progrès de l'informatique (HPC clusters, GPU). Aujourd'hui, les équipes de bio-informatique sont ainsi en mesure de générer une très grande quantité de données. En revanche, les outils visant à exploiter ces données (RMSD, RMSF, MSM, clustering) et à en extraire des informations pertinentes sur le plan biologique présentent certaines limites.

1.1 Protéines : structure, dynamique et fonction

1.1.1 Structure des protéines

Les protéines (du grec *prôteíon*, qui signifie "tout premier") sont des macromolécules présentes chez tous les êtres vivants. Intervenant dans de nombreux processus tels que la production d'énergie, le transfert de différentes molécules à travers la membrane des cellules (transport), la modulation de l'activité d'autres protéines (régulation), la capture de signaux extérieurs et le contrôle de leur transmission dans la cellule (signalisation), ou encore d'autres fonctions biochimiques comme la catalyse des réactions chimiques (enzymes). Les protéines peuvent être localisées à l'intérieur de la cellule, à l'extérieur ou dans la membrane plasmique.

La structure des protéines est décrite sur quatre niveaux : on parle des structures primaire, secondaire, tertiaire et quaternaire. Tout d'abord, une protéine est constituée d'une séquence d'*acides aminés*, reliés entre eux par des *liaisons peptidiques* dans un ordre précis. Chaque acide aminé comporte un groupement d'atomes constant formant la *chaîne principale*, et une partie variable appelée *chaîne latérale* ou *résidu*, déterminant la nature de l'acide aminé. Cette partie variable est liée à la chaîne principale au niveau de l'atome de carbone-α (C_α) de cette dernière par des liaisons covalentes. La nature des protéines est déterminée avant tout par leur séquence

d'acides aminés, laquelle constitue leur *structure primaire* : il s'agit d'une lecture "sans relief" de la protéine. Néanmoins, les acides aminés ayant des propriétés chimiques très diverses, leur disposition le long de la chaîne polypeptidique détermine l'arrangement spatial de la protéine.

Les atomes de C_α des acides aminés présentent une asymétrie en ce sens qu'ils possèdent les deux groupements *amine* (N-H) et *carboxyle* (C-O) permettant la liaison à l'atome N et à l'atome C de la liaison peptidique respectivement. Si la liaison peptidique en elle-même demeure rigide (on parle de *plan peptidique*), elle peut néanmoins faire l'objet de rotations au niveau de ses liaisons avec les atomes de C_α, décrites par les angles dihèdres Φ et Ψ correspondant respectivement aux liaisons N-C_α et C-C_α (voir Fig. 1.1)[28], ou bien de torsions au niveau du plan peptidique (angle ω). La *structure secondaire* décrit ces arrangements locaux lesquels résultent d'un repliement local de la protéine dû à des interactions stériques et électrostatiques, stabilisées par des liaisons hydrogène, grâce aux interactions faibles de type *van der Waals*(voir Chapitre 1 §1.3.1). On distingue plusieurs motifs structuraux caractéristiques observés en biologie : les hélices α, les brins β, les coudes γ, etc.

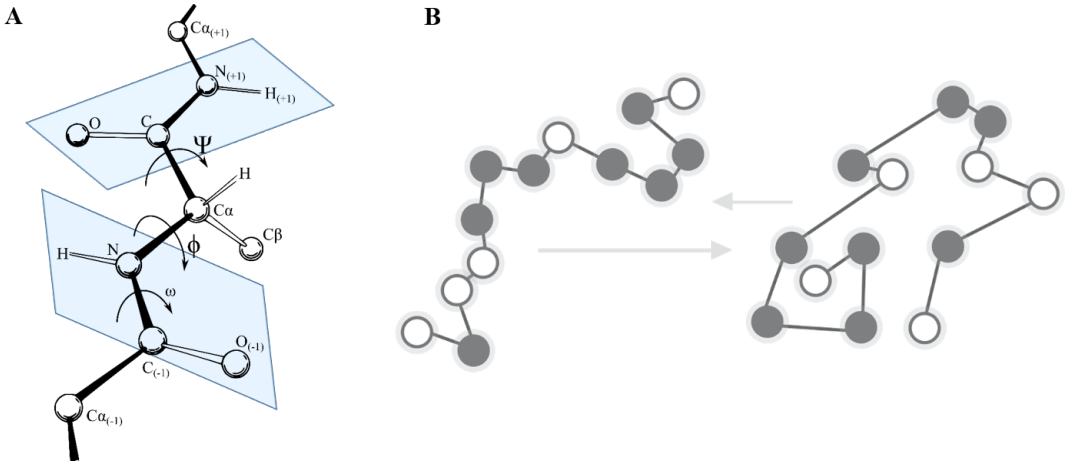

FIGURE 1.1 – Structure secondaire d'une liaison peptidique déterminée par ses trois angles Φ, Ψ et ω. A : angles dihèdres Φ et Ψ et angle ω ; B : différentes configurations possibles d'une chaîne peptidique.

La *structure tertiaire* décrit alors les repliements de la protéine toute entière que celle-ci peut adopter dans l'espace : on parle de *conformations spatiales*. Celles-ci correspondent à leur état natif qui détermine leur activité biologique. La structure tertiaire d'une protéine peut contenir plusieurs motifs structuraux secondaires comme des *bundles* ou des domaines *coiled-coil*.

La *structure quaternaire* décrit enfin l'assemblage de différentes sous-unités protéiques d'une même protéine ou de plusieurs protéines. La Fig. 1.2 ci-après résume les différents niveaux de structures.

1.1.2 Flexibilité protéique et état fonctionnel

Les protéines présentent une grande variété de mouvements internes, dont les temps caractéristiques vont de la fs (1 fs = 10^{-15} s) à la s. Les vibrations de hautes fréquences correspondent à des déformations au niveau de la longueur des liaisons (*stretching*) ou bien au niveau des angles de

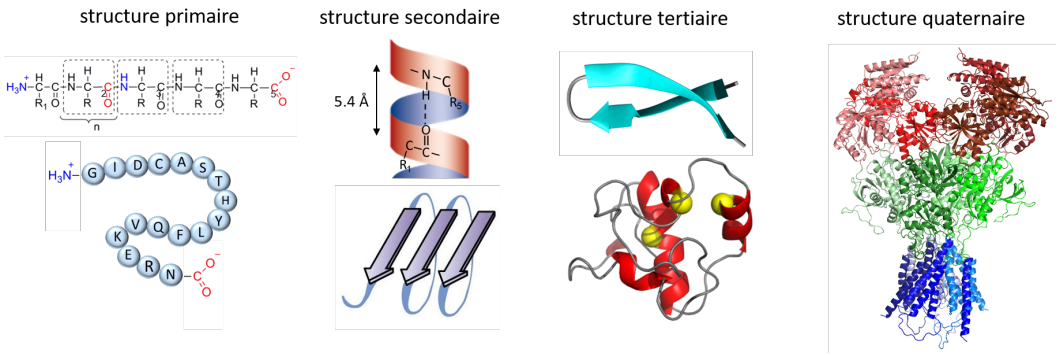

FIGURE 1.2 – Structure des protéines : primaire (lecture des acides aminés), secondaire (repliements locaux), tertiaire (repliements globaux) et quaternaire (complexe protéique).

liaisons (*bending*). Les rotations du squelette peptidique et des chaînes latérales autour des angles dihèdres cités précédemment, se produisent en des temps caractéristiques de l'ordre de 10^{-8} fs. Les mouvements de plus grande amplitude se réalisent quant à eux en des temps plus longs. Les combinaisons de ces mouvements à différentes échelles de temps sont à l'origine de la très grande *flexibilité conformationnelle* des protéines. Comme nous l'avons évoqué précédemment, ces repliements sont dûs à des forces d'interactions tout-atome, lesquelles peuvent être décrites via un *potentiel d'interaction* (voir Chapitre 1 §1.3.1).

Par ailleurs, la fixation d'un ligand sur une protéine peut également induire une transition globale de celle-ci d'une conformation à une autre, modifiant ainsi son affinité envers différents partenaires. Dans le cas d'une enzyme, l'adaptation des chaînes latérales met en évidence un *site actif* (ou *site d'interaction*) sur lequel un substrat viendra se fixer [24]. L'étude de la richesse conformationnelle des protéines flexibles demeure donc un moyen d'approcher de nombreux problèmes biologiques, d'expliquer divers phénomènes physiopathologiques, et d'en proposer des solutions sur le plan pharmacologique [50].

1.2 Exemples de problèmes en biologie structurale

Nous évoquons ici plusieurs types de problèmes biologiques en lien étroit avec la flexibilité des protéines et par conséquent, avec l'espace conformationnel engendré par les objets en jeu. Chacune de ces problématiques sera illustrée à l'aide d'une protéine précise.

1.2.1 Prédiction d'états enzymatiques : exemple de VKORC1

La prédiction d'états enzymatiques d'une protéine jouant un rôle dans l'apparition d'une pathologie peut s'avérer cruciale dans la conception des médicaments [14]. En effet, la détermination de conformations actives d'une telle protéine permet d'isoler une cible thérapeutique précise et de développer la molécule appropriée qui permettra d'inhiber le site actif au sein de la protéine.

Identifiée par Bell et Matschiner en 1970 [11], la protéine VKORC1 (*Vitamin K epOxide Reductase Complex subunit 1*) a été mise en évidence par l'administration du coumaphène, une molécule anticoagulante développée en 1943. Il a fallu attendre 2004 pour que la protéine VKORC1, initialement considérée comme un complexe multi-protéique, soit identifiée comme une protéine

de 163 acides aminés, située dans la membrane du réticulum endoplasmique (RE) [68, 46].

La protéine VKORC1 effectue le recyclage de la vitamine K, laquelle intervient dans de nombreux phénomènes biologiques tels que la métabolisation osseuse, les processus cancéreux, la réponse inflammatoire, le stress oxidatif, l'activité exocrine du pancréas, ou encore la coagulation sanguine [72]. La protéine VKORC1 fait ainsi l'objet de nombreuses études visant à développer des inhibiteurs anti-vitamine K (les AVKs), lesquels constituent alors des anticoagulants, notamment utilisés chez les patients présentant une pathologie cardio-vasculaire. La lutte contre les rongeurs constitue une autre application des AVKs : la mort de ces derniers survenant en effet très rapidement après l'administration d'un anti-coagulant en raison de leur très faible volume sanguin et d'une fréquence cardiaque élevée.

De manière simplifiée, la transformation biochimique de la vitamine K par la protéine VKORC1 s'effectue en deux étapes : tout d'abord d'une forme époxyde à une forme quinone, puis de la forme quinone à une forme hydroquinone [77]. Un mécanisme de réduction de la vitamine K proposé dans [31] suggère une suite de transformations chimiques impliquant les quatre résidus de cystéine (C) de VKORC1 (C43, C51, C132 et C135), induisant alors une réorganisation conformationnelle et permettant la transition d'un état inactif à un état actif, c'est-à-dire capable de transformer la vitamine K. La prédiction de tels états se heurte néanmoins à l'absence de données expérimentales quant à la structure de VKORC1. Plusieurs modèles théoriques 3D ont été proposés, avec des divergences notamment sur le nombre d'hélices trans-membranaires (TM), au nombre de trois [76] ou quatre [47]. En nous basant sur [20], nous considérerons le modèle de VKORC1 constitué de 4 hélices TMs (voir Fig. 1.3). Ce modèle, constitué également d'une petite hélice HH, présente une boucle luminale (L) reliant TM1 et TM2, ainsi que deux extrémités N- et C- terminales (très mobiles et négligées par la suite) et dont le site actif se situe au niveau des résidus C132 et C135.

En résumé, une bonne compréhension de la protéine quant à sa dynamique au niveau atomique permettra de prédire différents états enzymatiques, dont la connaissance constituera un atout clé pour la conception d'anticoagulants consistant à inhiber le site d'interaction de la vitamine K au sein de la protéine VKORC1.

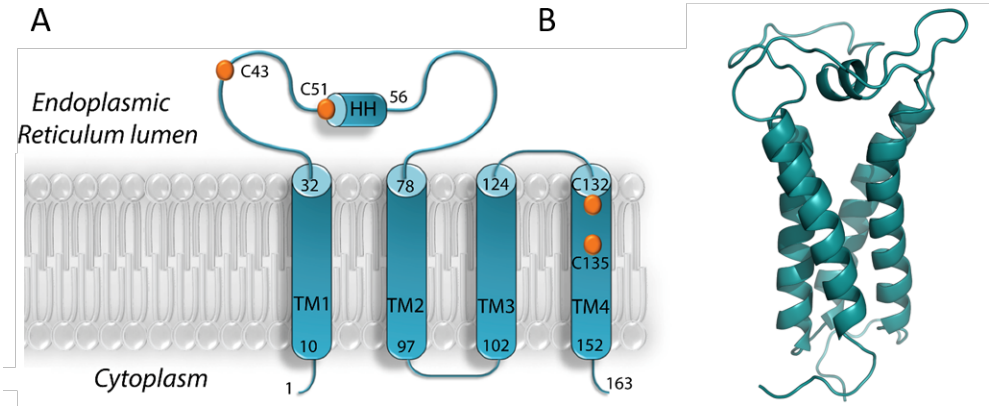

FIGURE 1.3 – A : modèle à 4 hélices de VKORC1 ; B : visualisation PyMol d'une structure 3D de VKORC1 prédite *in silico* [20].

1.2.2 Diversité conformationnelle d'un récepteur à activité tyrosine kinase (RTK) : exemple de KIT

Les récepteurs à activité tyrosine-kinase (RTKs) constituent une famille de récepteurs membranaires ayant la faculté de contrôler la transmission de messages dans la cellule à partir d'un signal extérieur [9]. Les RTKs sont composés d'un domaine extra-cellulaire (ECD) contenant plusieurs domaines (de type immunoglobuline, Ig) reliés par un domaine transmembranaire (TMD) au domaine cytoplasmique (CD) lequel comprend le domaine à activité kinase (KD). De manière simplifiée, suite à la fixation d'un ligand spécifique, on assiste à la dimérisation du récepteur, laquelle induit l'activation du domaine KD en une cascade de signalisations s'effectuant par des réactions de phosphorylation à partir de l'ATP [34]. Les ligands interagissant avec les récepteurs tyrosine-kinase agissent comme des régulateurs de la physiologie des cellules, notamment quant à leur croissance, leurs divisions, etc. On comprend ainsi l'importance de ces récepteurs quant à la régulation du développement d'une tumeur cancéreuse, auquel cas une fois encore, il s'agira de considérer un inhibiteur bloquant ces récepteurs afin de maîtriser la maladie [12].

Parmi ces récepteurs compte le récepteur KIT, dont la dérégulation est à l'origine de différents cancers et de maladies qui consistent en la prolifération anormale de certaines cellules [64][30]. KIT est composé de 5 sous-domaines d'interaction Ig, et son domaine cytoplasmique comprend une région juxtamembranaire (JMR) ainsi que le domaine à activité kinase. Ce dernier présente deux extrémités N-*lobe* et C-terminale, liées entre elles par le domaine à insertion kinase, KID, pour lequel les données expérimentales manquent et dont la structure demeure inconnue (voir Fig. 1.4)[9].

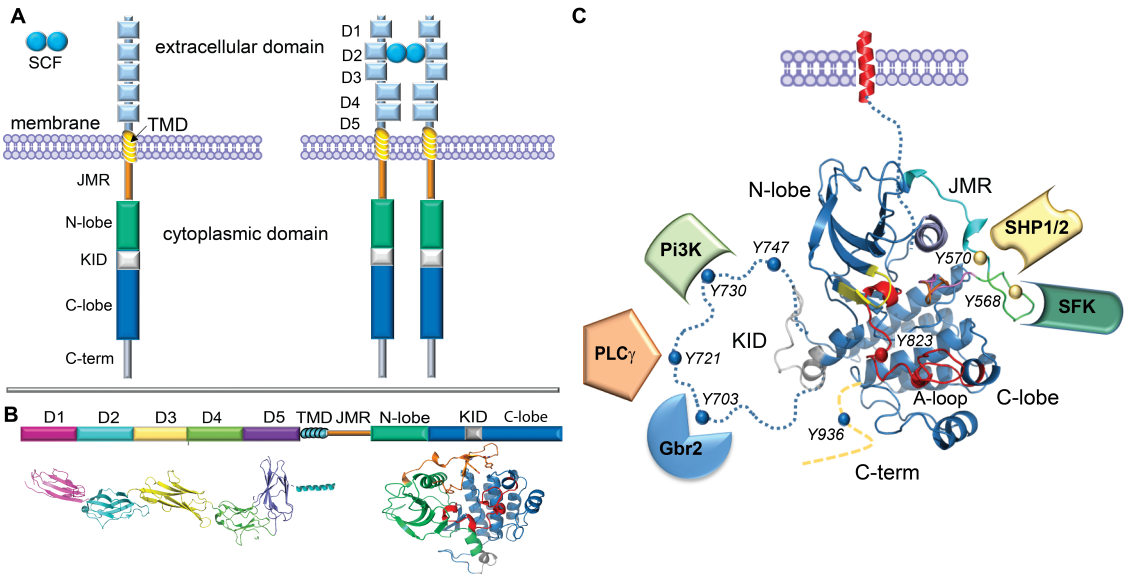

FIGURE 1.4 – Architecture et structure 3D du récepteur KIT. A : organisation de KIT sous forme monomérique (gauche) et dimérique (droite) en fonction de son interaction avec le ligand SCF ; B : domaines structuraux de KIT et leurs structures 3D caractérisées par rayons X ; C : domaine cytoplasmique de KIT dans l'état inactif. Les sous-domaines manquants, le domaine KID, l'extrémité C-terminale et la région JMR sont représentés en pointillés. Les résidus tyrosine (Y) de KID, JMR et C-terminale ainsi que les protéines qui reconnaissent spécifiquement les sites de phosphotyrosine sont également représentés [9].

A ce titre, les travaux de modélisation (*de novo* et *ab initio*) ont permis de retenir un modèle 3D du domaine KID et de C-terminale, de sorte qu'un modèle du domaine cytoplasmique complet dans lequel KID est associé au domaine KD a pu être construit. Les simulations de dynamique moléculaire de ce modèle ont alors été analysées à l'aide des outils classiques (RMSD, RMSF, ACP, modes normaux, corrélations croisées) [35], mais l'importante variabilité conformationnelle observée nécessite une analyse plus poussée.

Remarque 1. *Nous allons dans ce qui suit considérer différents sous-domaines de la molécule KIT complète : en effet, les domaines JMR et C-terminale montrant une importante flexibilité, il peut être bon de ne pas les inclure dans les objets étudiés. Nous résumons au sein du tableau suivant les différents domaines considérés associés aux résidus de C_α correspondants.*

	JMR	KD	KID	C-terminal
Résidus	1-34	35-142, 223-369	143-222	370-400

TABLEAU 1.1 – Différents sous-domaines de KIT considérés et résidus de C_α correspondants.

On parlera ainsi d'une trajectoire de KIT+KID pour désigner la protéine complète formée par la réunion de ces 4 sous-domaines, puis d'une trajectoire de KID seul, de KD+KID, et de KD seul pour désigner les sous-domaines décrits par le tableau précédent.

Remarque 2. *La numérotation précédente ne correspond pas à celle des atomes dans la lecture de la séquence protéique, mais simplement à l'ordre d'apparition des résidus de C_α au sein de cette dernière.*

1.2.3 Canaux ioniques : exemple du récepteur NMDA

Certaines protéines membranaires peuvent effectuer le transport d'ions (Na^+, Cl^-, Ca^{2+}, K^+) par l'ouverture d'un canal ionique contrôlé par les stimulus externes appliqués à une cellule dite *excitable*. Ces canaux sont de la plus haute importance en biologie, et interviennent dans de nombreux phénomènes, tels que la signalisation intra-cellulaire, la communication entre cellules, la contraction musculaire, etc. A ce titre, les cellules excitables et les canaux ioniques formés au niveau de leur membrane ont fait l'objet de nombreuses études visant à construire des modèles structuraux ainsi qu'à simuler des trajectoires dans le but d'observer la formation des canaux [52].

Les neurones comptent parmi ces cellules excitables, et constituent l'unité de base du système nerveux. Les récepteurs NMDA (*N-Methyl-D-Aspartate*) en sont des protéines membranaires participant à la communication synaptique entre ces neurones, et jouant un rôle de tout premier plan dans les mécanismes de potentialisation à long terme (LTP), et de dépression à long terme (LTD), c'est-à-dire de renforcement ou de réduction de la transmission synaptique, régissant les phénomènes d'apprentissage et de mémoire [67]. A ce titre, un dysfonctionnement des récepteurs NMDA a été constaté dans le cas de maladies neuro-dégénératives telles que les maladies d'Alzheimer, Parkinson, Huntington ou encore dans la prolifération de tumeurs cancéreuses [82, 2, 26, 45]. En particulier, les phénomènes de neurotoxicité se traduisent par une activité trop importante et continue des récepteurs NMDA [59].

Les récepteurs NMDA sont des tétramères constitués de deux chaînes obligatoires GluN1 et de deux chaînes GluN2 définissant les caractéristiques du canal. L'activation des récepteurs NMDA,

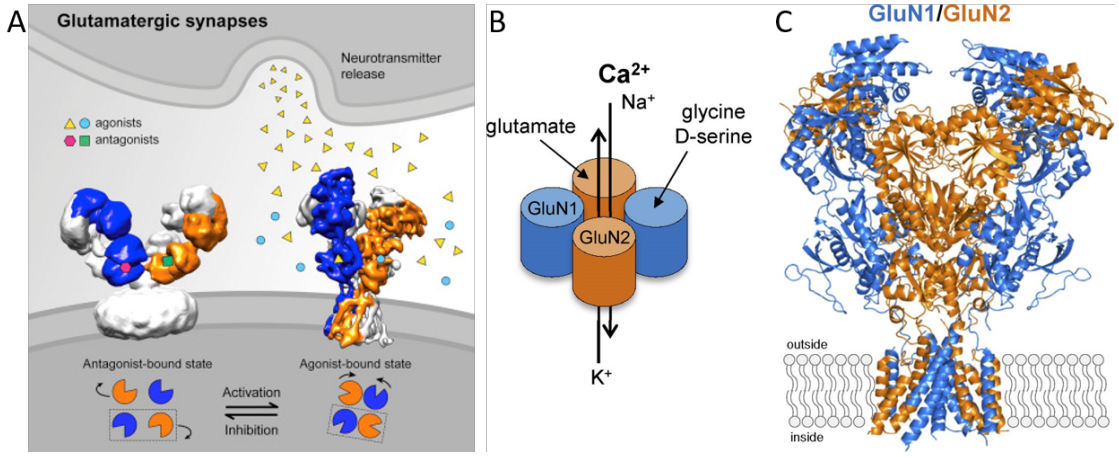

FIGURE 1.5 – Récepteur NMDA. A : représentation de l'activation du récepteur par la fixation de ses ligands agonistes (Glu et Gly/D-serine) ; B : passage de cations à travers le canal ; C : structure 3D d'un récepteur NMDA.

et donc l'ouverture du canal repose sur leur interaction avec deux partenaires : le glutamate et la glycine, la D-sérine pouvant se substituer à cette dernière (voir Fig. 1.5)[83][32].

Si le fonctionnement des récepteurs NMDA a été abordé sous l'angle de la dynamique moléculaire, par le biais notamment de simulations du récepteur avec et sans ligands, diverses questions restent en suspens [60]. En effet, nous ne savons pas si le mécanisme d'ouverture du canal ionique a pu être observé dans son intégralité ou bien si seule une ouverture partielle a pu être constatée, et la question d'un éventuel phénomène de périodicité concernant l'oscillation du récepteur entre deux états équilibrés demeure sans explications. Par ailleurs, de même que pour VKORC1 et pour KIT, les traitements des maladies en lien avec une activité déficiente ou excessive de ces récepteurs reposent sur un mécanisme d'inhibition, qui se traduit ici par le blocage du canal. Pour ces raisons, une étude approfondie des caractéristiques conformationnelles des récepteurs NMDA est requise.

1.3 Modélisation physique et simulation de la dynamique moléculaire

Comme nous venons de le voir à travers les exemples précédents, l'exploration de l'espace conformationnel d'une protéine peut se révéler capitale à maints égards. Cependant le comportement dynamique des protéines est très difficile à observer expérimentalement, voire impossible. De ce fait, le recours à la simulation numérique constitue une approche très repandue en biophysique. La *dynamique moléculaire*, et notamment la simulation de trajectoires dites de DM demeure donc un moyen de tout premier plan pour observer cette variabilité conformationnelle et améliorer notre compréhension du fonctionnement des objets simulés [25, 41].

Pour ce faire, la communauté des bio-informaticiens aborde généralement la question en deux étapes. Dans un premier temps, on procède à la simulation des trajectoires de DM afin d'acquérir un maximum de données, lesquelles correspondent à des suites de conformations adoptées par la protéine, évoluant dans son environnement naturel. Puis, dans un second temps, on cherche à exploiter ces données simulées à l'aide de divers outils et méthodes afin d'en extraire des informations

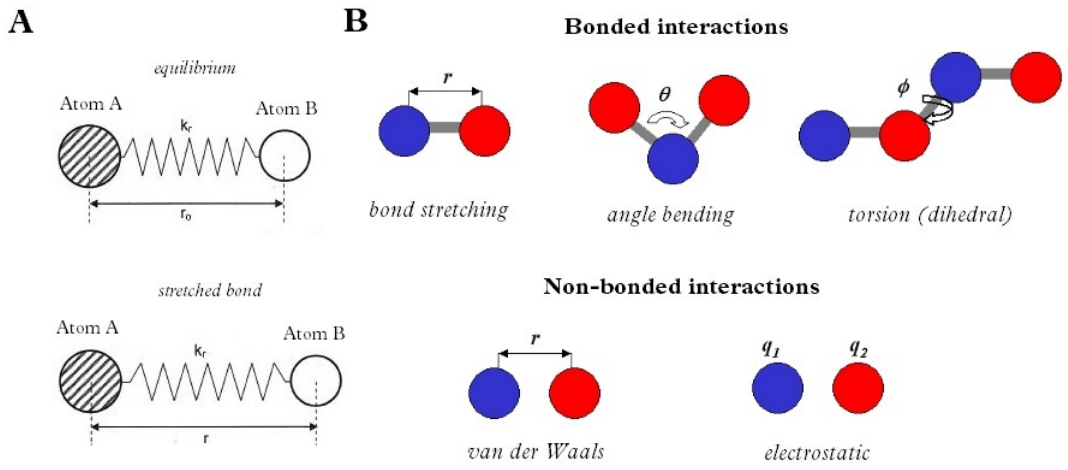

FIGURE 1.6 – A : modélisation d'une liaison covalente entre deux atomes A et B par un ressort de raideur k_r et de longueur à l'équilibre r_0 [8] ; B : représentation schématique des interactions covalentes (*bond stretching, angle bending* et *bond dihedral torsion*) et non-convalentes (*van der Waals* et *electrostatic*)[81].

pertinentes sur le plan biologique. Nous décrivons ici les étapes de productions des trajectoires ; les méthodes d'analyses seront quant à elles décrites plus loin (Chapitre 1 §1.4.2).

1.3.1 Modèle statique d'une protéine et champs de forces

Une première étape consiste à modéliser la structure d'une protéine dans son état statique. De tels modèles peuvent être envisagés directement par des méthodes expérimentales (cristallographie ou RMN), ou bien par homologie avec une protéine similaire à celle que l'on étudie, et dont la structure est déjà connue. A cet égard, des bases de données décrivant de nombreuses protéines via certaines caractéristiques (séquence des acides aminés, coordonnées des atomes dans l'espace, etc.) sont disponibles dans la PDB (*Protein Data Bank*). Néanmoins, certaines protéines ne présentent aucune homologie avec des structures 3D déjà caractérisées expérimentalement. Des méthodes *ab initio* (ou *de novo*) ont donc été développées pour pallier ce manque de données et générer leurs structures.

La structure d'une protéine pourra être représentée par un modèle mécanique dans lequel les liaisons chimiques entre atomes sont traduites par des ressorts (voir Fig. 1.6, A). On distingue deux types d'interactions entre atomes, correspondant aux liaisons covalentes d'une part (i.e. avec partage d'électrons), et aux liaisons non-covalentes d'autre part (i.e. sans partage d'électrons). Les premières interviennent au niveau de l'énergie potentielle de la protéine par des effets d'étirements de liaisons (*bond stretching*), de fléchissement de l'angle de deux liaisons dans un même plan (*angle bending*), ou de torsion, c'est-à-dire de rotation d'une liaison autour de son axe (*bond torsion*). Les interactions non-covalentes comprennent quant à elles les effets de "courte distance" (*van der Waals*) et de "longue distance" (*loi de Coulomb électrostatique*)(voir Fig. 1.6, B). Ces interactions non-covalentes présentent une énergie associée relativement faible. En revanche, étant très nombreuses au sein des protéines, la résultante de toutes les forces engendrées induit une énergie non-négligeable, assurant le maintien de la structure 3D des protéines.

De manière simplifiée, le potentiel d'une molécule peut s'écrire de la façon suivante :

$$V = \underbrace{\sum_{r} K_b(r - r_0)^2}_{V_{\text{bond}}} + \underbrace{\sum_{\theta} K_a(\theta - \theta_0)^2}_{V_{\text{angle}}} + \underbrace{\sum_{\Phi} \frac{V_n}{2} \cos(n\Phi - \gamma)}_{V_{\text{dihedral}}}$$
$$+ \underbrace{\sum_{i>j} \frac{q_i q_j}{\epsilon r_{ij}}}_{V_{\text{electrostatic}}} + \underbrace{\sum_{i>j} A_{ij} \left[(\frac{r^*_{ij}}{r_{ij}})^{12} - (\frac{r^*_{ij}}{r_{ij}})^{6} \right]}_{V_{\text{van der Waals}}}$$

avec K_b constante des forces de liaisons, r_0 distance interatomique à l'équilibre, θ_0 angle à l'équilibre, V_n constante de torsion associée à n (ordre de la série de Fourier), Φ angle dihèdre, q_i charge de l'atome i, ϵ constante diélectrique, r_{ij} distance entre les atomes i et j, A_{ij} profondeur du puits de potentiel approché (minimum de l'énergie), r^*_{ij} distance interatomique au minimum d'énergie.

1.3.2 Simulation de la dynamique moléculaire

Rappelons qu'une protéine présente une plasticité conformationnelle à la source de ses propriétés fonctionnelles comme par exemple la possibilité de s'associer à d'autres molécules (protéines, ADN, ARN, petites molécules). Or, les structures 3D ou modélisées *in silico* représentent un état statique d'une molécule, correspondant à une conformation moyenne (soit la plus forte probabilité d'observer chaque atome à un positionnement donné). Il est donc indispensable d'avoir accès à la dynamique conformationnelle à travers des méthodes de *simulation de la dynamique moléculaire*.

La première simulation de DM a été réalisée dans le but d'étudier des collisions élastiques entre des sphères solides par Alder et Wainwright en 1957 [4]. En 1964, Rahman publie des travaux visant à simuler l'argon liquide et l'eau, et utilise pour ce faire un potentiel continu (Lennard-Jones) tenant compte des nuages électroniques des atomes et de leurs comportements attractifs et répulsifs [66]. La première simulation de DM d'une dizaine de picosecondes (1 ps = 10^{-12} s) d'une protéine (d'environ 500 atomes) dans le vide a été effectuée par A. McCammon [53]. Depuis, les progrès technologiques et informatiques ont abouti à une croissance très significative de la taille des molécules étudiées et de la durée des simulations. Outre ces avantages, les simulations de DM actuelles prennent également en considération l'environnement de la molécule d'intérêt : molécules d'eau, ions, membrane lipidique, éventuel partenaire réactionnel (protéine, acide nucléique, petite molécule, etc.). Ainsi, des simulations de DM de ces systèmes complets et sophistiqués contenant plusieurs centaines de milliers (voire millions) d'atomes, sur des durées de plusieurs centaines de nanosecondes (1 ns = 10^{-9} s) ou de plusieurs microsecondes (1 µs = 10^{-6} s), sont actuellement classiques. De plus, des simulations de l'ordre de la milliseconde (1 ms = 10^{-3} s) ont pu être réalisées par D. Shaw sur *Anton*, supercalculateur conçu pour exécuter des simulations de DM assez longues permettant l'accès à des échelles de temps permettant d'observer des transitions conformationnelles pour des protéines de grande taille [70]. Ces trajectoires de DM sont générées à l'aide de logiciels tels que AMBER (*Assisted Model Building with Energy Refinement* [17]), CHARMM (*Chemistry at Harvard Macromolecular Mechanics* [15]), NAMD (*Nanoscale Molecular Dynamics* [63]), et GROMACS (*Groningen Machine for Chemical Simulations [73]*).

Nous résumons succinctement dans ce qui suit les différents aspects théoriques et pratiques conduisant à la simulation des trajectoires de DM en nous basant sur [7, 21, 1], et en prenant l'exemple de VKORC1 : ces étapes sont néanmoins similaires pour toutes les protéines envisagées.

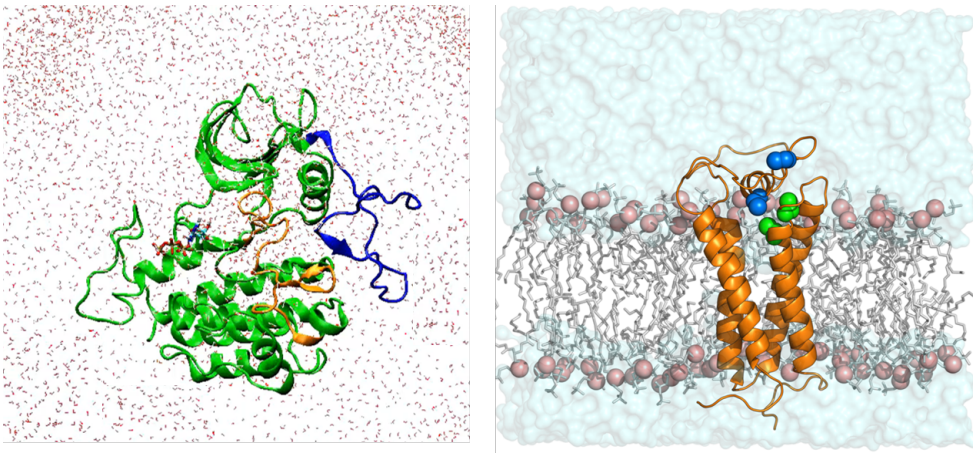

FIGURE 1.7 – Préparation de l'environnement des protéines pour la simulation de DM. Gauche : KIT entourée de molécules d'eau et d'ions négatifs ; Droite : VKORC1 insérée dans la membrane du reticulum endoplasmique, entourée de molécules d'eau et d'ions négatifs.

Préparation de l'environnement

On considère dans un premier temps une boîte de simulation avec des conditions de continuité aux limites, et contenant l'environnement de la protéine. Pour les protéines transmembranaires, un premier point consiste à modéliser la membrane. Par exemple, pour la protéine VKORC1, il s'agit de modéliser la membrane du reticulum endoplasmique laquelle se compose d'une bi-couche lipidique (voir Fig. 1.7)[19]. Cette étape peut-être réalisée grâce à MemBuilder (*Membrane builder* [29]). A la suite de quoi, on incorpore dans la boîte les molécules d'eau qui engloberont la membrane : on parle de *solvatation*. On procède ensuite à la minimisation de l'énergie potentielle du système composé de cette membrane et de son environnement aqueux, via un algorithme de type descente de gradient : cela permet de réajuster la structure 3D du système et ainsi d'opter pour une expérience plus proche de la réalité biologique. En pratique, ceci s'effectue par la réalisation de courtes simulations de DM du système en question. Enfin, on termine la préparation du milieu par des étapes de thermalisation et d'équilibration de pression.

Insertion de la protéine dans son milieu

On peut alors procéder à l'insertion de la protéine dans son milieu. Dans le cas de KIT il s'agit de molécules d'eau et d'ions négatifs pour contrebalancer la charge positive de la protéine. Il en est de même pour VKORC1, à ceci près que la membrane du reticulum endoplasmique s'ajoute à la constitution du milieu. On procède à nouveau aux manipulations précédentes pour calibrer le nouveau système : minimisation de l'énergie du système {eau+protéine+ions (+membrane)} via une descente de gradient (à l'aide d'un nouveau potentiel), thermalisation et équilibration de pression.

Les simulations précédentes ne permettent pas de visiter le paysage conformationnel, mais simplement d'atteindre un minimum local d'énergie potentielle en un minimum d'itérations, et n'ont été ainsi produites que dans le but de compléter le modèle statique de la protéine par la bonne intégration de cette dernière dans son environnement.

Production de la trajectoire de simulation

Nous pouvons désormais envisager de simuler le système sur des plages temporelles plus longues, cette fois-ci dans l'optique d'observer la variabilité conformationnelle de la protéine. On fournit au logiciel un état initial du système d'une part (positions et vitesses), et un champs de force d'autre part : dans notre cas, nous utilisons le champs de force tout-atome CHARMM22 [51] avec le modèle d'eau TIP3P [38]. Si q (resp. p) décrit la *position* (resp. le *moment*) du système total (i.e pour tous les atomes de la protéine et de son environnement), le coeur de la simulation est l'intégration des équations de Newton, c'est-à-dire des équations hamiltoniennes suivantes :

$$\begin{cases} \dot{q} &= \nabla_p H(q,p) \\ \dot{p} &= -\nabla_q H(q,p) \end{cases}$$

associées au hamitonien H (voir Chapitre 2 §2.1.2 (2.2)). Plus précisément, la production d'une trajectoire de DM est obtenue à l'aide d'un intégrateur symplectique : par défaut, nous considérons ici l'intégrateur Leap-Frog [33] de GROMACS) selon un pas de temps typique de $(\Delta t)_{\text{sim}} = 2\,\text{fs}$. Pour tout atome a de masse m_a du système total :

$$\begin{cases} p_a(t + \frac{1}{2}(\Delta t)_{\text{sim}}) &= p_a(t - \frac{1}{2}(\Delta t)_{\text{sim}}) + (\Delta t)_{\text{sim}} F_a(t) \\ q_a(t + (\Delta t)_{\text{sim}}) &= q_a(t) + (\Delta t)_{\text{sim}} \frac{p_a(t+\frac{1}{2}(\Delta t)_{\text{sim}})}{m_a} \end{cases}$$

où $F_a(t)$ désigne la résultante des forces conservatives appliquées à l'atome a à l'instant t. L'utilisateur fixe un nombre de pas à simuler $N_{\text{sim}} - 1$, de sorte à générer un total de N_{sim} conformations observées (comprenant la conformation initiale), pour une durée d'observation $T = (N_{\text{sim}} - 1)(\Delta t)_{\text{sim}}$. La durée d'observation T choisie résulte d'un compromis entre les ordres de grandeurs d'apparition des phénomènes étudiés pour un système précis, et les ressources à disposition en termes de capacité de calcul.

Remarque 3. *Dans notre cas, nous avons eu recours à différents calculateurs : CINES, IDRIS, GENCI, FUSION ou encore TopDyn.*

Remarque 4. *Le logiciel GROMACS dispose d'autres intégrateurs (Verlet-Stoermer [79], Runge-Kutta [16], Beeman [10], intégrateurs avec contraintes, etc), notamment de type Langevin invoquant des thermostats, lesquels traduisent l'interaction du système simulé avec son environnement. En revanche, si ce type de simulation n'est pas adapté au système global, lequel encode déjà le système protéique et son environnement, il ne l'est pas non plus dans le cas où l'on souhaiterait simuler la protéine seule. En effet, une telle simulation nécessiterait de connaître l'énergie libre dont le calcul est difficile d'accès (voir Chapitre 2 §2.1.3 (2.6)).*

Trajectoire de DM résultante

La mise en mouvement des atomes du système à partir de son état initial conduisent donc à l'observation d'une suite de N_{sim} états du système total. Classiquement, ceux-ci sont représentées par une matrice $X \in \mathcal{M}_{d_{\text{tot}}, N_{\text{sim}}}(\mathbb{R})$ où $d_{\text{tot}} = 3 \times n_{\text{tot}}$ avec n_{tot} désignant le nombre d'atomes du système total, protéine et environnement confondus.

Bien souvent, l'utilisateur décide en amont de la simulation de ne conserver qu'un certain nombre de conformations, correspondant à un échantillonage temporel régulier de la matrice précédente. C'est alors qu'il choisit un pas d'échantillonnage Δt, fournissant cette fois-ci $N \leq N_{\text{sim}}$

	Système total (protéine+environnement)	Protéine	Système d'intérêt
VKORC1	58273	2599	143
KIT avec KID	69089	6358	400
NMDA avec ligands	897188	51513	3278

TABLEAU 1.2 – Nombre d'atomes des différents systèmes biologiques considérés.

conformations. On note alors $\mathbf{X} = (x_i)_{1\leq i\leq N} \in \mathcal{M}_{d_{\text{tot}},N}(\mathbb{R})$ la matrice résultante. En pratique, le pas Δt est choisi de sorte à ce que l'on ait toujours exactement que $T = (N_{\text{sim}} - 1)(\Delta t)_{\text{sim}} = (N-1)\Delta t$ (i.e. aucune partie entière n'intervient).

D'autre part, on ne s'intéresse souvent en pratique qu'aux conformations de la protéine en elle-même, et ce faisant, on ne procédera qu'au stockage des lignes de la matrice $\mathbf{X}$ correspondant aux atomes choisis. Plus encore, pour alléger la dimension du problème et gagner en temps de calcul, il sera souvent commode de suivre l'évolution temporelle de la protéine par l'unique considération des atomes de C_α qu'elle comporte : en première approximation, ces résidus suffisent à rendre compte du positionnement de l'ensemble des résidus, diminuant ainsi de près de 20 fois le nombre d'atomes considérés. Dans toute la suite, on désignera ainsi par n le nombre d'atomes du *système d'intérêt*, lesquels seront notés $A = \{a_1, ..., a_n\}$. La dimension de l'espace ambiant correspondant sera notée $d = 3n$, et la trajectoire sera ainsi sauvegardée au sein d'une matrice à nouveau notée $\mathbf{X} \in \mathcal{M}_{d,N}(\mathbb{R})$. A titre d'exemples, le Tableau 1.2 résume la situation pour les systèmes biologiques évoqués précédemment.

Enfin, comme nous allons le voir, cette trajectoire pourra faire l'objet de différents traitements, tels que le recalage en rotation et en translation (recalage rigide), mais également la projection sur un sous-espace de dimension inférieure à d. La matrice $\mathbf{X}$ sera ainsi comprise comme résultant des étapes de production schématisées par la Fig. 1.8.

1.4 Traitement des trajectoires de simulation de DM

1.4.1 Pré-traitements avant analyse

Signalons ici deux pré-traitements à appliquer éventuellement à la trajectoire obtenue en sortie de simulation avant de la soumettre à toute méthode d'analyse : le *recalage en rotation et translation* (recalage rigide) et l'ACP *(Analyse en Composantes Principales)*.

Recalage en rotation et translation (recalage rigide)

Lorsqu'une trajectoire a été simulée, on veut pouvoir comparer les différentes conformations de la protéine entre elles. De sorte à ce que cette comparaison ne soit pas "bruitée" par des effets de translations et de rotations, il est bon de choisir une conformation de référence sur laquelle toute la trajectoire sera recalée : nous exploitons pour cela une méthode basée sur la SVD (*Singular Value Decomposition*) décrite dans [6].

Notons $x^{\text{réf}} = (x_a^{\text{réf}})_{a\in A}$ la conformation de référence choisie (rappel : A désigne ici l'ensemble des atomes d'intérêt uniquement), et $x = (x_a)_{a\in A}$ une conformation à recaler sur $x^{\text{réf}}$. Le problème

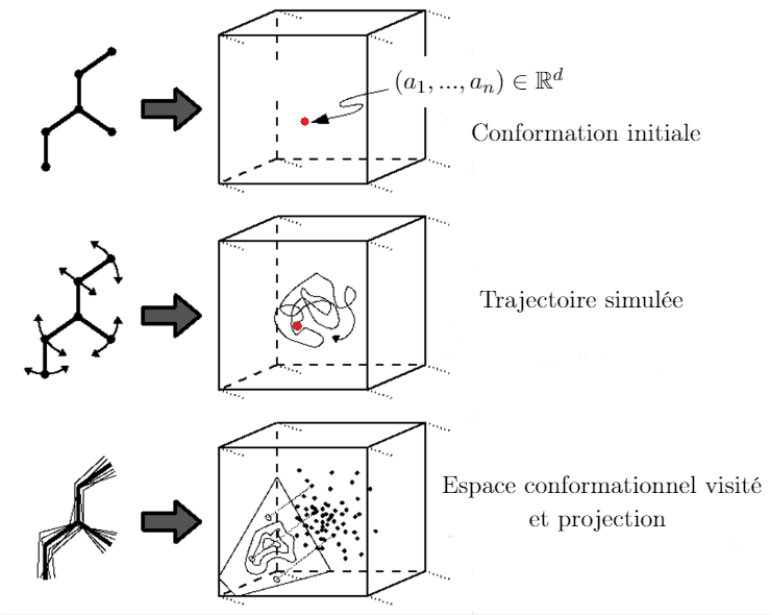

FIGURE 1.8 – Conformation initiale d'une protéine en 3D (représentée par un point de $\mathbb{R}^d$), mise en mouvement correspondant à une trajectoire simulée, et espace résultant avec projection.

revient à chercher la rotation $\Theta^* \in SO_3(\mathbb{R})$ et la translation $\tau^* \in \mathbb{R}^3$ vérifiant :

$$(\Theta^*, \tau^*) = \underset{(\Theta,\tau)\in SO_3(\mathbb{R})\times\mathbb{R}^3}{\operatorname{argmin}} J(\Theta, \tau) \triangleq \sum_{a\in A} m_a \|\Theta x_a + \tau - x_a^{\text{réf}}\|^2$$

où m_a représente la masse de l'atome a. Les détails de cette optimisation sont décrits dans l'Annexe A.2.1.

Analyse en composantes principales

Une fois le recalage effectué, on souhaite désormais effectuer une ACP (*Analyse en Composantes Principales*). Il s'agit de déterminer, par le calcul des vecteurs et valeurs propres de la matrice de covariance d'une trajectoire de simulation de DM, les directions privilégiées par le mouvement des atomes de la protéine, lesquelles seront les plus explicatives de la variabilité du système. On pourra alors projeter la trajectoire dans la base des k premiers vecteurs propres (par ordre décroissant des valeurs propres correspondantes). En plus de réduire la dimension des données, opérer cette ACP en dimension 2 ou 3 constitue un mode de compression des données offrant une visualisation commode à l'utilisateur. Nous décrivons plus précisément la méthode dans l'Annexe A.2.2 [37].

Remarque 5. *Pour comparer deux répliques ou plus, c'est-à-dire des trajectoires partant d'une même conformation initiale mais avec des vélocités différentes, il faudra considérer une conformation de référence sur laquelle toutes les trajectoires que nous souhaitons comparer soient recalées. Quant à la projection, on pourra au choisir des axes arbitraires sur lesquels toutes les trajectoires seront projetées, ou bien effectuer une ACP sur l'une des trajectoire et projeter toutes les autres sur les axes de cette dernière.*

1.4.2 Outils d'analyse

Une fois la trajectoire simulée, échantillonnée, recalée et éventuellement projetée à la guise de l'utilisateur, celui-ci procède maintenant à son exploitation dans le but d'étudier la variabilité conformationnelle de la protéine. Nous résumons ici quelques méthodes parmi les plus utilisées.

Root Mean Square Fluctuation/Deviation (RMSF/RMSD)

L'approche la plus répandue et constituant un protocole standard dans l'analyse des trajectoires est l'emploi des métriques de RMSF (*Root Mean Square Fluctuation of atomic positions*) et RMSD (*Root Mean Square Deviation of atomic positions*)[44], qui se proposent d'étudier respectivement la fluctuation du système au niveau de chaque atome par rapport à une position moyennée dans le temps de l'atome en question, et la déviation du système à chaque instant vis-à-vis d'une conformation de référence (souvent la conformation initiale ou la conformation moyenne). En notant n le nombre total d'atomes du système d'intérêt, $(a_i)_{1\leq i\leq n}$ ses atomes, et N le nombre total de conformations observées au pas Δt, on a les définitions suivantes, pour tout $(i,k) \in [\![1,n]\!] \times [\![1,N]\!]$:

$$\begin{cases} \mathrm{RMSF}(i) = \sqrt{\frac{1}{N}\sum_{k=1}^{N} \|r_{a_i}(k\Delta t) - \langle r_a \rangle\|^2} \\ \\ \mathrm{RMSD}(k\Delta t) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|r_{a_i}(k\Delta t) - r_{a_i}(0)\|^2} \end{cases}$$

où $\langle r_{a_i} \rangle = \frac{1}{N}\sum_{k=1}^{N} r_{a_i}(k\Delta t)$ et où les deux quantités sont calculées en *ångström* (1 Å $= 10^{-10}$ m). On pourra donc dans un premier temps calculer la RMSF de sorte à identifier les régions de la protéine explicatives de la diversité conformationnelle, mais aussi rejeter celles qui font "exploser" la RMSF et qui correspondent à des régions de très forte variabilité. En second lieu, la RMSD permettra de visualiser le niveau d'équilibration de la protéine : le bio-informaticien prendra ainsi appui sur le profil de RMSD afin d'en déduire ce qui lui apparaîtra comme étant un potentiel état d'équilibre.

Ci-dessous sont représentés les tracés de ces deux quantités pour cinq répliques de 100 ns de la protéine VKORC1. Après recalage par rapport à leur même conformation initiale, la RMSF a été calculée à partir des 163 résidus de C_α, et incite à ne conserver que les résidus 9 à 151 : les extrémités N- et C- terminales sont rejetées. La RMSD a ensuite été calculée pour ces résidus restants (voir Fig. 1.9)[19].

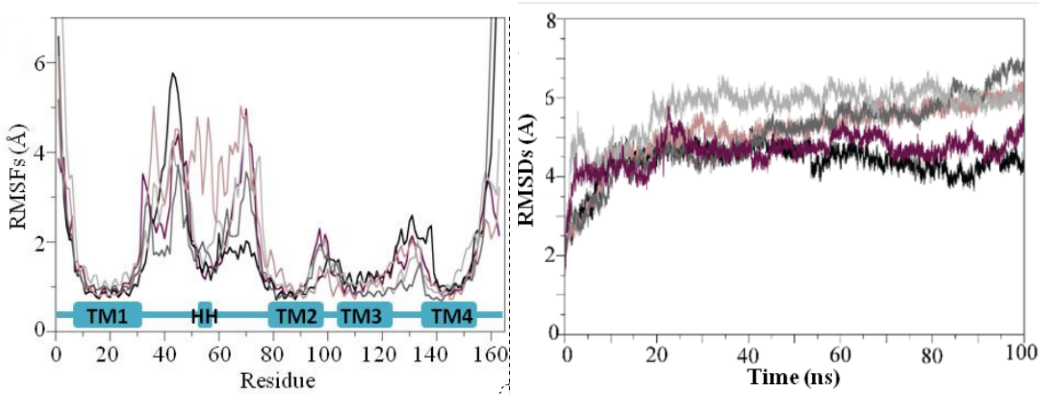

FIGURE 1.9 – Gauche : RMSF calculée sur cinq répliques de 100 ns de VKORC1 (résidus 1-163), et recalées par rapport à $t = 0$ ns ; Droite : RMSD calculée sur cinq répliques de 100 ns de VKORC1 (résidus 9-151), et recalées par rapport à $t = 0$ ns [19].

Algorithmes de clustering

Une autre approche, également très employée, consiste à appliquer un algorithme de clustering aux trajectoires simulées. De très nombreux algorithmes ont ainsi été plébiscités et employés pour l'analyse des trajectoires de DM, et ont notamment fait l'objet de diverses études comparatives vis-à-vis de leurs performances [39, 69, 78]. Nous faisons le choix de présenter ici très brièvement deux d'entre eux : l'algorithme de Jarvis-Patrick, car inclus dans GROMACS, et le *clustering spectral*, car constituant un premier pas vers une modélisation markovienne sur laquelle nous reviendrons.

L'algorithme de Jarvis-Patrick [36] est très couramment utilisé et consiste à regrouper des conformations dans une même classe si leurs listes respectives de plus proches voisins se recoupent suffisamment : nous résumons ici très brièvement son fonctionnement. Cette méthode étant d'essence purement géométrique, il est dans un premier temps indispensable de choisir la métrique appropriée : dans notre cas, la norme euclidienne sur $\mathbb{R}^d$ est un choix naturel. Deux paramètres doivent ensuite être fixés : K et $K_{\min}$, correspondant respectivement au nombre de plus proches voisins à considérer pour chaque conformation, et au nombre minimum de conformations que les listes des plus proches voisins de deux structures doivent posséder en commun afin d'être regroupées dans la même classe. Dans un premier temps, on établit les listes des K plus proches voisins de chaque conformation, puis chaque couple de conformations est affecté à la même classe dans les deux cas suivants : si l'un appartient à la liste des K plus proches voisins de l'autre, ou bien s'ils possèdent $K_{\min}$ plus proches voisins en commun. En appliquant l'algorithme plusieurs fois avec des valeurs croissantes de K, on partira de classes très fines pour se diriger vers un clustering plus grossier, ce qui permet d'obtenir une *hiérarchie*.

Le *clustering spectral* constitue une autre méthode de classification qui a également été abondamment employée en dynamique moléculaire [3]. La classification des données s'opère de la façon suivante [48]. Elle consiste en premier lieu à considérer les données comme un graphe de $\mathbb{R}^d$ et à résumer leurs proximités à l'aide d'une matrice de similarité W, avec par exemple, $w_{ij} = \exp(-\|x_i - x_j\|^2/2\epsilon)$, où $\epsilon > 0$. On définit ensuite la matrice des *degrés* notée $D = \mathrm{diag}(d_i)$ définie par $d_i = \sum_{j=1}^{N} w_{ij}$, puis le *laplacien de graphe* $L = D - W$ (matrice symétrique définie positive). En présence de clusters, on comprend que cette matrice, quitte à réordonner ses colonnes, sera "proche" d'une matrice diagonale par blocs $L = (L_1, ..., L_p)$. Or, chacun de ces blocs pourra

être repéré par un vecteur propre de L prenant la forme d'une indicatrice $\mathbb{1}_{L_i}$, et associé à une valeur propre proche de 0. Le clustering spectral revient alors à identifier les plus petites valeurs propres du laplacien de graphe, les vecteurs propres associés indiquant les clusters. Par ailleurs, de même que dans le cas de l'ACP, on peut montrer que :

$$\lambda_{\min}(L) = \min_{q \in \mathbb{R}^N : \|q\|^2 = 1} \{q^T L q\}$$

Par conséquent, la recherche de la plus petite valeur propre peut être vue sous l'angle de la minimisation du coût inter-classes suivant :

$$C(q) = q^T L q = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (q_i - q_j)^2 w_{ij}$$

Markov State Models (MSM)

Enfin, l'étude du paysage conformationnel peut également se faire via une modélisation de la trajectoire par une chaîne de Markov, ce qui permet d'ajouter une notion de dynamique absente des méthodes de clustering décrites précédemment. On peut alors considérer les données comme un graphe plongé dans $\mathcal{X}$, dont les états seront construits à partir d'une partition de $\mathcal{X}$, et dont les arêtes seront pondérées par des probabilités de transition. L'espace conformationnel visité $\mathcal{X}_N$ sera ainsi compris comme la réalisation d'une chaîne de Markov sur ce graphe : on parle de MSM (*Markov State Models*). Ce type de méthode a été largement employé dans le domaine de la dynamique moléculaire afin d'identifier et caractériser des états stables avec un point de vue statistique : nous nous référons dans ce qui suit à [75, 61, 71, 65].

En résumé, cette approche propose dans un premier temps de partitionner l'espace en des zones disjointes, les *macro-états*, qui sont donc des sous-ensembles de $\mathbb{R}^d$, lesquels peuvent s'obtenir par exemple par le calcul des RMSD de toutes les structures deux à deux. Sur le plan physique, on distingue en effet les micro-états, décrivant des conditions thermodynamiques spécifiques (en terme de température, pression, volume) qu'un système peut adopter avec une certaine probabilité, des macro-états, correspondant à la donnée d'une densité de probabilité sur les micro-états. A la suite de quoi, pour une trajectoire notée $\mathbf{X} = (x_i)_{1 \leq i \leq N}$ vivant dans $\mathbb{R}^d$, on affecte chacune des conformations à l'un des macro-états. On définit alors une chaîne de Markov sur les macro-états de la façon suivante à l'aide d'une matrice de comptage et d'une matrice de transition. Pour tout couple de macro-états A et B :

$$C_{A,B} = \#\{i \in [\![1, N]\!] / x_i \in A \text{ et } x_{i+1} \in B)\}$$

$$P(A, B) = \frac{C_{A,B}}{\#\{i \in [\![1, N]\!] / x_i \in A\}}$$

Dans le cas où certaines zones de l'espace seraient insuffisamment échantillonnées pour avoir une bonne estimation de la probabilité de transition, on pourra avoir recours à une procédure d'*adaptative sampling* : il s'agira de relancer de courtes simulations de DM partant de ces zones afin d'obtenir plus d'information quant au comportement de la trajectoire lorsqu'elle s'y trouve, et ainsi d'affiner le modèle. On pourra également réduire le nombre de macro-états en effectuant une ACP sur la matrice P.

Le modèle ainsi construit (différents tests de validations existent) permet de quantifier des zones de l'espace empruntées par la trajectoire et de les pondérer entre elles, et permet ainsi à l'utilisateur d'avoir une vue dynamique de l'espace conformationnel et d'isoler des macro-états interprétés alors comme des bassins conformationnels.

1.5 Limites des méthodes actuelles

1.5.1 Information dynamique

Les simulations de DM telles qu'elles ont été générées ici constituent des trajectoires ayant un sens *physique*, en ce que chaque système biologique a été simulé dans des conditions proches de la réalité et que chaque trajectoire possède un sens concret et qu'elle peut être interprétée comme un enchaînement de conformations que la protéine aurait pu adopter dans la réalité. De ce fait, les trajectoires considérées ici méritent d'être étudiées en elles-mêmes, plutôt que d'être perçues comme un simple moyen d'accéder au paysage conformationnel visité. Par conséquent, le présent travail s'oriente naturellement vers le parti pris d'une exploitation de l'*information dynamique* que recèle chacune des trajectoires simulées.

A cet égard, de nombreuses méthodes d'analyse des trajectoires de simulations de DM ignorent cet aspect. L'ACP observe les données comme un nuage de points en oubliant tout aspect dynamique. Les méthodes de clustering considèrent également les données sous cet angle : si une dynamique artificielle peut néanmoins être recréée sur les données via une matrice de transition, elle n'exploite pas pour autant le *lien* existant d'une itération faisant évoluer le processus de x_i à x_{i+1}. L'analyse par les modes normaux (NMA [5]), basée sur un modèle mécanique de ressorts (voir Chapitre 1 §1.3.1) permettant d'exhiber différents modes d'oscillations d'une protéine, oublie de même la dynamique de la trajectoire initiale.

De manière générale, et sans même évoquer l'efficacité de ces méthodes, nous constatons simplement qu'elles ignorent l'information dynamique, potentiellement très riche, que contiennent les simulations de trajectoires de DM, pour n'en garder que les conformations visitées sans se souvenir de leur ordre, et sans s'intéresser à la nature de leur dépendance, ne les exploitant ainsi que partiellement.

Remarque 6. *Un autre point de vue consiste à employer les simulations de DM dans le but principal d'étudier le paysage conformationnel relatif à un système biologique donné. Pour ce faire, diverses méthodes existantes consistent non pas à étudier des trajectoires physiques en elles-mêmes, mais à mimer un générateur d'états d'équilibre, se servant de la simulation de trajectoires de DM simplement comme moyen, quitte à ce que celles-ci s'éloignent éventuellement d'une réalité observable.*

1.5.2 Détection d'artefacts

La RMSD opère en quelque sorte une lecture de la trajectoire dans le temps au regard d'une métrique sur l'espace ambiant, ne niant ainsi pas complètement l'aspect dynamique des trajectoires. Cependant, la RMSD fait l'objet d'un autre écueil : si elle semble certes indiquer une plage de temps pendant laquelle le système a atteint un équilibre local, il est très difficile de juger de la pertinence de cet équilibre au regard du reste du paysage conformationnel. Tout d'abord, le

calcul des distances de RMSD masque les phénomènes angulaires en jeu. Par exemple, un système vivant à la surface d'une sphère dans $\mathbb{R}^3$ fournira une valeur de RMSD constante, malgré ses déplacements : une stabilisation de RMSD ne prouve donc en rien qu'un équilibre soit atteint. Par ailleurs, ce que l'utilisateur interprète comme un équilibre peut très bien correspondre à un phénomène très localisé au niveau d'une zone transitoire, et n'ayant ainsi rien à voir avec un équilibre plus global qu'il serait bon de retenir. Des vérifications sont possibles, mais demeurent longues et délicates, et consistent à étudier les conformations identifiées sur le plan structural afin de juger de leur pertinence du point de vue des interactions possibles avec des partenaires chimiques.

Les différents algorithmes de *clustering* évoqués, en plus de considérer les données comme un simple nuage de points, suggèrent également de *faux* clusters à l'utilisateur. Imaginons en effet appliquer un algorithme de ce type à la simulation d'un mouvement brownien dans le plan. L'algorithme va alors détecter des accumulations, et définir des clusters qu'il interprétera comme étant des maximums locaux de la densité de probabilité, c'est-à-dire des minimas locaux de l'énergie libre du système, et donc des conformations stables. Or, puisqu'il s'agit d'un mouvement brownien, il n'existe en réalité aucun maximum local de la densité de probabilité. Nous comprenons donc que toute méthode de classification uniquement basée sur des critères géométriques, ou bien dont les critères dynamiques découlant de l'estimation locale de l'énergie ou de la densité de probabilité sera source de ces même problèmes.

Cela pose également la question du prolongement des trajectoires, qui compte tenu de la lourdeur des simulations sur le plan informatique, est loin d'être anodine. En effet, il est impossible de savoir si la simulation mérite d'être poursuivie alors même que nous ne sommes pas en mesure de juger de la pertinence des bassins identifiés par RMSD ou *clustering*.

1.5.3 Problèmes de modélisation

A l'acquisition des données, un premier réflexe consiste à chercher un modèle adéquat duquel la trajectoire serait issue.

Comme nous le verrons, un certain nombre d'approximations concernant la dynamique du système global constitué de la protéine et de son environnement permettent d'aboutir à une équation de Langevin suramortie qui décrit l'évolution de la protéine seule dans le temps, par l'intermédiaire d'une fonction d'énergie libre (voir Chapitre 2 §2.1.3 (2.8)). Il est ainsi commun de rencontrer la modélisation de la trajectoire d'une protéine par un processus stochastique $X_t \in \mathbb{R}^d$ solution d'une *équation différentielle stochastique* (EDS)[57, 40]. De manière générale, il s'agit d'un processus régi par l'équation suivante, faisant intervenir un terme de dérive $\mu : \mathbb{R}^d \to \mathbb{R}^d$ et un terme de diffusion $\boldsymbol{\sigma} : \mathbb{R}^d \to \mathcal{M}_d(\mathbb{R})$:

$$dX_t = \mu(X_t)dt + \boldsymbol{\sigma}(X_t)dB_t$$

où B_t est un mouvement brownien d-dimensionnel. Bien que ce dernier cadre soit plus large que celui d'une équation de Langevin suramortie, la réalité physique que l'on mime par la simulation de trajectoires de DM est potentiellement bien plus complexe qu'un simple modèle comme celui-ci. Nous verrons en particulier qu'une telle modélisation dépend de l'échelle de temps adoptée avec laquelle on considère la trajectoire.

Comme nous l'avons vu précédemment, les MSM visent également à modéliser la trajectoire par une chaîne de Markov. D'une part, ce type de modèle ne se prête que très mal aux trajectoires

de dynamique moléculaire car l'aspect statistique qu'il présente suppose l'observation de retours récurrents au sein de certains puits de potentiel. Or, il semblerait que ces temps de récurrence soient trop grands (plusieurs µs) pour être observés en pratique lors des simulations de DM. Si les simulations gros grains (GG [42]) permettent de créer des trajectoires plus longues, celles-ci perdent en revanche en précision. A ce titre, le recours à l'*adaptative sampling* permet au processus de repartir des zones peu visitées pour de courtes simulations. Cependant, les nouvelles acquisitions de données constituent des étapes assez lourdes puisqu'elles obligent à repasser par la simulation de données par la dynamique moléculaire. D'autre part, la validation d'un modèle de Markov suppose que le passage à la limite sur la chaîne discrète corresponde bien au processus continu sous-jacent dont les données constituent un échantillon, or comme nous pourrons le voir ultérieurement cela n'a de sens que sous certaines conditions.

1.6 Plan de la thèse

Nous nous concentrerons dans un premier temps sur le problème de modélisation des trajectoires (Chapitre 2). Après avoir présenté le cadre conceptuel dans lequel notre travail se situe, nous verrons qu'il peut être naturel de comprendre les trajectoires à partir des équations de type Langevin dites *suramorties*. A la suite de quoi nous verrons dans quelle mesure il est possible de se rattacher à cette modélisation à partir d'une exploitation pragmatique des données. Nous présenterons ainsi un travail effectué sur les variations quadratiques des trajectoires permettant d'exhiber la distinction de deux régimes EDS/EDO (*équation différentielle ordinaire*) délimités par un pas d'échantillonnage pivot, et soumettant ainsi à condition toute éventuelle modélisation par un processus stochastique : nous en déduirons néanmoins une première approche permettant d'estimer le coefficient de diffusion d'un processus observé en régime EDS. Nous reviendrons ensuite sur la modélisation par chaîne de Markov, en montrant que si celle-ci n'est pertinente que sous certaines conditions vérifiées par le processus sous-jacent, il est néanmoins possible de construire une méthode de re-simulation rapide des trajectoires (visitant les mêmes états) à l'aide d'un tel modèle.

Nous proposons alors la construction d'un nouveau critère permettant la quantification des puits (Chapitre 3). Ce critère, κ, sans dimension et invariant par changements d'échelles de temps et d'espace, permet en outre de juger de la pertinence relative des puits les uns par rapport aux autres, mais également de manière absolue à l'aide d'ordres de grandeurs. De cette façon, nous enrichissons une approche de type RMSD ou clustering en ce sens que ce critère permettra de contourner d'éventuels artefacts. Nous présenterons alors la construction d'un nouvel algorithme de segmentation des trajectoires de simulations de DM basée sur ce critère. Cette nouvelle méthode emploie l'information dynamique des trajectoires en faisant intervenir un coefficient de diffusion estimé tel qu'il peut être envisagé conformément au Chapitre 2, ainsi qu'en proposant des lectures successives de la trajectoire partant de diverses conformations initiales. Nous appliquerons cet algorithme à VKORC1 et à KIT en considérant divers sous-domaines de cette dernière molécule (Chapitre 4).

Nous montrerons comment, à l'aide des informations déduites de l'application de l'algorithme de κ-segmentation à une trajectoire de simulation de DM, celle-ci peut être décrite par un modèle de trois puits (Chapitre 5). Nous exploiterons alors cette façon de procéder pour définir une méthode d'estimation des paramètres dynamiques (dérive et diffusion) d'une trajectoire observée

et modélisée par une EDS, et qui contourne les méthodes d'estimations basées sur la connaissance de l'énergie du sous-système consistué de la protéine, au profit là encore d'une exploitation du contenu dynamique de la trajectoire étudiée. Nous comparerons alors, pour VKORC1 et KIT, les résultats issus d'une estimation de la dérive via la dynamique telle que nous l'envisageons, à ceux que l'on obtient par le calcul du gradient d'une énergie estimée.

Enfin, dans une dernière partie (Chapitre 6), nous présenterons nos conclusions vis-à-vis des méthodes développées, ainsi que des perspectives d'approfondissement.

Chapitre 2

Modélisation stochastique de la simulation de la dynamique moléculaire

Nous abordons dans ce chapitre la question de la modélisation en dynamique moléculaire. Tout d'abord, nous présentons le cadre formel permettant d'appréhender la dynamique moléculaire à travers la notion d'espace conformationnel et les équations de Newton. Nous verrons alors qu'il découle de ce point de vue une modélisation des trajectoires par des EDS de Langevin. A la suite de quoi, par l'étude des variations quadratiques associées à une trajectoire de DM donnée, nous montrons que celle-ci peut-être comprise comme issue d'une EDS seulement pour des pas d'échantillonnage supérieurs à une valeur pivot ; pour des pas situés en dessous de cette valeur, on pourra considérer la trajectoire comme issue d'une EDO. En second lieu, nous reviendrons sur une forme de modélisation par chaîne de Markov sur les conformations observées. Nous verrons que si une telle modélisation n'a de sens que lorsque le processus sous-jacent répond à une équation de Langevin, il est néanmoins possible de l'exploiter afin de construire une méthode de re-simulation, fournissant de nouvelles trajectoires visitant les mêmes données observées, mais dans un ordre différent. Enfin, nous verrons dans quelle mesure il est finalement raisonnable ou non de considérer les trajectoires observées comme issues d'équations de type Langevin.

2.1 Contexte mathématique autour des dynamiques moléculaires

Nous présentons dans ce qui suit une manière d'aborder la dynamique moléculaire sous un angle plus mathématique. Nous formalisons dans un premier temps la notion d'espace conformationnel avant de préciser le point de vue hamiltonien régissant la dynamique des systèmes intégraux considérés. Enfin, nous nous pencherons sur la dynamique du sous-système constitué de la protéine seule, laquelle relève des dynamiques de Langevin, et nécessite de définir la notion d'énergie libre.

2.1.1 Espaces conformationnels

Au coeur des représentations, les entités élémentaires sont des atomes a, codés mathématiquement par leur position r_a dans l'espace ambiant tridimensionnel $\mathbb{R}^3$, ou bien dans un tore tridimensionnel noté $\mathbb{T}_3$ (obtenu en considérant les coordonnées *modulo* les longueurs selon les trois axes d'une boite de simulation rectangulaire). Une protéine constituée de n_p atomes décrite par l'ensemble $A^p = \{a_1^p, ..., a_{n_p}^p\}$, sera ainsi codée par un n_p-uplet $x = (r_a)_{a\in A^p} = (r_{a_i^p})_{1\leq i\leq n_p} \in \mathcal{X}$, avec $\mathcal{X} = \mathbb{R}^{3n_p}$ ou $\mathbb{T}_3^{n_p}$ et vit donc dans un espace de dimension $d_\mathcal{X} = 3n_p$: $\mathcal{X}$ est ainsi l'espace conformationnel associé à la protéine. De manière similaire, une conformation d'un environnement de n_e atomes notés $A^e = \{a_1^e, ..., a_{n_e}^e\}$ sera décrite par un point $y = (r_a)_{a\in A^e} = (r_{a_i^e})_{1\leq i\leq n_e} \in \mathcal{Y}$ où $\mathcal{Y} = \mathbb{R}^{3n_e}$ ou $\mathbb{T}_3^{n_e}$, vivant donc dans un espace de dimension $d_\mathcal{Y} = 3n_e$. Au final, une conformation $q = (x, y)$ du système intégral vit dans l'espace conformationnel produit suivant :

$$\mathcal{Q} = \mathcal{X} \times \mathcal{Y} \tag{2.1}$$

de dimension $d_{\text{tot}} = d_\mathcal{X} + d_\mathcal{Y} = 3(n_p + n_e)$. On peut désormais décrire la physique du système grâce à cet espace, et notamment définir le *potentiel d'interaction* noté $V : \mathcal{Q} \to \mathbb{R}$, défini entre tous les atomes a de l'ensemble $A^{\text{tot}} = A^p \cup A^e$ de chaque conformation $q = (r_a)_{a\in A^{\text{tot}}}$ du système total. En revanche, les questions les plus intéressantes relatives aux propriétés conformationnelles de la protéine relèvent de l'espace réduit $\mathcal{X}$ dans lequel la physique n'est cependant pas exprimée (elle est cependant facilement exprimable dans $\mathcal{Q}$).

Remarque 7. *Dans toute la suite, et pour plus de commodité, la position $r_a \in \mathbb{R}^3$ ou $\mathbb{T}_3$ de l'atome a d'une conformation $q \in \mathcal{Q}$ sera désormais notée q_a, de sorte que $q = (q_a)_{a\in A^{\text{tot}}}$.*

2.1.2 Espace des phases et dynamique hamiltonienne

Afin d'aborder les questions de dynamique, il est naturel pour la mécanique newtonienne qui sera considérée ici, d'introduire pour chaque atome $a \in A^{\text{tot}}$ du système, sa masse m_a, sa vitesse v_a, son *moment* $p_a = m_a v_a$, et la résultante F_a des forces conservatives qui lui sont appliquées. On notera alors $p = (p_a)_{a\in A^{\text{tot}}}$ la famille des moments vivant dans l'espace $\mathcal{P} = \mathbb{R}^{d_\mathcal{P}}$ ($d_\mathcal{P} = d_{\text{tot}}$), et $\mathcal{E}$ *l'espace des phases* défini par le produit :

$$\mathcal{E} \triangleq \mathcal{Q} \times \mathcal{P}$$

Cet espace des phases permet d'exprimer aisément la dynamique du système selon la mécanique newtonienne par l'introduction du *hamiltonien*, noté $H : \mathcal{E} \to \mathbb{R}$, et correspondant à la somme de l'énergie potentielle V et de l'énergie cinétique :

$$H(p, q) \triangleq V(q) + \sum_{a\in A} \frac{\|p_a\|^2}{2m_a} \tag{2.2}$$

L'équation fondamentale de la dynamique $m_a\ddot{q}_a = F_a \triangleq -\nabla_{q_a} V$, fournissant un système de deux équations différentielles du premier ordre portant sur les variations de vitesse et de position, peut être mise en lien avec le hamiltonien précédent en remarquant que pour tout $a \in A$:

$$\begin{cases} \nabla_{q_a} H(p, q) &= \nabla_{q_a} V = -\dot{p}_a \\ \nabla_{p_a} H(p, q) &= \frac{p_a}{m_a} = \dot{q}_a \end{cases}$$

L'expression hamiltonienne de l'équation fondamentale de la dynamique prend donc la forme compacte suivante sur $\mathcal{E}$:

$$\left\{ \begin{array}{rcl} \dot{p} & = & -\nabla_q H(p,q) \\ \dot{q} & = & \nabla_p H(p,q) \end{array} \right. \tag{2.3}$$

Lorsque l'on considère $\mathcal{Q}$ comme un système isolé, la dynamique hamiltonienne correspond aux trajectoires physiques du système sur lesquelles le hamiltonien H est conservé puisque $\dot{H} = \langle \nabla_p H, \dot{p} \rangle + \langle \nabla_q H, \dot{q} \rangle = 0$.

2.1.3 Dynamiques stochastiques

Dynamique de Langevin

Du point de vue de la dynamique moléculaire, les dynamiques stochastiques considérées trouvent leur source dans les travaux d'Einstein et de Langevin sur la théorie de la diffusion moléculaire [58] et du mouvement brownien. Le point de vue moderne est celui des *équations différentielles stochastiques* (EDS) dont le prototype est fourni, pour $\lambda, \beta > 0$, par le système suivant :

$$\left\{ \begin{array}{rcl} dp_t & = & -\nabla_{q_t} V dt - \lambda K p_t dt + \sqrt{2\lambda/\beta}\, dB_t \\ dq_t & = & K p_t dt \end{array} \right. \tag{2.4}$$

où $q \in \mathcal{Q}$ est l'état du système à l'instant donné, $p \in \mathcal{P}$ est le vecteur des moments, $K = \mathrm{diag}((\mathrm{I}_3/m_a)_{a \in A})$ la matrice diagonale par blocs avec I_3 la matrice identité sur $\mathbb{R}^3$, et B_t un mouvement brownien d-dimensionnel. Le facteur $\lambda K p_t$ est un facteur d'*amortissement* ou de *dissipation* (contrôlé par la valeur de λ) et $\sqrt{2\lambda/\beta} dB_t$ un terme d'excitation qui s'interprète comme une *force aléatoire* (agissant sur la dérivée du moment) ou encore *une agitation thermique*.

Cette dynamique de Langevin est associée au hamiltonien précédent, lequel se réécrit en fonction de K de la façon suivante :

$$H(p,q) = V(q) + \frac{p^T K p}{2}$$

Cela permet alors d'interpréter (2.4) comme une *perturbation* de la dynamique hamiltonienne (2.3) :

$$\left\{ \begin{array}{rcl} dp_t & = & -\nabla_{q_t} H(q_t,p_t) dt + \big(-\lambda \nabla_{p_t} H(q_t,p_t) dt + \sqrt{2\lambda/\beta}\, dB_t \big) \\ dq_t & = & \nabla_{p_t} H(q_t,p_t) dt \end{array} \right.$$

dont la mesure invariante sur l'espace des phases $\mathcal{E}$ s'écrit :

$$d\mu_\beta(q,p) \propto \exp(-\beta H(q,p)) dq dp$$

où $\beta = 1/k_B\theta$, avec θ la température du système, et k_B la constante de Boltzmann : on retrouve ainsi une distribution de Maxwell-Boltzmann à la température θ.

Lorsque $\beta = +\infty$ (i.e. température nulle), l'absence d'excitation entraîne une décroissance du hamiltonien (on a alors $\frac{dH}{dt} = -\lambda p_t^T K^2 p_t < 0$) et une convergence vers un minimum local q^* de V. Lorsque $\lambda = \beta^{-1} = 0$, la dynamique déterministe correspond à la version hamiltonienne classique des équations de Newton pour le système q dans le potentiel V, et pour laquelle le hamiltonien est conservé durant la trajectoire. Les dynamiques de Langevin, indexées par (λ, β), peuvent donc être interprétées comme une famille de perturbations stochastiques de la dynamique déterministe hamiltonienne.

Dynamique de Langevin suramortie

Lorsque la durée d'observation est très grande devant le temps caractéristique $1/\lambda$ de l'amortissement, le changement de variable $t \longleftarrow \lambda t$ s'impose et invite à considérer $q_t^\lambda = q_{\lambda t}$ et $B_t = \sqrt{\lambda} W_{t/\lambda}$ où W est un mouvement brownien standard. On obtient alors le nouveau système :

$$\left\{ \begin{array}{rcl} dp_t^\lambda & = & \lambda(-\nabla_{q_t^\lambda} V dt - \lambda K p_t^\lambda dt + \sqrt{2/\beta}\, dW_t) \\ dq_t^\lambda & = & \lambda K p_t^\lambda dt \end{array} \right.$$

Lorsque $\lambda \to +\infty$, ce système converge et la dynamique répond alors à une équation de Langevin dite *suramortie* :

$$dq_t^\infty = -\nabla_{q_t^\infty} V dt + \sqrt{2/\beta}\, dW_t \,. \tag{2.5}$$

Ceci peut se dériver formellement en remarquant que :

$$\begin{aligned} q_t^\lambda &= q_0^\lambda + \int_0^t \lambda K p_s^\lambda ds \\ &= q_0^\lambda + \int_0^t \lambda K e^{-\lambda^2 K s} \left(p_0^\lambda + \lambda \int_0^s e^{\lambda^2 K r} (-\nabla_{q_r^\lambda} V dr + \sqrt{2/\beta}\, dW_r) \right) ds \\ &= q_0^\lambda + \underbrace{(\mathrm{I}_{d_\mathcal{Q}} - e^{-\lambda^2 K t}) \frac{p_0^\lambda}{\lambda}}_{\to 0 \text{ lorsque } \lambda \to \infty} + \int_0^t \underbrace{(\mathrm{I}_{d_\mathcal{Q}} - e^{-\lambda^2 K (t-r)})}_{\to \mathrm{I}_{d_\mathcal{Q}} \text{ lorsque } \lambda \to \infty} \left(-\nabla_{q_r^\lambda} V dr + \sqrt{2\beta}\, dW_r \right) \end{aligned}$$

où $\mathrm{I}_{d_\mathcal{Q}}$ correspond à la matrice identité sur $\mathbb{R}^{d_\mathcal{Q}}$, et où la troisième égalité vient de l'application du théorème de Fubini.

Par construction l'équation de Langevin suramortie admet comme probabilité d'équilibre la marginale ν_β de μ_β sur q :

$$d\nu_\beta(q) = \frac{1}{Z_\beta} e^{-\beta V(q)} dq$$

où $Z_\beta = \int e^{-V(q)} dq$ est une constante de normalisation. Cela se vérifie directement en considérant le générateur infinitésimal $\mathcal{L}$ de l'EDS (2.5). Pour $f \in C_c^2(\mathcal{Q}, \mathbb{R})$:

$$\mathcal{L} f = -\langle \nabla f, \nabla V \rangle + \beta^{-1} \Delta f$$

Si $f = e^{-\beta V}$, alors f vérifie l'équation de Fokker-Planck :

$$\mathcal{L}^* f \triangleq \operatorname{div}(f \nabla V) + \beta^{-1} \Delta f = 0$$

puisque $\beta^{-1} \Delta f = \beta^{-1} \operatorname{div}(\nabla f)$ et $\beta^{-1} \nabla f = -f \nabla V$.

Énergie libre

Dans les dérivations précédentes, la dynamique est définie sur le système intégral, caractérisé par sa conformation $q \in \mathcal{Q}$ et son moment $p \in \mathcal{P}$. La mesure d'équilibre ν_β porte alors sur la distribution du système intégral $q = (x, y)$ qui englobe la conformation de la protéine $x \in \mathcal{X}$ et celle de l'environnement $y \in \mathcal{Y}$ (2.1), et non directement sur x. En particulier on ne peut pas parler de l'énergie potentielle $V_\mathcal{X}(x)$ de la conformation x de la même façon que cela peut être défini pour le système q, puisqu'il faut tenir compte des interactions avec les environnements *possibles* $y \in \mathcal{Y}$.

Pour pouvoir parler d'une énergie potentielle sur $\mathcal{X}$, il faut dans un premier temps considérer la mesure d'équilibre ν_β sur $\mathcal{Q}$, puis définir l'énergie d'une conformation x par rapport à la densité de probabilité de la loi marginale de ν_β sur $\mathcal{X}$. Plus précisément, en introduisant la projection :

$$\begin{array}{rccc} \xi : & \mathcal{Q} & \to & \mathcal{X} \\ & q = (x, y) & \mapsto & x \end{array}$$

on définit la marginale $\nu_\beta^{\mathcal{X}} \triangleq \xi\nu_\beta$ de ν_β sur $\mathcal{X}$, puis $U_{\mathcal{X}} : \mathcal{X} \to \mathbb{R}$ solution (définie à une constante additive près) de l'équation :

$$U_\beta^{\mathcal{X}} \triangleq -\frac{1}{\beta}\log(\frac{d\nu_\beta^{\mathcal{X}}}{dx}) + \text{Cte} \tag{2.6}$$

et qui vérifie donc :

$$d\nu_\beta^{\mathcal{X}} = \frac{e^{-\beta U_\beta^{\mathcal{X}}}}{Z_\beta^{\mathcal{X}}} dx$$

La fonction $U_\beta^{\mathcal{X}}$ est habituellement appelée *énergie libre* et nous noterons ici qu'elle est un objet complexe, qui dépend de la distribution ν_β (elle a donc un sens statistique) et du paramètre β. Notons par ailleurs que son calcul passe par un processus d'intégration puisque l'on a :

$$e^{-\beta U_\beta^{\mathcal{X}}(x)} = \int_{\mathcal{Y}} e^{-\beta V(x,y)} dy + \text{Cte} \tag{2.7}$$

Remarque 8. *On pourra se reporter à [43] pour une présentation des techniques actuelles de calcul de l'énergie libre dans le cas présenté ici et dans celui plus général où $\xi : \mathcal{Q} \to \mathbb{R}^m$ code un certain nombre de variables structurant la dynamique, souvent appelées* coordonnées de réactions *(voir Chapitre 2 §2.3.1).*

Un exemple très simple permet d'illustrer les liens entre le potentiel V et l'énergie libre $U_\beta^{\mathcal{X}}$, celui où le potentiel V s'écrit sous la forme :

$$V(x, y) = V^{\mathcal{X}}(x) + \frac{|y - \zeta(x)|^2}{2\sigma^2(x)}$$

avec $V^{\mathcal{X}} : \mathcal{X} \to \mathbb{R}$, $\zeta : \mathcal{X} \to \mathcal{Y}$ et $\sigma : \mathcal{X} \to \mathbb{R}_+$. A toute conformation x est associé un environnement optimal $y = \zeta(x)$ tel que les potentiels V et $V^{\mathcal{X}}$ coïncident. Par ailleurs, l'éloignement de l'environnement y par rapport à sa valeur optimale induit un écart quadratique entre les deux potentiels V et $V^{\mathcal{X}}$. Enfin le terme $\sigma^2(x)$ agit sur la singularité de l'optimum. Nous avons ainsi pour tout $x \in \mathcal{X}$ que :

$$V^{\mathcal{X}}(x) = \inf_{y \in \mathcal{Y}} V(x, y) = V(x, \zeta(x))$$

Dans ce cas, la loi conditionnelle $\nu_\beta(dy|x)$ de y sachant x sous ν_β correspond à une gaussienne de moyenne $\zeta(x)$ et de variance $\sigma^2(x)$. Un calcul direct partant de (2.6) donne :

$$U_\beta^{\mathcal{X}}(x) = V^{\mathcal{X}}(x) - \frac{n_e}{2\beta}\log(\sigma^2(x)) + \text{Cte}(\beta)$$

Ceci nous montre que lorsque $\beta^{-1} \to 0$ (température nulle), on a que $U_\infty^{\mathcal{X}} = V^{\mathcal{X}} + \text{Cte}$: c'est-à-dire que $U_\infty^{\mathcal{X}}$ correspond, à une constante additive près, au minimum du potentiel V à x fixé. Cependant, à $\beta^{-1} > 0$ fixé (température non-nulle), le terme $n_e \log(\sigma^2(x))/2\beta$ vient déformer le profil de l'énergie libre à température nulle par un terme qui dépend de façon croissante de la variance de la loi conditionnelle $\nu_\beta(dy|x)$. Plus la variance $\sigma^2(x)$ est élevée, plus la loi conditionnelle

de y sachant x est "étalée" autour de l'optimum (i.e. plus il existe d'environnements compatibles avec la conformation x), et plus l'énergie libre décroît. Le terme $n_e \log(\sigma^2(x))/2\beta$ peut ainsi s'interpréter comme un terme d'entropie de la loi conditionnelle venant contrebalancer la valeur de $V^{\mathcal{X}}$ dans le calcul de l'énergie libre.

États métastables et équilibres locaux

Dans l'approximation donnée par la dynamique de Langevin suramortie, nous avons donc que :

$$dq_t = -\nabla_{q_t} V dt + \sqrt{2/\beta}\, dW_t$$

Dans ce cas, le comportement du système à basse température est lié à la dynamique de circulation, dans le paysage d'énergie décrit par le potentiel V, entre des *états métastables* au voisinage de minimums locaux de V. La dynamique s'équilibre localement au niveau de ces minimums avant d'en sortir à proximité de points selles, et ce au bout d'un temps dépendant de la barrière d'énergie qui sépare deux états métastables communicants.

Comme nous l'avons vu, le comportement de x est plutôt lié à l'énergie libre $U_\beta^{\mathcal{X}}$, indirectement liée au potentiel V via la loi conditionnelle $\nu_\beta(dy|x)$. Ce lien dynamique peut se comprendre dans le cas où la dynamique sur y s'équilibre beaucoup plus rapidement que celle de x. En effet, considérons la dynamique précédente qui s'écrit :

$$\begin{cases} dx_t &= -\nabla_{x_t} V(x_t, y_t) dt + \sqrt{2/\beta}\, dW_t^{\mathcal{X}} \\ dy_t &= -\nabla_{y_t} V(x_t, y_t) dt + \sqrt{2/\beta}\, dW_t^{\mathcal{Y}} \end{cases}$$

Essayons alors de remplacer la loi de y_t par la loi conditionnelle $\nu_\beta(dy_t|x_t)$ de y_t sachant x_t sous ν_β, et correspondant à un équilibre local à x_t fixé. Intégrons alors, dans l'équation sur x_t, le gradient $\nabla_{x_t} V(x_t, y_t)$ par rapport à y_t selon cette loi conditionnelle. Ceci donne la nouvelle équation :

$$dx_t = -\left(\int_{\mathcal{Y}} \nabla_{x_t} V(x_t, y_t) \nu_\beta(dy_t|x_t) \right) dt + \sqrt{2/\beta}\, dW_t^{\mathcal{X}}$$

et puisque (2.7) donne $\nabla U_\beta^{\mathcal{X}}(x_t) = \int_{\mathcal{Y}} \nabla_{x_t} V(x_t, y_t) \nu_\beta(dy_t|x_t)$, on tire une nouvelle équation pour x_t qui ne dépend que de l'énergie libre :

$$dx_t = -\nabla_{x_t} U_\beta^{\mathcal{X}} dt + \sqrt{2/\beta}\, dW_t^{\mathcal{X}} \tag{2.8}$$

C'est une nouvelle dynamique de Langevin suramortie dans laquelle le potentiel V a été remplacé par l'énergie libre (qui n'intervient que par son gradient, et demeure donc insensible au choix d'une constante additive). L'analyse sur les états métastables doit être alors transposée du potentiel V à l'énergie libre $U_\beta^{\mathcal{X}}$.

Les états métastables s'organisent de façon hiérarchique et font classiquement apparaître plusieurs ordres de grandeurs caractéristiques en fonction de l'importance des barrières de potentiel à franchir (voir Fig. 2.1 et [22])

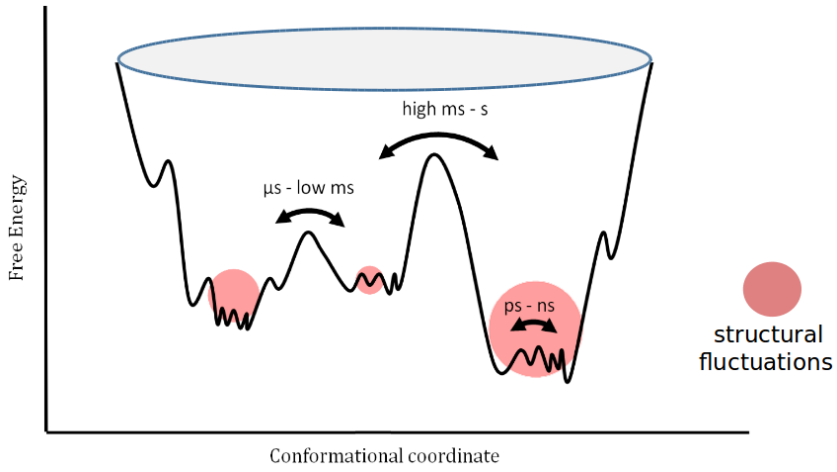

FIGURE 2.1 – Représentation d'une trajectoire de simulation de DM dans le repère de ses coordonnées de réactions mettant en évidence des bassins conformationnels et minimums d'énergie libre [22].

L'espace conformationnel $\mathcal{X}$ correspondant peut finalement être interprété du point de vue physique comme le *paysage énergétique* pouvant être balayé par la protéine, lequel est décrit par la mesure d'équilibre $\nu_\beta^{\mathcal{X}}$. La définition d'une *conformation stable* se traduit ainsi par un phénomène d'équilibration du système au niveau d'un minium local de l'énergie libre de la protéine. Puis, si $x_0 \in \mathcal{X}$ est une conformation stable, on peut alors définir un *bassin conformationnel* centré en x_0 comme étant un voisinage de ce minimum local d'énergie libre $U_\beta^{\mathcal{X}}$.

2.2 Modélisation par une EDS

A partir d'une trajectoire de simulation de DM, une première étape dans la modélisation consiste à retrouver une notion de température, laquelle correspond intuitivement aux *variations quadratiques* du processus. Nous rappelons ici la définition de cette notion, ainsi que quelques propriétés. Nous montrons en particulier comment celles-ci peuvent être employées afin d'estimer le coefficient de diffusion $\mathbf{D}$ d'un processus à partir d'une trajectoire observée. Espérant raisonnablement pouvoir faire correspondre les données observées à une dynamique stochastique, on se propose d'étudier dans un premier temps le profil des variations quadratiques associées à une EDS, puis dans un second temps de comparer la situation vis-à-vis des données.

On se place dans le cas d'un processus $X = (X_t)_{t\geq 0}$ solution d'une EDS. Plus formellement, nous considérons que nous disposons d'un espace de probabilité filtré $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$, sur lequel on dispose d'un $(\mathcal{F}_t)$-mouvement brownien standard $B = (B_t)_{t\geq 0}$ à valeurs dans $\mathbb{R}^d$. On considérera que X est solution de l'EDS :

$$dX_t = \mu(X_t)dt + \boldsymbol{\sigma}(X_t)dB_t \tag{2.9}$$

où $\mu : \mathbb{R}^d \to \mathbb{R}^d$ et $\boldsymbol{\sigma} : \mathbb{R}^d \to \mathcal{M}_d(\mathbb{R})$ sont deux fonctions bornées globalement lipschitziennes. Pour simplifier, on considérera souvent le cas où :

$$\boldsymbol{\sigma} = \sigma \mathrm{I}_d \quad \text{avec} \quad \sigma \geq 0 \tag{1bis}$$

Rappelons également que l'on considère en pratique un *échantillonnage* de ce processus continu, observé au pas de temps Δt, pour un total de N conformations observées de sorte à définir la matrice de $\mathcal{M}_{d,N}(\mathbb{R})$ notée $\mathbf{X} = (x_i)_{1 \leq i \leq N}$, avec $x_i \in \mathbb{R}^d$, telle que pour tout $i \in [\![1, N]\!]$:

$$x_i = X_{(i-1)\Delta t} \tag{2.10}$$

2.2.1 Variation quadratique

Version continue

Définition 1. (Variation quadratique) *Soit $X = (X_t)_{t \geq 0}$ un processus à valeurs dans $\mathbb{R}^d$. On dit que X admet une variation quadratique s'il existe un processus stochastique $\langle X \rangle = (\langle X \rangle_t)_{t \geq 0}$ tel que pour tout $t > 0$ et pour toute suite de subdivisions $(t_0^n, ..., t_{k(n)}^n)$ de l'intervalle $[0, t]$, où $(k_n)_n$ est une suite strictement croissante de $\mathbb{N}$ dans $\mathbb{N}^*$, et dont le pas $\Delta_n = \sup_{0 \leq i \leq k(n)-1} |t_{i+1}^n - t_i^n|$ tend vers 0 lorsque $n \to \infty$, on ait la convergence en probabilité suivante :*

$$\langle X \rangle_t = \lim_{n \to \infty} \sum_{i=0}^{k(n)-1} \| X_{t_{i+1}^n \wedge t} - X_{t_i^n \wedge t} \|^2 \tag{2.11}$$

où $a \wedge b = \min(a, b)$. L'appellation variation quadratique associée au processus X désigne alors le processus $\langle X \rangle$.

Remarque 9. *On peut notamment montrer que toute martingale continue bornée admet une variation quadratique aussi appelée crochet de la martingale [13, 27].*

Proposition 1. (Crochet comme variation quadratique) *Soit $X = (X_t)_{t \geq 0}$ un processus à valeurs dans $\mathbb{R}^d$ vérifiant* (2.9)*. Alors, X admet une variation quadratique $\langle X \rangle$ vérifiant :*

$$\langle X \rangle_t = \int_0^t \| \boldsymbol{\sigma}(X_s) \|_{HS}^2 ds$$

En particulier, on a dans le cas $\boldsymbol{\sigma} = \sigma I_d$ que :

$$\langle X \rangle_t = Dt \quad où \quad D = d\sigma^2 \tag{2.12}$$

Version discrète

Nous montrons ici comment les coefficients de diffusions $\mathbf{D}$ peuvent être estimés à l'aide d'une version discrète des variations quadratiques appliquées à des trajectoires de simulation de DM.

Définition 2. (Variation quadratique - version discrète) *Soit $\mathbf{X} = (x_i)_{1 \leq i \leq N}$ construite selon* (2.10)*. Pour un entier $i \leq N$, on appelle variation quadratique associée à $\mathbf{X}$ jusqu'à i, et on note $\langle \mathbf{X} \rangle_i$ la quantité suivante :*

$$\langle \mathbf{X} \rangle_i = \sum_{k=1}^{i-1} \| x_{k+1} - x_k \|^2$$

L'écriture (2.12) nous invite alors à estimer le coefficient de diffusion de la façon suivante :

Définition 3. (Coefficient de diffusion - version discrète) *Soit $\mathbf{X} = (x_i)_{1 \leq i \leq N}$ construite selon* (2.10)*. On appelle coefficient de diffusion discret la quantité :*

$$\mathbf{D} = \frac{1}{T} \langle \mathbf{X} \rangle_N$$

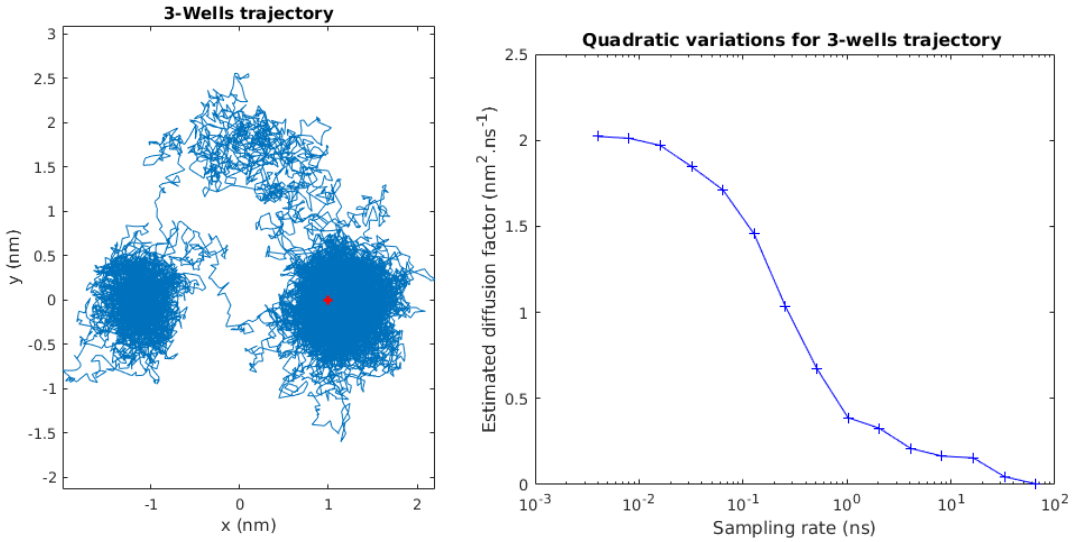

FIGURE 2.2 – Gauche : Trajectoire issue du modèle des trois puits, $T = 100$ ns, $\Delta t = 4$ ps, $\theta = 0.5$ nm^2·ns^{-1} ; Droite : Tracé de $(\Delta t)_m \mapsto \langle \mathbf{X}^m \rangle_m / T_m$ pour la trajectoire issue des 3 puits.

où $\langle \mathbf{X} \rangle_N = \sum_{k=1}^{N-1} \|x_{k+1} - x_k\|^2$.

Remarque 10. *En plus de ne pas garantir une estimation satisfaisante du coefficient de diffusion du fait d'un pas d'observation Δt potentiellement trop grossier, le calcul de $\mathbf{D}$, reposant sur la Proposition 1, suppose la validité d'un modèle d'EDS associé aux données, qui plus est valable à l'échelle de temps considérée. Nous verrons dans ce qui suit que cela n'est pas évident, et nécessite quelques précautions.*

2.2.2 Variations quadratiques d'une EDS

Nous considérons le phénomène de convergence de la variation quadratique vers le coefficient de diffusion (2.11) comme caractéristique d'une EDS, et souhaitons comparer ce profil avec celui observé sur des données. Pour observer cette convergence, considérons $\mathbf{X} = (x_i)_{1 \leq i \leq N}$ construite selon (2.10). On définit pour $m \in \mathbb{N}^*$ la matrice échantillonnée au pas $\Delta_m = \lfloor N/m \rfloor$ notée $\mathbf{X}^m$, et dont le pas temporel correspondant est noté $(\Delta t)_m = \lfloor T_m/m \rfloor$, avec $T_m = \lfloor N/m \rfloor m \times \Delta t$. On calcule alors la quantité suivante pour différentes valeurs de m :

$$\langle \mathbf{X}^m \rangle_m = \sum_{k=1}^{m-1} \|x_{k+1}^m - x_k^m\|^2$$

Considérons une EDS simulée sur le paysage des trois puits tiré de [55] (voir Annexe A.1.1) avec les valeurs $T = 100$ ns, $\theta = 0.5$ nm^2·ns^{-1}, $\Delta t = 4$ ps, pour un total de $N = 25000$ points. La Fig. 2.2 présente alors le tracé de $\langle \mathbf{X}^m \rangle_m / T_m$ pour différentes valeurs de $(\Delta t)_m$.

Lorsque $(\Delta t)_m \to 0$, on retrouve le phénomène de convergence vers le coefficient de diffusion attendu correspondant au processus continu, à savoir $D = 2d\theta = 2$ nm^2·ns^{-1}. Lorsque $(\Delta t)_m$ croît : on quitte le comportement brownien, et la décroissance traduit les effets de la dérive qui apparaissent. En effet, la trajectoire étant désormais contrainte par les puits, la taille des sauts quadratiques n'est plus linéaire en $(\Delta t)_m$, mais converge vers une valeur constante. Par conséquent,

lorsque $(\Delta t)_m \to \infty$, le calcul des variations quadratiques consiste à sommer un nombre décroissant de sauts, ce qui, combiné à la taille constante de ces derniers, fait bien décroître les variations quadratiques.

2.2.3 Variations quadratiques d'une trajectoire de simulation de la dynamique moléculaire

Nous nous proposons alors d'effectuer le précédent calcul de variations quadratiques à partir des données de simulation de DM dont nous disposons.

Remarque 11. *Comme nous le verrons plus tard (Chapitre 3), l'algorithme de segmentation des trajectoires de simulations de DM que nous présenterons utilisera les différents coefficients de diffusion* **D** *estimés ici, en guise d'input.*

Application à VKORC1

Deux répliques d'une durée $T = 100$ ps ont été simulées avec des vitesses initiales différentes au pas $(\Delta t)_{\text{sim}} = 2$ fs, et échantillonnées au pas $\Delta t = 20$ fs pour un total de $N = 5000$ conformations observées. Pour la conformation initiale a été choisie la dernière conformation de la Trajectoire 1 de VKORC1 (1 µs échantillonnée toutes les 5 ps, voir Annexe §A.1.2), par rapport à laquelle les données ont été recalées. Pour ces deux répliques, seuls les atomes de C_α 9 à 151 ont été conservés : on a ici $n = 143$, et $d = 429$.

Après calcul d'une ACP en deux dimensions effectuée à partir de la Trajectoire 1, nous projetons les deux répliques sur ces mêmes axes, puis nous en calculons les variations quadratiques associées (voir Fig. 2.3).

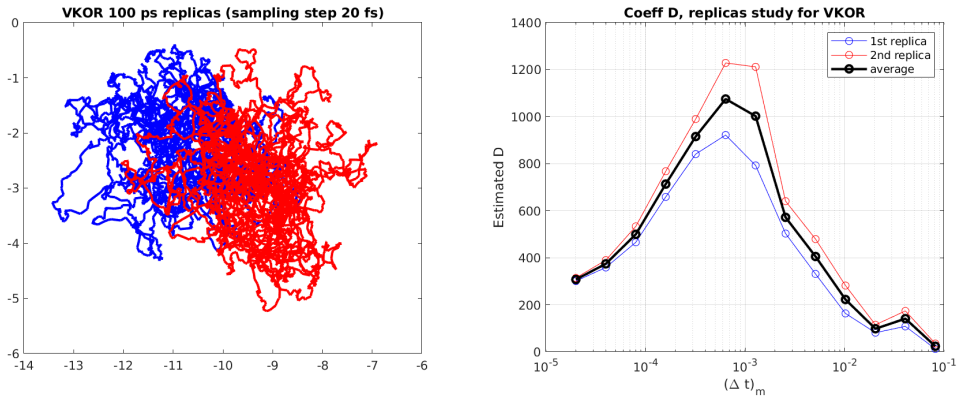

FIGURE 2.3 – Gauche : répliques de VKORC1 (C_α, 9-151) de 100 ps/20 fs, recalées par rapport à leur conformation initiale, et projetées sur les axes de l'ACP 2D de la trajectoire longue (Trajectoire 1). Bleu : réplique 1 ; Rouge : réplique 2. Droite : Estimation du coefficient de diffusion D : $(\Delta t)_m \mapsto \langle \mathbf{X}^m \rangle_m / T_m$ à partir de la variation quadratique.

Nous remarquons alors que l'on retrouve le même phénomène de décroissance lorsque $(\Delta t)_m \to \infty$. En revanche, au lieu d'observer la convergence vers le coefficient de diffusion, on constate plutôt la présence d'un optimum, que nous noterons $(\Delta t)^*$, en-deçà duquel on constate un régime de décroissance des variations quadratiques. Pour $(\Delta t)_m \geq (\Delta t)^*$, par analogie avec le phénomène

observé sur les trois puits, nous pouvons supposer la présence d'un terme de dérive, et donc supposer que la trajectoire puisse être décrite par une EDS faisant intervenir un terme de gradient, ainsi qu'un bruit lié à la température. Nous comprenons alors que le calcul des variations quadratiques pour le pas $(\Delta t)^*$ correspond à la plus fine estimation accessible du coefficient de diffusion à partir de ces données : en deçà, le régime EDS n'est plus valable. Ici, on identifie alors un pas temporel pivot $(\Delta t)^*_{\text{VKORC1}} = 1$ ps, auquel correspond une diffusion estimée $\mathbf{D}_{\text{VKORC1}} = 1000$ nm^2·ns^{-1}.

Que se passe-t-il alors pour $(\Delta t)_m \leq (\Delta t)^*$? On considérera que pour cette plage temporelle, la trajectoire puisse être décrite par une fonction déterministe de classe $\mathcal{C}^1$. En effet, rappelons que les données globales sont générées de manière déterministe via l'intégration des équations hamiltoniennes, et que par conséquent il est vraisemblable que la dynamique restreinte au système d'intérêt corresponde, pour des pas de temps très faibles, à une fonction lisse. D'autre part, l'hypothèse est cohérente du fait que pour toute fonction $f \in \mathcal{C}^1([0,T],\mathbb{R})$, en subdivisant $[0,T]$ avec un pas T/m où m est un entier positif on obtient facilement de par la majoration donnée par Cauchy-Schwarz :

$$(f(t^m_{i+1}) - f(t^m_i))^2 \leq \left(\int_{t^m_i}^{t^m_{i+1}} f'(t)dt \right)^2 \leq (t^m_{i+1} - t^m_i) \int_{t^m_i}^{t^m_{i+1}} f'(t)^2 dt$$

si bien que lorsque $m \to \infty$:

$$\frac{1}{T} \sum_{k=0}^{m-1} (f(t^m_{k+1}) - f(t^m_k))^2 \leq \frac{1}{m} \int_0^T (f'(t))^2 dt \to 0$$

Simulons alors une nouvelle réplique, cette fois-ci d'une durée de 10 ps mais échantillonnée toutes les 2 fs. Si la trajectoire peut d'un point de vue macroscopique présenter un caractère diffusif (voir Fig. 2.4, gauche), on constate en revanche l'émergence d'un comportement plus lisse au fur et à mesure que l'observateur adopte un point de vue microscopique (voir Fig. 2.4, droite) : à une telle échelle, il semble en effet qu'une notion de tangente émerge, nous invitant à interpréter le processus comme l'approximation d'une fonction lisse.

En résumé, s'intéresser aux variations quadratiques nous a permis de montrer qu'il était possible de considérer la trajectoire observée avec différents points de vues selon le pas de temps d'échantillonnage choisi $(\Delta t)_m$. On pourra voir la trajectoire comme étant la solution d'une EDS, ou bien celle d'une EDO déterministe, selon que le pas d'échantillonnage utilisé soit supérieur ou inférieur à la valeur $(\Delta t)^*$.

Application à KIT avec KID

Trois répliques de $T = 100$ ps ont été simulées avec des vitesses initiales différentes au pas $(\Delta t)_{\text{sim}} = 2$ fs, et échantillonnées au pas $\Delta t = 20$ fs pour un total de $N = 5000$ conformations observées. Toutes trois ont pour conformation initiale la dernière conformation de la Trajectoire KIT+KID (2 µs pour la protéine KIT intégrale avec le domaine KID, voir Annexe §A.1.2), par rapport à laquelle elles ont été recalées. Pour ces trois répliques, seuls les atomes de C_α ont été conservés : on a ici $n = 400$, et $d = 1200$.

Après calcul d'une ACP en deux dimensions effectuée à partir de la Trajectoire KIT+KID (2 µs échantillonnée toutes les 10 ps), nous projetons les trois répliques sur ces mêmes axes, puis nous en calculons les variations quadratiques associées (voir Fig. 2.5).

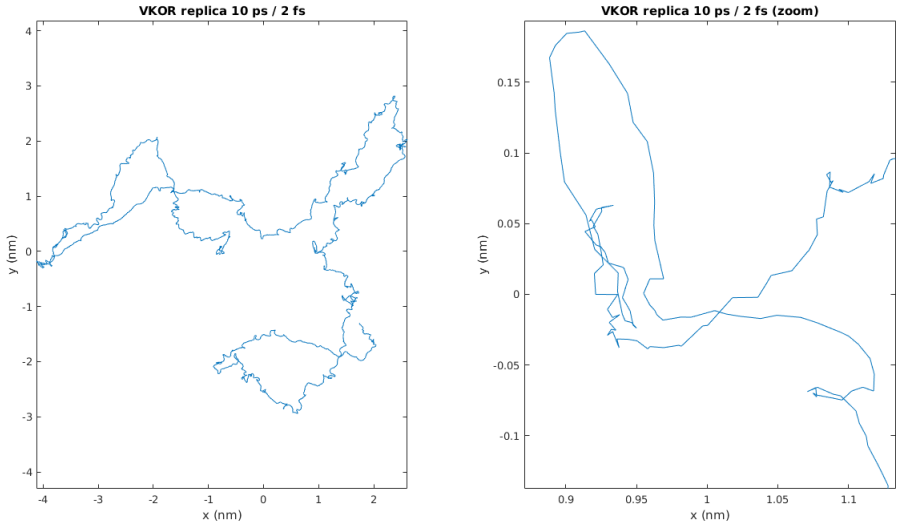

FIGURE 2.4 – ACP en dimension 2 d'une trajectoire de VKORC1 ($T = 10$ ps et $\Delta t = 2$ fs) recalée par rapport à sa conformation initiale. Aspects aléatoire (gauche) et $\mathcal{C}^1$ (droite) perceptibles suivant le pas temporel d'observation.

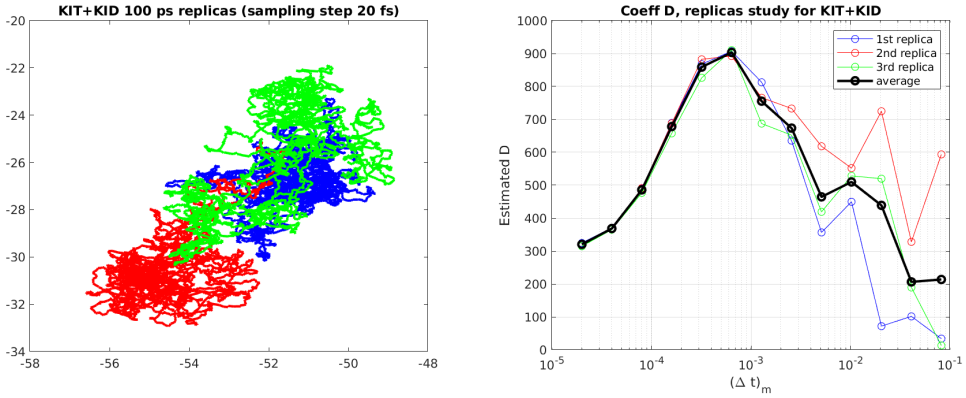

FIGURE 2.5 – Gauche : répliques de KIT avec KID (atomes C_α, 100 ps / 20 fs) recalées par rapport à leur conformation initiale, et projetées sur les axes de l'ACP 2D de la trajectoire longue. Bleu : réplique 1 ; Rouge : réplique 2 ; Vert : réplique 3 ; Droite : Estimation du coefficient de diffusion $D : (\Delta t)_m \mapsto \langle \mathbf{X}^m \rangle_m / T_m$ à partir de la variation quadratique.

On identifie alors un pas temporel pivot $(\Delta t)^*_{\text{KIT+KID}} = 1$ ps, auquel correspond une diffusion estimée $\mathbf{D}_{\text{KIT+KID}} = 900$ nm^2·ns^{-1}.

Application à KID seul

On effectue les mêmes calculs, mais cette fois-ci appliqués au domaine KID seul. Pour ce faire, on restreint les trois répliques précédentes aux résidus 143 à 222 (on a alors $n = 80$ et $d = 240$), puis on effectue de même les opérations de recalage par rapport à la conformation initiale, d'ACP et de calcul des variations quadratiques. Nous obtenons alors les résultats suivants (voir Fig. 2.6).

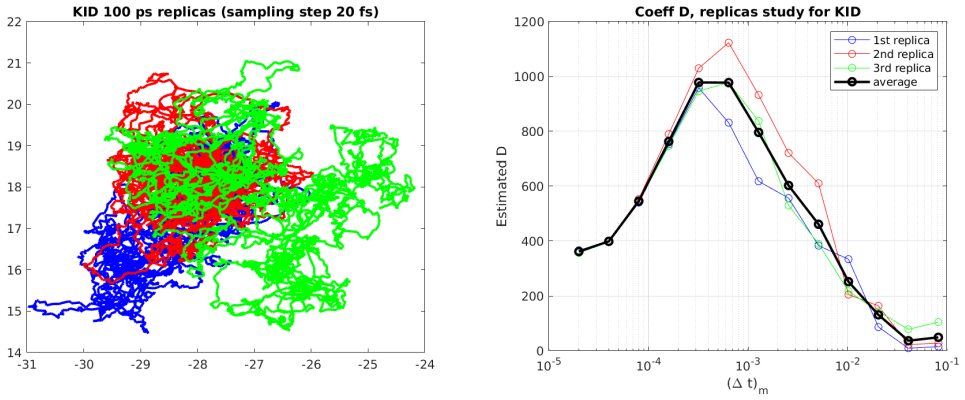

FIGURE 2.6 – Répliques de KID seul (C_α, 100 ps/20 fs) recalées par rapport à leur conformation initiale, et projetées sur les axes de l'ACP 2D de la trajectoire longue. Bleu : réplique 1 ; Rouge : réplique 2 ; Vert : réplique 3. Droite : Estimation du coefficient de diffusion $D : (\Delta t)_m \mapsto \langle \mathbf{X}^m \rangle_m / T_m$ à partir de la variation quadratique.

On identifie alors un pas temporel pivot $(\Delta t)^*_{\text{KID}} = 1$ ps, auquel correspond une diffusion estimée $\mathbf{D}_{\text{KID}} = 1000$ nm^2·ns^{-1}.

Application à KD avec KID

On effectue les mêmes calculs, mais cette fois-ci appliqués au domaine KD (i.e. KIT sans le domaine JMR) avec KID. Pour ce faire, on restreint les trois répliques de KIT avec KID aux résidus 35 à 369 (on a alors $n = 335$ et $d = 1005$), puis on effectue de même les opérations de recalage par rapport à la conformation initiale, d'ACP sur les axes de la trajectoire KD+KID longue, et de calcul des variations quadratiques. Nous obtenons alors les résultats suivants (voir Fig. 2.7).

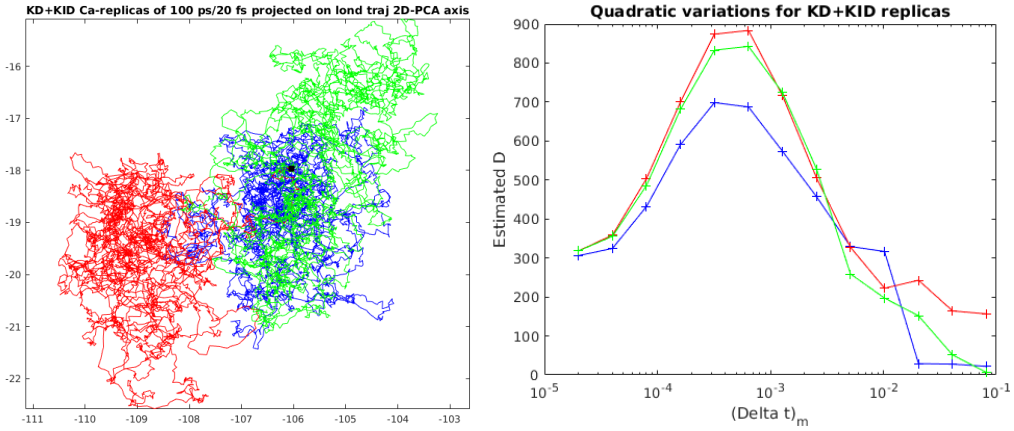

FIGURE 2.7 – Répliques de KD avec KID (C_α, 100 ps/20 fs) recalées par rapport à leur conformation initiale, et projetées sur les axes de l'ACP 2D de la trajectoire longue. Bleu : réplique 1 ; Rouge : réplique 2 ; Vert : réplique 3. Droite : Estimation du coefficient de diffusion $D : (\Delta t)_m \mapsto \langle \mathbf{X}^m \rangle_m / T_m$ à partir de la variation quadratique

On identifie alors un pas temporel pivot $(\Delta t)^*_{\text{KD+KID}} = 1$ ps, auquel correspond une diffusion

estimée $\mathbf{D}_{\mathrm{KD+KID}} = 750$ nm^2·ns^{-1}.

Application à KD seul

On effectue enfin les mêmes calculs, mais cette fois-ci appliqués au domaine KD seul. Pour ce faire, on restreint les trois répliques de KIT avec KID aux résidus 35 à 142 puis 223 à 369 (on a alors $n = 255$ et $d = 765$), puis on effectue de même les opérations de recalage par rapport à la conformation initiale, d'ACP sur les axes de la trajectoire KD longue, et de calcul des variations quadratiques. Nous obtenons alors les résultats suivants (voir Fig. 2.8).

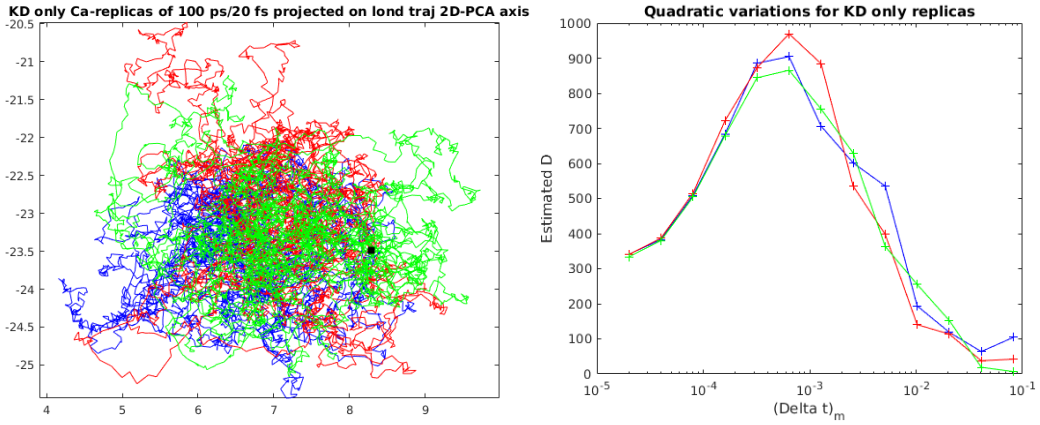

FIGURE 2.8 – Répliques de KD (C_α, 100 ps/20 fs) recalées par rapport à leur conformation initiale, et projetées sur les axes de l'ACP 2D de la trajectoire longue. Bleu : réplique 1 ; Rouge : réplique 2 ; Vert : réplique 3. Droite : Estimation du coefficient de diffusion D : $(\Delta t)_m \mapsto \langle \mathbf{X}^m \rangle_m / T_m$ à partir de la variation quadratique.

On identifie alors un pas temporel pivot $(\Delta t)^*_{\mathrm{KD}} = 1$ ps, auquel correspond une diffusion estimée $\mathbf{D}_{\mathrm{KD}} = 900$ nm^2·ns^{-1}.

Résumé

Le tableau suivant présente les valeurs identifiées des coefficients de diffusion et des pas de temps pivot pour les différentes molécules étudiées.

	$\mathbf{D}$ (nm^2·ns^{-1})	$(\Delta t)^*$ (ps)
VKORC1 (Traj. 1)	1000	1
VKORC1 (Traj. 2)	1000	1
KIT+KID	900	1
KID	1000	1
KD+KID	750	1
KD	900	1

TABLEAU 2.1 – Valeurs de $\mathbf{D}$ et $(\Delta t)^*$ pour VKORC1, KIT+KID, KID, KD+KID et KD.

2.3 Modélisation par chaîne de Markov

Rappelons que si les modèles markoviens, comme les MSM, offrent un point de vue statistique à l'étude, celui-ci ne se justifie pleinement que lorsqu'un nombre suffisant de retours au sein des puits est observé. Cela n'est possible qu'à des échelles de temps nettement supérieures aux durées des trajectoires simulées (de l'ordre de la ms, voire de la s). A cet obstacle près, ce type d'approche permettrait néanmoins d'envisager divers tests statistiques destinés à quantifier la profondeur des puits, ou encore à caractériser le passage d'un puits à un autre, et à ce titre, mérite notre attention.

Nous allons dans un premier temps définir une dynamique artificielle sur les données, via la construction d'une chaîne de Markov homogène ; un résultat de convergence nous permettra, dans certains cas, de considérer la chaîne ainsi construite comme la discrétisation du phénomène continu sous-jacent. Cela nous permettra alors d'envisager une méthode de re-simulation du processus observé par la chaîne discrète : nous verrons qu'un certain nombre de réglages seront nécessaires de manière à créer une nouvelle trajectoire en cohérence avec l'observation. Cette dernière revisitera les états observés mais dans un ordre différent, et on l'interprétera comme une nouvelle réalisation du processus que nous aurions pu observer.

2.3.1 Construction du modèle

Chaîne de Markov

Soit $\mathbf{X} = (x_i)_{1\leq i\leq N}$ construite selon (2.10), pour laquelle on note $\mathcal{X}_N = \{x_1, ..., x_N\}$ l'espace conformationnel visité. Pour toutes conformations $x, y \in \mathcal{X}_N$, on peut définir $\mathbf{W}_\epsilon^{\mathcal{X}_N}$, pour un $\epsilon > 0$ par :

$$\mathbf{W}_\epsilon^{\mathcal{X}_N}(x,y) \triangleq \exp\left(-\frac{\|x-y\|^2}{2\epsilon}\right)$$

De cette façon, l'espace conformationnel visité $\mathcal{X}_N$ peut être considéré comme un graphe plongé dans $\mathcal{X} = \mathbb{R}^d$, et pondéré par $\mathbf{W}_\epsilon^{\mathcal{X}_N}$. On peut alors définir la matrice de transition associée à $\epsilon > 0$, dont chaque ligne constitue une loi de probabilité discrète sur le graphe. On a pour tous $x, y \in \mathcal{X}_N$:

$$\mathbf{P}_\epsilon^N(x,y) \triangleq \frac{\mathbf{W}_\epsilon^{\mathcal{X}_N}(x,y)}{\sum_{z\in\mathcal{X}_N}\mathbf{W}_\epsilon^{\mathcal{X}_N}(x,z)}$$

On considère enfin le processus $\mathbf{X}_\epsilon^N$, à temps discret sur $\mathbb{N}$, et à espace discret sur $\mathcal{X}_N$ régi par cette matrice de transition. Pour tout $k \in \mathbb{N}$, on a ainsi que :

$$\mathbb{P}[\mathbf{X}_\epsilon^N(k) = y | \mathbf{X}_\epsilon^N(k-1) = x] = \mathbb{P}[\mathbf{X}_\epsilon^N(1) = y | \mathbf{X}_\epsilon^N(0) = x] = \mathbf{P}_\epsilon^N(x,y)$$

Remarque 12. *La question du choix de ϵ sera traitée au paragraphe 2.3.2.*

Remarque 13. *Le choix de la fonction décroissante dans la construction de $\mathbf{W}_\epsilon^{\mathcal{X}_N}$ n'influera pas sur la procédure de re-simulation, car à ϵ petit, seul le développement de Taylor nous importe.*

Constatons que l'on peut, à partir des données, construire toute une *famille* de chaînes de Markov à partir des $(\mathbf{P}_\epsilon^N)_{\epsilon>0}$. En particulier, nous avons le :

Lemme 1. *Soit $(x_i)_{1\leq i\leq N}$ dans $\mathbb{R}^d$, et soit $\mathbf{P}_\epsilon^N$ construite comme précédemment. Alors, pour tout couple $x, y \in \mathcal{X}_N$:*

$$\mathbf{P}_\epsilon^N(x,y) \underset{\epsilon\to 0}{\longrightarrow} \delta_{xy}$$

Démonstration. En effet, on a pour tout $y \neq x$, on a $\mathbf{P}_\epsilon^N(x,y)/\mathbf{P}^N(x,x) = e^{-\|x-y\|^2/2\epsilon} \to 0$, ce qui donne immédiatement le résultat. □

Ce lemme nous montre en particulier que lorsque $\epsilon \to 0$, la chaîne de Markov se fige, mais que dès lors que ϵ devient > 0, on engendre une dynamique sur les données.

Variables lentes

De manière à simplifier le suivi du processus, nous proposons de le considérer non plus comme à valeur dans l'espace $\mathbb{R}^d$, mais de le voir à travers les valeurs qu'une fonction $q : \mathcal{X} \mapsto \mathbb{R}$ renvoie. Dans le cas d'un espace conformationnel visité $\mathcal{X}_N$, on considérera simplement la restriction de q à cet espace. On se propose alors, à partir de la chaîne de Markov construite précédemment, de considérer le nouveau processus, toujours à temps et espace discrets, cette fois-ci de $\mathbb{N}$ dans $\mathbb{R}$: le processus $(q(\mathbf{X}_\epsilon^N(k)))_{k\in\mathbb{N}}$.

En se basant sur le lemme précédent, il est légitime pour un ϵ fixé de caractériser le comportement de la chaîne $\mathbf{X}_\epsilon^N$ construite, en comparant la variation de la chaîne, partant d'une conformation de départ x, en une unité de temps 1, à la valeur de ϵ. Finalement, on s'intéresse pour tout $x \in \mathcal{X}_N$ à la quantité :

$$(\Delta_\epsilon q)(x) = \frac{1}{\epsilon}\left(\mathbb{E}_x[q(\mathbf{X}_\epsilon^N(1))] - q(x)\right)$$

Or, nous avons que :

$$\mathbb{E}_x[q(\mathbf{X}_\epsilon^N(1))] = \sum_{y\in\mathcal{X}_N} \mathbf{P}_\epsilon^N(x,y)q(y) = (\mathbf{P}_\epsilon^N q)(x)$$

Par conséquent :

$$(\Delta_\epsilon q)(x) = \frac{1}{\epsilon}\left(\mathbb{E}_x[q(\mathbf{X}_\epsilon^N(1))] - q(x)\right) = \frac{1}{\epsilon}\left((\mathbf{P}_\epsilon^N - I)q(x)\right)$$

Dans le cas où le processus $(\mathbf{X}_\epsilon^N(k))_{k\in\mathbb{N}}$ fait l'objet de captures successives dans différents puits d'une durée significative à l'échelle de la durée totale d'observation, on cherchera une fonction q permettant de rendre compte de cette équilibration en temps long : on parlera alors de *variable lente*, ou encore de *coordonnée de réaction*. Cette variable sera ainsi telle qu'elle encodera le comportement du processus sous-jacent.

L'identification de variables lentes constitue donc un moyen efficace pour résumer la dynamique d'un système à travers un nombre restreint de variables, lesquelles rendent compte des phénomènes d'équilibration du processus, mais permettent également de réduire la dimension d'un problème [56, 23]. Ces variables correspondent à des fonctions propres du générateur infinitésimal associé au processus continu, formant ainsi une *diffusion map* qui offre une représentation du processus en basse dimension. Dans un cas discret, elles peuvent être approximées à partir des données empiriques par le calcul des vecteurs propres associés au *laplacien* de la chaîne de Markov construite comme précédemment.

Remarque 14. *Dans toute la suite, nous considérerons les quantités discrètes comme des restrictions des quantités continues à l'espace conformationnel visité $\mathcal{X}_N$. En effet, une étude approfondie,*

sortant du cadre de ce manuscrit, a été menée concernant la convergence des différents objets évoqués lorsque $N \to \infty$ [49]. Nous nous contentons ici de résumer les versions discrète et continue de chacun d'entre eux au sein du tableau suivant :

	Discret	Continu
Espace	$\mathcal{X}_N$	$\mathcal{X}$
Similarité	$\mathbf{W}_\epsilon^N$	W_ϵ
Noyau	$\mathbf{P}_\epsilon^N$	P_ϵ

TABLEAU 2.2 – Correspondance continu-discret

Soit donc $\mathbf{P}_\epsilon^N : \mathcal{X}^2 \to [0,1]$ la matrice de transition associée à la chaîne de Markov irréductible et réversible construite sur les données empiriques et μ son unique mesure invariante. Pour $f, g : \mathcal{X} \to \mathbb{R}$, on définit le produit scalaire $\langle f, g\rangle_\mu = \sum_{x\in\mathcal{X}} f(x)g(x)\mu(x)$. En considérant le laplacien de la chaîne $L \triangleq \mathbf{P}_\epsilon^N - \mathrm{I}_d$ et utilisant la réversibilité de la chaîne ($\mu(x)\mathbf{P}_\epsilon^N(x,y) = \mu(y)\mathbf{P}_\epsilon^N(y,x)$), on montre que :

$$\langle -Lf, f\rangle_\mu = \frac{1}{2}\sum_{x,y\in\mathcal{X}} \mu(x)\mathbf{P}_\epsilon^N(x,y)(f(x)-f(y))^2 \geq 0\,. \tag{2.13}$$

Les valeurs propres de L sont donc positives, et donc celles de $\mathbf{P}_\epsilon^N$ comprises dans l'intervalle $[-1,1]$. Nous voyons que la valeur propre 1 de $\mathbf{P}_\epsilon^N$ est associée à la fonction constante. La recherche de variables lentes revient donc à l'identification des fonctions propres $f \in (\mathbb{R}\mathbb{1})^\perp$ minimisant le "coût" (2.13), c'est-à-dire des fonctions propres associées aux plus petites valeurs propres non-nulles de L, i.e. aux valeurs propres de $\mathbf{P}_\epsilon^N$ les plus proches de 1 par valeurs négatives.

En effet, soit $f \in (\mathbb{R}\mathbb{1})^\perp$ une fonction propre de $\mathbf{P}_\epsilon^N$ de valeur propre $\lambda \in [-1,1[$. Nous caractérisons la "lenteur" de la variable f à travers la variance de son estimateur empirique $\hat{f}_N = \frac{1}{N}\sum_{0\leq k<N} f(X_k)$ (on sait par ailleurs que $\mathbb{E}_\mu[\hat{f}_N] = 0$). On s'intéresse ainsi à la quantité :

$$\mathbb{V}\mathrm{ar}[\hat{f}_N] = \frac{1}{N^2}\sum_{0\leq k,l<N} \mathbb{E}_\mu[f(X_k)f(X_l)]$$

Par stationnarité, on a que :

$$\mathbb{V}\mathrm{ar}[\hat{f}_N] = \frac{1}{N}\|f\|_\mu^2 + \frac{2}{N^2}\sum_{0\leq k,l<N} \mathbb{E}_\mu[f(X_0)f(X_{l-k})] = \frac{1}{N}\|f\|_\mu^2 + \frac{1}{N^2}\langle f, (\mathbf{P}_\epsilon^N)^{l-k} f\rangle_\mu$$

En posant $S_N \triangleq \mathrm{I}_d + 2\sum_{1<k<N} \frac{N-k}{N}(\mathbf{P}_\epsilon^N)^k$, on a que :

$$\mathbb{V}\mathrm{ar}(\hat{f}_N) = \frac{1}{N}\langle S_N f, f\rangle_\mu$$

Mais alors, on peut écrire que :

$$\begin{aligned} S_N f &= \left(1 + 2\sum_{0<k<N} \frac{N-k}{N}\lambda^k\right) f \\ &\underset{N\to\infty}{=} \left(1 + 2\frac{\lambda}{1-\lambda}\right) f + o(1/N) \end{aligned}$$

En supposant alors que $\langle f, f\rangle_\mu = 1$ (i.e.$\mathbb{V}\mathrm{ar}[f] = \mathbb{E}_\mu[f^2] = 1$), on a que :

$$\mathbb{V}\mathrm{ar}[\hat{f}_N] \underset{N\to\infty}{\longrightarrow} \mathrm{diag}\left(\frac{1+\lambda}{1-\lambda}\right)$$

Ayant alors que $\frac{1+u}{1-u} \underset{u\to 1^-}{\longrightarrow} +\infty$, nous comprenons que plus la valeur propre λ de f est proche de 1, plus la variance de l'estimateur associé est grande (équilibration lente). On retrouve par là que les variables lentes correspondent à des fonctions propres dont les valeurs propres sont proches de 1.

Processus sous-jacent

Si l'étude du générateur infinitésimal du processus sous-jacent constitue un point clé dans l'identification des variables lentes, nous nous y intéressons ici non pas dans le cadre des objectifs de Coifman et al. cités précédemment, mais dans le but de caractériser le processus sous-jacent à la chaîne de Markov discrète introduite. Le passage du discret au continu correspond aux passages à la limite suivants :

$$\begin{cases} N \longrightarrow \infty & \text{(espace continu)} \\ \epsilon \longrightarrow 0 & \text{(temps continu)} \end{cases}$$

Si le premier passage à la limite a été évoqué précédemment, reste à étudier la quantité suivante :

$$\frac{\partial q(x,t)}{\partial t} = \lim_{\epsilon\to 0} \big[\frac{P_\epsilon - I}{\epsilon}\big] q(x)$$

On a alors la :

Proposition 2. *Supposons que la loi sous-jacente s'écrive sous la forme suivante (distribution de Maxwell-Boltzmann) :*

$$\mu(x) = \frac{1}{Z_\theta} \exp\Big(- \frac{U(x)}{\theta}\Big)$$

où $\theta > 0$ est un paramètre de température. Alors, la dynamique créée à partir des données fournit que pour tout $x \in \mathcal{X}$:

$$\lim_{\epsilon\to 0} \big[\frac{P_\epsilon - I}{\epsilon}\big] q(x) = (\mathcal{L}_b q)(x)$$

où $\mathcal{L}_b$ est l'opérateur de Fokker-Planck rétrograde (backward) défini par :

$$(\mathcal{L}_b q)(x) = \frac{\Delta q(x)}{2} - \frac{1}{\theta}\langle \nabla q(x), \nabla U(x)\rangle$$

et dont la loi stationnaire est donnée, pour tout $x \in \mathcal{X}$, par :

$$\mu_\infty(x) = \frac{1}{Z_\theta} \exp\Big(- 2\frac{U(x)}{\theta}\Big)$$

Démonstration. Voir Annexe §A.3.1 □

On constate donc qu'en supposant que les données soient issues d'une distribution de type Maxwell-Boltzmann, on parvient quasiment à retrouver la loi sous-jacente à la limite, à un facteur de température près.

2.3.2 Méthode de re-simulation

La constatation précédente nous invite à envisager une procédure de re-simulation du processus sous-jacent par la chaîne de Markov discrète sur les données observées : nous avons en effet dans l'idée que pour N grand, la Proposition 2 garantisse, à un facteur de température près, la

reconstruction de la bonne loi stationnaire. En mettant en place une telle méthode, nous serions en mesure en générer rapidement d'autres trajectoires dans le même paysage conformationnel local.

Nous venons de voir que la loi limite construite à partir des données s'approchait de la loi stationnaire sous-jacente, et ce par l'intermédiaire d'un opérateur de Fokker-Planck correspondant à une dynamique bien précise. En effet, le générateur infinitésimal du processus limite s'écrit :

$$(\mathcal{L}_\infty q)(x) = \frac{\Delta q(x)}{2} - \frac{1}{\theta}\langle \nabla q(x), \nabla U(x)\rangle$$

et correspond à l'EDS suivante :

$$dX_t^\infty = -\frac{1}{\theta}\nabla U(X_t^\infty)dt + dB_t$$

Nous comprenons donc que la procédure de re-simulation ne sera pertinente que pour des données issues d'une EDS de ce type, c'est-à-dire, des équations de Langevin suramorties évoquées précédemment (voir Chapitre 2 §2.1.3). En partant d'une telle équation, nous serons ainsi certains que l'opérateur limite aura une réalité physique, ainsi que les variables lentes qui en découleront.

Définition-Proposition 1. (dynamique de Langevin) *Soit X_t la conformation dans laquelle se trouve la protéine à l'instant t. On dit que la dynamique de ce système est de Langevin si l'on a :*

$$dX_t = -\nabla U(X_t)dt + \sqrt{2\theta}dB_t$$

Par ailleurs, la mesure d'équilibre associée est la distribution de Maxwell-Boltzmann suivante :

$$\mu(x) = \frac{1}{Z_\theta}\exp\left(-\frac{U(x)}{\theta}\right)$$

Démonstration. Déjà vu (voir Chapitre 2 §2.1.3). □

Par conséquent, lorsque la dynamique est de type Langevin, nous pouvons espérer construire une re-simulation cohérente avec la dynamique sous-jacente des données observées. L'EDS limite construite est la suivante :

$$dX_t^\infty = -\frac{1}{\theta}\nabla U(X_t^\infty)dt + dB_t$$

Et on remarque qu'en faisant le changement de variable $t \leftarrow t/\theta$:

$$d\tilde{X}_t^\infty = -\nabla U(\tilde{X}_t^\infty)dt + \sqrt{\theta}dB_t$$

L'opérateur rétrograde est alors le suivant :

$$\begin{aligned}(\tilde{\mathcal{L}}_\infty q)(x) &= \frac{\theta}{2}\Delta q(x) - \langle \nabla q(x), \nabla U(x)\rangle \\ &= \theta(\mathcal{L}_\infty q)(x)\end{aligned}$$

On retrouve ainsi, tant dans l'expression des EDS que des générateurs infinitésimaux, le problème de température évoqué précédemment.

Correspondance discret-continu

Notons $(\mathbf{X}_k^N)_{k\in\mathbb{N}}$ la chaîne de Markov construite à partir de la matrice $\mathbf{P}_\epsilon^N$. Nous cherchons à préciser le pas de temps physique auquel correspond un pas ϵ de la chaîne de Markov qui simule la dynamique créée à partir des données en prenant en compte le changement de variable temporel. D'une part on a que :

$$\begin{aligned}\mathbb{E}\big[q(\mathbf{X}_\epsilon^N)|\mathbf{X}_0^N = x\big] - q(x) &= [(\mathbf{P}_\epsilon^N - I)q](x)\\ &\approx [(P_\epsilon - I)q](x)\\ &\approx \epsilon(\mathcal{L}_\infty q)(x)\end{aligned}$$

D'autre part, il existe un pas de temps Δt tel que :

$$\begin{aligned}\mathbb{E}\big[q(\mathbf{X}_\epsilon^N)|\mathbf{X}_0^N = x\big] - q(x) &\approx \mathbb{E}\big[q(\tilde{X}_{\Delta t}^\infty)|\tilde{X}_0^\infty = x\big] - q(x)\\ &\approx \Delta t(\tilde{\mathcal{L}}_\infty q)(x)\\ &\approx \Delta t\theta(\mathcal{L}_\infty q)(x)\end{aligned}$$

Par identification, nous en déduisons que :

$$\epsilon \approx \theta\Delta t$$

Par conséquent, un pas de temps ϵ de la chaîne discrète correspond à un pas de temps $\Delta t = \epsilon/\theta$ de la dynamique de $(\tilde{X}_t^\infty)_{t\geq 0}$. Pour tout $k \in \mathbb{N}$:

$$\mathbf{X}_{k\epsilon}^N = \tilde{X}_{k\times\frac{\epsilon}{\theta}}^\infty$$

Par ailleurs, nous pouvons à l'aide de ce pas temporel déterminer le nombre d'itérations nécessaires à la re-simulation d'une plage temporelle de la bonne durée : en effet, si le processus est observé pendant un intervalle de temps $[0, T]$, alors le nombre d'itérations K de la re-simulation discrète par la chaîne $(\mathbf{X}_k^N)_{k\in\mathbb{N}}$ correspondant à un intervalle de temps équivalent devra vérifier la relation :

$$K\frac{\epsilon}{\theta} = T \quad \text{i.e.} \quad K = T\frac{\theta}{\epsilon}$$

Nous allons illustrer ce problème de température à l'aide du modèle des trois puits. On simule alors un échantillon de N observations en partant d'un point $x_0 \in \mathbb{R}^2$ choisi dans le domaine contenant les puits et d'un pas temporel de simulation $(\Delta t)_{\text{sim}}$ selon la récurrence suivante. Pour tout i :

$$x_{i+1} = x_i - \nabla U(x_i)(\Delta t)_{\text{sim}} + \sqrt{2\theta(\Delta t)_{\text{sim}}} \times n_i$$

où n_i est tirée suivant $\mathcal{N}(0_{\mathbb{R}^2}, \mathrm{I}_2)$. On obtient donc une matrice d'observations $\mathbf{X}$, à partir de laquelle on édifie, pour une valeur ϵ, une matrice de transition $\mathbf{P}_\epsilon^N$, à laquelle on fait correspondre une marche discrète de re-simulation $\mathbf{X}_\epsilon^N$.

D'une part, le nombre K d'itérations à effectuer pour re-simuler une durée d'observation T doit être tel que $K = T\theta/\epsilon$: ainsi on re-simule bien la bonne durée d'observation. D'autre part, chaque pas de la marche ainsi créée correspondant à un temps physique $(\Delta t)_{\text{resim}} = \epsilon/\theta$, on veillera à construire au préalable la matrice $\mathbf{P}_\epsilon^N$ avec un paramètre $\epsilon = \theta(\Delta t)_{\text{sim}}$: de cette façon on aura bien $(\Delta t)_{\text{resim}} = (\Delta t)_{\text{sim}}$, et donc l'échantillonnage temporel sera le même ($N = K$).

Voici l'algorithme correspondant :

Algorithm 1 Re-simulation : les trois puits

Entrée Fonction d'énergie U_c

- Choix des paramètres c et θ
- Choix de T, N et calcul de $\Delta t = T/N$
- Choix de x_0 et construction de $\mathbf{X}$
- Calcul de $\epsilon = \theta \Delta t$
- Construction de $\mathbf{W}_\epsilon^{\mathcal{X}_N}$ et $\mathbf{P}_\epsilon^N$
- Détermination de K
- Choix d'un $\mathbf{X}_0^N$ et simulation de la chaîne $\mathbf{X}_\epsilon^N$

Sortie Tracé des observations $\mathbf{X}$ et de la trajectoire re-simulée $\mathbf{X}^N$

Voici une illustration de ce que fournit l'algorithme pour les paramètres suivants : $\theta = 0.5$ nm^2·ns^{-1}, $T = 500$ ns, $N = 5000$, $\Delta t = 0.1$ ns, $\epsilon = 0.05$, et $K = 5000$.

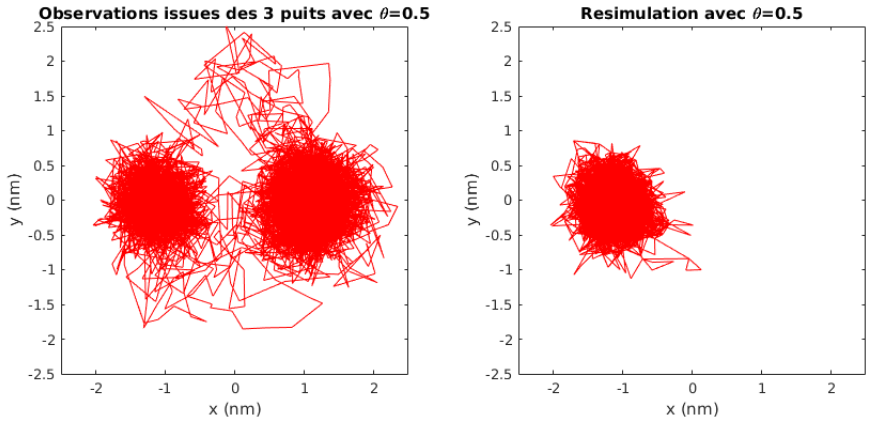

FIGURE 2.9 – Gauche : observations sur issues du paysage des trois puits à $\theta = 0.5$ nm^2·ns^{-1} ; Droite : resimulation à la même température.

Si le processus resimulé semble bien être de même nature que les observations, nous constatons néanmoins que le paysage visité est restreint. En effet, la division de la température par un facteur 2 se traduit par l'incapacité du processus à quitter un premier bassin d'énergie, alors qu'il en existe d'autres : il est donc indispensable de corriger ce facteur de température sans quoi une partie du paysage est inaccessible.

Re-simulation corrigée

Pour prendre en compte le phénomène de "refroidissement" constaté précédemment, nous souhaitons alors construire différemment la matrice de transition P_ϵ de sorte que l'on ait cette fois-ci à la limite :

$$\lim_{\epsilon \to 0} \Big[\frac{P_\epsilon - I}{\epsilon}\Big] q(x) = \frac{\Delta q(x)}{2} - \frac{1}{2\theta} \langle \nabla q(x), \nabla U(x) \rangle$$

c'est-à-dire la même formule que précédemment, mais avec une température multipliée par un facteur 2. Ceci est possible à l'aide du résultat suivant :

Proposition 3. (dynamique corrigée) *Soit $\alpha > 0$. Soit la fonction d'adjacence suivante :*

$$\forall x, y \in \mathcal{X}, W_\epsilon^{(\alpha)}(x,y) = \frac{W_\epsilon(x,y)}{(d_\epsilon(x))^\alpha (d_\epsilon(y))^\alpha}$$

avec :

$$d_\epsilon(x) = \frac{1}{(2\pi\epsilon)^{d/2}} \int_{y\in\mathcal{X}} W_\epsilon(x,y)\mu(y)dy$$

Soit le noyau de transition :

$$\forall x, y \in \mathcal{X}, P_\epsilon^{(\alpha)}(x,y) = \frac{W_\epsilon^{(\alpha)}(x,y)}{d_\epsilon^{(\alpha)}(x)}$$

avec :

$$d_\epsilon^{(\alpha)}(x) = \int_{y\in\mathcal{X}} W_\epsilon^{(\alpha)}(x,y)\mu(y)dy$$

Alors, on a finalement que :

$$\lim_{\epsilon\to 0} \big[\frac{P_\epsilon^{(\alpha)} - I}{\epsilon}\big] q(x) = \frac{\Delta q(x)}{2} - \frac{1-\alpha}{\theta}\langle \nabla q(x), \nabla U(x)\rangle$$

En particulier, pour $\alpha = 1/2$:

$$\lim_{\epsilon\to 0} \big[\frac{P_\epsilon^{(1/2)} - I}{\epsilon}\big] q(x) = \frac{\Delta q(x)}{2} - \frac{1}{2\theta}\langle \nabla q(x), \nabla U(x)\rangle$$

Démonstration. Voir Annexe A.3.2. □

En construisant la matrice de transition avec un paramètre $\alpha = 1/2$, nous avons donc l'opérateur rétrograde :

$$(\mathcal{L}_\infty^{(\frac{1}{2})} q)(x) = \frac{\Delta q(x)}{2} - \frac{1}{2\theta}\langle \nabla q(x), \nabla U(x)\rangle$$

Cette équation correspond à la dynamique :

$$dX_t^{\infty,(\frac{1}{2})} = -\frac{1}{2\theta}\nabla U(X_t^{\infty,(\frac{1}{2})})dt + dB_t$$

Et avec le changement temporel $t \leftarrow t/2\theta$, on retombe bien sur l'équation que nous souhaitions représenter :

$$d\tilde{X}_t^{\infty,(\frac{1}{2})} = -\nabla U(\tilde{X}_t^{\infty,(\frac{1}{2})})dt + \sqrt{2\theta}dB_t$$

et dont l'opérateur rétrograde associé est :

$$\begin{aligned}(\tilde{\mathcal{L}}_\infty^{(\frac{1}{2})} q)(x) &= \theta\Delta q(x) - \langle \nabla q(x), \nabla U(x)\rangle \\ &= 2\theta(\mathcal{L}_\infty^{(\frac{1}{2})} q)(x)\end{aligned}$$

Nous allons donc pouvoir re-simuler la dynamique cible à l'aide d'une chaîne discrète créée à partir d'une matrice de transition $\mathbf{P}_\epsilon^{N,(\frac{1}{2})}$, dont nous notons $(\mathbf{X}_k^{N,(\frac{1}{2})})_{k\in\mathbb{N}}$ la chaîne de Markov associée. D'une part on a que :

$$\begin{aligned}\mathbb{E}\big[q(\mathbf{X}_\epsilon^{N,(\frac{1}{2})})|\mathbf{X}_0^{N,(\frac{1}{2})} = x_i\big] - q(i) &= [(\mathbf{P}_\epsilon^{N,(\frac{1}{2})} - I)q][i] \\ &\approx [(\mathbf{P}_\epsilon^{(\frac{1}{2})} - I)q](x_i) \\ &\approx \epsilon(\mathcal{L}_\infty^{(\frac{1}{2})} q)(x_i)\end{aligned}$$

D'autre part, il existe un pas de temps Δt tel que :

$$\begin{aligned}\mathbb{E}\big[q(\mathbf{X}_\epsilon^{N,(\frac{1}{2})})|\mathbf{X}_0^{N,(\frac{1}{2})} = x_i\big] - q(i) &\approx \mathbb{E}\big[q(\tilde{X}_{\Delta t}^{\infty,(\frac{1}{2})})|\tilde{X}_0^{\infty,(\frac{1}{2})} = x_i\big] - q(x_i)\\ &\approx \Delta t(\tilde{\mathcal{L}}_\infty^{(\frac{1}{2})}q)(x_i)\\ &\approx 2\theta\Delta t(\mathcal{L}_\infty^{(\frac{1}{2})}q)(x_i)\end{aligned}$$

Par identification, nous en déduisons cette fois-ci que :

$$\epsilon \approx 2\beta\Delta t$$

Dans ce cas, un pas de temps ϵ de la chaîne discrète correspond à un pas de temps $\Delta t = \epsilon/2\theta$ de la dynamique de $(\tilde{X}_t^{\infty,(\frac{1}{2})})_{t\geq 0}$. Pour tout $k \in \mathbb{N}$:

$$\mathbf{X}_{k\epsilon}^{N,(\frac{1}{2})} = \tilde{X}_{k\times\frac{\epsilon}{2\theta}}^{\infty,(\frac{1}{2})}$$

De même que précédemment, on veillera à ce que d'une part la re-simulation corresponde bien à la durée d'observation T en définissant un nombre d'itérations K par :

$$K = 2\frac{T\theta}{\epsilon}$$

D'autre part, on s'assurera que lorsqu'on simule les observations avec un pas de temps fixé $(\Delta t)_{\text{sim}}$, alors le pas de temps ϵ la chaîne discrète corresponde à un temps physique $(\Delta t)_{\text{resim}} = (\Delta t)_{\text{sim}}$, ce qui est obtenu en construisant initialement la matrice de transition avec un ϵ défini par :

$$\epsilon = 2\theta\Delta t$$

L'algorithme est alors quasiment le même que précédemment, mais en construisant cette fois-ci les quantités corrigées. Signalons que pour rendre convaincante la comparaison avec la re-simulation précédente, nous sommes partis des mêmes observations (les deux algorithmes ont été appliqués au même jeu d'observations).

Algorithm 2 Re-simulation corrigée : les trois puits

Entrée Fonction d'énergie U_c

- Choix des paramètres c et θ
- Calcul de Z_β et tracé de la densité de probabilité f
- Choix de T, N
- Calculs de $\Delta t = T/N$ et $\epsilon = 2\theta\Delta t$
- Choix de x_0 et construction de $\mathbf{X}$, $\mathbf{W}_\epsilon^{N,(\frac{1}{2})}$ et $\mathbf{P}_\epsilon^{N,(\frac{1}{2})}$
- Choix de $x_0^{N,(\frac{1}{2})}$, détermination de K et construction de $\mathbf{X}^{N,(\frac{1}{2})}$

Sortie Tracé des observations $\mathbf{X}$ et de la trajectoire re-simulée $\mathbf{X}^{N,(\frac{1}{2})}$

Voici ce que l'on obtient avec $\theta = 0.5$ nm$^2\cdot$ns^{-1}, $T = 500$ ns, $N = 5000$, $\Delta t = 0.1$ ns, $\epsilon = 0.1$, et $K = 5000$:

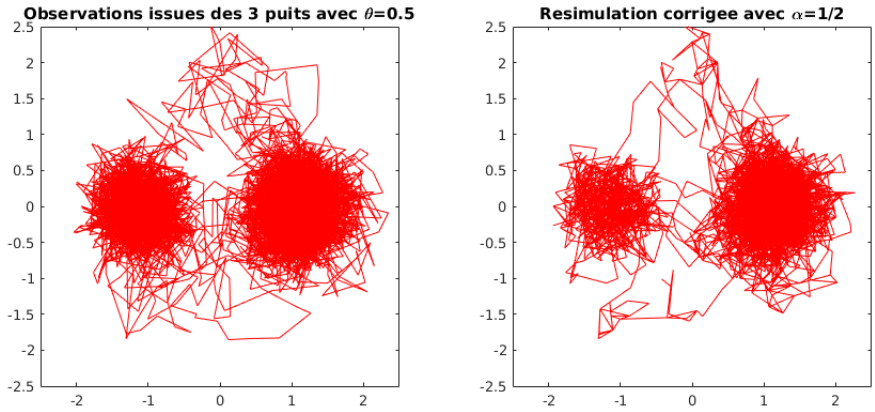

FIGURE 2.10 – Gauche : observations issus du paysages des trois puits à $\theta = 0.5$ nm^2·ns^{-1}. Droite : resimulation corrigée avec $\alpha = 1/2$.

La correction du noyau de transition permet bien de rétablir la température sous-jacente : la re-simulation est pertinente.

2.4 Bilan : Langevin ou non Langevin ?

La validité de l'approximation par un modèle de Langevin suramorti semble supportée par un certain nombre de développements théoriques dont certains ont été abordés au Chapitre 2 §2.1. Cependant, lorsque nous étudions non plus des trajectoires théoriques, mais des trajectoires obtenues en production par simulation de la dynamique moléculaire sur des macromolécules complexes comme celles qui nous intéressent, il est légitime de s'interroger sur l'adéquation des modèles avec les données. Si nous avons vu que l'approximation des trajectoires par une EDS était largement supportée par notre étude sur la variation quadratique pour des trajectoires simulées après projection, la plupart des développements théoriques sont plus spécifiques et ne s'appuient pas seulement sur un modèle d'EDS pour une dérive quelconque du type :

$$dX_t = \mu(X_t)dt + \sqrt{2\theta}dB_t$$

mais plus précisément pour une dérive $\mu(x) = -\nabla U(x)$ associée à un potentiel U que l'on peut assimiler à l'énergie libre (qui dépend de la température θ mais nous ne ferons pas apparaître ici cette dépendance) i.e :

$$dX_t = -\nabla U(X_t)dt + \sqrt{2\theta}dB_t$$

pour $\theta = k_B T$. Dans ce dernier cas, la densité ρ_θ de la mesure invariante est donnée par $\rho_\theta(x) \propto \exp(-U(x)/\theta)$ si bien que la donnée de la mesure invariante permet de retrouver la dérive par :

$$\nabla U(x) = \theta \nabla \log(\rho_\theta)(x)$$

Il s'agit d'un résultat bien spécifique des dynamiques de Langevin suramorties. Pour une dérive plus générale, le lien entre la densité de la mesure invariante (en supposant qu'elle existe) et la dérive μ n'est plus univoque du fait que l'application $\mu \mapsto \rho$ *n'est pas injective.*

Pour s'en convaincre facilement, il suffit de considérer le cas des diffusions sur $\mathbb{R}^2$ et de considérer pour $\alpha : \mathbb{R} \to \mathbb{R}$ C^1 bornée :

$$\mu_\alpha(x) = -\nabla U(x) + \alpha(U(x))\nabla^\perp U(x)$$

où pour $\nabla U(x) = (u_1, u_2) \in \mathbb{R}^2$, $\nabla^\perp U(x) = (-u_2, u_1)$ désigne la rotation du vecteur $\nabla U(x)$ dans le sens direct, si bien que $\langle \nabla U(x), \nabla^\perp U(x) \rangle = 0$.

On vérifie que si L^*_α désigne l'adjoint du générateur infinitésimal donné ici par :

$$L^*_\alpha g \triangleq -\mathrm{div}(g\mu_\alpha) + \theta\Delta g = L^*_0 - \mathrm{div}(g\alpha(U)\nabla^T U)$$

on a :

$$L^*_\alpha \rho_\theta = 0$$

et donc ρ_θ *reste la densité de la mesure invariante.*

En effet, nous avons que $L^*_\alpha \rho_\theta = L_0\rho_\theta - \mathrm{div}(\rho_\theta \alpha(U)\nabla^\perp)$ avec $L_0\rho_\theta = 0$ puisque ρ_θ est la densité d'équilibre de la dynamique de Langevin pour $\alpha = 0$. De plus :

$$\begin{aligned}\mathrm{div}(\rho_\theta\alpha(U)\nabla^\perp) &= \langle \nabla\rho_\theta + \rho_\theta\alpha'(U)\nabla U, \nabla^\perp U\rangle + \rho_\theta\alpha(U)\underbrace{\mathrm{div}(\nabla^\perp U)}_{0} \\ &= (-1/\theta + \alpha'(U))\rho_\theta \underbrace{\langle \nabla U, \nabla^\perp U\rangle}_{=0} = 0\end{aligned}$$

On a donc montré que toutes les dynamiques associées à la famille de champs de dérives $(\mu_\alpha)_\alpha$ ont même mesure invariante. Il n'est donc pas possible *a priori* de s'appuyer seulement sur la mesure invariante pour retrouver la dérive. C'est pourquoi l'approche proposée par Coifman et al. [56, 23] ne saurait être appliquée aveuglément dans le cas de trajectoires de dynamiques quelconques. A ce stade, nous pouvons donc au mieux modéliser le processus sous-jacent à nos données par un régime d'EDS sans précision sur la nature du terme de dérive μ, et ce pour des pas de temps $\Delta t \geq (\Delta t)^*$.

Chapitre 3

Algorithme de κ-segmentation

Ce chapitre est dédié à la présentation d'une nouvelle méthode de segmentation des trajectoires de dynamique moléculaire, palliant certains défauts des outils actuels (RMSD, RMSF, clustering), et notamment la détection d'artefacts. Nous présenterons dans un premier temps la définition d'une nouvelle quantité notée κ et appelée nombre de tours, qui vise à quantifier la profondeur des puits de manière absolue. Par la suite, nous présenterons un algorithme permettant une segmentation automatique des trajectoires de DM basée sur cette quantité, et nous le testerons sur des exemples de bases, que seront le mouvement brownien et le modèle des trois puits.

3.1 Quantification des puits et critère du nombre de tours κ

Nous construisons dans cette section le critère du nombre de tours κ à partir de l'analyse de l'évolution en fonction du temps t du rayon de la plus petite boule contenant la trajectoire jusqu'à l'instant t. Nous mettons ensuite en évidence quelques propriétés remarquables de κ, concernant notamment son comportement pour une trajectoire brownienne. Nous proposerons alors la caractérisation d'un puits par sa valeur de κ, laquelle nécessite de définir les instants d'accès et de sortie d'un puits. Enfin, nous présenterons le cas d'une trajectoire discrète, et les quantités associées par analogie avec le cas continu.

Dans ce chapitre, nous modéliserons le signal à traiter comme la solution $X = (X_t)_{t\geq 0}$ d'une EDS. Plus formellement, nous considérons un espace de probabilité filtré $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$, sur lequel on dispose d'un $(\mathcal{F}_t)$-mouvement brownien standard $B = (B_t)_{t\geq 0}$ à valeurs dans $\mathbb{R}^d$. On considérera que X est solution de l'EDS :

$$dX_t = \mu(X_t)dt + \boldsymbol{\sigma}(X_t)dB_t \tag{3.1}$$

où $\mu : \mathbb{R}^d \to \mathbb{R}^d$ et $\boldsymbol{\sigma} : \mathbb{R}^d \to \mathcal{M}_d(\mathbb{R})$ sont deux fonctions bornées globalement lipschitziennes.

3.1.1 Rayon maximum

Soit donc $X = (X_t)_{t\geq 0}$ le processus à valeurs dans $\mathbb{R}^d$ décrivant l'évolution d'un système d'intérêt telle que la trajectoire conformationnelle d'une protéine. La première caractéristique permettant de détecter l'existence d'un équilibre local sur un segment temporel $[s, t]$ est donnée par le suivi de la dynamique radiale $\|X_u - X_s\|$ à partir d'une configuration de référence. Nous

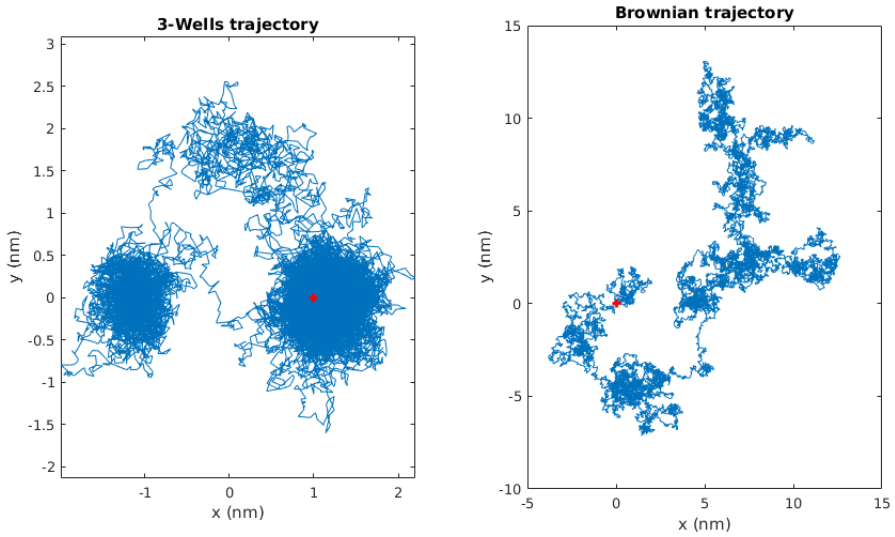

FIGURE 3.1 – Gauche : trajectoire issue du modèle de trois puits, $T = 100\,\text{ps}$, $\Delta t = 4\,\text{fs}$, $\theta = 0.5\,\text{nm}^2{\cdot}\text{ns}^{-1}$; Droite : trajectoire brownienne, $T = 100\,\text{ps}$, $\Delta t = 4\,\text{fs}$; Croix rouges : conformations initiales.

suggérons alors de suivre l'évolution du *rayon maximum* associé au processus, défini de la façon suivante :

Définition 4. (Rayon maximum) *Soit $X = (X_t)_{t\geq 0}$ vérifiant (3.1). Pour tous $0 \leq s \leq t$, on appelle rayon maximum de X au temps t partant de s, noté $R_{\max}(s,t)$, la variable aléatoire :*

$$R_{\max}(s,t) \triangleq \max_{s\leq u\leq t} \{\|X_u - X_s\|\}\,.$$

En guise d'exemple, étudions le profil de ce rayon maximum à l'aide de deux trajectoires (observées à un pas assez fin de sorte à ce qu'elles puissent être considérées comme continues) : une EDS établie sur le paysage des trois puits (voir Annexe A.1.1), et une trajectoire brownienne dans $\mathbb{R}^2$ (voir Fig. 3.1).

$$\begin{cases} dX_t = -\nabla U_c(X_t)dt + \sigma dB_t \quad \text{(trois puits)} \\ \\ dX_t = dB_t \quad \text{(mouvement brownien)} \end{cases}$$

où la fonction U_c est tirée de [55], $(B_t)_{t\geq 0}$ désigne un mouvement brownien bi-dimensionnel, et $\sigma = \sqrt{2\theta}$, avec $\theta = 0.5\ \text{nm}^2{\cdot}\text{ns}^{-1}$ correspondant à la température du processus.

Au regard de la Fig. 3.2, nous comprenons qu'en présence de puits, cette variable présentera plusieurs paliers de saturation traduisant le piégeage local du processus au sein d'un puits : le rayon maximum cesse de croître, et la durée du palier correspond approximativement à une réalisation du temps de sortie du puits. A l'inverse, si le paysage énergétique ne présente aucun relief, comme c'est le cas en présence d'un mouvement brownien, alors ce rayon maximum n'aura de cesse de croître, en présentant des paliers d'une durée de plus en plus grande. Formellement, pour un mouvement brownien bi-dimensionnel $(B_t)_{t\geq 0}$ et $r > 0$, en notant $\tau_r = \inf\{t \geq 0 : \|B_t\| \geq r\}$, on a que τ_r est une fonction croissante de r à $\omega \in \Omega$ fixé. En effet, pour $r \leq r'$, on a que $\{t \geq 0 : \|B_t\| \geq r\} \subset \{t \geq 0 : \|B_t\| \geq r'\}$, donc $\tau_r \leq \tau_{r'}$. Plus précisément, on a que $\tau_r \underset{r\to\infty}{\longrightarrow} \infty$ (p.s).

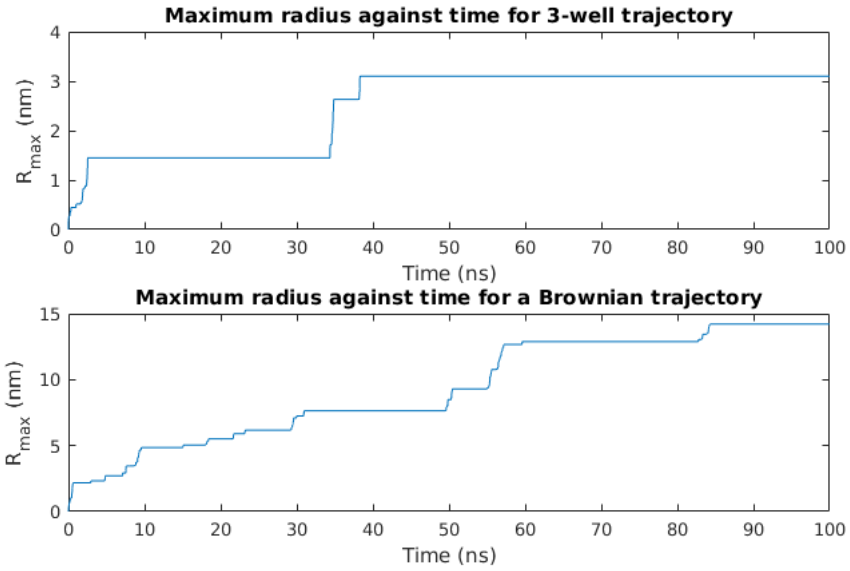

FIGURE 3.2 – Haut : profil de $t \to R_{\max}(0,t)$ dans le cas issu des trois puits ; Bas : profil de $t \to R_{\max}(0,t)$ dans le cas brownien.

3.1.2 Définition du critère κ

Jusqu'ici, juger de l'existence et de la pertinence d'équilibres locaux à la seule lecture de l'évolution de $R_{\max}$ est difficile. En effet, bien que la considération de $R_{\max}$ enrichisse l'approche par RMSD en ce sens qu'il est désormais possible de faire varier le point de départ, ces deux outils à eux-seuls présentent le même inconvénient, à savoir que constater une stagnation de $R_{\max}$ ou de la RMSD ne suffit pas à conclure quant à l'observation d'un phénomène de stabilité significatif. Il nous faut compléter la définition du rayon maximum afin d'obtenir un critère plus précis.

L'idée principale consiste alors à quantifier la profondeur d'un puits par le quotient de la distance parcourue par le processus pendant la durée de visite du puits *s'il était un mouvement brownien* (c'est-à-dire par pure diffusion, avec $\mu \equiv 0$), par le rayon maximum introduit précédemment, i.e. la distance maximum effectivement parcourue pendant ce même temps par le processus *contraint par le paysage énergétique* (c'est-à-dire en tenant compte des effets de dérive).

Définition 5. (Critère κ) *Soit $X = (X_t)_{t\geq 0}$ vérifiant (3.1).*

1. *Pour $0 \leq s \leq t$, on note $\kappa(s,t)$ la variable aléatoire relative à la trajectoire X prise entre les instants s et t définie par :*

$$\kappa(s,t) \triangleq \frac{\langle X\rangle_t - \langle X\rangle_s}{{R_{\max}}^2(s,t)} \tag{3.2}$$

où $\langle X\rangle_u \triangleq \int_0^u \|\boldsymbol{\sigma}(X_r)\|_{HS}^2 dr$ pour $u \geq 0$ est le crochet du processus X (ici pour $A = (A_{ij})_{1\leq i,j\leq d} \in \mathcal{M}_d(\mathbb{R})$, $\|A\|_{HS}^2 \triangleq \sum_{i,j} A_{ij}^2$ désigne le carré de la norme de Hilbert-Schmidt).

2. *Dans le cas où $\boldsymbol{\sigma} = \sigma \mathrm{I}_d$, on a :*

$$\kappa(s,t) \triangleq \frac{D(t-s)}{{R_{\max}}^2(s,t)} \tag{3.3}$$

f où $D \triangleq \sigma^2 d$ est le coefficient constant de diffusion du processus.

Remarque 15. *Dans la suite, nous nous placerons le plus souvent dans le cas 2. qui inclut le cas d'une diffusion de Langevin suramortie. En effet, ce dernier cas correspond d'une part au cadre conceptuel le plus naturel (voir Chapitre 2 §2.1.3), et d'autre part au seul cas utilisable en pratique puisqu'il est peu réaliste de chercher à estimer le champ matriciel $\boldsymbol{\sigma}$. En revanche, nous avons bien vu que l'estimation de D pouvait s'obtenir par le calcul de la variation quadratique de la trajectoire $(X_t)_{t\geq 0}$ (voir Chapitre 2 §2.2.3).*

A $s \geq 0$ fixé, une croissance linéaire de $t \mapsto \kappa(s,t)$ traduira donc le piégeage local au sein d'un puits conséquence de la stagnation de $R_{\max}$ (dénominateur) alors que le temps continue de croître linéairement (numérateur). A l'inverse, une chute brutale du nombre de tours indiquera une sortie de puits, conséquence de l'augmentation de $R_{\max}$ faisant diminuer κ après que cette dernière quantité ait atteint un maximum local. Nous illustrons ces phénomènes à l'aide des deux trajectoires évoquées précédemment (voir Fig. 3.3).

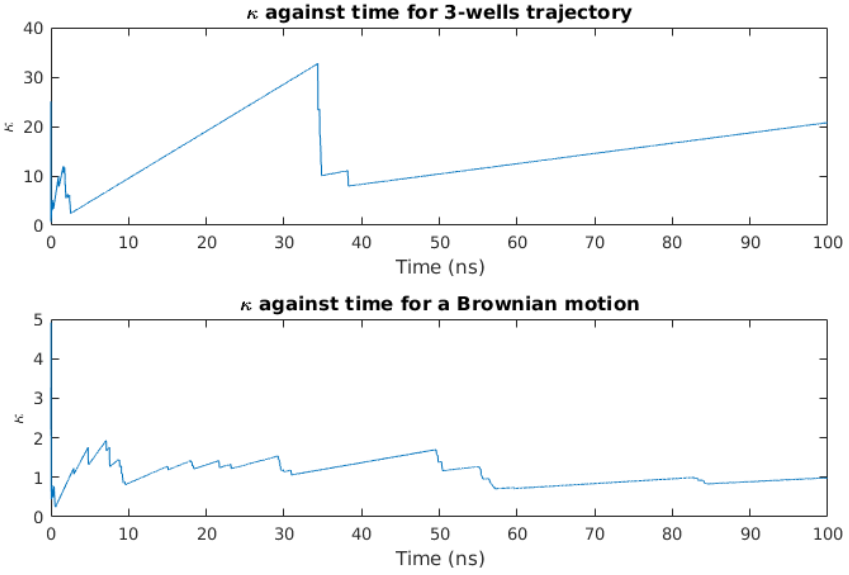

FIGURE 3.3 – Haut : profil de κ pour la trajectoire issue des trois puits ; Bas : profil de κ pour la trajectoire brownienne.

3.1.3 Propriétés de κ

Nous mettons ici en évidence quelques propriétés de κ, notamment son invariance par changement d'échelle (temps et espace) ainsi que quelques phénomènes asymptotiques remarquables dans le cas où $(X_t)_{t\geq 0}$ est un mouvement brownien ($\mu \equiv 0$).

Invariance par changement d'échelle temps-espace

Proposition 4. *(Invariance par changement d'échelle temps-espace) Soit $X = (X_t)_{t\geq 0}$ vérifiant (3.1) et Y_t le processus défini par : $Y_t = \rho X_{t/\alpha}$, avec $\rho, \alpha > 0$. Alors, en notant κ_X et κ_Y les valeurs de κ associées aux processus X et Y respectivement, on a que pour tous $0 \leq s \leq t$:*

$$\kappa_Y(s,t) = \kappa_X(s/\alpha, t/\alpha) \ p.s. \tag{3.4}$$

Démonstration. Par définition de Y on a $\langle Y\rangle_t = \rho^2\langle X\rangle_{t/\alpha}$. Par ailleurs, en notant $R_{X,\max}$ et $R_{Y,\max}$ les rayons maximums associés à X et Y respectivement, on a que :

$$R^2_{Y,\max}(s,t) = \rho^2 R^2_{X,\max}(s/\alpha, t/\alpha)$$

Par suite on a que :

$$\kappa_Y(s,t) = \frac{\langle Y\rangle_t - \langle Y\rangle_s}{R^2_{Y,\max}(s,t)} = \frac{\rho^2(\langle X\rangle_{t/\alpha} - \langle X\rangle_{s/\alpha})}{\rho^2 R^2_{X,\max}(s/\alpha, t/\alpha)} = \kappa_X(s/\alpha, t/\alpha)\ p.s.$$

□

Cette propriété rend robuste la quantification des puits par le critère κ, lequel demeure donc insensible aux changements d'échelles d'observation d'un processus. Par ailleurs, compte tenu des manipulations que nous effectuons sur les données telles que l'ACP, cette propriété assure que le critère κ garde un sens après projection.

Comportements asymptotiques (cas brownien)

Comme nous allons le voir, les propriétés d'invariance de $\kappa(s,t)$ induisent un certain nombre de conséquences concernant le comportement asymptotique de $t \to \kappa(s,t)$ lorsque $t \to s^+$ ou $t \to +\infty$.

Proposition 5 (Comportements asymptotiques pour un mouvement brownien)**.** *Soit $B = (B_t)_{t\geq 0}$ un mouvement brownien à valeurs dans $\mathbb{R}^d$. Alors on a les propositions suivantes.*

1. *Pour tous $0 \leq s \leq t$, $\kappa(s,t) \overset{loi}{=} \kappa_d$, où κ_d est la loi indépendante de s et t et définie sur $\mathbb{R}_+$ comme la loi de la variable :*
$$d/\sup_{0\leq t\leq 1} \|B_t\|^2$$
2. *Pour $s \geq 0$ fixé, on a :*
$$\liminf_{t\to s+} \kappa(s,t) = 0 \ \textit{et}\ \limsup_{t\to s+} \kappa(s,t) = +\infty$$
3. *Pour $s \geq 0$ fixé, on a :*
$$\liminf_{t\to +\infty} \kappa(s,t) = 0 \ \textit{et}\ \limsup_{t\to +\infty} \kappa(s,t) = +\infty$$

Démonstration. 1. Soit $\alpha > 0$. Considérons le changement d'échelle $t \leftarrow t/\alpha$ et le processus associé $B^{(\alpha)}_t = \sqrt{\alpha} B_{t/\alpha}$, lequel reste un mouvement brownien. Nous avons alors que pour tous $0 \leq s \leq t$:

$$\kappa(s,t) \overset{\text{loi}}{=} \kappa(0, t-s) \overset{\text{loi}}{=} \frac{d(t-s)}{\sup\limits_{0<u\leq t-s} \|B^{(\alpha)}_u\|^2} \overset{\text{loi}}{=} \frac{d(t-s)}{\alpha \sup\limits_{0\leq r\leq (t-s)/\alpha} \|B_r\|^2} \overset{\text{loi}}{=} \kappa(0, (t-s)/\alpha)$$

Quitte à prendre $\alpha = t - s$, on a que pour tous $0 \leq s \leq t$, $\kappa(s,t) \overset{\text{loi}}{=} \kappa(0,1) \triangleq \kappa_d$, qui correspond à la loi de la variable $d/\sup\limits_{0\leq t\leq 1} \|B_t\|^2$.

2. Comme la propriété ne dépend que de $B_{s+h} - B_s$ pour $h \geq 0$, il suffit de la montrer pour $s = 0$. En introduisant la filtration $\mathcal{F}_t \triangleq \sigma(B_t \mid t \geq 0)$ et $\mathcal{F}_{0^+} \triangleq \cap_{t>0}\mathcal{F}_t$, on a par la loi du 0-1 de Blumenthal que $\mathcal{F}_{0^+}$ est triviale. Or, $\kappa^*(0,0) \triangleq \limsup_{t\to 0+} \kappa(0,t)$ et $\kappa_*(0,0) \triangleq \liminf_{t\to 0^+} \kappa(0,t)$ est $\mathcal{F}_{0^+}$ mesurable puisque $\kappa(0,t) \in \mathcal{F}_t$ pour tout $t > 0$. Par suite, il existe deux constantes $C_*, C^* \in [0, +\infty]$ telles que presque sûrement, $\kappa_*(0,0) = C_*$ et $\kappa^*(0,0) = C^*$. Supposons que $C_* > 0$. On a alors par convergence dominée que $\lim_{t\to 0+} \mathbb{P}[\kappa(0,t) \leq C_*/2] = 0$ puisque $\lim_{t\to 0^+} \mathbf{1}_{\kappa(0,t)\leq C_*/2} = 0$. Or, ceci est absurde car d'après le point 1., on a $\mathbb{P}[\kappa(0,t) \leq C_*/2] = \mathbb{P}[\kappa_d \leq C_*/2] > 0$ pour tout $t > 0$. De même, si $C^* < +\infty$, alors $\lim_{t\to 0^+} \mathbb{P}[\kappa(0,t) \geq 2C^*] = 0$ ce qui contredit $\mathbb{P}[\kappa(0,t) \geq 2C^*] = \mathbb{P}[\kappa_d \geq 2C^*] > 0$ pour tout $t > 0$.
3. La preuve peut se faire en utilisant la loi du 0-1 de Kolmogorov. En notant $X_n \triangleq (B_t - B_n)_{t\in[n,n+1[}$, on remarque que les X_n sont i.i.d., que $\mathcal{F}_n^+ \triangleq \sigma(X_l \mid l \geq n) = \sigma(B_t - B_n \mid t \geq n)$. Par suite, la tribu $\mathcal{F}_\infty^+ = \cap_{n\geq 0}\mathcal{F}_n^+$ est triviale. Or, comme p.s. on a pour $n \geq s$

$$\lim_{t\to+\infty} \frac{\sup_{s\leq u\leq t} \|B_u - B_s\|^2}{\sup_{n\leq u\leq t} \|B_u - B_n\|^2} = 1 \tag{3.5}$$

on déduit que $\kappa^*(s, +\infty) \triangleq \limsup_{t\to+\infty} \kappa(s,t)$ et $\kappa_*(s, +\infty) \triangleq \liminf_{t\to+\infty} \kappa(s,t)$ sont $\mathcal{F}_\infty^+$-mesurables. Comme la tribu est triviale, il existe $D_*,\ D^* \in [0, +\infty]$ telles que presque sûrement, $\kappa^*(s, +\infty) = D^*$ et $\kappa_*(s, +\infty) = D_*$. On peut procéder alors comme pour le point précédent. Supposons que $D_* > 0$. Alors, par convergence dominée, on a que $\lim_{t\to+\infty} \mathbb{P}[\kappa(s,t) \leq D_*/2] = 0$ puisque $\lim_{t\to+\infty} \mathbf{1}_{\kappa(s,t)\leq D_*/2} = 0$. Or, ceci est absurde car d'après le point 1., on a $\mathbb{P}[\kappa(s,t) \leq D_*/2] = \mathbb{P}[\kappa_d \leq D_*/2] > 0$ pour tout $t > 0$. On montre de façon similaire que $D^* = +\infty$ ce qui donne le résultat.

□

Remarque 16. *1. Nous pouvons mettre en avant une certaine symétrie entre le comportement en 0 et en $+\infty$ en faisant un changement de temps $h \to e^h$. En effet, si l'on définit pour $h \in \mathbb{R}$ le processus $Y_h \triangleq e^{-h}B_{e^{2h}}$, alors par calcul d'Itô, $M_h \triangleq Y_h + \int_0^h Y_u du$ est une martingale continue dont le crochet est $2h$, et donc (à un scaling par $\sqrt{2}$ près) un mouvement brownien β_h sur $\mathbb{R}$ tel que $\beta_0 = \sqrt{2}B_1$. On déduit alors que :*

$$dY_h = -Y_h dh + \sqrt{2}d\beta_h \tag{3.6}$$

est donc que Y_h est un processus d'Orstein-Uhlenbeck ou encore une diffusion de Langevin suramortie pour le potentiel harmonique $U(x) = |x|^2/2$ de mesure invariante la loi $\mathcal{N}(0, \mathrm{I}_d)$. C'est donc un processus stationnaire pour lequel on a :

$$\kappa(0,t) \leq \frac{d \times t}{\|B_t\|^2} = \frac{d}{\|Y_{\log(t)/2}\|^2} \tag{3.7}$$

et l'on voit apparaître clairement pour le majorant de $\kappa(0,t)$ donné par $\|Y_{\log(t)/2}\|^{-2}$ une symétrie entre $t \in]0,1]$ et $t \in [1, +\infty[$, une fois passé à l'échelle logarithmique sur le temps.

2. *Il est à noter ici que le comportement en temps long de $t \to \kappa(s,t)$ n'est pas le plus pertinent pour nos études, qui se feront toujours sur des intervalles de temps $[0,T]$ avec T fini. De fait, en utilisant la propriété d'invariance par renormalisation du temps, nous pouvons renormaliser le temps d'observation $[0,T]$ à $[0,1]$. Ainsi, c'est le comportement pour t au voisinage de s qui demande une étude spécifique que nous réalisons dans le paragraphe suivant.*

Comportement en s^+ de κ (cas brownien)

Le résultat précédent montre donc que pour une trajectoire brownienne, $t \to \kappa(s,t)$ peut prendre des valeurs arbitrairement grandes dans tout demi-voisinage $[s, s+\epsilon[$ à droite de s. Pour-

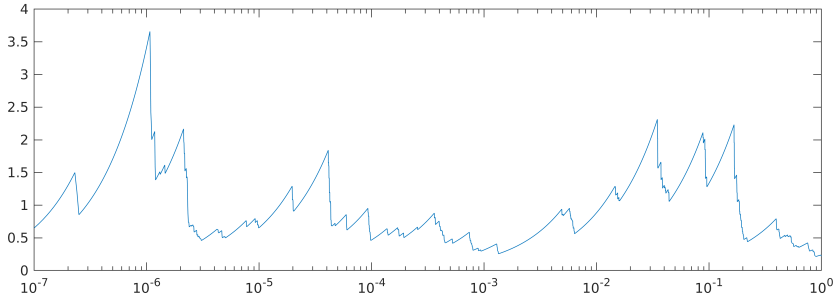

FIGURE 3.4 – Trajectoire $t \to \kappa(0,t)$ pour $t \in [0,1]$. En abscisse le temps t utilise une échelle logarithmique qui permet d'illustrer le comportement au voisinage de 0. La courbe est obtenue en simulant le mouvement brownien sur $[0,1]$ comme une marche aléatoire de pas gaussien $\mathcal{N}(0, 1/\sqrt{N})$ pour $N = 10^8$.

tant, la détection de puits passera par la recherche, pour différentes valeurs de s, de maximums (locaux) de $t \to \kappa(s,t)$. L'existence d'une infinité de maximums locaux à droite de s prenant des valeurs arbitrairement grandes semble être en contradiction avec notre objectif. Cette contradiction apparente peut être levée par une étude plus approfondie du comportement de $\kappa(0,t)$ sur un intervalle $[0,T]$.

La première remarque est que de par la propriété d'invariance (3.4), nous pouvons normaliser le temps $t \to t' \triangleq t/T$ et nous ramener au cas $T = 1$. La question centrale est alors l'étude de la variable aléatoire :

$$\kappa^*(s_0) \triangleq \sup_{t \in [s_0,1]} \kappa(0,t) = \sup_{t \in [s_0,1]} \frac{d \times t}{\sup_{u \in [0,t]} \|B_u\|^2} \tag{3.8}$$

Le point clé est que si $\kappa^*(s_0) \to +\infty$ lorsque $t \to 0^+$, la valeur de $\kappa^*(s_0)$ décroît très vite avec s_0. Nous pouvons illustrer ce comportement par le résultat théorique suivant.

Proposition 6. *Pour tout $d \geq 1$, tout $\alpha > 0$ et tout $\epsilon > 0$, il existe $A > 0$ tel qu'avec probabilité $1-\alpha$ on ait pour tout $t \in]0,1]$:*

$$\kappa(0,t) \leq A \log(\frac{2}{t})^{\frac{2}{d}+\epsilon} \tag{3.9}$$

et en particulier pour tout $s_0 \in]0,1[$:

$$\kappa^*(s_0) \leq A \log(\frac{2}{s_0})^{\frac{2}{d}+\epsilon} \tag{3.10}$$

Démonstration. La preuve de ce résultat s'obtient facilement une fois remarqué que, pour tout $n \geq 0$ et pour tout $t \in]2^{-(n+1)}, 2^{-n}]$, on a $\kappa(0,t) \leq \frac{d \times t}{\|B_{2^{-(n+1)}}\|^2} \leq 2\frac{d \times 2^{-(n+1)}}{\|B_{2^{-(n+1)}}\|^2}$. Par suite, comme la propriété de scaling du mouvement brownien donne $B_s/\sqrt{s} \overset{\text{loi}}{=} B_1$, on déduit pour $M_n \triangleq \sup_{t \in]2^{-(n+1)}, 2^{-n}]}\{\kappa(0,t)\}$ que :

$$\mathbb{P}[M_n > z_n] \leq \mathbb{P}\left[\|B_1\| < \left(\frac{2d}{z_n}\right)^{1/2}\right] \tag{3.11}$$

Comme il existe $c > 0$ tel que l'on ait $\mathbb{P}[\|B_1\| < r] \leq cr^d$ pour $r > 0$, on déduit que pour a assez grand et $z_n = a(n+1)^{2/d+\epsilon}$ on a :

$$\sum_{n\geq 0} \mathbb{P}[M_n > z_n] \leq \sum_{n\geq 0} c\left(\frac{2d}{a(n+1)^{2/d+\epsilon}}\right)^{d/2} \leq \alpha \tag{3.12}$$

On obtient alors le résultat en posant $A = a/\log(2)^{2/d+\epsilon}$ et en remarquant pour tout $t \in]2^{-(n+1)}, 2^{-n}]$, on a $(n+1) \leq \log(2/t)/\log(2)$. □

Remarque 17. *On peut par le même procédé de preuve (découpage sur les intervalles $]2^n, 2^{n+1}]$) établir la proposition suivante :*

Proposition 7. *Pour tout $d \geq 1$, tout $\alpha > 0$ et tout $\epsilon > 0$, il existe $A > 0$ tel que avec probabilité $1-\alpha$ on ait pour tout $t \in [1, \infty[$*

$$\kappa(0,t) \leq A\log(2t)^{\frac{2}{d}+\epsilon} \tag{3.13}$$

Ce résultat est le pendant en $+\infty$ du résultat en 0. En particulier, on voit que les valeurs de $\kappa(0,t)$ doivent rester sous une enveloppe à croissance très lente. Rappelons à nouveau cependant que par la normalisation du temps, nous pouvons nous ramener à l'étude au comportement de κ dans $t \in]0,1]$.

Les estimées théoriques peuvent être complétées par une étude empirique. Pour cela nous reportons ci-dessous les histogrammes de $\kappa^*(s_0)$ pour trois valeurs de s_0 proches de 0.

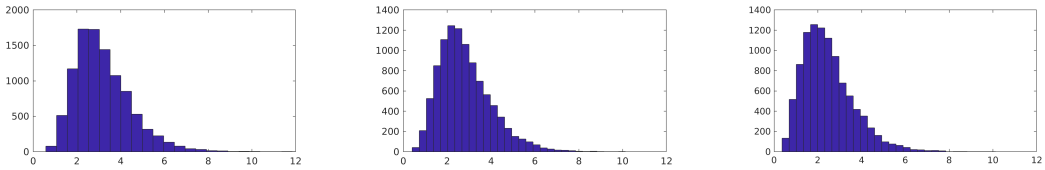

FIGURE 3.5 – Histogramme de $\kappa^*(s_0)$, calculé à partir d'un n-échantillon de taille $n = 10000$, pour les valeurs de $s_0 = 10^{-4}$ (gauche), $s_0 = 10^{-3}$ (milieu) et $s_0 = 10^{-2}$ dans le cas $d = 2$.

Le fait dominant dans la Fig. 3.5 est que même pour une valeur de s_0 très proche de 0, par exemple $s_0 = 10^{-4}$, le sup de $\kappa(0,t)$ sur l'intervalle $[s_0, 1]$ prend des valeurs en dessous de 10.

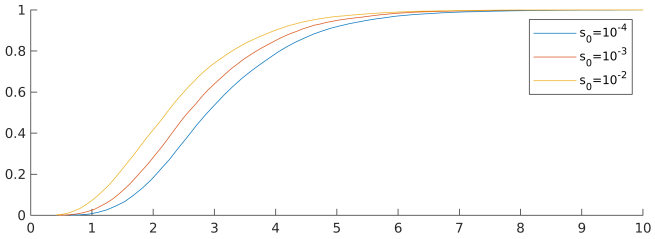

FIGURE 3.6 – Fonction de répartition de $\kappa^*(s_0)$, calculée à partir d'un n-échantillon de taille $n = 10000$, pour les valeurs de $s_0 \in \{10^{-4}, 10^{-3}, 10^{-2}\}$ $(d = 2)$.

En particulier, le quantile empirique $\hat{q}_n(0.99)$ au dernier centile sur notre n-échantillon donne :

s_0	10^{-4}	10^{-3}	10^{-2}
$\hat{q}_n(0.99)$	6.95	6.40	5.97

La retombée pratique de ce résultat est qu'il est très peu probable, en l'absence de puits, i.e. dans le cas d'un mouvement brownien, de trouver des valeurs de $\kappa(0,t) \geq 10$ pour des $t \in [s_0, 1]$. En revenant à la situation non-normalisée sur $[0,T]$, cela nous dit qu'une fois éliminée la portion des $t \in [0, s_0 T]$, il est très peu probable que le max de $\kappa(0,t)$ pour $t \in [s_0 T, T]$ dépasse la valeur 10. La valeur $s_0 = 10^{-4}$ donne pour une simulation typique de l'ordre de la µs la nécessité d'éliminer de l'analyse une portion de l'ordre de 100 ps, ce qui est très largement en dessous des temps caractéristiques de sortie des états d'équilibres qui sont susceptibles de nous intéresser.

3.1.4 Instant d'accès, instant de sortie et temps de sortie

On se propose ici de caractériser un puits par une valeur de κ, ce qui nécessite donc de définir les valeurs de s et t en conséquence. Plus précisément, on cherche le couple (s,t) qui sera en mesure de maximiser la valeur de κ caractérisant un puits. A cet égard, nous montrons dans ce qui suit que la valeur de $\kappa(s,t)$ est sensible au choix de s.

Tout d'abord, une façon de décrire un puits est de considérer une boule fermée de $\mathbb{R}^d$ centrée en x_0, et de rayon R_0, notée $\mathcal{B}(x_0, R_0) = \{x \in \mathbb{R}^d : \|x - x_0\| \leq R_0\}$. On considère alors l'ensemble $\mathcal{S} = \{s_0 > 0 : X_{s_0} \in \overset{\circ}{\mathcal{B}}(x_0, R_0)\}$, qui en tant qu'ouvert de $\mathbb{R}_+$, peut s'écrire comme une union dénombrable d'intervalles ouverts. Pour $s_0 \in \mathcal{S}$, on peut donc définir les instants :

$$\begin{cases} s^* = \sup\{s \leq s_0 : \|X_s - x_0\| \geq R_0\} \\ t^* = \inf\{t \geq s_0 : \|X_t - x_0\| \geq R_0\} \end{cases} \tag{3.14}$$

de sorte que pour tout $u \in]s^*, t^*[$ on a que $X_u \in \overset{\circ}{\mathcal{B}}(x_0, R_0)$. La donnée d'un $s_0 \in \mathcal{S}$ permet ainsi de définir une *excursion* au sein du puits, délimitée par son *instant initial* s^* et son *instant final* t^*. Cette excursion peut être *courte* ou *longue*, selon le choix de s_0. Dans le premier cas, la trajectoire reste proche du bord du puits pendant une courte durée avant d'en ressortir (la trajectoire ne "tombe" pas au fond du puits). Dans le second cas, la trajectoire se dirige résolument vers le centre x_0 du puits, et n'en sort qu'après une durée correspondant à une réalisation du temps de sortie du puits (la trajectoire "tombe" au fond du puits). La Fig. 3.7 illustre ces deux types d'excursions possibles à l'aide d'une même EDS simulée sur le paysage des trois puits, avec $x_0 = (-1, 0)$, $R_0 = 1$ nm, $T = 25$ ns, et $\Delta t = 0.2$ ps. Dans le premier cas, on choisit s_0 tel que $X_{s_0} = (-1, 3/4)$ (gauche), tandis que dans le second cas, on choisit s_0 correspondant à l'instant auquel le processus se trouve le plus proche du centre x_0 du puits : $s_0 = \text{argmin}_{t \geq 0} \|X_t - x_0\|$ (droite).

Dans la suite, nous nous intéresserons plus spécifiquement aux excursions longues, lesquelles visitent donc une grande partie du puits de par les nombreuses tentatives de sorties effectuées (à basse température).

Remarque 18. *Si nous avons ici défini un puits comme une région de l'espace délimitée par une boule, il est possible d'envisager une définition invoquant la notion d'énergie libre (voir Chapitre 2 §2.1.3). Plus précisément, et en reprenant les notations du chapitre précédent, on peut définir des niveaux de puits délimités par des lignes de niveaux $\mathcal{L}_u = \{x \in \mathbb{R}^d : U_\beta^{\mathcal{X}}(x) = u\}$, pour un niveau d'énergie $u > 0$. L'ensemble $\mathcal{X}_u = \{x \in \mathbb{R}^d : U_\beta^{\mathcal{X}}(x) \leq u\}$ des conformations d'énergie libre*

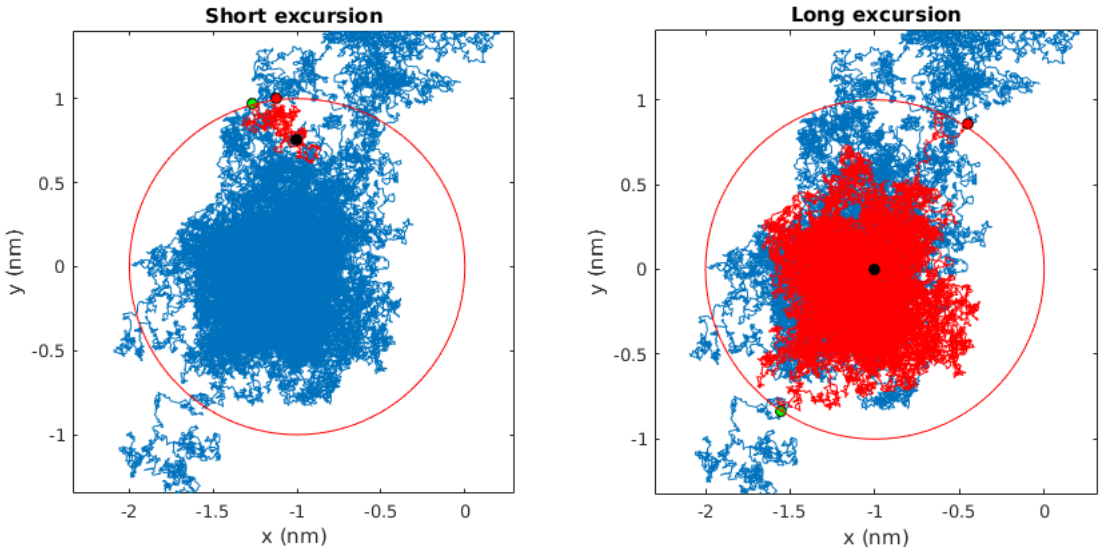

FIGURE 3.7 – Excursions courte et longue sur une trajectoire issue du modèle de trois puits partant de $(-2,-1)$ avec $\theta = 0.5$ nm^2·ns^{-1}, $T = 25$ ns et $\Delta t = 0.2$ ps. Bleu : trajectoire totale ; Point noir : conformation X_{s_0} la plus proche du centre x_0 ; Point vert : première conformation hors puits en rétrogradant depuis X_{s_0} ; Point rouge ; première conformation hors puits en progressant depuis X_{s_0}. Gauche : s_0 tel que $X_{s_0} = (-1, 3/4)$; Droite : $s_0 = \text{argmin}_{t\geq 0}\|X_t - x_0\|$.

inférieure à u peut alors s'écrire comme une réunion de parties connexes de $\mathbb{R}^d$ *définissant les puits (qui ne sont plus alors des boules) de niveau u. L'étude de l'organisation de ces ensembles connexes vis-à-vis de l'inclusion en fonction du niveau u est au coeur des analyses du paysage d'énergie libre (lorsque celle-ci est disponible) sous l'angle des fonctions de Morse et des diagrammes de persistance [18]. Dans les situations d'usage dans notre travail, la taille des systèmes considérés ne permet pas* a priori *une accession directe à l'énergie libre.*

Revenons à la caractérisation d'un puits par le critère κ. Dans l'idée qu'à t fixé le choix de s, et plus spécifiquement la proximité de la conformation X_s vis-à-vis du centre du puits, influe sur la maximisation de $\kappa(s,t)$, nous allons faire varier l'instant s et étudier son impact sur le profil de $t \to \kappa(s,t)$. Pour ce faire, nous considérons une famille de boules $(\mathcal{B}_\alpha(x_0, R_0))_{\alpha>0}$, où chaque boule est définie par $\mathcal{B}_\alpha(x_0, R_0) = \{x \in \mathbb{R}^d : \|x - x_0\| \leq R_\alpha\}$, avec $R_\alpha = \alpha R_0$. Pour un $s_0 \in \mathcal{S}$, et pour $\alpha > 0$, on définit alors :

$$s^*_\alpha = \sup\{s \leq s_0 : \|X_s - x_0\| \geq R_\alpha\}$$

En reprenant l'EDS précédente simulée sur le modèle des trois puits et le puits de centre $x_0 = (-1, 0)$ et de rayon $R_0 = 1$ nm, la Fig. 3.8 représente les différentes boules $(\mathcal{B}_\alpha(x_0, R_0))_{\alpha>0}$ pour $\alpha \in \{0.15, 0.25, 0.5, 0.75, 1\}$ (gauche) et les profils $t \to \kappa(s^*_\alpha, t)$ correspondants (droite), calculés, conformément à ce qui a été vu au Chapitre 3 §3.1.3, pour des $t \in [s^*_\alpha + \epsilon, t^* + \epsilon]$ de sorte à bannir d'éventuelles valeurs trop extrêmes de $\kappa(s,t)$ au voisinage de s^+, et à observer le comportement de $\kappa(s^*_\alpha, t)$ pour des valeurs de t légèrement supérieures à t^*. Dans les exemples qui suivent, on a $s_0 = \text{argmin}_{t\geq 0}\|X_t - x_0\|$ et $\epsilon = 2\%(t^* - s^*)$.

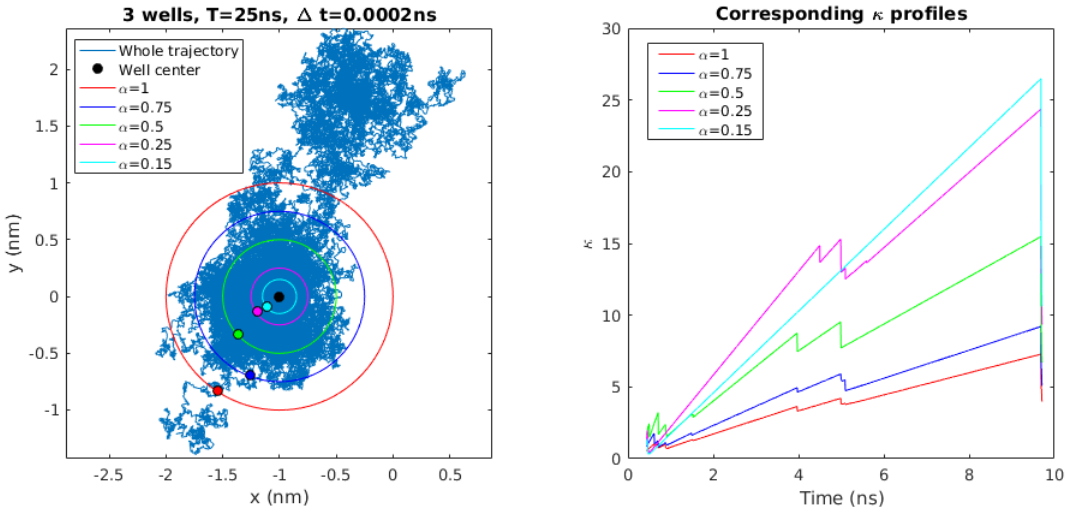

FIGURE 3.8 – Droite : tracé des boules fermées $\mathcal{B}_\alpha(x_0, R_0)$ et des $x_{s^*_\alpha}$ correspondants ; Gauche : profils des $\alpha \mapsto \kappa(s^*_\alpha, t)$, pour $t \in [s^*_\alpha + \epsilon, t^* + \epsilon]$.

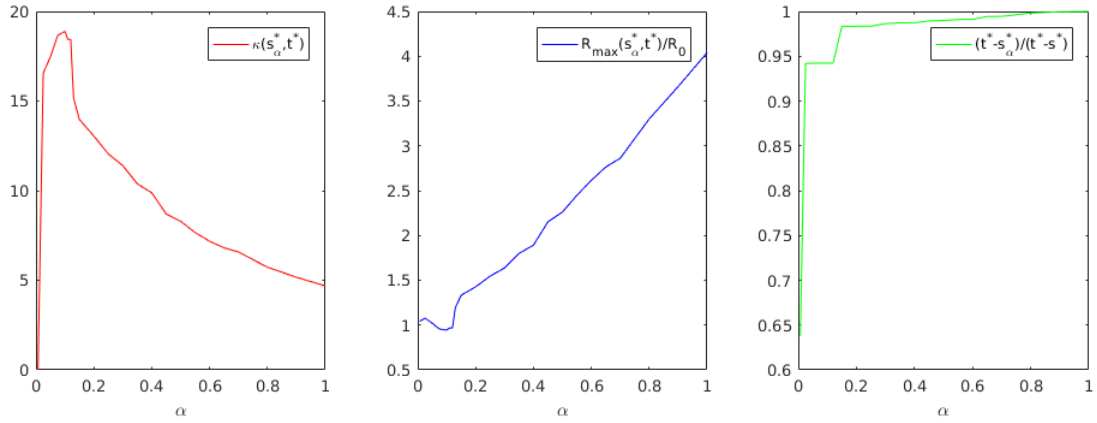

FIGURE 3.9 – Gauche : profil de $\alpha \mapsto \kappa(s^*_\alpha, t^*)$; Centre : profil de $\alpha \mapsto R_{\max}(s^*_\alpha, t^*)/R_0$; Droite : profil de $\alpha \mapsto (t^* - s^*_\alpha)/(t^* - s^*)$.

On constate que la valeur de $\kappa(s^*_\alpha, t^*)$ (Fig 3.9, gauche) croît lorsque $\alpha \to 0$, mais qu'elle décroît au voisinage de zéro. En effet, malgré un rayon $R_{\max}(s^*_\alpha, t^*)$ demeurant proche de R_0 au voisinage de 0 (Fig 3.9, centre), nous voyons que lorsque la localisation autour de x_0 devient trop sévère, la durée $t^* - s^*_\alpha$ décroît fortement (Fig 3.9, droite). En effet, plus α est proche de 0, plus l'instant s^*_α est rapidement identifié par la lecture arrière de la trajectoire depuis x_{s_0}, ce qui induit une perte de temps dans la lecture avant jusqu'à t^*, impactant alors l'optimisation de $\kappa(s^*_\alpha, t^*)$. Dès lors qu'est atteinte, dans cette lecture à l'envers, la portion de la trajectoire correspondant à la chute du processus au sein du puits, l'augmentation de α n'induit qu'une très faible croissance de la durée $t^* - s^*_\alpha$. Ceci traduit la rapidité de la chute du processus, lequel se dirige résolument vers le fond du puits.

Remarque 19. *Nous reprenons ici la même démarche à l'aide d'un nouveau processus, évoluant toujours dans le paysage des trois puits, mais de température plus basse* $\theta = 0.2$ $\mathrm{nm}^2 \cdot \mathrm{ns}^{-1}$ *(contre*

$\theta = 0.5$ nm^2·ns^{-1} *précédemment). Nous obtenons alors les résultats suivants :*

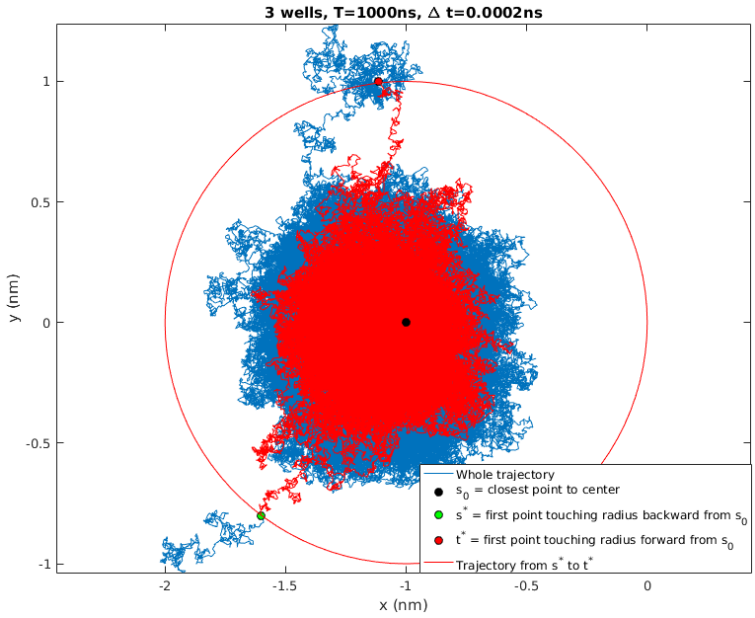

FIGURE 3.10 – Trajectoire issue du modèle de trois puits partant de $(-2,-1)$ avec $\theta = 0.2$ nm^2·ns^{-1}, $T = 1$ µs et $\Delta t = 0.2$ ps. Bleu : trajectoire totale ; Point noir : conformation X_{s_0} la plus proche du centre x_0 ; Point vert : première conformation hors puits en rétrogradant depuis X_{s_0} ; Point rouge ; première conformation hors puits en progressant depuis X_{s_0}. Choix de $s_0 = \text{argmin}_{t\geq 0}\|X_t - x_0\|$.

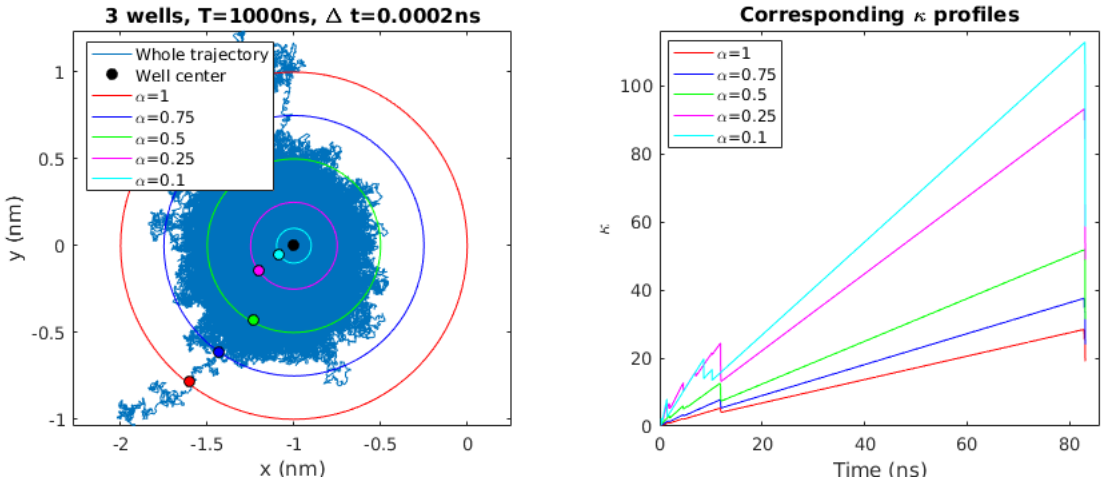

FIGURE 3.11 – Droite : tracé des boules fermées $\mathcal{B}_\alpha(x_0, R_0)$ et des $x_{s^*_\alpha}$ correspondants ; Gauche : profils des $\alpha \mapsto \kappa(s^*_\alpha, t)$, pour $t \in [s^*_\alpha + \epsilon, t^* + \epsilon]$.

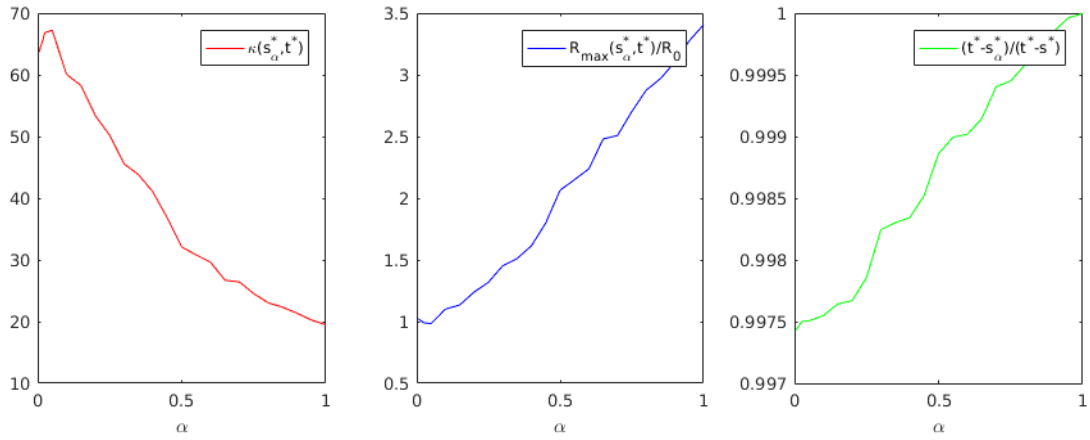

FIGURE 3.12 – Gauche : profil de $\alpha \mapsto \kappa(s^*_\alpha, t^*)$; Centre : profil de $\alpha \mapsto R_{\max}(s^*_\alpha, t^*)/R_0$; Droite : profil de $\alpha \mapsto (t^* - s^*_\alpha)/(t^* - s^*)$.

Les tracés suivants traduisent alors le fait qu'à basse température, le terme de diffusion devient négligeable, de sorte que l'évolution du processus approche le comportement limite décrit par l'équation : $dX_t = \mu(X_t)dt$ (i.e. pour $\boldsymbol{\sigma} \equiv 0$). D'une part, la chute du processus au sein du puits, qui n'est alors plus freinée par les effets de diffusion, s'effectue beaucoup plus rapidement à l'échelle de la durée totale de l'excursion $t^* - s^*$: en effet, même pour un voisinage très restreint autour du centre du puits, la durée $t^* - s^*_\alpha$ représente plus de 99% de la durée totale $t^* - s^*$ (voir Fig. 3.12, gauche). D'autre part, le temps $t^* - s^*_\alpha$ augmente considérablement passant de 10 ns *($\theta = 0.5$ nm^2·ns^{-1}, Fig. 3.9, droite) à plus de* 80 ns *($\theta = 0.2$ nm^2·ns^{-1}, Fig. 3.12, droite).*

Nous pouvons alors définir les instants d'accès, de sortie et le temps de sortie de la façon suivante.

Définition 6. (Instant d'accès, instant de sortie et temps de sortie) *Soit $X = (X_t)_{t\geq 0}$ vérifiant* (3.1)*, $\mathcal{B}(x_0, R_0) = \{x \in \mathbb{R}^d : \|x - x_0\| \leq R_0\}$ une boule fermée candidate pour représenter un puits, $s_0 = \text{argmin}_{t\geq 0}\|X_t - x_0\|$, (s^*, t^*) définis par* (3.14)*, et $\epsilon = \gamma(t^* - s^*)$, avec $\gamma \in]0, 1[$. On appelle instant d'accès au puits et instant de sortie du puits, et on note respectivement $\tilde{s}$ et $\tilde{t}$ les variables définies par :*

$$\begin{cases} \tilde{s} = s^*_{\tilde{\alpha}} \quad où \quad \tilde{\alpha} = \underset{\alpha>0}{\text{argmax}}\{\kappa(s^*_\alpha, t^*)\} \\ \tilde{t} = \underset{t\in[\tilde{s}+\epsilon, t^*]}{\text{argmax}}\{\kappa(\tilde{s}, t)\} \end{cases}$$

On définit alors le temps de sortie du puits par la variable :

$$\widetilde{T} = \tilde{t} - \tilde{s}$$

Remarque 20. *Les instants (s^*, t^*) et $(\tilde{s}, \tilde{t})$ diffèrent donc en ce sens que les premiers se réfèrent à une notion géométrique de franchissement de frontière, alors que les seconds correspondent à des instants optimaux au sens d'une maximisation locale de κ. Si t^* et $\tilde{t}$ indiquent des conformations à la fois proche dans l'espace et dans le temps, les instants s^* et $\tilde{s}$ renvoient quant à eux des conformations éloignées dans l'espace (d'une distance très proche de R_0), mais très proches dans le temps (et d'autant plus proches que la température est basse).*

Définition 7. (Centre d'un puits) *Soit $X = (X_t)_{t\geq 0}$ vérifiant (3.1), $\mathcal{B}(x_0, R_0) = \{x \in \mathbb{R}^d : \|x - x_0\| \leq R_0\}$ une boule fermée candidate pour représenter un puits, et $\tilde{s}$ l'instant d'accès au puits. On définit le centre du puits par la conformation $X_{\tilde{s}}$.*

Remarque 21. *La conformation donnée par l'instant d'accès dans le puits, et celle donnée par le centre géométrique du puits correspondent donc, au regard de κ, à une seule et même notion.*

Le puits étudié est ainsi caractérisé par la valeur κ_0 définie par :

$$\kappa_0 = \kappa(\tilde{s}, \tilde{t}) = \frac{D \times (\tilde{t} - \tilde{s})}{{R_{\max}}^2(\tilde{s}, \tilde{t})}$$

Quitte à diviser κ_0 au numérateur et au dénominateur par D, nous pouvons considérer l'écriture suivante :

$$\kappa_0 = \frac{\widetilde{T}}{(\Delta t)_{R_0}}$$

où $(\Delta t)_{R_0}$ correspond au temps nécessaire au processus pour parcourir une distance de R_0 par pure diffusion : $D \times (\Delta t)_{R_0} = R_0^2$. D'un certain point de vue, on compte alors le nombre de visites du puits que le processus aura effectuées avant de le quitter pendant le temps de sortie $\widetilde{T}$, ou encore le nombre de tentatives de sorties infructueuses pendant ce temps : d'où l'appellation *nombre de tours* (en anglais, *laps number*).

A ce stade, le critère κ constitue une variable sans dimension permettant de quantifier la profondeur des puits par l'atteinte de maximums locaux de $\kappa(s, t)$. Cette quantification est absolue et permet, par la considération d'ordres de grandeurs, de juger de la profondeur des puits, de rejeter les artefacts et de comparer les puits entre eux.

Remarque 22. *Comme nous le verrons plus tard (Chapitre 5 §5.1), la notion de nombre de tours pourra être directement mise en correspondance avec la notion de hauteur d'un puits via la représentation équivalente de tout processus dans le paysage des trois puits.*

3.1.5 Discrétisation

En pratique, l'étude d'une trajectoire de simulation de DM de durée T correspond à l'observation d'une *unique réalisation* du processus $X = (X_t)_{t\in[0,T]}$ à valeurs dans $\mathbb{R}^d$ (c'est-à-dire que l'on travaille à $\omega \in \Omega$ fixé), laquelle est de surcroît sujette à un *échantillonnage* selon un pas de temps régulier Δt, duquel résulte un total de N conformations observées. L'objet étudié sera donc une matrice de $\mathcal{M}_{d,N}(\mathbb{R})$ notée $\mathbf{X} = (x_i)_{1\leq i\leq N}$, avec $x_i \in \mathbb{R}^d$, telle que pour tout $i \in [\![1, N]\!]$:

$$x_i = X_{(i-1)\Delta t} \tag{3.15}$$

Remarque 23. *Ce que nous considérons formellement comme le processus à temps continu $X = (X_t)_{t\in[0,T]}$ correspond en réalité à une trajectoire discrète et déterministe, mais finement simulée par GROMACS, avec un pas de simulation d'une valeur typique de $(\Delta t)_{\text{sim}} = 2$ fs, et un nombre de pas à simuler $N_{\text{sim}} - 1$ défini par l'utilisateur, engendrant un total de N_{sim} conformations observées (comprenant la conformation initiale), pour une durée d'observation $T = (N_{\text{sim}} - 1)(\Delta t)_{\text{sim}}$. Ce dernier fixe également un pas d'échantillonnage Δt, de sorte à récupérer en fin de simulation une version échantillonnée de la trajectoire, fournissant cette fois-ci N conformations : c'est ce qui correspond ici à la matrice $\mathbf{X} = (x_i)_{1\leq i\leq N}$. En pratique, le pas Δt est choisi de sorte à ce que*

l'on ait toujours exactement que $T = (N_{\text{sim}} - 1)(\Delta t)_{\text{sim}} = (N-1)\Delta t$ *(i.e. aucune partie entière n'intervient).*

Nous proposons alors dans ce qui suit de passer en revue les notions introduites précédemment dans leurs équivalents discrets.

Définition 8. (Rayon maximum - version discrète) *Soit* $\mathbf{X} = (x_i)_{1\leq i\leq N}$ *construite selon* (3.15)*. Pour* $i \in [\![1, N]\!]$ *et* $j \geq i$*, on appelle rayon maximum de* $\mathbf{X}$ *en* j *partant de* i*, et on note* $R_{\max}(i, j)$ *la quantité :*

$$R_{\max}(i,j) = \sup_{k\in[\![i,j]\!]} \{\|x_k - x_i\|\}$$

Dans le cas continu, le nombre de tours fait intervenir un coefficient de diffusion aisément défini par $D = \sigma^2 d$. En pratique, la version discrète de κ nécessite une estimation de D à partir de la discrétisation observée, notée $\mathbf{D}$ (voir Chapitre 2 §2.2.3). Cette quantité sera dans ce suit considérée comme une donnée.

Définition 9. (Nombre de tours - version discrète) *Soit* $\mathbf{X} = (x_i)_{1\leq i\leq N}$ *construite selon* (3.15)*. Pour* $i \in [\![1, N]\!]$ *et* $j \geq i$*, on appelle nombre de tours de* $\mathbf{X}$ *entre* i *et* i*, et on note* $\kappa(i, j)$ *la quantité :*

$$\kappa(i,j) = \frac{(t_j - t_i) \times \mathbf{D}}{{R_{\max}}^2(i,j)}$$

où $t_k = (k-1)\Delta t$*.*

Remarque 24. *Notons que le pas d'échantillonnage* Δt *n'influe pas sur le calcul du nombre de tours discret.*

Remarque 25. *On gardera à l'esprit que le calcul des nombres de tours en dimension* $k \leq d$ *nécessite le calcul des variations quadratiques pour l'estimation du coefficient de diffusion à partir d'une projection en* k *dimensions par ACP de la trajectoire initiale.*

Définissons alors les instants d'entrée et de sortie d'un puits par analogie avec le cas continu. De même que précédemment, on définit d'abord $i_0 = \text{argmin}_{i\geq 0}\|x_i - x_0\|$. On définit ensuite i^* et j^* par :

$$\begin{cases} i^* = \sup\{i \leq i_0 : \|x_i - x_0\| \geq R_0\} \\ \\ j^* = \inf\{j \geq i_0 : \|x_j - x_0\| \geq R_0\} \end{cases} \tag{3.16}$$

de sorte que pour tout $k \in [\![i^*, j^*]\!]$ on ait que $x_k \in \mathcal{B}(x_0, R_0)$. La Fig. 3.13 illustre cette idée à l'aide de EDS précédente (voir Fig. 3.7) sur le paysage des trois puits, avec $x_0 = (-1, 0)$ et $R_0 = 1$ nm, cette fois-ci dans une version sous-échantillonnée ($T = 25$ ns observées toutes les $\Delta t = 6$ ps).

De même que précédemment, on considère pour tout $\alpha > 0$, l'instant :

$$i^*_\alpha = \sup\{i \leq i_0 : \|x_i - x_0\| \geq R_\alpha\}$$

La Fig. 3.14 représente les différentes boules $(\mathcal{B}_\alpha(x_0, R_0))_{\alpha>0}$ pour $\alpha \in \{0.1, 0.25, 0.5, 0.75, 1\}$ (gauche) et les profils $t \to \kappa(i^*_\alpha, j)$ correspondants (droite), calculés pour des $j \in [i^*_\alpha + \epsilon, j^* + \epsilon]$, avec $\epsilon = 2\%\Delta t(j^* - i^*)$.

Nous retrouvons alors le comportement des courbes de la Fig. 3.9. Définissons alors les quantités précédentes dans leurs équivalents discrets :

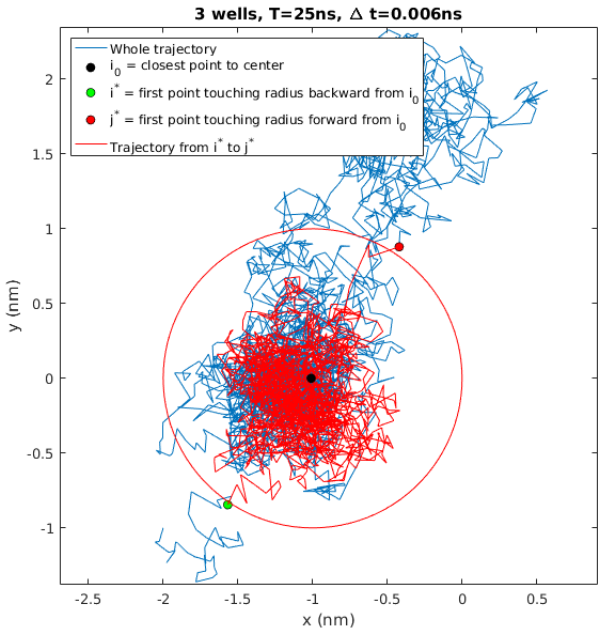

FIGURE 3.13 – Trajectoire issue du modèle des trois puits partant de $(-2,-1)$ avec $\theta = 0.5$ nm^2·ns^{-1}, $T = 25$ ns et $\Delta t = 6$ ps. Bleu : trajectoire totale ; Point noir : conformation x_{i_0} la plus proche du centre x_0 ; Point vert : première conformation hors puits en rétrogradant depuis x_{i_0} ; Point rouge ; première conformation hors puits en progressant depuis x_{i_0}.

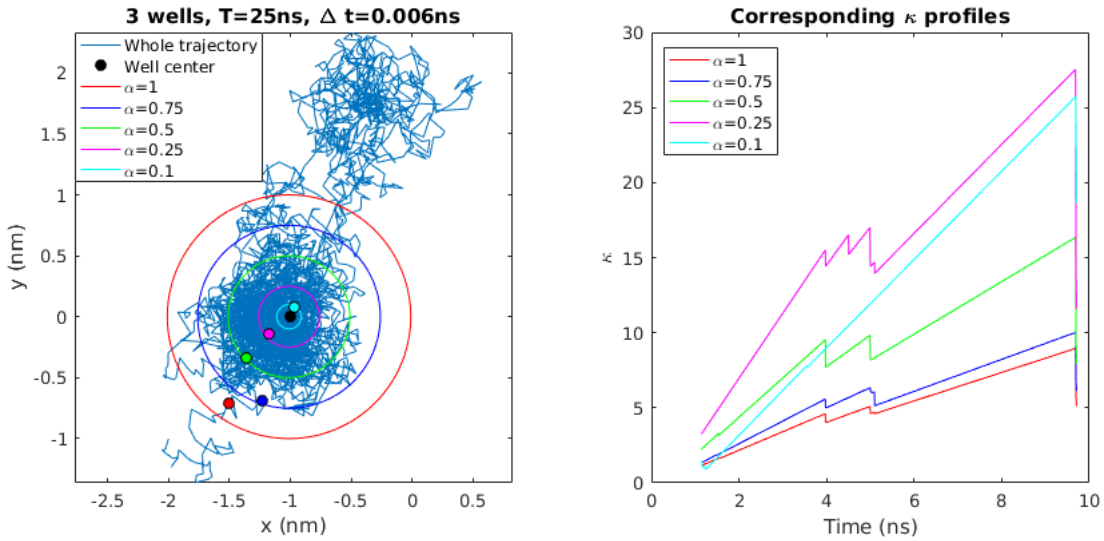

FIGURE 3.14 – Droite : tracé des boules fermées $\mathcal{B}(x_0, R_\alpha)$ et des $x_{i^*_\alpha}$ correspondants ; Gauche : profils des $\alpha \mapsto \kappa(i^*_\alpha, t)$, pour $t \in [i^*_\alpha + \epsilon, j^* + \epsilon]$.

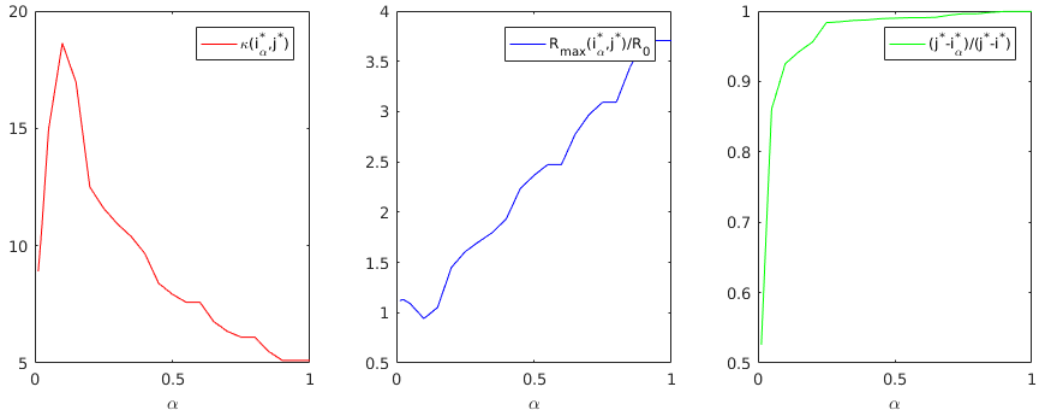

FIGURE 3.15 – Gauche : profil de $\alpha \mapsto \kappa(i^*_\alpha, j^*)$; Centre : profil de $\alpha \mapsto R_{\max}(i^*_\alpha, j^*)/R_0$; Droite : profil de $\alpha \mapsto (j^* - i^*_\alpha)/(j^* - i^*)$.

Définition 10. (Instant d'accès, instant de sortie et temps de sortie - version discrète) *Soit* $\mathbf{X} = (x_i)_{1\leq i\leq N}$ *construite selon* (3.15), $\mathcal{B}(x_0, R_0) = \{x \in \mathbb{R}^d : \|x - x_0\| \leq R_0\}$ *une boule fermée candidate pour représenter un puits,* $s_0 = \text{argmin}_{t\geq 0}\|X_t - x_0\|$, (i^*, j^*) *définis par* (3.16), *et* $\epsilon = \gamma\Delta t(j^* - i^*)$ *avec* $\gamma \in]0, 1[$. *On appelle instant d'accès au puits et instant de sortie du puits, et on note respectivement* $\tilde{\imath}$ *et* $\tilde{\jmath}$ *les quantités définies par :*

$$\begin{cases} \tilde{\imath} = i^*_{\tilde{\alpha}} \quad où \quad \tilde{\alpha} = \underset{\alpha>0}{\text{argmax}}\{\kappa(i^*_\alpha, j^*)\} \\ \tilde{\jmath} = \underset{j\in[\tilde{\imath}+\epsilon, j^*]}{\text{argmax}}\{\kappa(\tilde{\imath}, j)\} \end{cases}$$

On définit également le temps de sortie du puits par la quantité :

$$\widetilde{\mathbf{T}} = (\tilde{\jmath} - \tilde{\imath})\Delta t$$

Définition 11. (Centre d'un puits - version discrète) *Soit* $\mathbf{X} = (x_i)_{1\leq i\leq N}$ *construite selon* (3.15), $\mathcal{B}(x_0, R_0) = \{x \in \mathbb{R}^d : \|x - x_0\| \leq R_0\}$ *une boule fermée candidate pour représenter un puits, et* $\tilde{\imath}$ *l'instant d'accès au puits. On définit le centre du puits par la conformation* $x_{\tilde{\imath}}$.

Remarque 26. *Nous résumons la correspondance continu-discret par le tableau suivant :*

	Continu	Discret
Observations	$X = (X_t)_{t\geq 0}$	$\mathbf{X} = (x_i)_{1\leq i\leq N}$
Rayon maximum	$R_{\max}(s, t)$	$R_{\max}(i, j)$
Coefficient de diffusion	$D = \sigma^2 d$	$\mathbf{D}$
Nombre de tours	$\kappa(s, t)$	$\kappa(i, j)$
Instants d'accès/sortie	$(\tilde{s}, \tilde{t})$	$(\tilde{\imath}, \tilde{\jmath})$
Temps de sortie	$\widetilde{T}$	$\widetilde{\mathbf{T}}$
Centre	$X_{\tilde{s}}$	$x_{\tilde{\imath}}$

TABLEAU 3.1 – Correspondance continu-discret

3.2 Algorithme de κ-segmentation

Nous désirons alors construire un algorithme de segmentation des trajectoires de simulation de DM, basé sur le critère du nombre de tours κ et baptisé *algorithme de κ-segmentation*, dont le but est de fournir en sortie les différents puits identifiés, ainsi que certaines caractéristiques associées à chacun d'entre eux (centre, rayon, instant d'accès, instant de sortie, temps de sortie).

3.2.1 Matrice du nombre de tours

Définition 12. (Matrice du nombre de tours) *Soit* $\mathbf{X} = (x_i)_{1\leq i\leq N}$ *construite selon* (3.15). *On appelle matrice du nombre de tours associée à* $\mathbf{X}$, *et on note* $\kappa = (\kappa_{ij})_{ij}$ *la matrice de* $\mathcal{M}_N(\mathbb{R})$ *définie par :*

$$\forall (i,j) \in [\![1,N]\!]^2, \kappa_{ij} = \begin{cases} \kappa(i,j) & si \quad j \geq i \\ 0 & sinon \end{cases}$$

Pour une trajectoire de 100 ns observée au pas $\Delta t = 4$ ps ($N = 25000$ conformations) sur le paysage des 3 puits à la température $\theta = 0.5$ nm^2·ns^{-1} en dimension $d = 2$, nous obtenons la matrice de nombre de tours suivante :

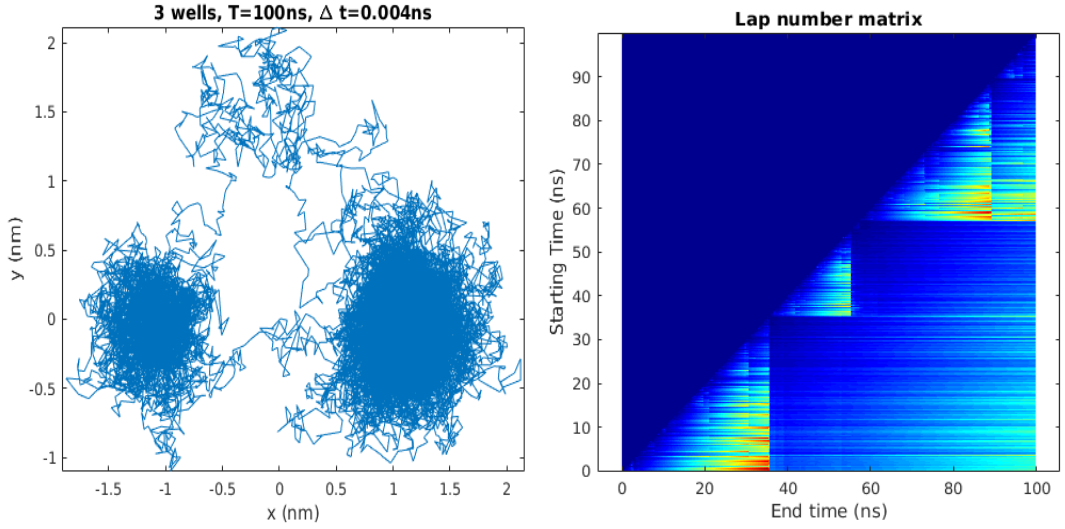

FIGURE 3.16 – Gauche : trajectoire simulée sur le paysage des trois puits avec $T = 100$ ns et $\Delta t = 4$ ps. Droite : matrice du nombre de tours associée. Ordonnées : temps initiaux (ns). Abscisses : temps finaux (ns).

Nous souhaitons identifier les différents puits à partir d'une *lecture automatique* de cette matrice. Il s'agit alors principalement d'identifier les instants d'accès et de sortie des puits, c'est-à-dire identifier des indices $\tilde{\imath}$ et $\tilde{\jmath}$.

D'une part, nous avons vu qu'il était nécessaire, à un instant de sortie j fixé, d'effectuer une optimisation sur l'instant d'accès i. Nous proposons alors de lire la matrice par *bandes horizontales*. Chaque bande permettra ainsi d'étudier une gamme de conformations de départ, définie par une conformation initiale $l_0 \in [\![1,N]\!]$ (i.e. une ligne de la matrice κ) et une hauteur $h \in [\![1,N]\!]$. Ces bandes, notées κ^B, correspondent donc à des sous-matrices de κ de $\mathcal{M}_{h,N-l_0+1}(\mathbb{R})$ (les bandes ne

lisent pas le triangle nul de la matrice du nombre de tours) définies pour tout $(i,j) \in [\![1,h]\!] \times [\![1, N-l_0+1]\!]$ par :

$$\kappa^B_{ij} = \kappa_{i+l_0-1, j+l_0-1}$$

Pour l_0 et h fixés, il s'agit donc de calculer les profils de nombre de tours correspondant aux portions de trajectoires $\mathbf{Z}_k \in \mathcal{M}_{d,N-l_0+1-k}(\mathbb{R})$ données par les colonnes l_0+k à N de $\mathbf{X}$, pour $k \in [\![0, h-1]\!]$. La lecture de la matrice κ se fera par l'incrémentation de l_0, induisant une lecture verticale et ascendante de κ par les bandes horizontales κ^B.

D'autre part, nous avons vu que pour un instant d'accès i fixé, une optimisation globale sur tous les $j \geq i$ était maladroite. Nous proposons alors, afin de traduire l'optimisation locale que nous souhaitons effectuer, de parcourir chaque bande par la progression d'une *fenêtre* de hauteur h et de largeur w, notée κ^F. Chaque fenêtre est définie par une conformation de départ $c_0 \in [\![1, N-l_0+1]\!]$ (i.e. une colonne de κ^B) permettant de lire les w conformations faisant suite à c_0 et correspondant à une sous-matrice de κ^B de $\mathcal{M}_{h,w}(\mathbb{R})$ définie pour tout $(i,j) \in [\![1,h]\!] \times [\![1,w]\!]$ par :

$$\kappa^F_{ij} = \kappa^B_{i,j+l_0-1}$$

La lecture d'une bande se fera ainsi par l'incrémentation de c_0, induisant une lecture horizontale et de gauche à droite de la bande κ^B par les fenêtres κ^F. On prendra soin d'engendrer à chaque étape une zone de chevauchement de sorte à lier les fenêtres les unes aux autres (voir Chapitre 3 §3.2.2, Critère no. 3)

3.2.2 Description de l'algorithme

Commençons par le parcours d'une bande κ^B de conformation initiale l_0 et de hauteur h. On commence par calculer tous les rayons maximums de cette bande à savoir la matrice $R^B_{\max}$ de $\mathcal{M}_{h,N-l_0+1}(\mathbb{R})$ telle que pour tout $(i,j) \in [\![1,h]\!] \times [\![1, N-l_0+1]\!]$:

$$R^B_{\max,ij} = R_{\max}(i+l_0-1, j+l_0-1)$$

Pour une fenêtre donnée κ^F, de conformation initiale c_0 et de largeur w, on calcule le maximum du nombre de tours sur celle-ci, que l'on note $\tilde{\kappa}^F$, qui fournit provisoirement un instant d'accès, et que l'on repère par ses indices au sein de la fenêtre notés $(\tilde{\imath}_F, \tilde{\jmath}_F)$ de sorte que $\tilde{\kappa}^F = \kappa^F_{\tilde{\imath}_F, \tilde{\jmath}_F}$. On désire alors soumettre ce maximum à différents critères afin de savoir s'il correspond à une sortie de puits.

Remarque 27. *On veillera à considérer les indices suivants selon que l'on souhaite repérer $\tilde{\kappa}^F$ dans le contexte de la fenêtre κ^F, de la bande κ^B, ou de la matrice κ.*

	Fenêtre κ^F	Bande κ^B	Matrice κ
i	$\tilde{\imath}_F$	$\tilde{\imath}_B = \tilde{\imath}_F$	$\tilde{\imath} = \tilde{\imath}_B + l_0 - 1$
j	$\tilde{\jmath}_F$	$\tilde{\jmath}_B = \tilde{\jmath}_F + c_0 - 1$	$\tilde{\jmath} = \tilde{\jmath}_B + l_0 - 1$

TABLEAU 3.2 – Repérage du maximum $\tilde{\kappa}^F$ dans la fenêtre κ^F, la bande κ^B, où de la matrice κ

Remarque 28. *Les indices $\tilde{\imath}$ et $\tilde{\jmath}$ calculés ici ne correspondent pas stricto sensu aux notions de temps d'accès et de temps de sortie telles qu'elles ont été définies au Chapitre 3 §3.1.4, lesquelles*

reposent sur la connaissance a priori des centres et rayons de chaque puits. Cette méthode cherche à approcher cette définition théorique des indices d'une manière plus pratique et algorithmique.

Critère no. 1 : On souhaite rejeter des puits dont le temps de sortie $\widetilde{\mathbf{T}}$ est trop court. Ces cas correspondent à des maximums du nombre de tours détectés après un très petit nombre de pas partant de la conformation initiale. Pour comprendre ce phénomène, il est bon de revenir au processus continu, et d'étudier son comportement au voisinage de zéro par changement d'échelle, lequel se rapproche du mouvement brownien. Plus précisément, on a la :

Proposition 8. *(Comportement au voisinage de 0) Soit $X = (X_t)_{t\geq 0}$ vérifiant (3.1) et tel que $X_0 = B_0 = 0_{\mathbb{R}^d}$. Pour $\alpha > 0$, en notant $X_t^{(\alpha)} = \sqrt{\alpha} X_{t/\alpha}$ et $B_t^{(\alpha)} = \sqrt{\alpha} B_{t/\alpha}$, on a que :*

$$\sup_{0\leq t\leq 1} \{\|X_t^{(\alpha)} - B_t^{(\alpha)}\|\} \leq \frac{\|\mu\|_\infty}{\sqrt{\alpha}}$$

où $\|\mu\|_\infty = \sup_{\mathbb{R}^d}\{\|\mu\|\}$.

Démonstration. Le processus $X_t^{(\alpha)}$ est régi par l'équation suivante :

$$dX_t^{(\alpha)} = \frac{1}{\sqrt{\alpha}}\mu(X_t^{(\alpha)}/\sqrt{\alpha})dt + dB_t^{(\alpha)}$$

Mais alors, puisque $X_0^{(\alpha)} = B_0^{(\alpha)} = 0_{\mathbb{R}^d}$, nous avons que :

$$X_t^{(\alpha)} = B_t^{(\alpha)} + \frac{1}{\sqrt{\alpha}} \int_0^t \mu(X_s)ds$$

Puis :

$$\|X_t^{(\alpha)} - B_t^{(\alpha)}\| \leq \frac{\|\mu\|_\infty}{\sqrt{\alpha}} t$$

d'où le résultat en prenant le sup sur $t \in [0, 1]$. □

Par conséquent, le comportement d'un tel processus $(X_t)_{t\geq 0}$ au voisinage de zéro ($\alpha \to \infty$) correspond à celui d'un mouvement brownien. Mais alors, rappelons-nous que dans ce cas, la loi de $\kappa(s, t)$ est indépendante du choix de s et t, et correspond à la loi de la variable $d/\sup_{0\leq t\leq 1} \|B_t\|^2$. Il n'est alors pas impossible d'observer des réalisations du processus X fournissant des valeurs importantes de $\kappa(s, t)$ au voisinage de zéro.

En revenant à la version discrète, malgré le travail à $\omega \in \Omega$ fixé, l'occurrence de points très brillants sur la diagonale de la matrice κ dans le cas d'un mouvement brownien illustre cette possibilité de réalisations extrêmes du nombre de tours (voir Fig. 3.19). Pour écarter ce problème et éviter la fausse détection, on définit une variable $N_{\min}$ correspondant au nombre minimum de conformations que l'on souhaite observer au sein d'un puits, et que l'on confronte au nombre d'observations du puits déterminé. On définit une variable booléenne de rejet $\mathbf{B}_1$ correspondant à ce critère.

$$(\tilde{\jmath}_F + c_0 - 1) - \tilde{\imath}_F \leq N_{\min} \Longrightarrow \mathbf{B}_1 = 1$$

Critère no. 2 : On souhaite par ailleurs rejeter tout puits dont le maximum du nombre de tours serait trop faible (rappelons en effet que le nombre de tours fournit des ordres de grandeurs

absolus auxquels il est raisonnable de comparer les quantités calculées). On considère alors un seuil $\kappa_{\min}$, en deçà duquel le nombre de tours sera perçu comme non-significatif, et le puits rejeté. En définissant de même une variable booléenne $\mathbf{B}_2$:

$$\tilde{\kappa}^F \leq \kappa_{\min} \Longrightarrow \mathbf{B}_2 = 1$$

Critère no. 3 : Nous voulons ici rejeter tout maximum qui serait trop proche du bord à droite de la fenêtre. En effet, cherchant à distinguer les phénomènes d'excursions des véritables sorties de puits, nous souhaitons disposer d'un nombre de pas post-maximum suffisamment élevé de sorte à ce que l'éventuel retour du processus au sein du puits après son excursion soit visible. En cas de rejet, le chevauchement des fenêtres sera calibré de sorte à ce que ce maximum soit tout de même pris en compte dans la fenêtre suivante. En notant K ce nombre de pas (nous le définirons plus précisément ultérieurement), le critère prend la forme suivante :

$$\tilde{\jmath}_F + K > h \Longrightarrow \mathbf{B}_3 = 1$$

Critère no. 4 : Nous souhaitons être en mesure de distinguer une réelle sortie de puits d'une simple excursion. On se propose pour ce faire de procéder à un suivi temporel non plus des rayons maximum $j \mapsto R_{\max}(i,j)$, mais des simples rayons $j \mapsto R(i,j) = \|x_j - x_i\|$ partant de la conformation de départ choisie, dont on étudiera les fluctuations, et que l'on stocke au sein d'une matrice $R^B \in \mathcal{M}_{h,N-l_0+1}(\mathbb{R})$, telle que $R^B_{ij} = R(i,j)$. En effet, on souhaite détecter dans une zone post-maximum tout éventuel retour du rayon en deçà d'une certaine fraction $\gamma \in]0,1[$ du rayon atteint en sortie de puits. Le critère prend la forme suivante :

$$\sum_{k=1}^{K} \mathbb{1}\{R(\tilde{\imath}_B, \tilde{\jmath}_B + k) \leq \gamma \times R_{\max}(\tilde{\imath}_B, \tilde{\jmath}_B)\} > 0 \Longrightarrow \mathbf{B}_4 = 1$$

Critère no. 5 : Enfin, nous désirons rejeter tout puits correspondant à une zone transitoire ou contenant une zone transitoire importante. En pratique, ces puits sont caractérisés par un rayon maximum important : en effet, la zone transitoire correspond à un comportement erratique du processus, lequel induit un élargissement du rayon maximum. En définissant ainsi un rayon maximum tolérable $\rho > 0$, le critère s'écrit :

$$R_{\max}(\tilde{\imath}_B, \tilde{\jmath}_B) > \rho \ \Longrightarrow \mathbf{B}_5 = 1$$

Remarque 29. *Le Critère no. 4 était initialement conçu à partir d'une étude du nombre de tours post-maximum, et plus précisément du nombre de chutes du nombre de tours. En effet, une simple excursion au bord du puits, ou légèrement en dehors, avant d'y plonger à nouveau, se traduit par une unique chute du nombre de tours, c'est-à-dire par une augmentation ponctuelle et brutale du rayon maximum. En revanche, une véritable sortie de puits se traduira par une succession d'augmentations du rayon maximum, et donc une chute "en cascade" du nombre de tours. En effet, d'un point de vue probabiliste, la sortie de puits correspond à un évènement rare : conditionnellement à sa réalisation, il est fort probable que la trajectoire s'éloigne résolument du puits après en être sortie, ce qui crée ce phénomène de chutes successives du nombre de tours (Fig. 3.17).*

En revanche, dans des cas moins nets que celui-ci, la simple lecture du profil du nombre de tours ne permet pas de savoir si le processus est revenu dans le puits (excursion terminée) après

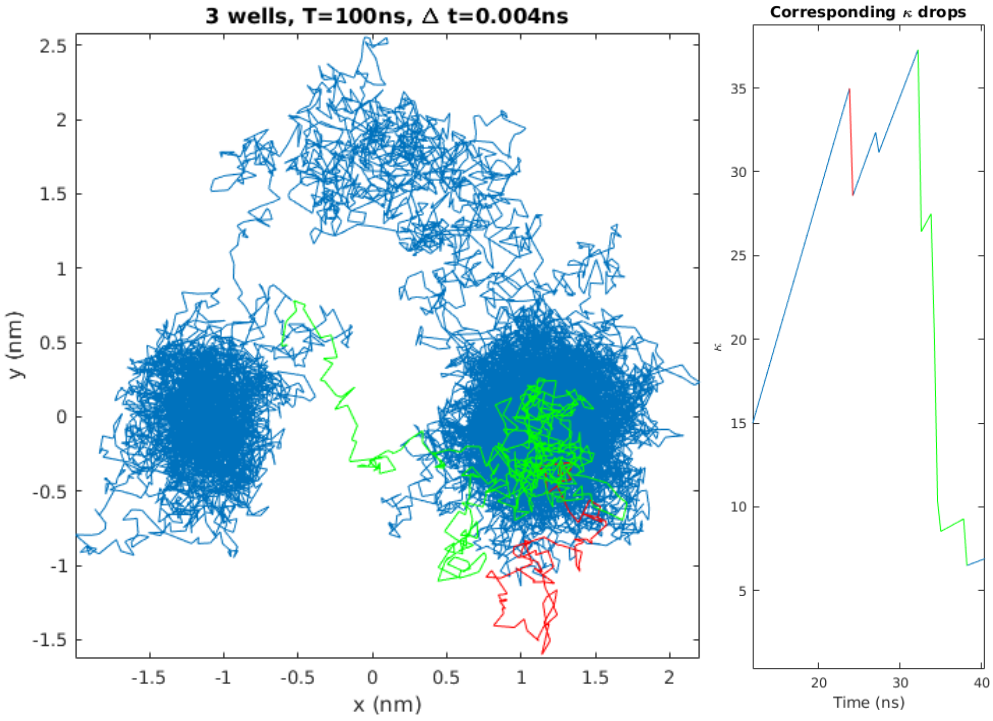

FIGURE 3.17 – Gauche : sortie réelle (vert) et excursion (rouge) du puits (1,0) d'une trajectoire simulée sur le paysage des trois puits avec $T = 100$ ns et $\Delta t = 4$ ps. Droite : profils de κ correspondant à la sortie réelle (chutes multiples, vert) et à l'excursion (chute simple, rouge).

une chute simple du nombre de tours, ou bien s'il poursuit un mouvement erratique à l'extérieur du puits. Le suivi du rayon au détriment de celui du nombre de tours permet de lever cette imprécision.

Remarque 30. *Le Critère no. 4 ne permet pas à l'algorithme d'adopter un point de vue multi-échelle. En effet, le passage au sein d'un même "grand" puits, d'un premier sous-puits à un second sous-puits pourra éventuellement être perçu comme une sortie de puits si le processus ne revient pas assez tôt dans le sous-puits initial. L'algorithme va donc distinguer deux puits différents, alors que ceux-ci pourraient être volontiers regroupés en un seul et même puits.*

Revenons au parcours de la bande. Après la lecture d'une fenêtre, on souhaite prendre une décision quant au maximum local étudié, laquelle se traduit par une variable $\mathbf{B}$ à valeurs dans $\{-1, 0, 1\}$ et exprimant respectivement le rejet du puits ($\mathbf{B} = -1$), une visite du puits toujours en cours ($\mathbf{B} = 0$), ou bien l'acceptation du puits ($\mathbf{B} = 1$). Trois cas se présentent :

1. Si $\mathbf{B}_5 = 1$ (zone transitoire), alors la plage de conformations de départ devra être incrémentée en actualisant l_0 de sorte à se rapprocher du futur puits. On affecte alors à la variable $\mathbf{B}$ la valeur $\mathbf{B} = -1$ et on sort de la bande.
2. Si $\mathbf{B}_5 = 0$, mais que l'un des $\mathbf{B}_i = 1$ pour $1 \leq i \leq 4$, alors le maximum de la fenêtre n'est pas satisfaisant et on doit passer à la fenêtre suivante en incrémentant $c_0 \leftarrow c_0 + w - K + 1$. Si cela est possible, on poursuit la lecture de la bande. Sinon, c'est que la bande a été parcourue en entier sans succès : en pratique ce phénomène se produit en fin de trajectoire, et traduit une exploration de puits toujours en cours (bien souvent c'est le critère $\mathbf{B}_3$ qui empêche l'acceptation du puits, et qui montre que la visite n'est pas terminée), dont on

stocke les caractéristiques. On affecte alors à la variable $\mathbf{B}$ la valeur $\mathbf{B} = 0$.

3. Enfin, si aucun des critères de rejet n'a été enclenché, nous sommes en présence d'un puits, dont on stocke les caractéristiques, et on sort de la bande. On affecte alors à la variable $\mathbf{B}$ la valeur $\mathbf{B} = 1$.

Dans tous les cas, une fois l'étude de la bande terminée, les indices d'accès et de sortie sont stockés dans leur contexte global, c'est-à-dire non plus $(\tilde{\imath}_F, \tilde{\jmath}_F)$, mais $(\tilde{\imath}, \tilde{\jmath})$.

Dans un deuxième temps, on gère alors le parcours de la matrice. Suite au parcours d'une bande, deux cas se présentent :

1. Si la lecture de la bande a été interrompue par la détection d'une zone transitoire, alors on incrémente $l_0 \leftarrow l_0 + h - 1$, et on parcourt la bande suivante. Si celle-ci sort du cadre de la matrice, l'algorithme s'arrête.
2. Sinon, c'est que le parcours de la bande a vu émerger un puits, que l'on a stocké. Si $\mathbf{B} = 0$, un puits est en cours de détection et nous sommes arrivés à la fin de la trajectoire par les dernières plages de conformations de départ possibles, et l'algorithme s'arrête. Si $\mathbf{B} = 1$, un puits a été détecté et on incrémente $l_0 \leftarrow l_0 + \tilde{\jmath} - 1$ (la nouvelle bande débute à l'instant de sortie du puits précédent). Si $l_0 + h - 1 \leq N$, on poursuit l'algorithme, sinon celui-ci s'arrête.

Outre des sorties graphiques, l'algorithme fournit principalement en sortie une structure W au sein de laquelle seront stockés les puits, ainsi que différents attributs d'intérêt : nombre de tours, indice d'accès (i.e. indice de centre), indice de sortie, durée et rayon.

Remarque 31. *La lecture du triangle inférieur gauche de la matrice du nombre de tours induit une complexité de calcul de $N \times h$. La version de l'algorithme décrite ici stocke en mémoire les rayons calculés pour la colonne à parcourir de taille $h \times (N - l_0 + 1)$ (une version améliorée existe, permettant de stocker uniquement les rayons sur la taille d'une fenêtre de taille $h \times w$).*

3.2.3 Calibration des paramètres

Nous proposons ici une première approche pour automatiser la calibration des paramètres. Une visualisation de la trajectoire projetée en dimension 2 par ACP permet de définir les paramètres de distances. On fixe ainsi ρ, nécessaire au Critère no. 5.

Rappelons ensuite que les Critères no. 3 et no. 4 dépendent de K, que nous avons précédemment défini comme étant le nombre de pas post-maximum suffisant à observer un retour du processus dans le puits. Ce paramètre est donc étroitement lié à la diffusion du processus : en notant $\overline{R}$ le rayon moyen des puits observés sur la projection en dimension 2 par ACP, et δ le nombre de pas nécessaire au parcours d'une distance de $\overline{R}$, on a :

$$\delta = \mathrm{E}\left(\frac{\overline{R}^2}{\mathbf{D} \times \Delta t}\right)$$

où E désigne la fonction partie entière. Le calcul de cette quantité permet alors de définir les paramètres suivants :

— $h =$ quelques δ : de cette façon, on s'assure d'avoir une plage de conformations de départ assez grande pour couvrir un puits, et donc d'inclure un point d'accès bien centré.

— $K =$ quelques δ : on couvre alors un nombre de points suffisant post-maximum pour observer le retour d'une excursion. Cette valeur correspondra également à la zone de chevauchement des fenêtres. De cette façon, un maximum rejeté car trop proche du bord sera considéré par la fenêtre suivante.

— $w =$ quelques K : assez grand pour détecter une sortie de puits.

On définit ensuite $\kappa_{\min}$ et $N_{\min}$, qui avec $\overline{R}$ et ρ sont les seuls paramètres définis manuellement par l'utilisateur.

3.2.4 Pseudo-code

Algorithm 3 Explore_strip

Entrée $\mathbf{X}$, Δt, $\mathbf{D}$, l_0, w, γ, $\kappa_{\min}$, K, $N_{\min}$, ρ

(1) Initialisation : Construction des matrices R^B, ${R_{\max}}^B$, κ^B de $\mathcal{M}_{h,N-l_0+1}(\mathbb{R})$ et initialisation de la fenêtre $c_0 \leftarrow 1$

(2) Recherche d'un maximum local :

- Construction de $\kappa^F \in \mathcal{M}_{h,w}(\mathbb{R})$
- $\tilde{\kappa}^F \leftarrow \max\{\kappa^F_{ij} \mid (i,j) \in [\![1,h]\!] \times [\![1,w]\!]\}$
- $(\tilde{\imath}_F, \tilde{\jmath}_F) \leftarrow \text{argmax}\{\kappa^F_{ij} \mid (i,j) \in [\![1,h]\!] \times [\![1,w]\!]\}$

(3) Tests de validation :

- Si $(\tilde{\jmath}_F + c_0 - 1) - \tilde{\imath}_F \leq N_{\min}$, alors $\mathbf{B}_1 = 1$
- Si $\tilde{\kappa}^F < \kappa_{\min}$, alors $\mathbf{B}_2 = 1$
- Si $\tilde{\jmath}_F + K > h$, alors $\mathbf{B}_3 = 1$
- Si $\sum_{k=1}^{K} \mathbb{1}\{R(\tilde{\imath}_B, \tilde{\jmath}_B + k) \leq \gamma \times R_{\max}(\tilde{\imath}_B, \tilde{\jmath}_B)\} > 0$, alors $\mathbf{B}_4 = 1$
- Si $R_{\max}(\tilde{\imath}_B, \tilde{\jmath}_B) > \rho$, alors $\mathbf{B}_5 = 1$

(4) Décision finale :

- Si $\mathbf{B}_5 = 1$, alors $B = -1$ *(partie transitoire détectée et sortie de l'algorithme)*
- Si $\mathbf{B}_5 = 0$ et $\mathbf{B}_i = 1$ pour un $i \leq 1 \leq 4$, alors $c_0 \leftarrow c_0 + w - K + 1$
  - Si $c_0 + w > N - l_0 + 1$, alors $B = 0$ *(puits en cours de détection et sortie de l'algorithme)*
  - Sinon, retour à (2)
- Si $\forall\ 1 \leq i \leq 5, \mathbf{B}_i = 0$, alors $B = 1$ *(puits détecté et sortie de l'algorithme)*

Sortie Matrices R^B, ${R_{\max}}^B$, κ^B et structure S contenant les valeurs de : $\mathbf{B}$ (statut de détection), $\tilde{\kappa}^F$ (nombre de tours), $\tilde{\imath}$ (indice d'accès et centre), $\tilde{\jmath}$ (indice de sortie), ${R_{\max}}^2(\tilde{\imath}_B, \tilde{\jmath}_B)$ (rayon du puits), $\widetilde{\mathbf{T}} = (\tilde{\imath} - \tilde{\jmath})\Delta t$ (temps de sortie).

L'algorithme de segmentation final peut donc se résumer par le pseudo-code suivant, et l'illustration qui suit :

Algorithm 4 κ-segmentation

Entrée $\mathbf{X}$, Δt, $\mathbf{D}$

(1) Initialisation : Choix des paramètres $\overline{R}$, $\kappa_{\min}$, $N_{\min}$ et α, calcul de K, w, h et initialisation à $c_0 = 1$, $n_{\mathrm{T}} = 1$ et $n_{\mathrm{W}} = 1$

(2) Parcours d'une bande :

$$[R^B, {R_{\max}}^B, \kappa^B, S] = \text{explore\_strip}(\mathbf{X}, \Delta t, \mathbf{D}, l_0, w, \gamma, \kappa_{\min}, K, N_{\min}, \rho)$$

(3) Stockage :

- Si $\mathbf{B} = -1$ *(partie transitoire détectée)*, alors $T(n_{\mathrm{T}}) \leftarrow S$, $n_{\mathrm{T}} \leftarrow n_{\mathrm{T}} + 1$ et $l_0 \leftarrow l_0 + h - 1$
  - Si $l_0 + h - 1 \leq N$, alors retour à (2)
  - Sinon, sortie de l'algorithme
- Sinon *(puits détecté ou en cours de détection)*, $W(n_{\mathrm{W}}) \leftarrow S$, $n_{\mathrm{W}} \leftarrow n_{\mathrm{W}} + 1$ et $l_0 \leftarrow \tilde{\jmath} + 1$
  - Si $l_0 + h - 1 \leq N$, alors retour à (2)
  - Sinon, sortie de l'algorithme

Sortie Structures T et W.

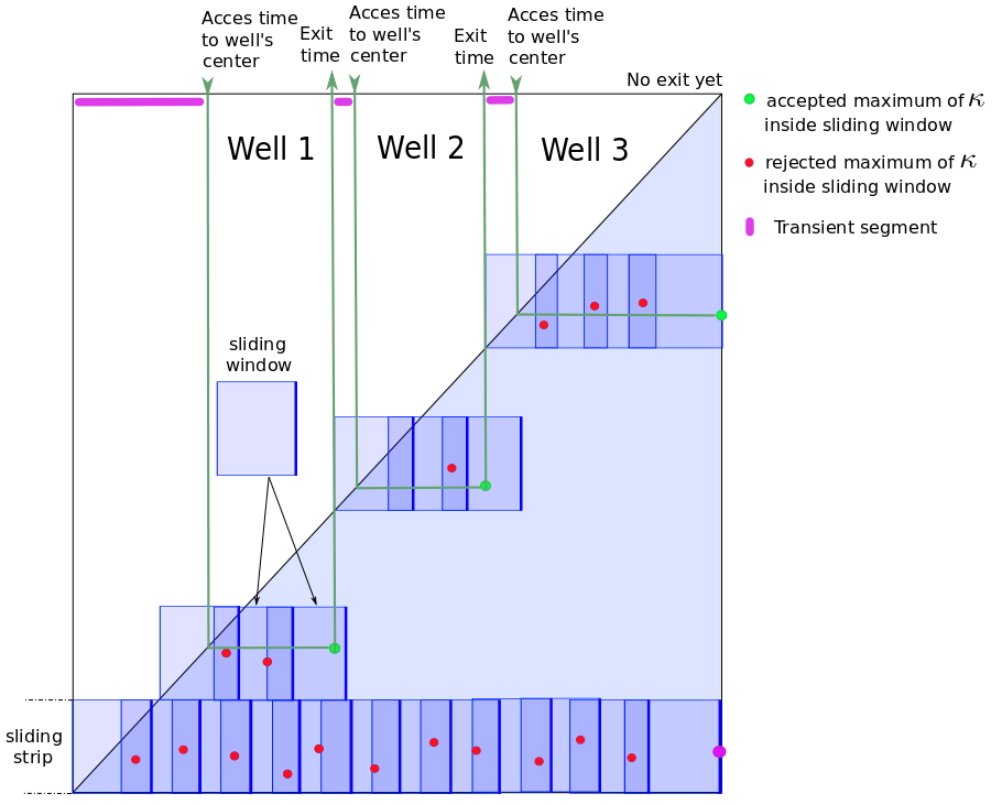

FIGURE 3.18 – Illustration du fonctionnement de l'algorithme.

3.3 Premiers exemples

3.3.1 Mouvement brownien

Nous appliquons ici l'algorithme à la trajectoire brownienne évoquée précédemment, d'une durée de 100 ns au pas de temps $\Delta t = 4$ ps à l'aide des paramètres suivants :

ρ (nm)	$\overline{R}$ (nm)	$\kappa_{\min}$	$N_{\min}$	δ	h	K	w
2	1	10	40	500	5δ	4δ	$3K$

TABLEAU 3.3 – Paramètres pour l'étude du mouvement brownien.

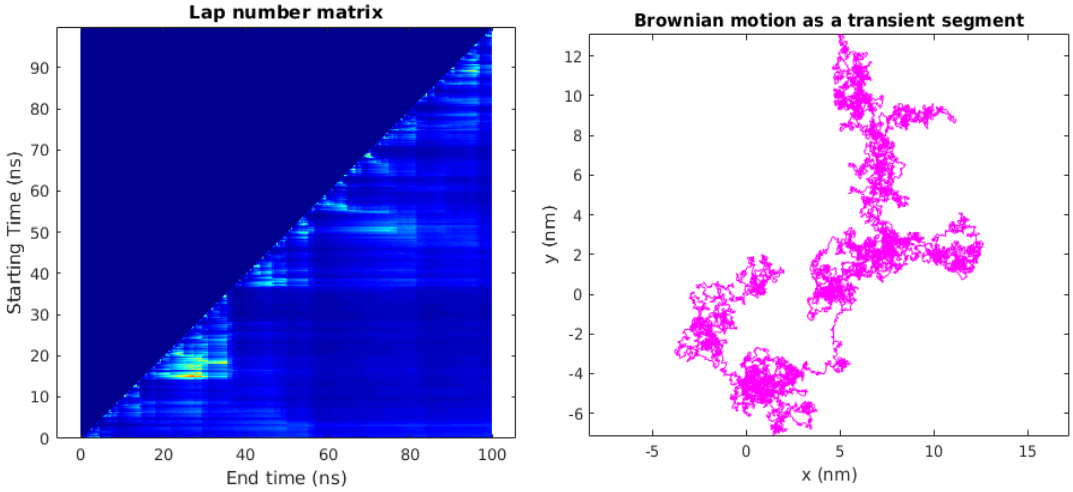

FIGURE 3.19 – Gauche : Matrice du nombre de tours pour la trajectoire brownienne. Droite : κ-segmentation effectuée sur la trajectoire brownienne. Magenta : segment transitoire.

Nous voyons alors que l'algorithme ne tombe pas dans le piège consistant à considérer la moindre accumulation comme un maximum de la densité de probabilité : la trajectoire entière est considérée comme un segment transitoire.

3.3.2 Modèle de trois puits

Nous appliquons ici l'algorithme à une trajectoire issue du modèle de trois puits, d'une durée de 100 ns au pas de temps $\Delta t = 4$ ps, d'une température $\theta = 0.5$ nm^2·ns^{-1}. Voici le résultat produit par l'algorithme à l'aide des paramètres suivants :

ρ (nm)	$\overline{R}$ (nm)	$\kappa_{\min}$	$N_{\min}$	δ	h	K	w
2	1	15	40	125	4δ	3δ	$2K$

TABLEAU 3.4 – Paramètres pour l'étude d'un modèle de trois puits.

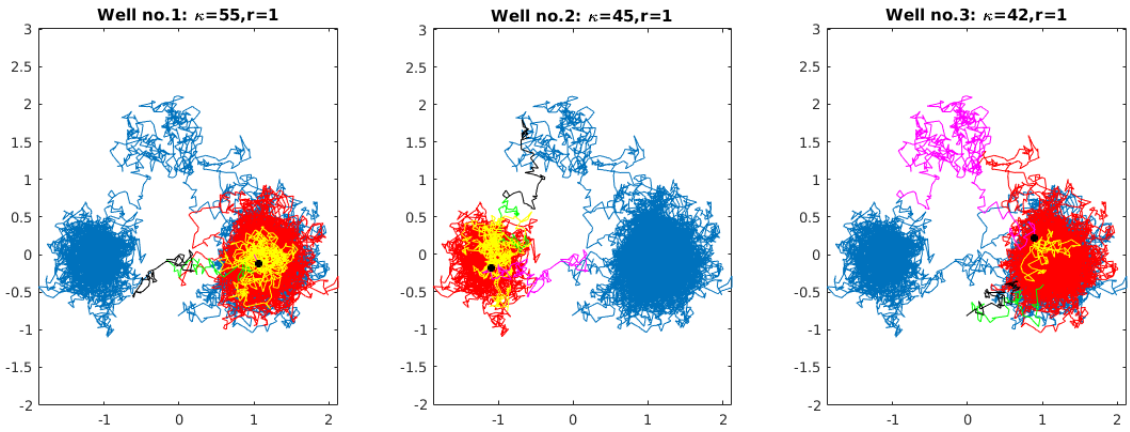

FIGURE 3.20 – κ-segmentation effectuée sur la trajectoire issue du modèle de trois puits. Magenta : partie transitoire ; Rouge : puits ; Jaune : conformations de départ suivant le point d'accès ; Vert : segment avant sortie du puits ; Noir : segment après sortie du puits.

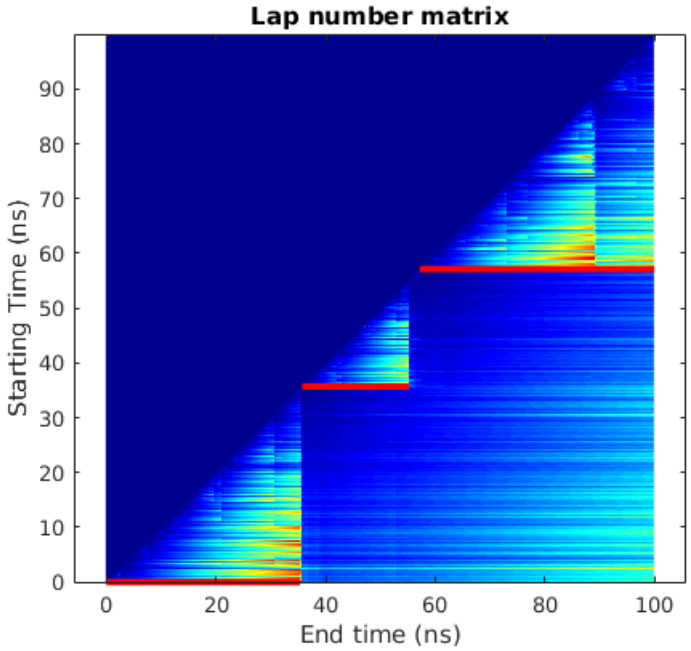

FIGURE 3.21 – κ-segmentation de la trajectoire issue des trois puits représentée sur sa matrice du nombre de tours (segments rouges $[\![\tilde{\imath}, \tilde{\jmath}]\!]$). Ordonnées : instant d'accès (ns) ; Abscisses : instant de sortie (ns).

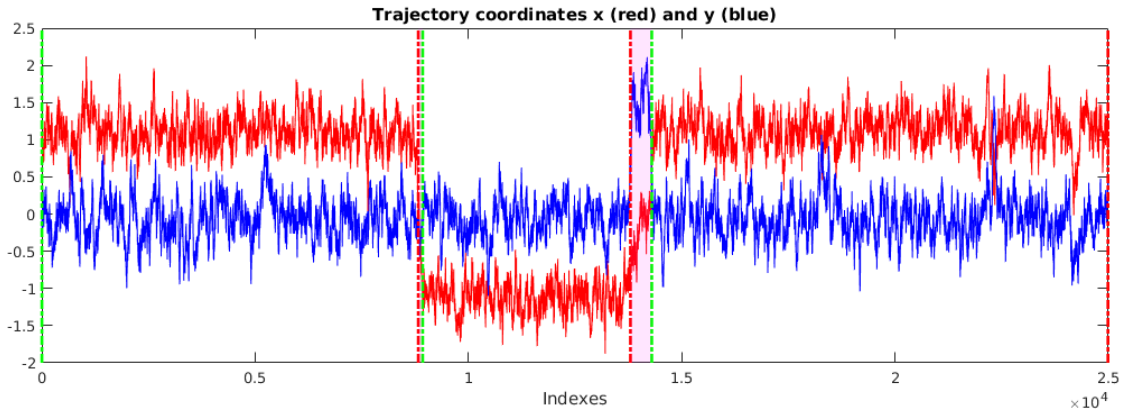

FIGURE 3.22 – κ-segmentation de la trajectoire issue des trois puits représentée sur le profil de ses coordonnées. Ordonnées : valeurs des coordonnées (nm) ; Abscisses : indices des conformations (*frames* de la trajectoire). Rouge : x ; Bleu : y. Pointillés verts : instants d'accès $\tilde{\imath}$; Pointillés rouges : instants de sortie $\tilde{\jmath}$; Magenta : zones transitoires.

On constate ici que l'algorithme parvient à juger convenablement de la profondeur des puits : si les deux puits centrés en $(1, 0)$ et $(-1, 0)$ sont identifiés comme des puits, celui centré en $(0, 5/3)$, dont nous savons qu'il est d'une importance significativement moindre que celles des deux autres puits, est perçu comme une partie transitoire.

3.4 Influences des paramètres

3.4.1 Exemples sur le modèle de trois puits

En guise de premiers exemples, mais également pour mieux comprendre les rouages de l'algorithme, nous nous proposons ici de faire varier certains paramètres de sorte à évaluer leur impact vis-à-vis des résultats obtenus.

Conformément aux légendes précédentes, pour chacun des tracés suivants, on représentera les parties transitoires (magenta), les zones de puits (rouge), les plages de conformations de départ (jaune), les instants précédant les sorties de puits (vert), ceux suivant les sorties de puits (noir) et les centres de puits (croix noires).

Influence de ρ

On considère trois cas différents. Les deux premiers cas sont étudiés sur une seule et même trajectoire simulée sur les trois puits avec $c = 0.1$, $\theta = 0.5$ nm^2·ns^{-1}, $d = 2$, $T = 100$ ns, $\Delta t = 4$ ps et $N = 25000$. Pour le second cas, seule la température change : $\theta = 0.6$ nm^2·ns^{-1}. Les paramètres de l'algorithme sont les suivants :

	ρ (nm)	$\overline{R}$ (nm)	$\kappa_{\min}$	$N_{\min}$	δ	h	K	w
Cas 1	$\rho_1 = 1.5$	1	10	40	125	5δ	4δ	$3K$
Cas 2	$\rho_2 = 3$	1	10	40	125	5δ	4δ	$3K$
Cas 3	$\rho_3 = 1$	1	10	40	105	5δ	4δ	$3K$

TABLEAU 3.5 – Paramètres pour l'étude de l'influence de ρ.

Dans le premier cas (voir Fig. 3.23, $\rho_1 = 1.5$ nm), on constate la détection d'une zone transitoire : l'élection, dans cette gamme de conformations de départ, d'un centre de puits induirait en effet la détection d'un puits trop grand, de rayon > 1.5 nm. L'algorithme préfère considérer la gamme suivante en incrémentant la conformation de départ ($l_0 = l_0 + h - 1$), ce qui permet à la nouvelle gamme de considérer des points plus proches du nouveau puits, voire dans notre cas qui en font partie. Si une partie transitoire demeure (voir cadrant Well no. 4), une identification raisonnable du centre du puits est possible, permettant ainsi un nombre de tours plus élevé, et un rayon acceptable.

A contrario, dans le second cas (voir Fig. 3.24, $\rho_2 = 3$ nm), l'algorithme étant plus tolérant sur la taille des puits, il va tout de même élire en tant que centre la conformation la "moins pire" de la gamme, i.e. la plus proche possible du puits tout en restant dans la gamme. Rejetée dans le cas précédent pour cause d'un rayon trop grand, cette fois-ci, ce centre offre aux yeux de l'algorithme un puits raisonnable du point du vue du rayon, lequel atteint ici 2 nm. Néanmoins, constatons évidemment que le centre choisi est précisément décentré vis-à-vis du puits réel, ce qui induit inévitablement une moins bonne saturation du nombre de tours, lequel n'atteint que la valeur 4 dans ce cas, contre 13 précédemment. Pour la trajectoire précédente, nous montrons que le rejet d'un rayon trop grand peut avoir lieu lorsque la gamme de points de départ est trop éloignée du futur puits. Mais il se peut aussi que ce rejet ait lieu même si la gamme de points recouvre déjà le puits : bien souvent, l'algorithme doit en fait gérer des excursions de puits. Distinguant l'excursion de la réelle sortie, l'algorithme évite l'artefact, mais cela engendre un rayon cumulé trop important. L'algorithme va alors incrémenter la plage de départ, jusqu'à ce que les excursions gênantes puissent être dépassées : outre la possibilité de se rapprocher d'un puits imminent comme c'était le cas pour le premier cas, l'incrémentation de la gamme de départ permet donc également d'ôter du puits des points qui seraient trop extrêmes.

A cet égard, le dernier exemple (voir Fig. 3.25, $\rho_3 = 1$ nm), simulé sur une trajectoire différente, est tout à fait parlant. La trajectoire présentant une forte excursion sur la droite du premier puits, la gêne de l'algorithme est palpable, et celui-ci détecte le puits, dès lors qu'il est en mesure de se délester de cette excursion.

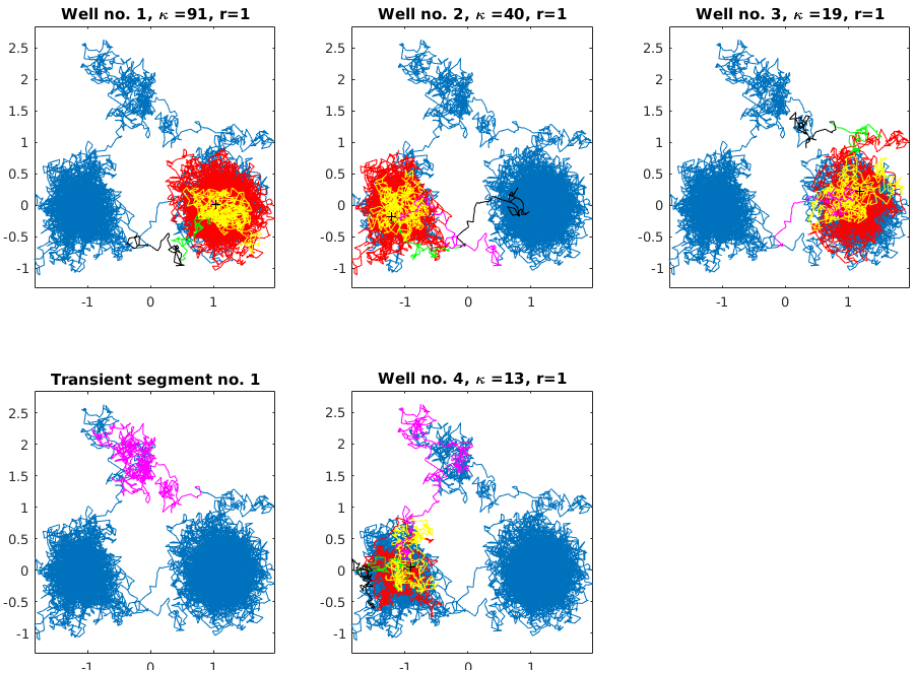

FIGURE 3.23 – Résultat pour $\rho_1 = 1.5$ nm.

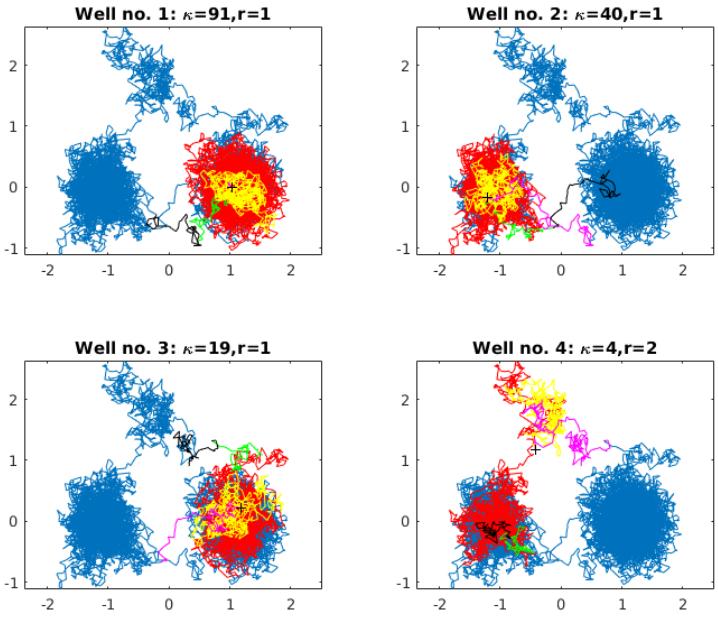

FIGURE 3.24 – Résultat pour $\rho_2 = 3$ nm.

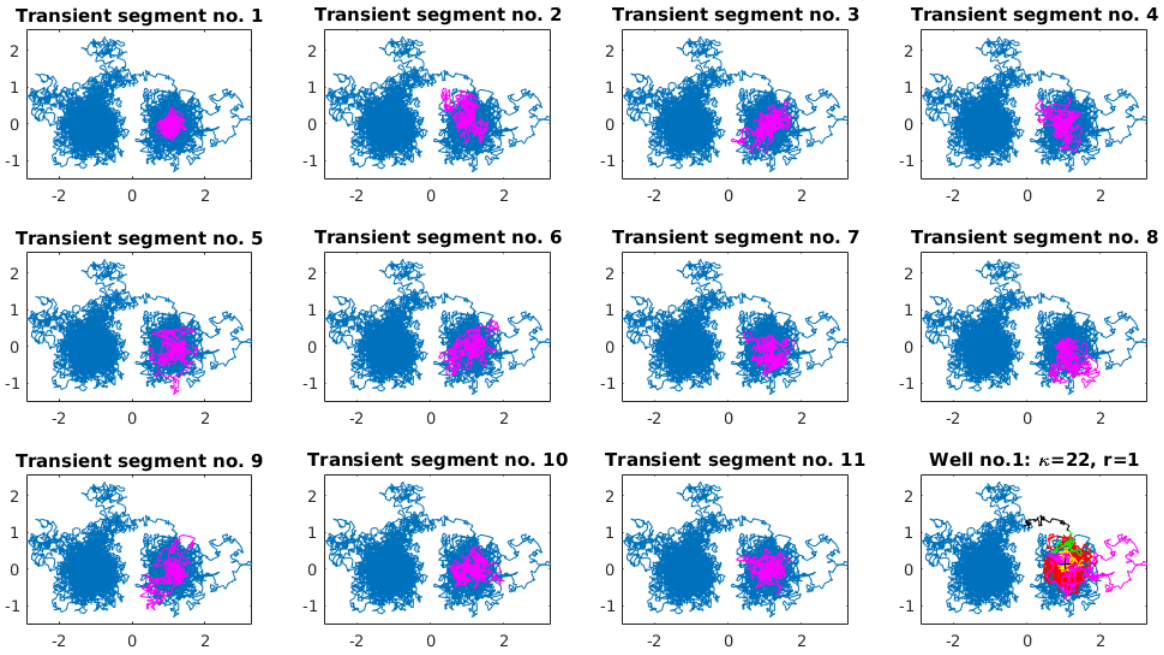

FIGURE 3.25 – Résultat pour $\rho_3 = 1$ nm.

Influence de $\kappa_{\min}$

On considère ici deux cas étudiés sur la même trajectoire simulée sur les trois puits, avec comme paramètres : $c = 0.1$, $\theta = 0.5$ nm^2·ns^{-1}, $d = 2$, $T = 100$ ns, $\Delta t = 4$ ps et $N = 25000$. Les paramètres de l'algorithme sont les suivants :

	ρ (nm)	$\overline{R}$ (nm)	$\kappa_{\min}$	$N_{\min}$	δ	h	K	w
Cas 1	2	1	$\kappa_{\min,1} = 5$	40	125	5δ	4δ	$3K$
Cas 2	2	1	$\kappa_{\min,2} = 10$	40	125	5δ	4δ	$3K$

TABLEAU 3.6 – Paramètres pour l'étude de l'influence de $\kappa_{\min}$.

Dans le premier cas (voir Fig. 3.26, $\kappa_{\min,1} = 5$), l'algorithme étudiera des maximums locaux du nombre de tours à partir de la valeur 5. Pourvu que les autres critères de rejets ne s'enclenchent pas, l'algorithme peut éventuellement fournir en sortie des puits dont le nombre de tours caractéristique est assez faible : c'est le cas du puits supérieur du paysage des trois puits, identifié ici avec une valeur du nombre de tours de 12, donc faible, et dont nous savons par ailleurs qu'il s'agit là d'un puits bien plus insignifiant que les deux autres.

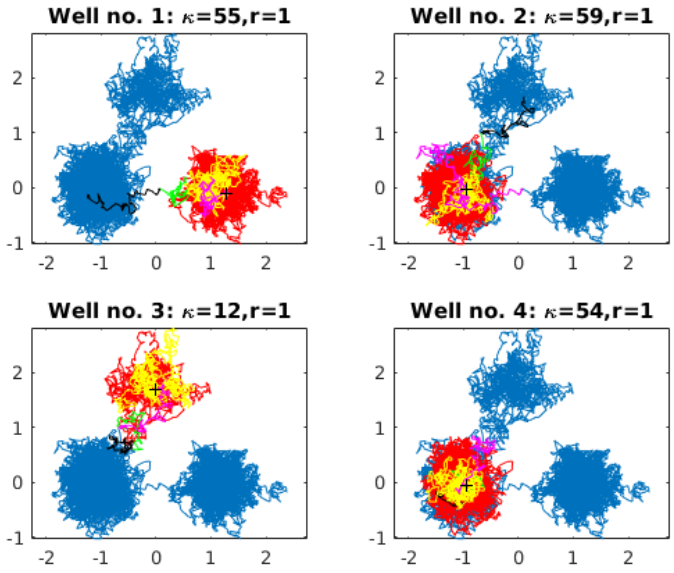

FIGURE 3.26 – Résultat pour $\kappa_{\min,1} = 5$.

A l'inverse, lorsque l'on augmente la valeur de $\kappa_{\min}$ dans le second cas (voir Fig. 3.27, $\kappa_{\min,2} = 10$), l'algorithme devient plus sévère, et refuse ainsi de considérer le puits supérieur du paysage : il identifie à la place des zones transitoires.

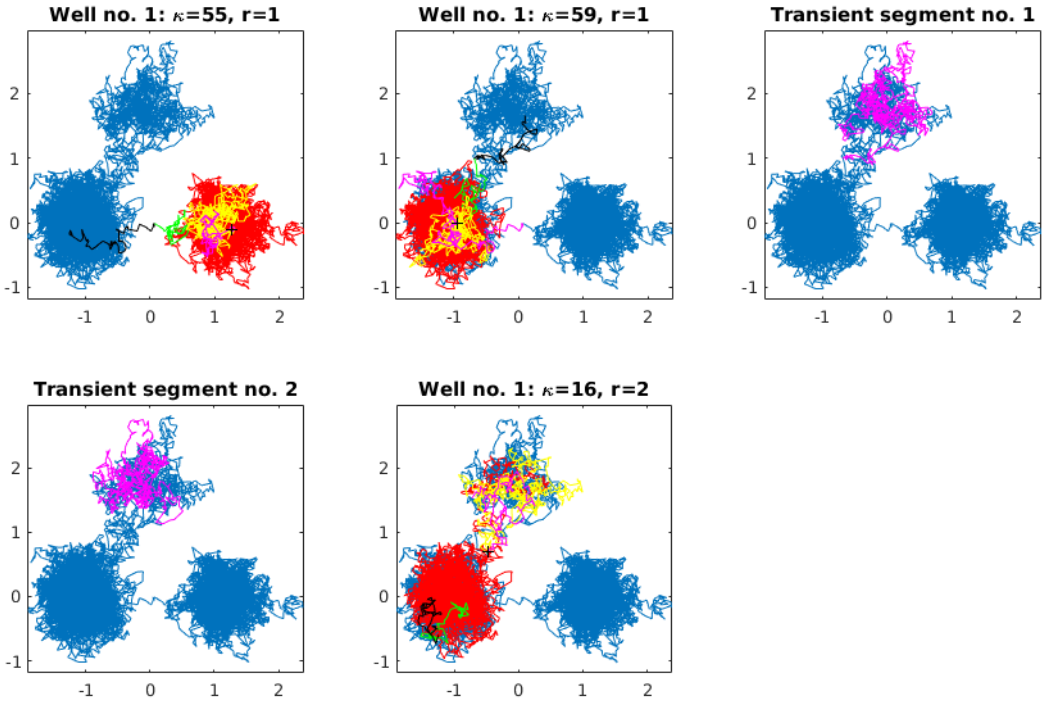

FIGURE 3.27 – Résultat pour $\kappa_{\min,2} = 10$.

En effet, alors qu'une bande est parcourue partant d'une gamme de conformations de départ, la haute valeur de $\kappa_{\min,2}$ refusant les maximums < 10 qui se présentent, l'algorithme est contraint

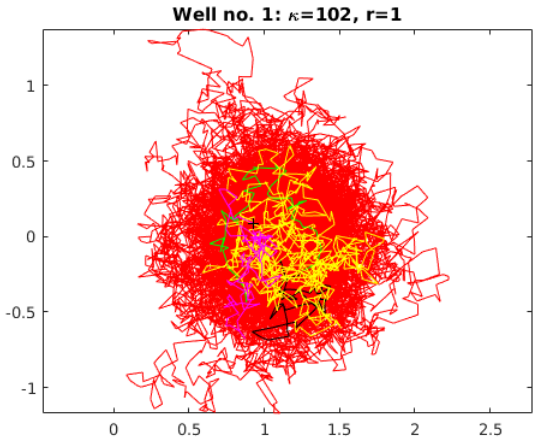

FIGURE 3.28 – Résultat pour $K_1 = 125$.

d'aller chercher des maximums locaux satisfaisants "plus loin" dans la trajectoire. Ce faisant, le rayon du puits finalement identifié devient trop grand, et le critère de rejet via ρ s'enclenche : une zone transitoire est détectée. Remarquons par ailleurs que le dernier puits est identifié avec une moins bonne valeur du nombre de tours que dans le cas précédent (16 au lieu de 54). En effet, dans le premier cas, la sortie du puits supérieur étant identifiée, la gamme de conformations suivante recouvre mieux le futur puits : le centre défini est plus raisonnable, et la valeur du nombre de tours augmente. Dans le second cas, la sortie du puits supérieur n'est pas identifiée. Par conséquent, la gamme de conformations de départ est incrémentée jusqu'à ce qu'un de ses points les "moins éloignés" du futur puits permette d'éviter le rejet selon ρ. Cette dite gamme recouvre alors moins bien le futur puits, le centre est mal choisi, et le nombre de tours sature mal.

Influence de K

On considère ici deux cas étudiés sur la même trajectoire simulée sur les trois puits, avec comme paramètres : $c = 0.1$, $\theta = 0.5$ nm^2·ns^{-1}, $d = 2$, $T = 100$ ns, $\Delta t = 4$ ps et $N = 25000$. Les paramètres de l'algorithme sont les suivants :

	ρ (nm)	$\overline{R}$ (nm)	$\kappa_{\min}$	$N_{\min}$	δ	h	K	w
Cas 1	2	1	15	40	125	5δ	$K_1 = \delta = 125$	$3K_1$
Cas 2	2	1	15	40	125	5δ	$K_2 = K_1 - 30 = 95$	$3K_1$

TABLEAU 3.7 – Paramètres pour l'étude de l'influence de K.

Dans le premier cas, l'algorithme dispose toujours d'une marge suffisante post-maximum pour étudier le retour du processus dans le puits, même si le maximum se trouve très proche de la zone limite (voir Fig. 3.28, $K_1 = 125$). Ainsi, il distingue bien l'excursion, visible à partir de 0.5 ns post-maximum, des réelles sorties de puits.

Dans le second cas, l'algorithme va en revanche prendre pour argent comptant de simples excursions (voir Fig. 3.29 et 3.30, $K_2 = 95$). En effet, en gardant le même δ mais en réduisant K, nous réduisons la zone d'étude du rayon post-maximum du nombre de tours à 0.38 ns. Par

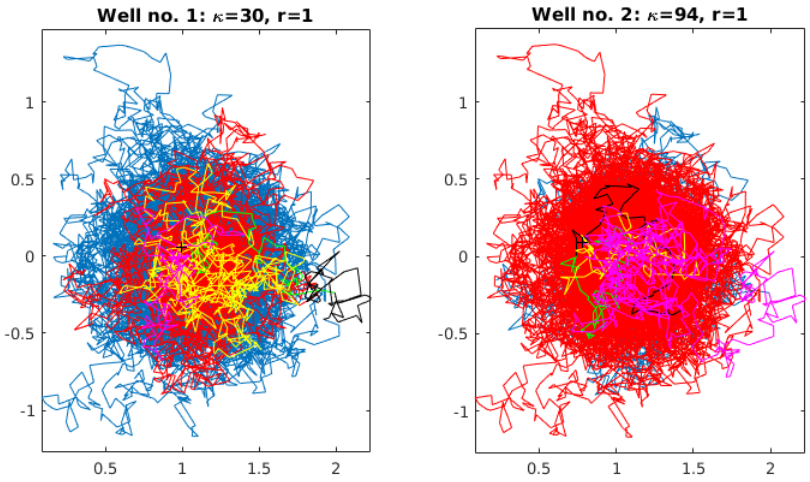

FIGURE 3.29 – Résultat pour $K_2 = 95$.

conséquent, l'algorithme n'a pas assez de temps pour constater un retour du rayon à une valeur "normale", et on détecte un faux puits.

Influence de h

On considère ici deux cas étudiés sur la même trajectoire simulée sur les trois puits, avec comme paramètres : $c = 0.1$, $\theta = 0.5$ nm^2·ns^{-1}, $d = 2$, $T = 100$ ns, $\Delta t = 4$ ps et $N = 25000$. Les paramètres de l'algorithme sont les suivants :

	ρ (nm)	$\overline{R}$ (nm)	$\kappa_{\min}$	$N_{\min}$	δ	h	K	w
Cas 1	2	1	15	40	125	$h_1 = 5\delta = 625$	4δ	$3K$
Cas 2	2	1	15	40	125	$h_2 = 100$	4δ	$3K$

TABLEAU 3.8 – Paramètres pour l'étude de l'influence de h.

Dans le premier cas (voir Fig. 3.31, $h_1 = 625$), la gamme de points de départ (jaune) est assez large pour qu'un centre de puits puisse être trouvé. Quand elle ne l'est pas, et que l'algorithme doit poursuivre trop loin dans la lecture de la trajectoire jusqu'à trouver un maximum, alors rejeté par ρ, un faible nombre d'incrémentation de la bande est suffisant pour qu'un centre raisonnable puisse être identifié : c'est le cas ici, avec un seul segment transitoire, et un centre choisi dès la bande suivante. Notons néanmoins que le point choisi n'est pas centré au mieux. Pour assurer l'identification d'un point bien centré (et donc d'une meilleure saturation du nombre de tours), il sera ainsi recommandé de choisir une valeur de h assez grande.

Dans le second cas (voir Fig. 3.32, $h_2 = 100$), les étapes transitoires se multiplient car la bande est trop étroite, et on n'envisage pas assez de points de départ. Quand on y parvient enfin, le choix est tout de même plus restreint, et on obtient une moins bonne saturation du nombre de tours pour le dernier puits (16 contre 33).

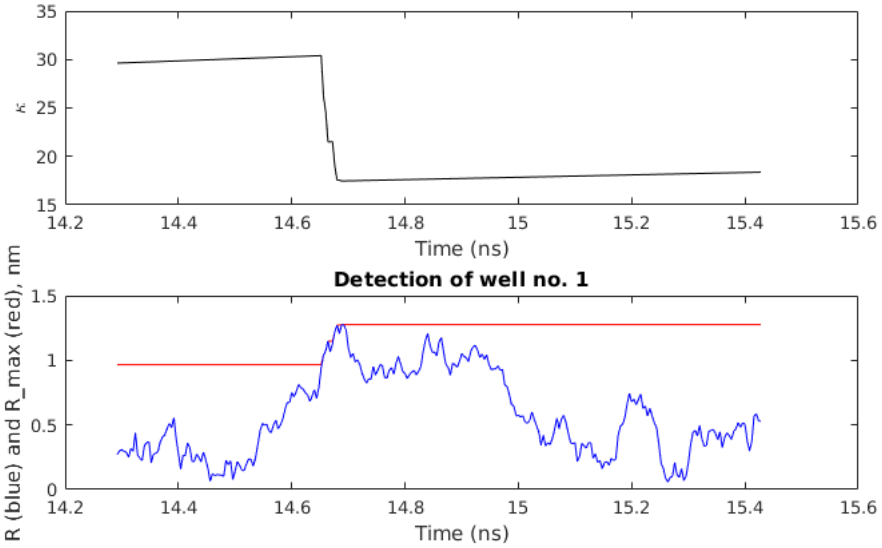

FIGURE 3.30 – $K_2 = 95$. Haut : profil de κ au voisinage de $\tilde{\jmath}$; Bas : profils de R (rouge) et $R_{\max}$ (bleu) au voisinage de $\tilde{\jmath}$.

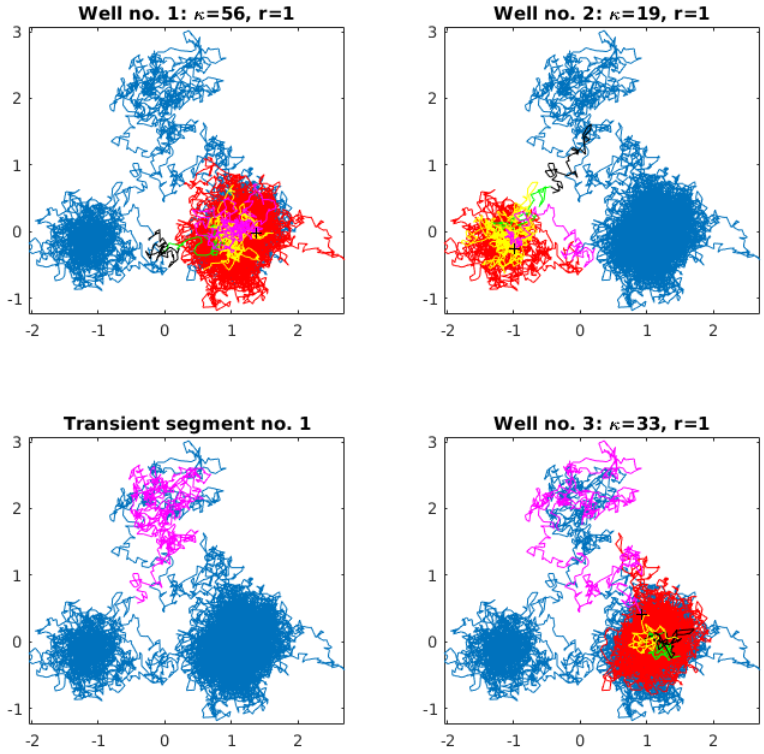

FIGURE 3.31 – Résultat pour $h_1 = 625$.

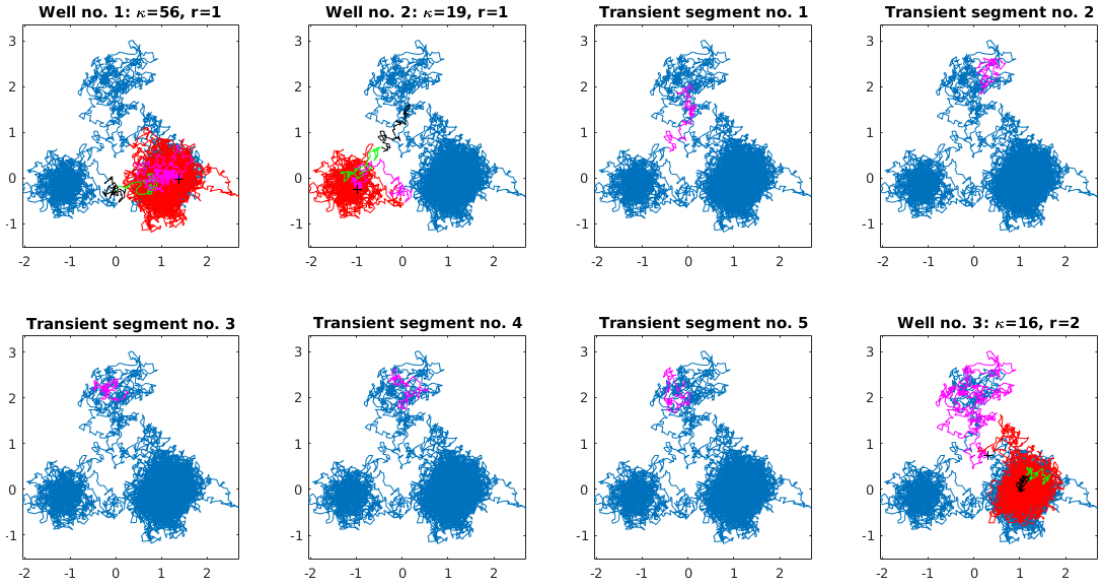

FIGURE 3.32 – Résultat pour $h_2 = 100$.

3.4.2 Bilan

En pratique, l'utilisateur sera souvent confronté de manière récurrente à certains enjeux. Bien souvent, la segmentation lui paraîtra trop fine, ou trop grossière, ne rendant ainsi pas bien compte de la segmentation qu'il souhaiterait obtenir s'il s'en tenait à une lecture de la matrice du nombre de tours. Concrètement, l'utilisateur souhaitera ainsi réunir, ou diviser des puits. Pour ce faire, il pourra jouer sur ρ (voir Critère no. 5) ou sur $\kappa_{\min}$ (voir Critère no. 2) de sorte à rendre l'algorithme plus ou moins sévère quant au rayon maximum et au nombre de tours minimum tolérables respectivement pour définir un puits. De manière similaire, il pourra agir quant à la détection de fausses excursions en faisant varier K (voir Critère no. 3) mais aussi γ (voir Critère no. 4).

Chapitre 4

Application de la κ-segmentation à VKORC1 et KIT

Nous présentons dans ce chapitre les résultats obtenus par l'application de l'algorithme de κ-segmentation aux trajectoires de simulations de DM pour VKORC1 et KIT (voir Chapitre 1 §1.2).

4.1 Application à VKORC1

Nous présentons ici les résultats que fournit la κ-segmentation appliquée à deux répliques de VKORC1 (même conformation initiale, mais vitesses initiales différentes) recalées par rapport à leur conformation initiale. Toutes deux ont été simulées au pas $(\Delta t)_{\text{sim}} = 2$ fs, et observées au pas $\Delta t = 5$ ps, pendant une durée totale de $T = 1$ µs. On s'intéresse en fin de simulation aux résidus 9 à 151 de C_α, de sorte à disposer d'une matrice $\mathbf{X} = (\mathbf{X}_i)_{1 \leq i \leq N} \in \mathcal{M}_{d,N}(\mathbb{R})$ avec $d = 3 \times 143$ et $N = 200000$ conformations observées. Par ailleurs, l'étude des variations quadratiques associées fournissent un coefficient de diffusion estimé en dimension 2 de $\mathbf{D}_{\text{VKORC1}} = 1000$ nm^2·ns^{-1} (voir Chapitre 2 §2.2.3). Dans toute la suite, nous nommerons Trajectoire 1 et Trajectoire 2 ces deux trajectoires. La Fig. 4.1 représente ces deux trajectoires sous leurs formes projetées suite à une ACP de dimension 2.

Avant toute chose, remarquons qu'aucune de ces deux trajectoires ne semble visiter plusieurs fois les mêmes zones de l'espace : à cette échelle de temps, le processus visite toujours de nouvelles régions. Par ailleurs, comme le montre la Fig. 4.2, bien que les conformations initiales soient identiques pour ces deux trajectoires, la simple considération de vitesses initiales différentes induit la visite de régions très différentes de l'espace. L'accumulation des trajectoires superposées traduit en réalité leurs évolutions dans des espaces orthogonaux à ceux sur lesquels s'effectuent les projections.

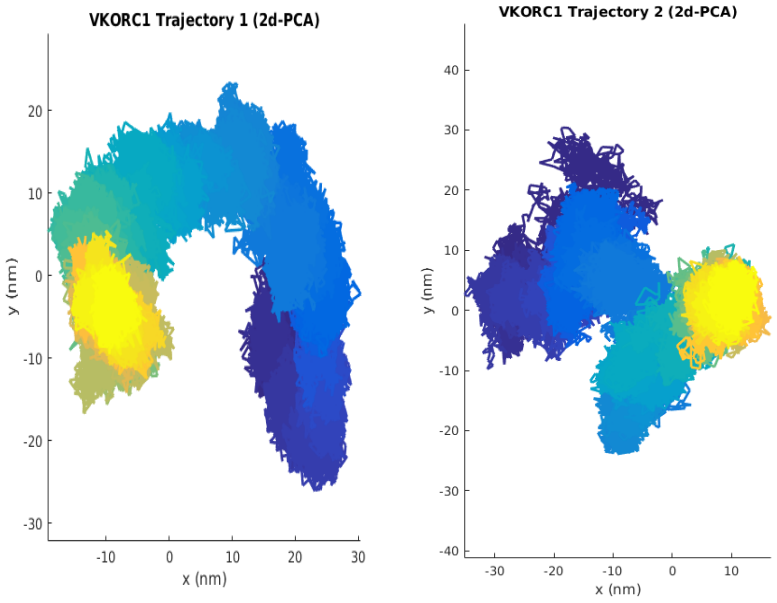

FIGURE 4.1 – Gauche : Trajectoire 1 ; Droite : Trajectoire 2. Toutes deux sont recalées et projetées en dimension 2 par ACP. Bleu → Jaune : Suivi temporel.

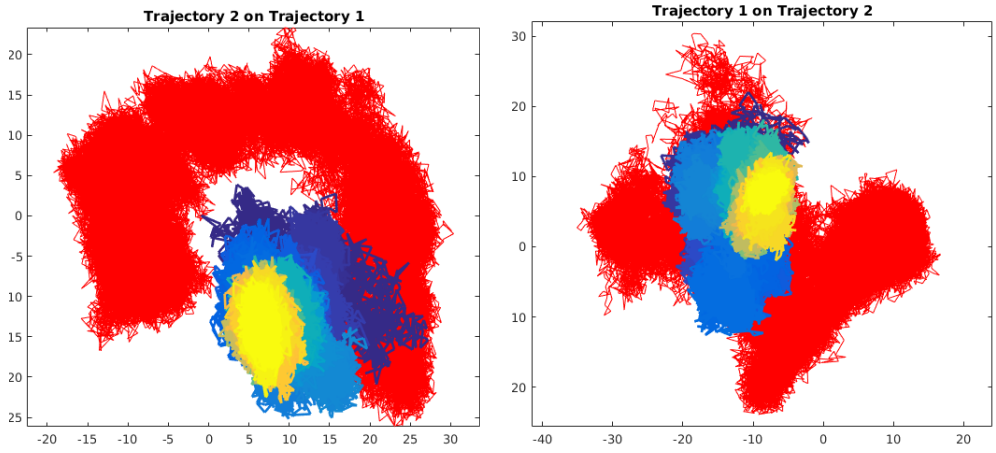

FIGURE 4.2 – Gauche : Trajectoire 2 (bleu → jaune) projetée par ACP sur la Trajectoire 1 (rouge) ; Droite : Trajectoire 1 (bleu → jaune) projetée par ACP sur la Trajectoire 2 (rouge).

4.1.1 Trajectoire 1 (1 µs/5 ps)

ρ (nm)	$\overline{R}$ (nm)	$\kappa_{\min}$	$N_{\min}$	δ	h	K	w
25	10	50	40	20	20δ	5δ	$5K$

TABLEAU 4.1 – Paramètres pour la κ-segmentation de la Trajectoire 1.

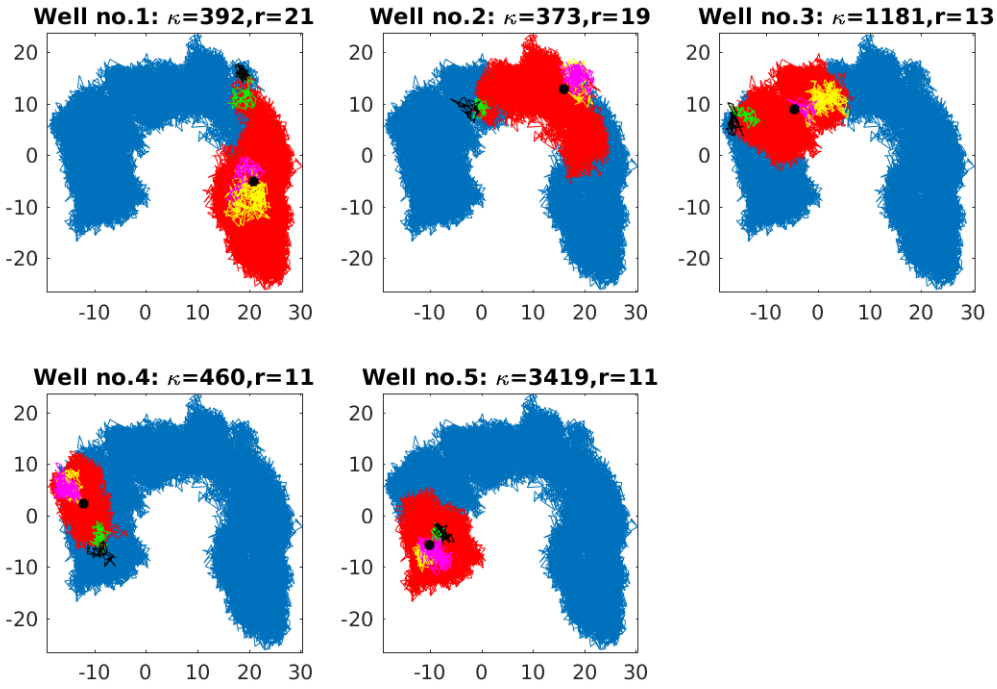

FIGURE 4.3 – κ-segmentation effectuée sur la Trajectoire 1. Pour chaque puits sont indiqués le nombre de tours associé $\kappa = \kappa(\tilde{\imath}, \tilde{\jmath})$ et le rayon correspondant $r = R_{\max}(\tilde{\imath}, \tilde{\jmath})$ (nm). Magenta : segments transitoires ; Rouge : zones de puits ; Jaune : plages de conformations de départ ; Vert : instants précédant les sorties de puits ; Noir : instants suivant les sorties de puits ; Croix noires : centres des puits.

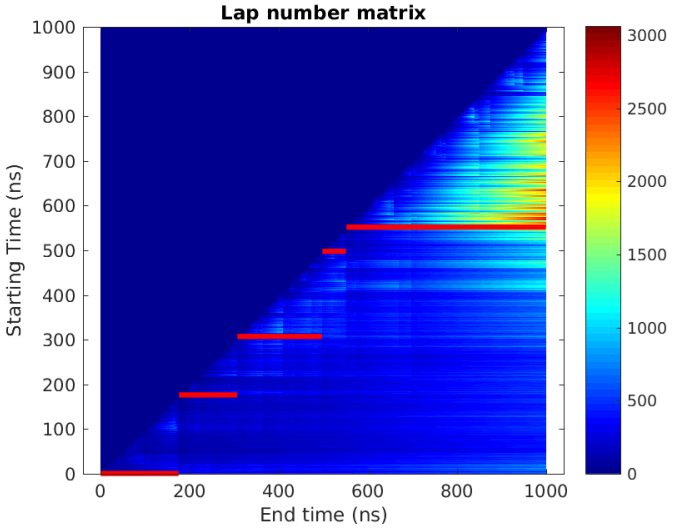

FIGURE 4.4 – κ-segmentation de la Trajectoire 1 représentée sur sa matrice du nombre de tours (segments rouges $[\![\tilde{\imath}, \tilde{\jmath}]\!]$). Ordonnées : instant d'accès (ns) ; Abscisses : instant de sortie (ns).

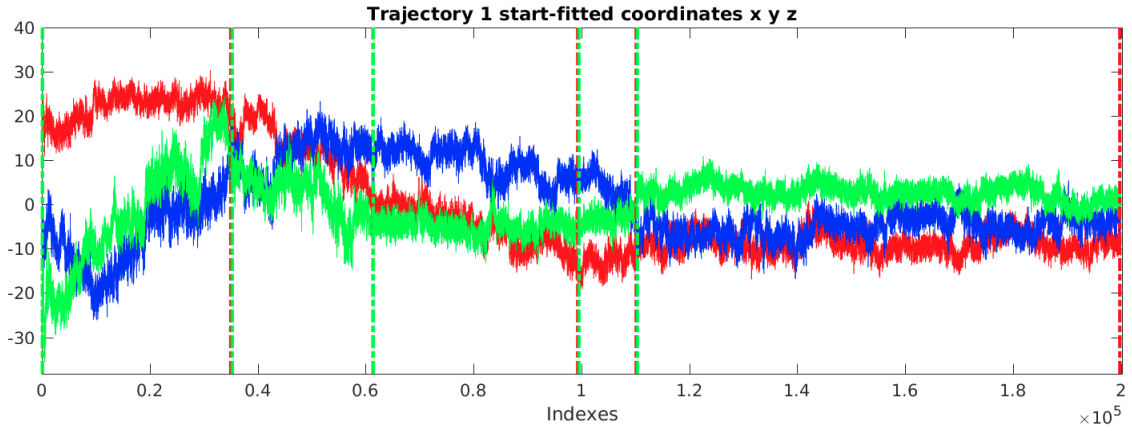

FIGURE 4.5 – κ-segmentation de la Trajectoire 1 représentée sur le profil de ses coordonnées en dimension 3 (ACP). Ordonnées : valeurs des coordonnées (nm) ; Abscisses : indices des conformations (*frames* de la trajectoire). Rouge : x ; Bleu : y ; Vert : z. Pointillés verts : instants d'accès $\tilde{\imath}$; Pointillés rouges : instants de sortie $\tilde{\jmath}$.

Remarquons immédiatement que la κ-segmentation fait émerger des nombres de tours très nettement supérieurs aux valeurs observées jusqu'alors : cette situation réelle est significativement différente de celle des trois puits. Plus précisément, la segmentation met ici en évidence le dernier puits trouvé, lequel permet d'atteindre un nombre de tours de $\kappa(\tilde{\imath},\tilde{\jmath}) = 3419$, pour un rayon $R_{\max}(\tilde{\imath},\tilde{\jmath}) = 11$ nm et un temps de sortie $\widetilde{\mathbf{T}} = 447$ ns. En guise de première confirmation de la stabilité de ce puits, nous proposons de simuler une réplique de 120 ns observée toutes les 1 ns, partant de la conformation correspondant au centre du puits, mais avec des vitesses différentes. Nous recalons cette réplique par rapport à la conformation initiale de la Trajectoire 1, et nous la projetons sur les axes de l'ACP en deux dimensions de cette dernière (voir Fig. 4.6).

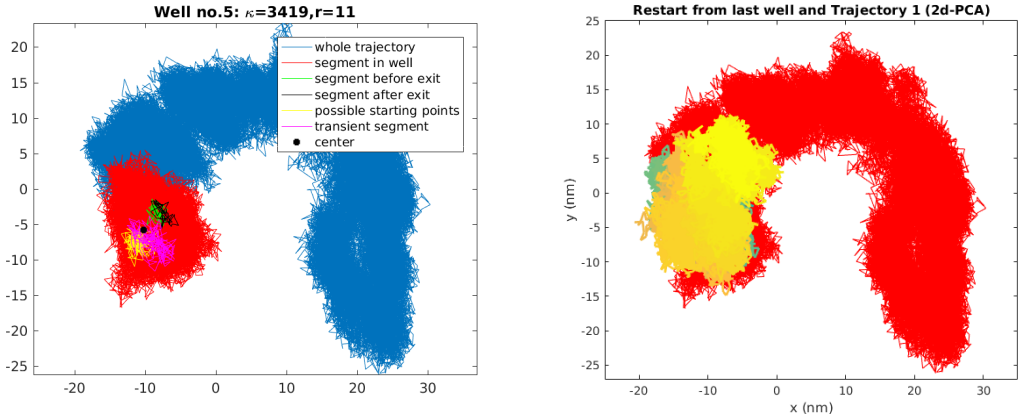

FIGURE 4.6 – Gauche : Dernier puits de la Trajectoire 1 : $\kappa(\tilde{\imath},\tilde{\jmath}) = 3419$, $R_{\max}(\tilde{\imath},\tilde{\jmath}) = 11$ nm, $\widetilde{\mathbf{T}} = 447$ ns. Droite : Projection de la réplique effectuée à partir du dernier puits de la Trajectoire 1 sur les axes d'ACP de la Trajectoire 1. Rouge : Trajectoire 1 ; Bleu → Jaune : Suivi temporel de la réplique.

On observe alors que la réplique reste globalement dans le puits identifié. Si l'on compare cette situation avec celle des Trajectoire 1 et 2, lesquelles constituent également des répliques mais qui

se dirigent immédiatement après leur départ vers des mondes très différents, on comprend que le phénomène observé est remarquable.

4.1.2 Trajectoire 2 (1 µs/5 ps)

ρ (nm)	$\overline{R}$ (nm)	$\kappa_{\min}$	$N_{\min}$	δ	h	K	w
18	10	50	40	20	400	10δ	$10K$

TABLEAU 4.2 – Paramètres pour la κ-segmentation de la Trajectoire 2.

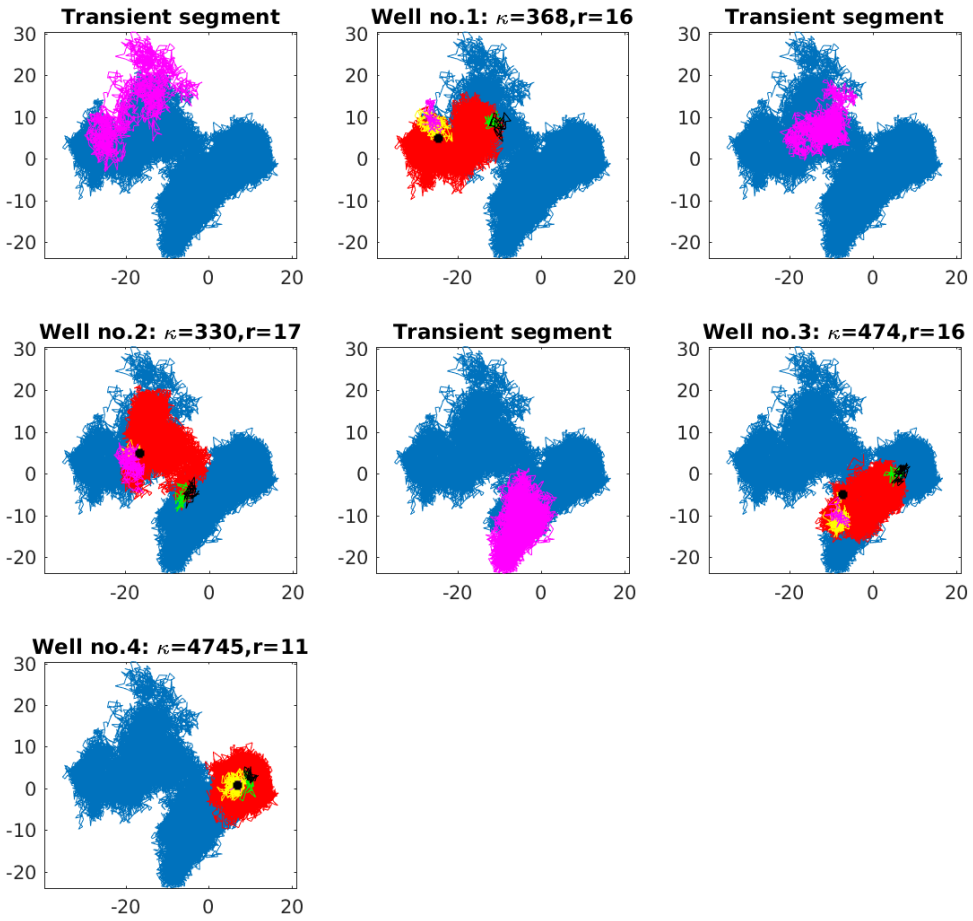

FIGURE 4.7 – κ-segmentation effectuée sur la Trajectoire 2. Pour chaque puits sont indiqués le nombre de tours associé $\kappa = \kappa(\tilde{\imath}, \tilde{\jmath})$ et le rayon correspondant $r = R_{\max}(\tilde{\imath}, \tilde{\jmath})$ (nm). Magenta : segments transitoires ; Rouge : zones de puits ; Jaune : plages de conformations de départ ; Vert : instants précédant les sorties de puits ; Noir : instants suivant les sorties de puits ; Croix noires : centres des puits.

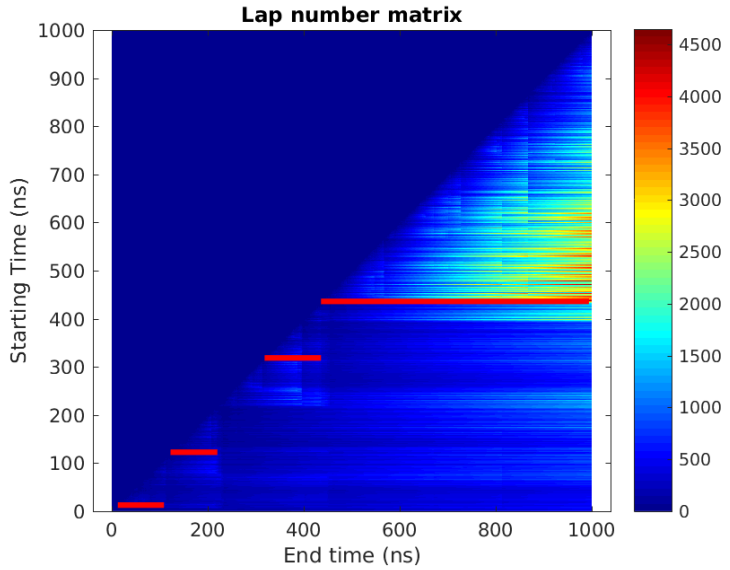

FIGURE 4.8 – κ-segmentation de la Trajectoire 2 représentée sur sa matrice du nombre de tours (segments rouges $[\![\tilde{\imath}, \tilde{\jmath}]\!]$). Ordonnées : instant d'accès (ns) ; Abscisses : instant de sortie (ns).

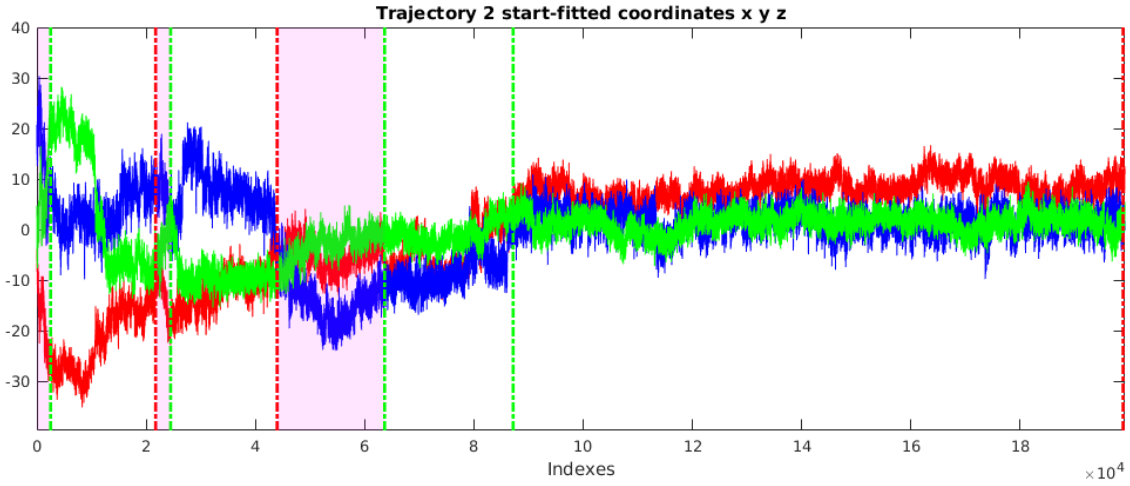

FIGURE 4.9 – κ-segmentation de la Trajectoire 2 représentée sur le profil de ses coordonnées en dimension 3 (ACP). Ordonnées : valeurs des coordonnées (nm) ; Abscisses : indices des conformations (*frames* de la trajectoire). Rouge : x ; Bleu : y ; Vert : z. Pointillés verts : instants d'accès $\tilde{\imath}$; Pointillés rouges : instants de sortie $\tilde{\jmath}$; Magenta : zones transitoires.

La κ-segmentation met à nouveau en évidence le dernier puits trouvé, lequel permet d'atteindre un nombre de tours de $\kappa(\tilde{\imath}, \tilde{\jmath}) = 4745$, pour un rayon $R_{\max}(\tilde{\imath}, \tilde{\jmath}) = 11$ nm et un temps de sortie $\widetilde{\mathbf{T}} = 558$ ns. Nous simulons alors une réplique de 80 ns observée toutes les 1 ns, partant de la conformation correspondant au centre du puits, mais avec des vitesses différentes. Nous recalons cette réplique par rapport à la conformation initiale de la Trajectoire 2, et nous la projetons sur les axes de l'ACP en deux dimensions de cette dernière (voir Fig. 4.10).

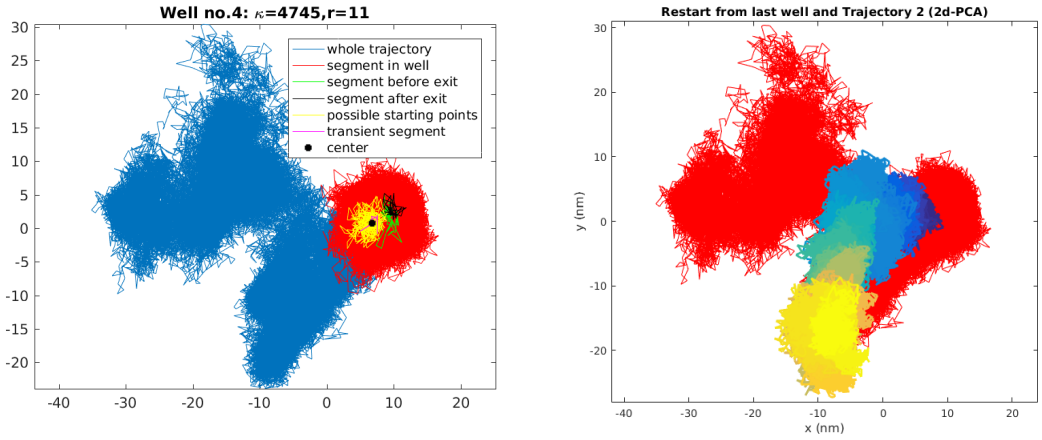

FIGURE 4.10 – Droite : Dernier puits de la Trajectoire 2 : $\kappa(\tilde{\imath},\tilde{\jmath}) = 4745$, $R_{\max}(\tilde{\imath},\tilde{\jmath}) = 11$ nm, $\widetilde{\mathbf{T}} = 558$ ns. Gauche : Projection de la réplique effectuée à partir du dernier puits de la Trajectoire 2 sur les axes de l'ACP de la Trajectoire 2. Rouge : Trajectoire 2 ; Bleu → Jaune : Suivi temporel de la réplique.

Nous faisons alors deux constats. Tout d'abord, bien que la réalisation du temps de sortie du dernier puits observé sur la Trajectoire 2 soit très grand, laissant ainsi à penser que la réplique y resterait confinée, celle-ci en sort immédiatement après quelques ns seulement. En considérant $\mathcal{B}(x_0, R_0) = \{x \in \mathbb{R}^d : \|x - x_0\| \leq R_0\}$ une boule fermée pouvant représenter ce puits, ce phénomène peut s'expliquer par la variance de la variable aléatoire $\tau_0 = \inf_{t \geq t_0} \{\|X_t - x_0\| \geq R_0\}$, où t_0 représente l'instant de départ de la réplique. En effet, en supposant que cette variable aléatoire suive une loi exponentielle de paramètre $\lambda > 0$ (hypothèse raisonnable à basse température), on a que $\mathbb{V}\mathrm{ar}(\tau_0) = 1/\lambda^2$ et donc un écart type du même ordre que l'espérance $\frac{1}{\lambda} = \mathbb{E}(\tau_0)$.

Outre le phénomène de sortie "prématurée", nous constatons que le processus semble rebrousser chemin, en ce sens qu'il se dirige vers la zone de l'espace conformationnel visitée en amont du dernier puits. Malgré les nombres de tours apparents plus faibles dans cette zone qu'en fin de trajectoire, nous comprenons que la variance du temps de sortie influe de manière importante sur les valeurs de κ obtenues en sortie de segmentation, pouvant potentiellement faire varier ces dernières de plusieurs ordres de grandeurs pour un même puits. En effet, sous l'hypothèse d'un temps de sortie suivant une loi exponentielle de paramètre $\lambda > 0$, représentons l'histogramme du quotient de deux lois X_1 et X_2 suivant une telle loi. La loi quotient résultante, indépendante de λ, fournit le tracé de la Fig. 4.11.

Cette figure nous montre ainsi que deux réalisations indépendantes du temps de sortie exponentielle d'un même puits sont à comprendre à un facteur multiplicatif près. Plus précisément, la loi de ce rapport correspond à une loi de type Pareto de densité $f(x) = 1/(1+x)^2$ pour $x \geq 0$ et de fonction de répartition $F(x) = 1 - 1/(1+x)$. En effet, si $R = U/V$ où U et V sont deux variables aléatoires indépendantes de loi géométrique de paramètre 1, le changement de variable $(u, v) \to (r = u/v, v)$ donne la densité $f(r, v) = ve^{-(1+r)v}$ pour la loi du couple (R, V) et par intégration en V la densité $f(r) = 1/(1+r)^2$ pour la variable R. Par suite, comme le rapport R de deux réalisations indépendantes d'un temps de sortie est compris avec probabilité $1 - \alpha$ entre les quantiles d'ordre $\alpha/2$ et $1 - \alpha/2$ on a :

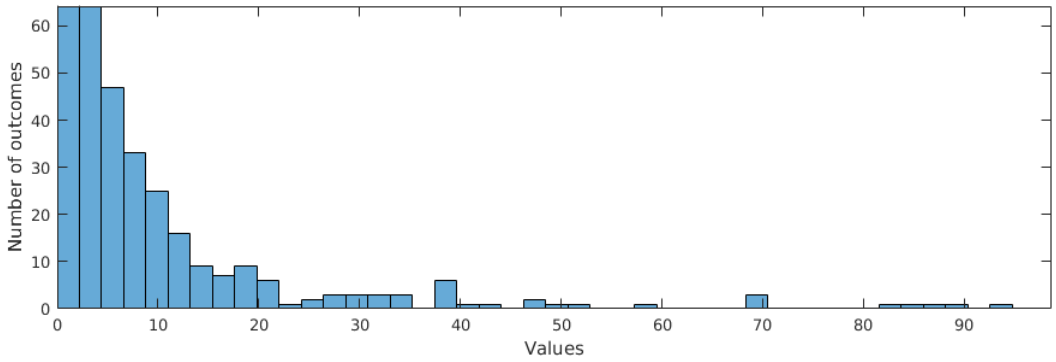

FIGURE 4.11 – Histogramme du quotient de deux lois exponentielles de paramètre 1 calculé à partir de $p = 1000$ réalisations de la variable $U = \log(u_1)/\log(u_2)$, avec u_1, u_2 suivant une loi uniforme sur $[0, 1]$.

$$\mathbb{P}(q_{\alpha/2} \leq R \leq q_{1-\alpha/2}) = 1 - \alpha$$

avec $q_\alpha = 1/(1-\alpha) - 1$. Pour donner un ordre d'idée, on a pour $\alpha = 0.05$, on a que $\mathbb{P}(1/39 \leq R \leq 39) = 0.95$ ce qui confirme que la valeur de κ estimée sur une réalisation doit être comprise à un ordre de grandeur près.

Au regard de cette remarque, il est impossible d'affirmer que la différence entre les nombres de tours observés pour les deux derniers puits de la Trajectoires 2, à savoir $\kappa_3 = 474$, et $\kappa_4 = 4745$, soit significative, ni que ces deux puits soient totalement dissociables.

Pour résumer, la sortie du dernier puits en un temps très court après le lancement de la réplique ne discrédite pas son importance, mais peut s'expliquer par la variance de la loi du temps de sortie. D'autre part, le non-retour dans ce puits et la visite du puits précédent n'induisent pas non plus le rejet du dernier puits, du fait que les valeurs de κ sont à considérer à des ordres de grandeurs près. Enfin, il est avant tout prometteur de constater que le processus ne quitte pas ce dernier puits pour visiter des zones totalement différentes de l'espace, mais reste dans des zones déjà visitées. Ceci nous invite à penser qu'à partir de la µs, la protéine VKORC1 est potentiellement en mesure d'avoir visité une grande partie de son espace conformationnel envisageable, et que des phénomènes de retours émergent. Il se pourrait alors qu'en simulant un nombre assez grand de trajectoires de VKORC1 d'une durée $T > 1$ µs, des méthodes relevant des MSM soient légitimes.

4.2 Application à KIT

De même que précédemment, nous distinguons différents sous-domaines de la molécule KIT complète.

4.2.1 Trajectoire de KIT avec KID (2 µs/10 ps)

On commence par appliquer l'algorithme de κ-segmentation à la trajectoire du système KIT complet, d'une durée de $T = 2$ µs, simulée toutes les 2 fs et observée au pas $\Delta t = 10$ ps. On isole les $n = 400$ atomes de C_α de sorte à disposer d'une matrice $\mathbf{X} = (\mathbf{X}_i)_{1 \leq i \leq N} \in \mathcal{M}_{d,N}(\mathbb{R})$, avec

$d = 3 \times 400$ avec $N = 200000$ conformations observées. Pour ce système, on a en dimension 2 que $\mathbf{D}_{\text{KIT+KID}} = 900$ nm^2·ns^{-1} (voir Chapitre 2 §2.2.3). Dans toute la suite, cette trajectoire sera nommée Trajectoire KIT+KID.

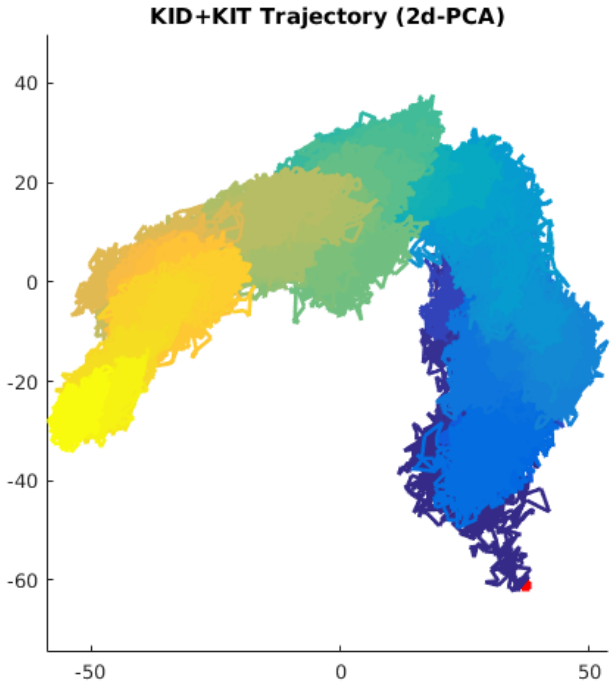

FIGURE 4.12 – Trajectoire KIT+KID recalée et projetée en dimension 2 par ACP. Bleu → Jaune : Suivi temporel.

ρ (nm)	$\overline{R}$ (nm)	$\kappa_{\min}$	$N_{\min}$	δ	h	K	w
40	30	250	25	100	500	4δ	$3K$

TABLEAU 4.3 – Paramètres pour la κ-segmentation de la Trajectoire KIT+KID.

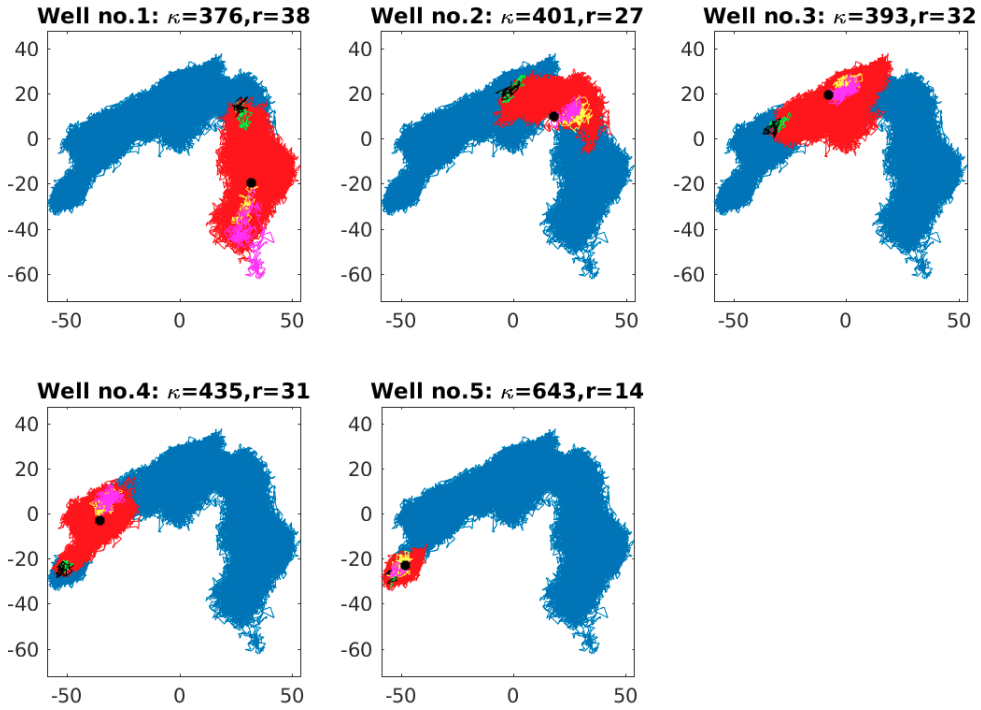

FIGURE 4.13 – κ-segmentation effectuée sur la Trajectoire KIT+KID. Pour chaque puits sont indiqués le nombre de tours associé $\kappa = \kappa(\tilde{\imath}, \tilde{\jmath})$ et le rayon correspondant $r = R_{\max}(\tilde{\imath}, \tilde{\jmath})$ (nm). Magenta : segments transitoires ; Rouge : zones de puits ; Jaune : plages de conformations de départ ; Vert : instants précédant les sorties de puits ; Noir : instants suivant les sorties de puits ; Croix noires : centres des puits.

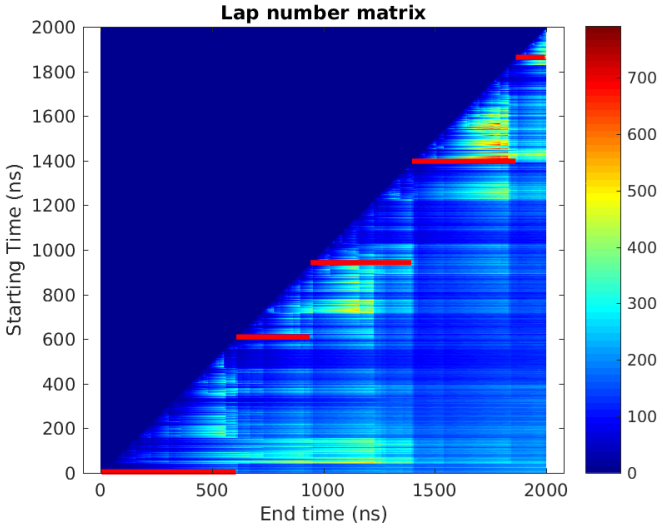

FIGURE 4.14 – κ-segmentation de la Trajectoire KIT+KID représentée sur sa matrice du nombre de tours (segments rouges $[\![\tilde{\imath}, \tilde{\jmath}]\!]$). Ordonnées : instant d'accès (ns) ; Abscisses : instant de sortie (ns).

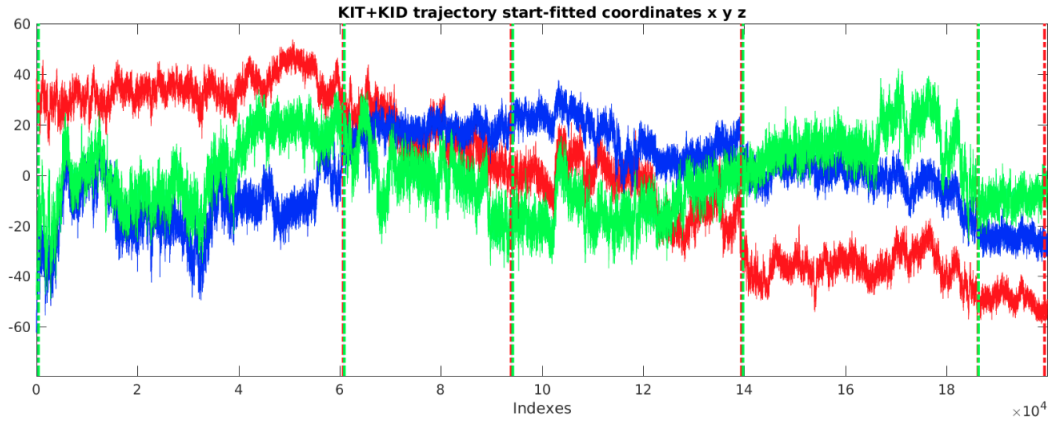

FIGURE 4.15 – κ-segmentation de la Trajectoire KIT+KID représentée sur le profil de ses coordonnées en dimension 3 (ACP). Ordonnées : valeurs des coordonnées (nm) ; Abscisses : indices des conformations (*frames* de la trajectoire). Rouge : x ; Bleu : y ; Vert : z. Pointillés verts : instants d'accès $\tilde{\imath}$; Pointillés rouges : instants de sortie $\tilde{\jmath}$; Magenta : zones transitoires.

La κ-segmentation met en avant le dernier puits de nombre de tours $\kappa = 643$, de rayon $R_{\max} = 14$ nm et de temps de sortie $\widetilde{\mathbf{T}} = 130$ ns.

4.2.2 Trajectoire de KID seul (2 µs/10 ps)

On applique la κ-segmentation à une trajectoire du domaine KID seul, qui correspond à la restriction de la trajectoire précédente aux atomes correspondant à KID, c'est-à-dire aux résidus 143 à 222 pour un total de $n = 80$ atomes de C_α. Cette trajectoire est ainsi d'une durée de $T = 2$ µs observée toutes les $\Delta t = 10$ ps de sorte que l'on dispose d'une matrice $\mathbf{X} = (\mathbf{X}_i)_{1 \leq i \leq N} \in \mathcal{M}_{d,N}(\mathbb{R})$, avec $d = 3 \times 80$ et $N = 200000$ conformations observées. Pour ce système, on a en dimension 2 que $\mathbf{D}_{\text{KID}} = 1000$ nm^2·ns^{-1} (voir Chapitre 2 §2.2.3). Dans toute la suite, cette trajectoire sera nommée Trajectoire KID.

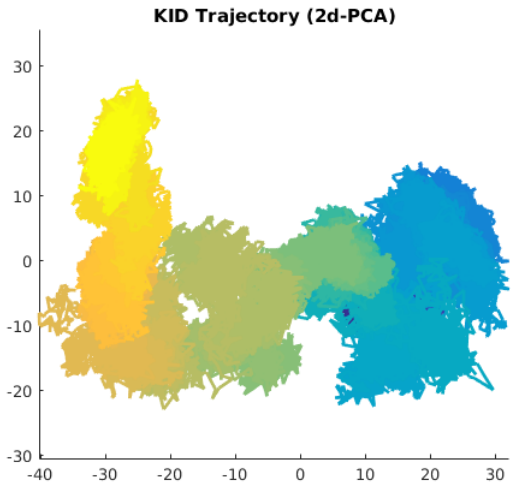

FIGURE 4.16 – Trajectoire KID recalée et projetée en dimension 2 par ACP. Bleu → Jaune : Suivi temporel.

ρ (nm)	$\overline{R}$ (nm)	$\kappa_{\min}$	$N_{\min}$	δ	h	K	w
20	18	800	25	33	16000	1500	5000

TABLEAU 4.4 – Paramètres pour la κ-segmentation de la Trajectoire KID.

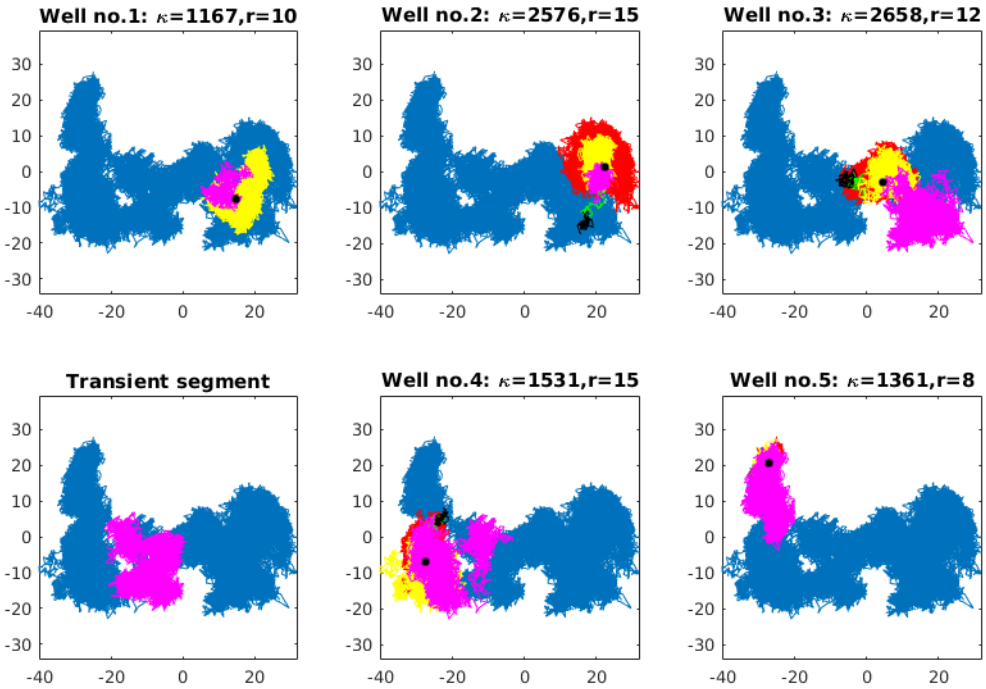

FIGURE 4.17 – κ-segmentation effectuée sur la Trajectoire KID. Pour chaque puits sont indiqués le nombre de tours associé $\kappa = \kappa(\tilde{\imath}, \tilde{\jmath})$ et le rayon correspondant $r = R_{\max}(\tilde{\imath}, \tilde{\jmath})$ (nm). Magenta : segments transitoires ; Rouge : zones de puits ; Jaune : plages de conformations de départ ; Vert : instants précédant les sorties de puits ; Noir : instants suivant les sorties de puits ; Croix noires : centres des puits.

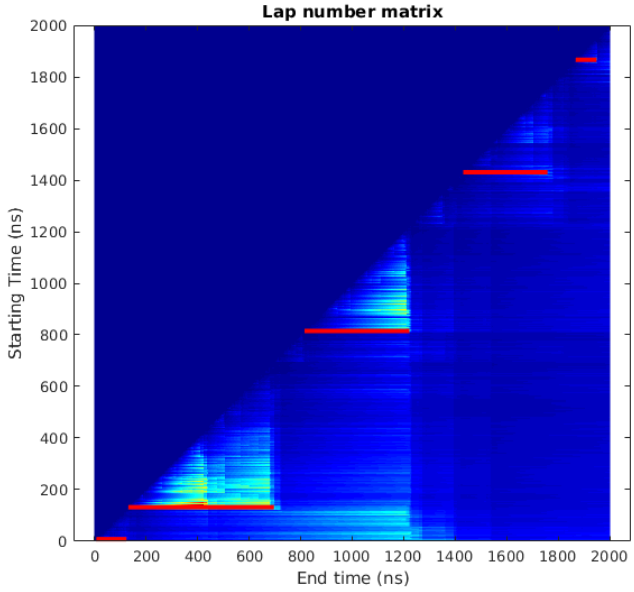

FIGURE 4.18 – κ-segmentation de la Trajectoire KID représentée sur sa matrice du nombre de tours (segments rouges $[\![\tilde{\imath}, \tilde{\jmath}]\!]$). Ordonnées : instant d'accès (ns) ; Abscisses : instant de sortie (ns).

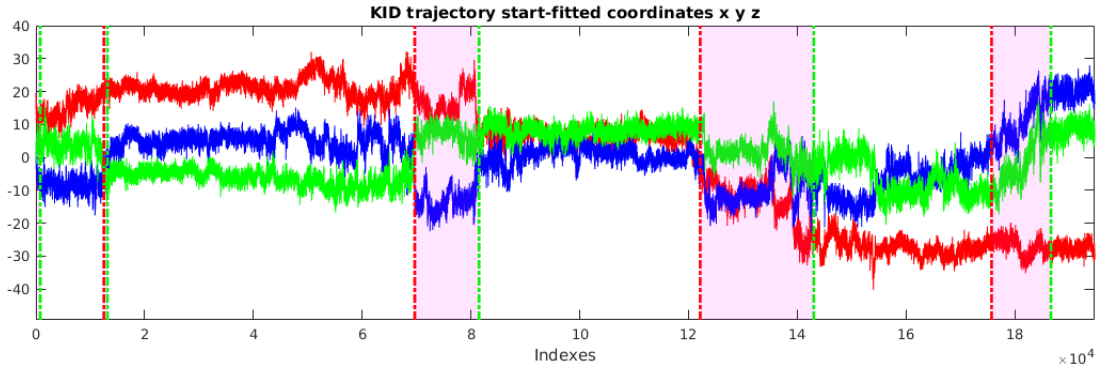

FIGURE 4.19 – κ-segmentation de la Trajectoire KID représentée sur le profil de ses coordonnées en dimension 3 (ACP). Ordonnées : valeurs des coordonnées (nm) ; Abscisses : indices des conformations (*frames* de la trajectoire). Rouge : x ; Bleu : y ; Vert : z. Pointillés verts : instants d'accès $\tilde{\imath}$; Pointillés rouges : instants de sortie $\tilde{\jmath}$.

La κ-segmentation met dans un premier temps en évidence le troisième puits avec une valeur du nombre de tours de $\kappa_3 = 3059$. Nous observons également une transition globale très nette vers les $x < 0$ (voir Fig. 4.19, courbe rouge).

4.2.3 Trajectoire de KD avec KID (2 µs/10 ps)

On applique la κ-segmentation à une trajectoire du domaine KD (c'est-à-dire de KIT sans le domaine JMR) avec KID, qui correspond à la restriction de la trajectoire de KIT+KID aux résidus 35 à 369 pour un total de $n = 335$ atomes de C_α. Cette trajectoire est ainsi d'une durée de $T = 2$ µs observée toutes les $\Delta t = 10$ ps de sorte que l'on dispose d'une matrice $\mathbf{X} = (\mathbf{X}_i)_{1 \leq i \leq N} \in \mathcal{M}_{d,N}(\mathbb{R})$,

avec $d = 3 \times 335$ et $N = 200000$ conformations observées. Pour ce système, on a en dimension 2 que $\mathbf{D}_{\text{KD+KID}} = 750$ nm^2·ns^{-1} (voir Chapitre 2 §2.2.3). Dans toute la suite, cette trajectoire sera nommée Trajectoire KD+KID.

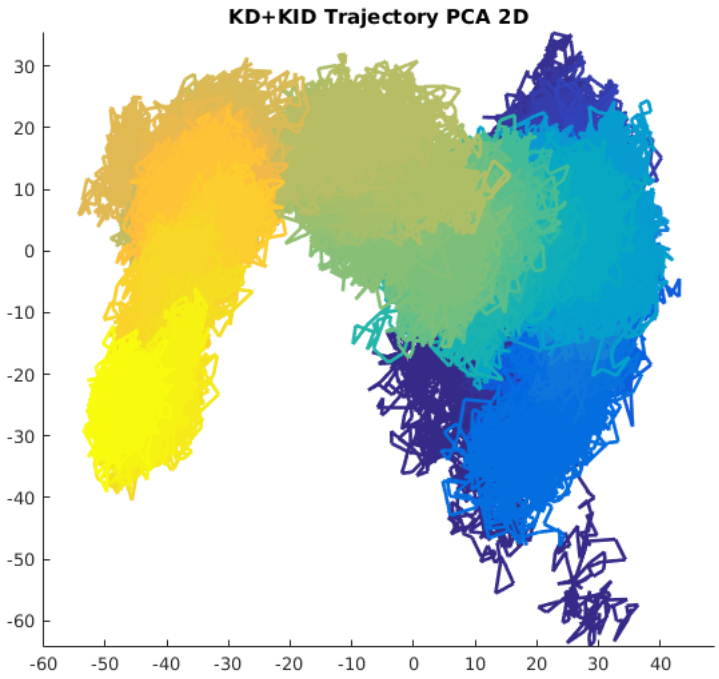

FIGURE 4.20 – Trajectoire KD+KID recalée et projetée en dimension 2 par ACP. Bleu → Jaune : Suivi temporel.

ρ (nm)	$\overline{R}$ (nm)	$\kappa_{\min}$	$N_{\min}$	δ	h	K	w
35	20	250	25	54	270	4δ	$3K$

TABLEAU 4.5 – Paramètres pour la κ-segmentation de la Trajectoire KD+KID.

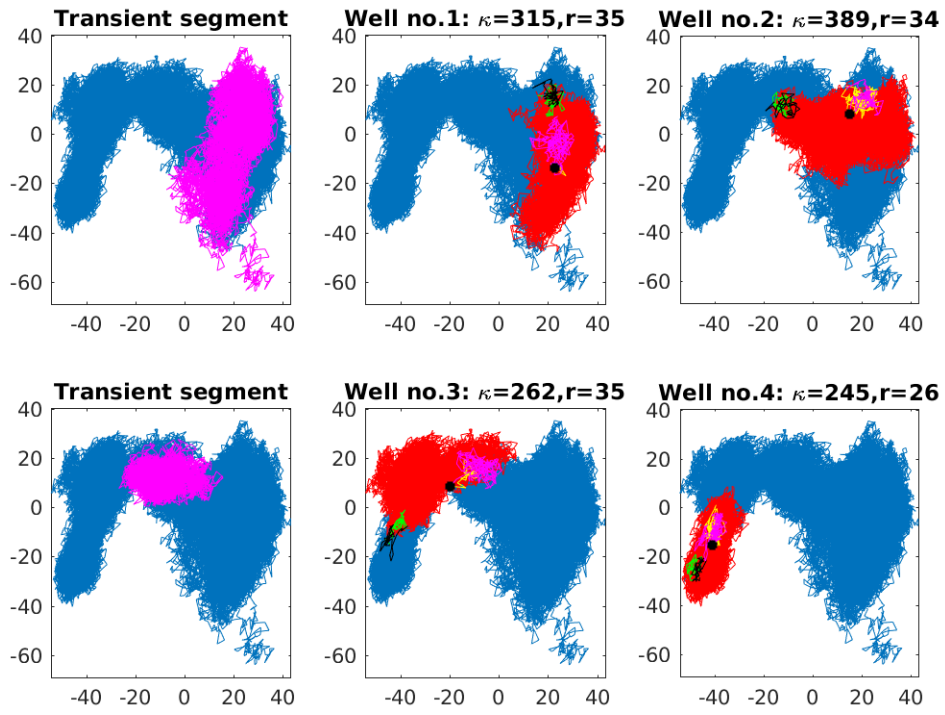

FIGURE 4.21 – κ-segmentation effectuée sur la Trajectoire KD+KID. Pour chaque puits sont indiqués le nombre de tours associé $\kappa = \kappa(\tilde{\imath}, \tilde{\jmath})$ et le rayon correspondant $r = R_{\max}(\tilde{\imath}, \tilde{\jmath})$ (nm). Magenta : segments transitoires ; Rouge : zones de puits ; Jaune : plages de conformations de départ ; Vert : instants précédant les sorties de puits ; Noir : instants suivant les sorties de puits ; Croix noires : centres des puits.

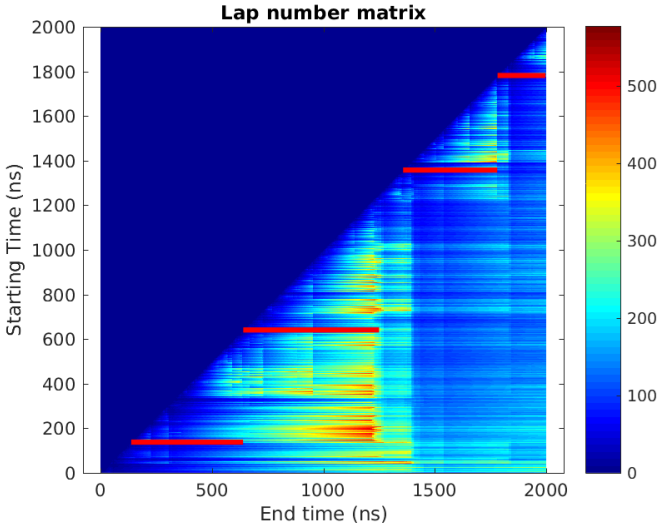

FIGURE 4.22 – κ-segmentation de la Trajectoire KD+KID représentée sur sa matrice du nombre de tours (segments rouges $[\![\tilde{\imath}, \tilde{\jmath}]\!]$). Ordonnées : instant d'accès (ns) ; Abscisses : instant de sortie (ns).

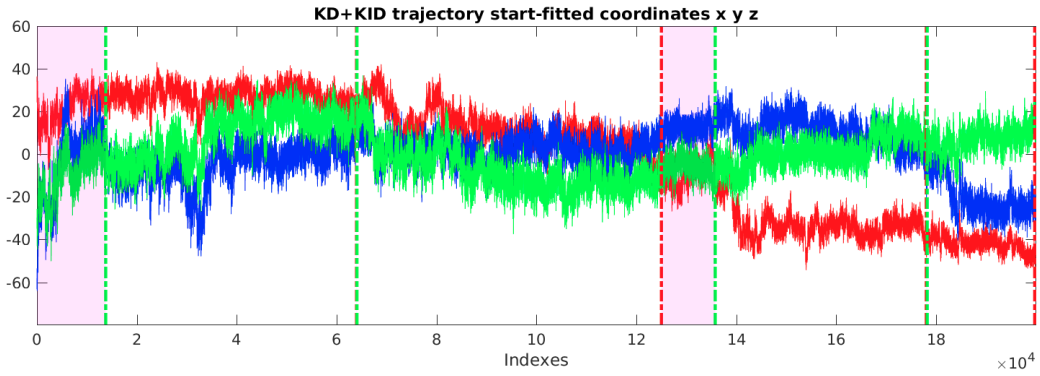

FIGURE 4.23 – κ-segmentation de la Trajectoire KD+KID représentée sur le profil de ses coordonnées en dimension 3 (ACP). Ordonnées : valeurs des coordonnées (nm) ; Abscisses : indices des conformations (*frames* de la trajectoire). Rouge : x ; Bleu : y ; Vert : z. Pointillés verts : instants d'accès $\tilde{\imath}$; Pointillés rouges : instants de sortie $\tilde{\jmath}$.

4.2.4 Trajectoire de KD seul (2 µs/10 ps)

On applique la κ-segmentation à une trajectoire du domaine KD seul, qui correspond à la restriction de la trajectoire de KIT+KID aux résidus 35 à 142 puis 223 à 369 pour un total de $n = 255$ atomes de C_α. Cette trajectoire est ainsi d'une durée de $T = 2$ µs observée toutes les $\Delta t = 10$ ps de sorte que l'on dispose d'une matrice $\mathbf{X} = (\mathbf{X}_i)_{1 \leq i \leq N} \in \mathcal{M}_{d,N}(\mathbb{R})$, avec $d = 3 \times 255$ et $N = 200000$ conformations observées. Pour ce système, on a en dimension 2 que $\mathbf{D}_{\text{KD}} = 900$ nm^2·ns^{-1} (voir Chapitre 2 §2.2.3). Dans toute la suite, cette trajectoire sera nommée Trajectoire KD.

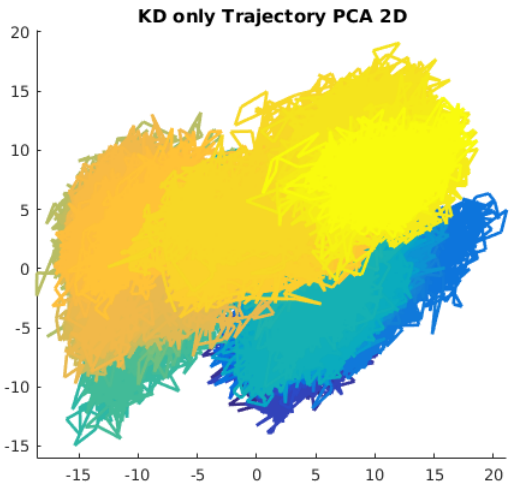

FIGURE 4.24 – Trajectoire KD recalée et projetée en dimension 2 par ACP. Bleu → Jaune : Suivi temporel.

ρ (nm)	$\overline{R}$ (nm)	$\kappa_{\min}$	$N_{\min}$	δ	h	K	w
25	15	2000	40	25	10000	5δ	$3K$

TABLEAU 4.6 – Paramètres pour la κ-segmentation de la Trajectoire KD

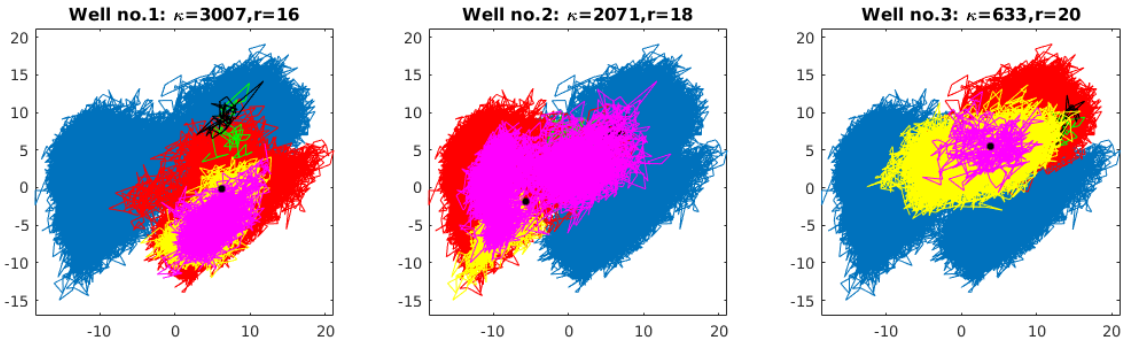

FIGURE 4.25 – κ-segmentation effectuée sur la Trajectoire KD. Pour chaque puits sont indiqués le nombre de tours associé $\kappa = \kappa(\tilde{\imath}, \tilde{\jmath})$ et le rayon correspondant $r = R_{\max}(\tilde{\imath}, \tilde{\jmath})$ (nm). Magenta : segments transitoires ; Rouge : zones de puits ; Jaune : plages de conformations de départ ; Vert : instants précédant les sorties de puits ; Noir : instants suivant les sorties de puits ; Croix noires : centres des puits.

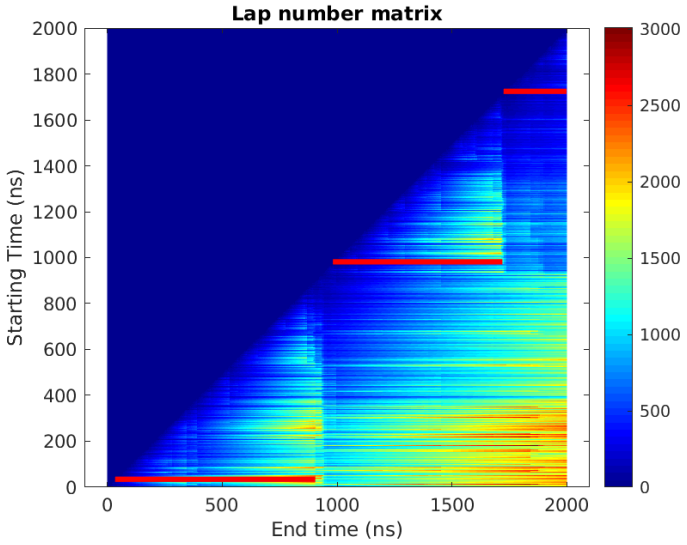

FIGURE 4.26 – κ-segmentation de la Trajectoire KD représentée sur sa matrice du nombre de tours (segments rouges $[\![\tilde{\imath}, \tilde{\jmath}]\!]$). Ordonnées : instant d'accès (ns) ; Abscisses : instant de sortie (ns).

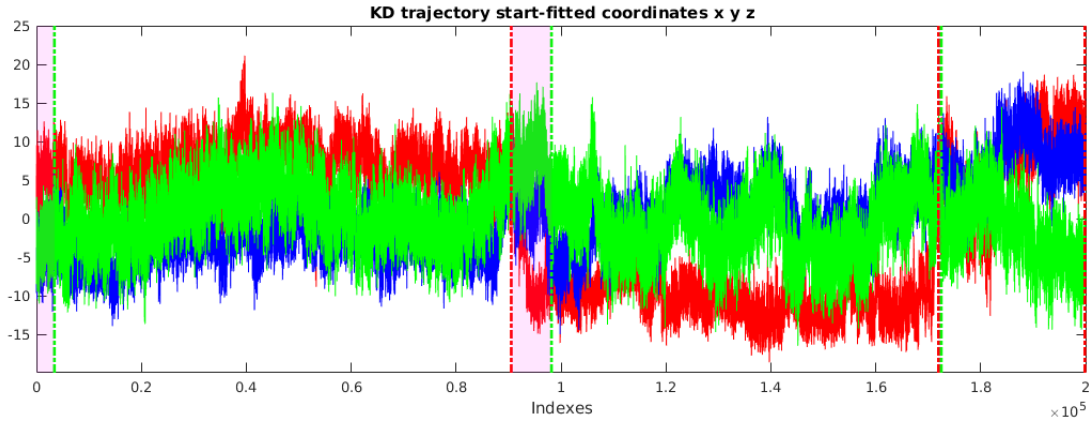

FIGURE 4.27 – κ-segmentation de la Trajectoire KD représentée sur le profil de ses coordonnées en dimension 3 (ACP). Ordonnées : valeurs des coordonnées (nm) ; Abscisses : indices des conformations (*frames* de la trajectoire). Rouge : x ; Bleu : y ; Vert : z. Pointillés verts : instants d'accès $\tilde{\imath}$; Pointillés rouges : instants de sortie $\tilde{\jmath}$.

4.3 Bilan

Nous voyons à travers ces exemples la faculté de la κ-segmentation à offrir une analyse très fine mettant en évidence de nombreux phénomènes de stabilisations locales, lesquels demeureraient invisibles si l'on appliquait un algorithme de clustering classique. Dans des cas complexes, la κ-segmentation constitue donc un outil exploratoire permettant d'isoler certaines portions d'une trajectoire qui paraissent dignes d'intérêt.

On peut alors envisager de compléter l'étude par une visualisation (par exemple sous VMD ou PyMol) de certaines conformations, ou de certains segments tout entiers comme des zones transitoires. En effet, l'algorithme de κ-segmentation est à considérer pour une *gamme* de paramètres, et non pour un unique jeu de paramètres qui serait optimal. A cet égard, il est indispensable de prendre en compte les enjeux biologiques propres à chaque système étudié, et d'affiner ou non la segmentation en conséquence.

Chapitre 5

Équivalent trois puits et estimation de paramètres dynamiques

Nous proposons au cours de ce chapitre un moyen de représenter toute trajectoire de simulation de DM par son équivalent dans le paysage des trois puits. Nous pourrons nous servir de cette représentation dans le but d'édifier une méthode d'estimation des termes de dérive et de diffusion du processus sous-jacent en régime EDS. Nous appliquerons ensuite ces estimateurs aux trajectoires de simulations de DM dont nous disposons, et comparerons les résultats obtenus concernant l'estimation de la dérive par rapport à une estimation basée sur le calcul du gradient de l'énergie estimée.

5.1 Modélisation de la simulation de DM par un modèle de trois puits

Nous proposons dans cette partie une approche consistant à modéliser tout processus par un modèle de trois puits. Cette modélisation s'effectue par la considération des trois paramètres suivants : la température θ, le paramètre d'échelle spatiale ρ, et le paramètre d'échelle temporelle α. On considère donc la *famille* de processus $(Y_t^{(\theta,\rho,\alpha)})_{\theta,\rho,\alpha\in\mathbb{R}^*}$, vérifiant :

$$Y_t^{(\theta,\rho,\alpha)} = \rho X_{t/\alpha}^{(\theta)}$$

où $X_t^{(\theta)}$ correspond quant à lui au processus canonique de température θ sur le paysage des trois puits. Il vérifie donc :

$$dX_t^{(\theta)} = -\nabla U_c(X_t)dt + \sqrt{2\theta}dB_t$$

La dynamique d'un tel processus est alors régie par l'EDS suivante :

$$dY_t^{(\theta,\rho,\alpha)} = -\frac{\rho}{\alpha}\nabla U_c\left(\frac{Y_t}{\rho}\right)dt + \sqrt{2\theta\frac{\rho^2}{\alpha}}dB_t$$

Nous cherchons alors à calibrer ces paramètres de sorte à ce qu'ils correspondent à la réalité observée sur les données. En nous basant sur une κ-segmentation effectuée sur des données $\mathbf{X} = (x_i)_{1\leq i\leq N}$, nous pouvons élire un puits de référence, dont on notera ici $\mathbf{T}_0$ le temps de sortie,

$\mathbf{R}_0$ le rayon et κ_0 le nombre de tours. Par ailleurs, cette segmentation aura été effectuée à l'aide d'un coefficient de diffusion $\mathbf{D}_0 = 2d\theta_0$ identifié par une étude des variations quadratiques associées au aux données. On cherche alors à identifier le triplet $(\theta_{\text{éq}}, \rho_{\text{éq}}, \alpha_{\text{éq}})$, tel que les quantités associées $(T_{\text{éq}}, D_{\text{éq}}, R_{\text{éq}})$ coïncident avec le triplet $(\mathbf{T}_0, \mathbf{D}_0, \mathbf{R}_0)$ observé sur les données, c'est-à-dire à résoudre le système :

$$\begin{cases} T_{\text{éq}} = \mathbf{T}_0 \\ D_{\text{éq}} = \mathbf{D}_0 \\ R_{\text{éq}} = \mathbf{R}_0 \end{cases}$$

ou, de manière équivalente, le système suivant :

$$\begin{cases} \kappa_{\text{éq}} = \kappa_0 \\ D_{\text{éq}} = \mathbf{D}_0 \\ R_{\text{éq}} = \mathbf{R}_0 \end{cases}$$

Considérons désormais ce nouveau système. Par construction nous avons que $R_{\text{éq}} = \rho_{\text{éq}}\mathbf{R}_{3P}$ et $D_{\text{éq}} = 2\theta_{\text{éq}} d\rho_{\text{éq}}^2/\alpha_{\text{éq}}$, avec $\mathbf{R}_{3P} = 1$, rayon caractéristique sur le paysage des trois puits canonique. On identifie alors les paramètres de calibration ainsi :

$$\begin{cases} \kappa_{\text{éq}}(\theta_{\text{éq}}) = \kappa_0 \\ \rho_{\text{éq}} = \mathbf{R}_0 \\ \alpha_{\text{éq}} = 2\theta_{\text{éq}} d\rho_{\text{éq}}^2/\mathbf{D}_0 \end{cases} \tag{5.1}$$

Pour résoudre la première équation, on se propose d'effectuer une série de simulations dans le but d'estimer les temps de sortie d'un même puits pour différentes températures, et d'identifier celle qui permet d'atteindre le nombre de tours κ_0. Ayant constaté précédemment que le calcul du nombre de tours était invariant par changements d'échelles de temps et d'espace (voir Chapitre 3 §3.1.3), nous pouvons ainsi, pour trouver la température équivalente $\theta_{\text{éq}}$, nous contenter de simuler le processus canonique $X^{(\theta)}$ pour diverses températures. Nous cherchons alors, partant du point $x_0 = (1, 0)$, à trouver la température $\theta_{\text{éq}}$ permettant de retrouver le nombre de tours observé sur les données. Pour une température θ donnée, et en posant $D = 2\theta d$, le temps passé dans le bassin considéré s'écrit :

$$\psi(D) = \mathbb{E}_\theta[\tau_{\mathcal{B}(x_0,1)}]$$

où $\tau_{\mathcal{B}(x_0,1)}$ est la variable aléatoire correspondant au temps de sortie du puits centré en x_0 et de rayon $\mathbf{R}_{3P} = 1$. Et on cherche donc $D_{\text{éq}}$ de sorte que :

$$D_{\text{éq}}\psi(D_{\text{éq}}) = \kappa_0$$

Dans le but d'identifier la valeur recherchée, nous nous proposons alors d'implémenter un tracé de la courbe $D \longmapsto D\psi(D)$. Pour chaque valeur de D, on souhaite simuler en partant du fond d'un puits, de manière à estimer la quantité $\psi(D)$. Quel pas $(\Delta t)_{\text{sim}}$ doit-on utiliser ? Il faudra simuler à une échelle de temps suffisamment fine de sorte que d'éventuelles sorties de puits ne soient pas masquées par une simulation trop grossière (la qualité de l'estimation du temps de sortie en serait altérée). On choisira un δ tel que $\delta \ll \mathbf{R}_{3P} = 1$, de sorte à définir :

$$(\Delta t)_{\text{sim}} = \frac{\delta^2}{D}$$

Nous devons ensuite choisir un nombre de points à simuler qui permette d'observer la sortie du puits : on choisira une valeur très élevée notée $N_{\max}$ permettant l'observation des sorties de

puits. On arrêtera la simulation dès lors que le processus se sera éloigné d'une distance strictement supérieure à $\mathbf{R}_{3P} = 1$ du centre du puits, ou bien que le nombre maximal d'itérations sera atteint : dans ce dernier cas, si le processus n'est toujours pas sorti du puits, on ignorera la réalisation effectuée et on incrémentera un compteur de saturation N_{sat} pour détecter la possibilité d'une censure.

On effectuera, toujours pour chaque température d'un ensemble choisi noté Θ, un nombre N_{exp} de ces simulations, afin de disposer d'un grand nombre de réalisations du temps de sortie, que l'on moyennera : en effet, rappelons que dans le cas d'un temps de sortie suivant une loi exponentielle de paramètre $\lambda > 0$, celle-ci présente une variance $1/\lambda^2$. Il ne restera alors qu'à multiplier ce temps moyen par le quotient $D/\mathbf{R}_{3P}^2 = D$ de sorte à obtenir le nombre de tours.

Nous parvenons à l'algorithme suivant :

Algorithm 5 Temps de sortie et nombre de tours

Entrée $\mathbf{R}_{3P}$, δ, x_0, $\Theta = (\theta_p)_p$, $N_{\max}$, N_{exp}

Boucle externe Pour chaque température $\theta \in \Theta$

Initialisation $D \leftarrow 2d\theta$, $(\Delta t)_{\text{sim}} = \delta/D$, $k \leftarrow 1$, $0 \leftarrow N_{\text{sat}}$

Boucle interne (Simulation) Tant que $k \leq N_{\text{exp}}$:

- $i \leftarrow 0$, $x \leftarrow x_0$
- tant que $(\|x - x_0\| \leq R_{\text{éq}})$ et $(i < N_{\max})$
  - $x \leftarrow x - \nabla U_c(x)(\Delta t)_{\text{sim}} + \sqrt{2\theta(\Delta t)_{\text{sim}}} \times n$ avec $n \sim \mathcal{N}(0_{\mathbb{R}^2}, \mathrm{I}_2)$.
  - $i \leftarrow i + 1$
- Si $\|x - x_0\| > \mathbf{R}_{3P}$, alors :
  - $\psi_k(D) \leftarrow i \times (\Delta t)_{\text{sim}}$
  - $k \leftarrow k + 1$
- Sinon $N_{\text{sat}} \leftarrow N_{\text{sat}} + 1$

Calcul du temps de sortie et du nombre de tours

- $\hat{\psi}(D) \leftarrow \frac{1}{N_{\text{exp}}} \sum_{k=1}^{N_{\text{exp}}} \psi_k(D)$ et $\kappa(\hat{D}) \leftarrow D\hat{\psi}(D)/\mathbf{R}_{3P}^2$

Sortie Tracé des nombres de tours $D \longmapsto \hat{\kappa}(D)$

Nous reportons Fig. 5.1 les estimations fournies par l'algorithme pour $\Theta = \{0.2 + k \times 0.01 : 0 \leq k \leq 30\}$, $N_{\text{exp}} = 100$, $N_{\max} = 5 \times 10^6$ et $\delta = 0.1$. Par ailleurs, on reporte une interpolation (courbe rouge en pointillés) de nos estimées du temps de sortie moyen en fonction de la température par une régression aux moindres carrés de la forme :

$$\log(\mathbb{E}_\theta[\tau_{B(x_0,1)}]) \simeq \log(\psi(D) = a + b/\theta \tag{5.2}$$

i.e. $\psi(D) \simeq \exp(a + b/\theta)$ ou encore $D\phi(D) \simeq D\exp(a + 4bd/D)$.

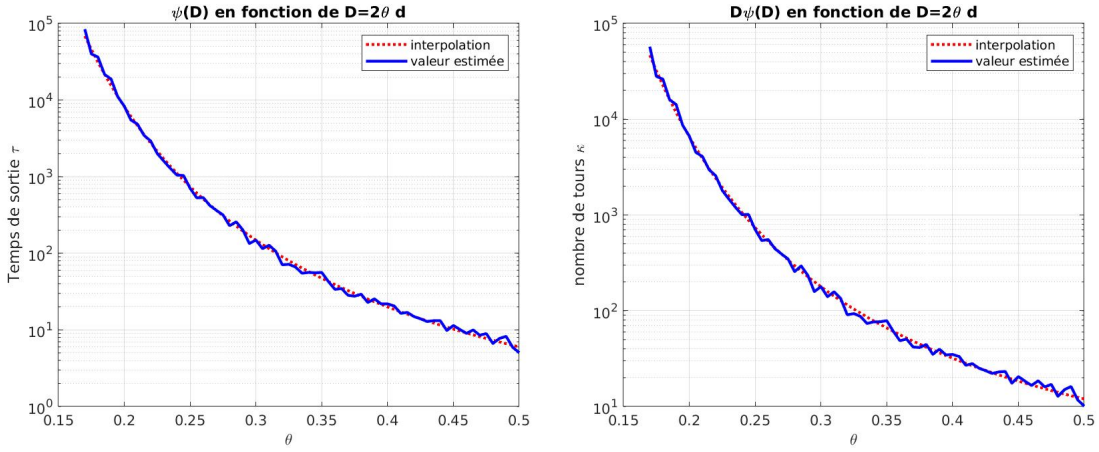

FIGURE 5.1 – Gauche : tracé de $\theta \mapsto \psi(\theta)$ (temps de sortie en fonction de la température) ; Droite : tracé de $\theta \mapsto 2d\theta\psi(\theta)$ (nombre de tours en fonction de la température) ; Pour les valeurs choisies des paramètres, aucune censure n'a été observée ($N_{\text{sat}} = 0$).

Grâce à ces tracés, il est aisé de résoudre l'équation $D\psi(D) = \kappa_0$, pour un κ_0 calculé sur les données observées. Il suffit en effet de confronter la valeur de κ_0 que l'on souhaite retrouver, avec l'interpolation de la Fig. 5.1 de sorte à trouver la température équivalente $\theta_{\text{éq}}$, à laquelle on sait associer la diffusion $D_{\text{éq}} = 2\theta_{\text{éq}}d$.

Remarque 32. *Le caractère exponentiel à basse température de la dépendance du nombre de tours en fonction de $1/\theta$ nous montre que le nombre de tours correspond, dans le paysage des trois puits, à une notion de profondeur de puits liée au choix de la température.*

Remarque 33. *Nous aurions pu également résoudre l'équation invoquant le temps de sortie plutôt que le nombre de tours, auquel cas nous aurions dû identifier la température telle que $T_{\text{éq}} = \mathbf{T}_0/\alpha_{\text{éq}}$.*

Nous sommes alors en mesure de terminer la résolution du système (5.1), et pouvons disposer des trois paramètres $(\theta_{\text{éq}}, \rho_{\text{éq}}, \alpha_{\text{éq}})$. Ainsi, pour créer un modèle de trois puits équivalent à une trajectoire quelconque de durée T et observée au pas Δt, il suffit de simuler le processus canonique $(X^{(\theta_{\text{éq}})})_t$ pendant une durée $T_{\text{éq}} = T/\alpha_{\text{éq}}$ au pas $(\Delta t)_{\text{éq}} = \Delta t/\alpha_{\text{éq}}$, puis de normaliser la trajectoire par $\rho_{\text{éq}}$: de cette façon, on simule bien le processus équivalent $Y^{(\theta_{\text{éq}},\rho_{\text{éq}},\alpha_{\text{éq}})}$.

5.2 Méthode d'estimation de dérive et diffusion

Nous proposons dans cette section de mettre au point des estimateurs de dérive et de diffusion d'un processus sous-jacent à partir d'un jeu de données observé. Si cela est envisageable dans le cas d'un régime EDS, nous montrerons en revanche que la situation est plus délicate dans un cas concret, et qu'un recours à une modélisation par le modèle des trois puits permet de solutionner le problème.

5.2.1 Cas du régime EDS

Nous choisissons donc de nous placer dans le cas d'un processus à valeurs dans $\mathbb{R}^d$ solution de l'EDS :

$$dX_t = \mu(X_t)dt + \sigma(X_t)dB_t \tag{5.3}$$

où $x \mapsto \mu(x) \in \mathbb{R}^d$ et $x \mapsto \sigma(x) = \sqrt{2\theta(x)}$ sont C^1 bornées et Lipschitz, avec $\theta(x) > 0$ représentant une température locale, et B_t un mouvement brownien d-dimensionnel.

Pour $x \in \mathcal{X}$ et $h > 0$ on a que :

$$\begin{cases} \mathbb{E}[(X_h - X_0)/h \ \mid X_0 = x] = \mu(x) + o(1) \\ \mathbb{E}[\|X_h - X_0\|^2/h \mid X_0 = x] = d\sigma(x)^2 + o(1) = D(x) + o(1) \end{cases}$$

où $D(x) = \sigma(x)^2 d = 2\theta(x)d$ est le coefficient de diffusion local du processus.

Dans le cas d'observations discrètes (ce qui sera le cas des trajectoires de simulation de DM), comme nous n'avons que peu d'information sur μ, nous nous tournons vers une approche non-paramétrique par noyau de type Nadaraya-Watson [54, 80] comme introduite dans [74] dans le cas 1D. L'idée, pour l'estimation de la dérive $\mu(x)$, est d'utiliser le fait que le processus est autonome puis d'approximer l'espérance $\mathbb{E}[(X_h - X_0)/h \ \mid X_0 = x]$ au voisinage d'un point x par une moyenne locale pondérée par un noyau des accroissements $(X_{(j+1)\Delta t} - X_{j\Delta t})/\Delta t$ pour $X_{j\Delta t}$ dans un voisinage de x. On procède de même pour l'estimation de $D(x)$ en calculant des moyennes locales pondérées au voisinage de x de $\|X_{(j+1)\Delta t} - X_{j\Delta t}\|^2/\Delta t$.

On considère un noyau gaussien, et une trajectoire $x_i = X_{i\Delta t}$ pour $0 \leq i \leq N-1$ échantillonnant la trajectoire sur $[0, T]$ en N points (on a alors $\Delta t = T/(N-1)$. En reprenant alors les notations du Chapitre 2 §2.3.1, on définit, pour $\epsilon > 0$, la matrice d'adjacence et la matrice de probabilité par :

$$\begin{cases} \mathbf{W}_\epsilon^N[i,j] = \exp(-\|x_i - x_j\|^2/2\epsilon) \\ \\ \mathbf{P}_\epsilon^N[i,j] = \mathbf{W}_\epsilon^N[i,j] / \sum_{k=1}^N \mathbf{W}_\epsilon^N[i,k] \end{cases}$$

On aboutit aux estimateurs suivants pour $x_i \triangleq X_{i\Delta t}$:

$$\begin{cases} \hat{\mu}(x_i) = \sum_{j=1}^{N-1} \mathbf{P}_\epsilon^N[i,j](x_{j+1} - x_j)/\Delta t \\ \\ \widehat{D}(x_i) = \sum_{j=1}^{N-1} \mathbf{P}_\epsilon^N[i,j]\|x_{j+1} - x_j\|^2/\Delta t \end{cases}$$

Dans la suite, pour illustrer le comportement en fonction de Δt des estimations, on considérera la possibilité d'un sous-échantillonage en $m \leq N$ points. Par conséquent, on se propose d'introduire comme paramètre le nombre de points m, et de considérer des matrices régulièrement échantillonnées notées $\mathbf{X}^m$ (voir Chapitre 2 §2.2.2). On précise alors la définition des estimateurs de la façon suivante. Pour tout x_i :

$$\begin{cases} \hat{\mu}_m(x_i) = \sum_{j=1}^{m-1} \mathbf{P}_\epsilon^m[i,j](x_{j+1} - x_j)/\Delta t_m \\ \\ \widehat{D}_m(x_i) = \sum_{j=1}^{m-1} \mathbf{P}_\epsilon^m[i,j]\|x_{j+1} - x_j\|^2/\Delta t_m \end{cases} \tag{5.4}$$

avec $\Delta t_m = T/(m-1)$.

Remarque 34. *Les estimateurs ne dépendent pas du temps. On considérera en effet que seule la conformation courante détermine la dérive ainsi que la température.*

La formulation (5.4) est un cas particulier, en les données, de la formule générale définissant les estimations en tout point x :

$$\begin{cases} \hat{\mu}_m(x) = \sum_{j=1}^{m-1} \frac{K_\epsilon(x-x_j)}{\sum_{k=1}^{m-1} K_\epsilon(x-x_k)} (x_{j+1} - x_j)/\Delta t_m \\ \\ \widehat{D}_m(x) = \sum_{j=1}^{m-1} \frac{K_\epsilon(x-x_j)}{\sum_{k=1}^{m-1} K_\epsilon(x-x_k)} \|x_{j+1} - x_j\|^2/\Delta t_m \end{cases} \tag{5.5}$$

où $K_\epsilon(y) = \exp(-\|y\|^2/2\epsilon)$.

Dans les formules précédentes, nous voyons que nos estimateurs dépendent de deux paramètres : le nombre m de données choisies, et le paramètre ϵ localisant l'estimation. De manière à quantifier l'influence de ces deux paramètres sur la qualité de l'estimation, on se propose de définir les *écarts types empiriques* sur les données associés aux estimateurs :

$$\begin{cases} \mathrm{RMSD}_{\hat{\mu}}(m,\epsilon) = (\frac{1}{m} \sum_{k=1}^{m} \|\hat{\mu}_m(x_k) - \mu(x_k)\|^2)^{1/2} \\ \\ \mathrm{RMSD}_{\hat{D}}(m,\epsilon) = (\frac{1}{m} \sum_{k=1}^{m} |\hat{D}_m(x_k) - D(x_k)|^2)^{1/2} \end{cases}$$

Étudions le profil de ces écarts types à l'aide d'un exemple simulé sur le paysage des trois puits, avec $T = 100$ ns, $\Delta t = 0.02$ ns et $\theta = 0.5$ nm^2.ns^{-1} (voir Fig. 5.2).

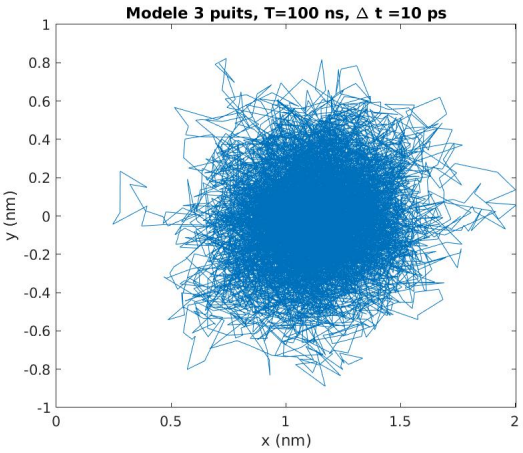

FIGURE 5.2 – Simulation d'une trajectoire issue d'un modèle de trois puits avec $\theta = 0.4$ nm^2·ns^{-1}, $T = 100$ ns, $\Delta t = 10$ ps. Ici la trajectoire reste dans un seul des deux puits principaux sur la durée de la simulation.

On calcule alors les estimées et les écarts types associés pour différentes valeurs de m et ϵ.

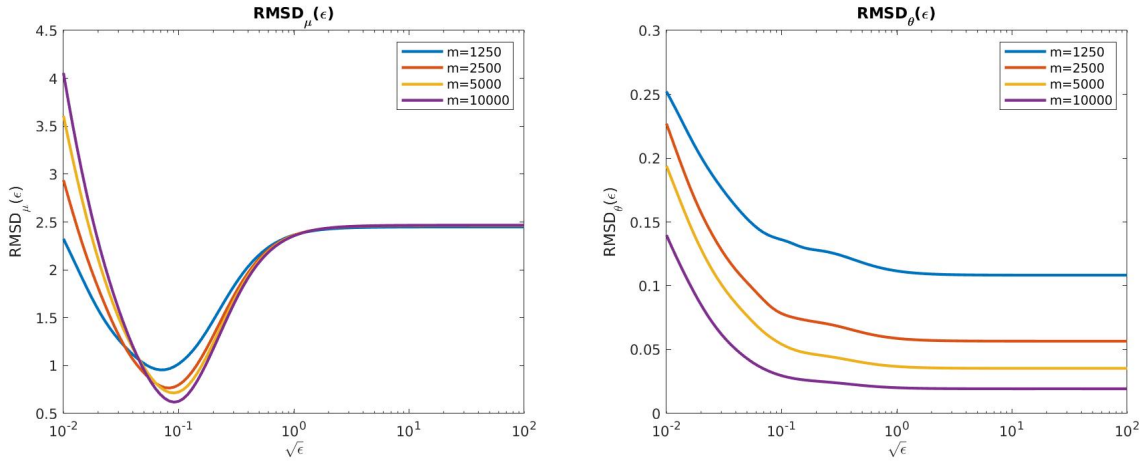

FIGURE 5.3 – Écarts types empiriques associés à l'estimateur $\hat{\mu}_m$ (gauche) et à l'estimateur $\hat{\theta}_m$ (droite) en fonction de $\sqrt{\epsilon}$ et pour diverses valeurs de m.

Pour la dérive, on met en évidence l'existence d'une valeur optimale de ϵ^* qui minimise l'écart type à m fixé dans un compromis classique entre le biais (qui augmente avec ϵ) et la variance (qui diminue avec ϵ). On peut constater que le choix de cette valeur influe assez largement sur la précision du résultat. Sans grande surprise non plus, la précision du résultat augmente avec m. Quant à la diffusion, comme pour le cas de μ, la précision augmente avec la valeur de m. Cependant, la température de notre modèle ne dépendant pas de x, la précision augmente ici toujours avec ϵ puisque θ n'est pas locale.

Procédons alors au calcul des estimées avec les paramètres trouvés précédemment. On prendra pour la dérive $(m_\mu, \sqrt{\epsilon_\mu}) = (5 \times 10^3, 7.9 \times 10^{-2})$ et pour la température $(m_\theta, \sqrt{\epsilon_\theta}) = (5 \times 10^3, 10)$. On commence par l'estimation de la dérive en les données (voir Fig. 5.4).

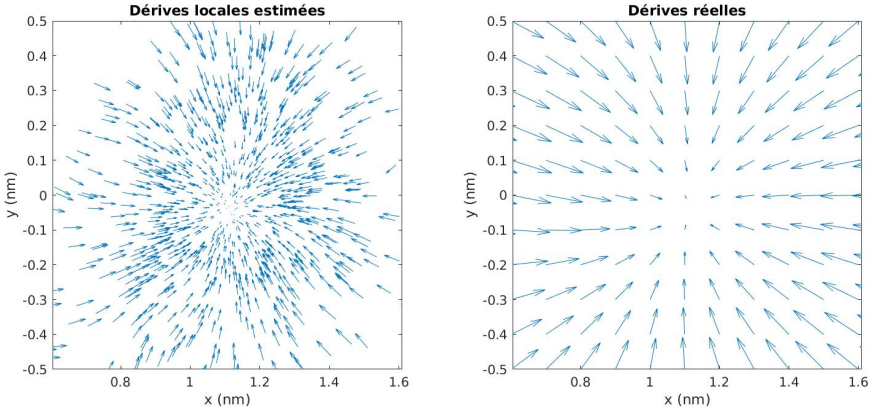

FIGURE 5.4 – Droite : $\hat{\mu}_m(x_i)$ pour $m = 5 \times 10^3$ au voisinage du centre du puits (après sous-échantillonnage tous les 5 points pour plus de visibilité) ; Gauche : $\mu(x)$ calculé dans la même zone sur une grille régulière.

On retrouve tout à fait l'allure générale de la dérive au sein du puits visité avec toutefois des

erreurs légèrement visibles sur la direction du gradient en bord de puits. Considérons maintenant l'estimée de la température $\hat{\theta}_m(x) \triangleq \hat{D}_m(x)/2d$:

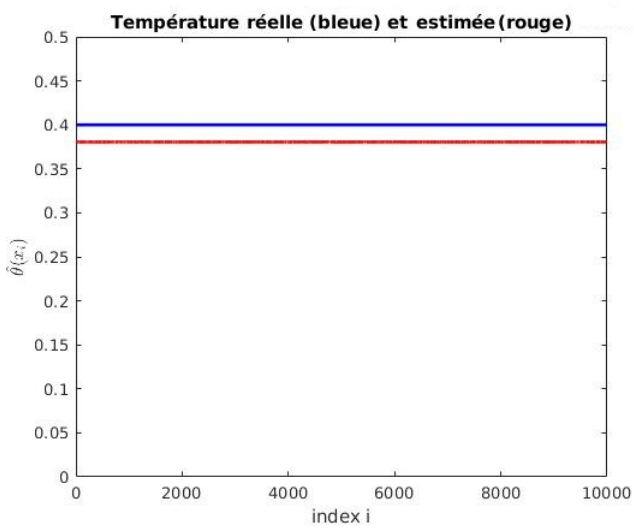

FIGURE 5.5 – Bleu : $\theta(x_i)$ (vraie valeur) ; Rouge : $\hat{\theta}_m(x_i)$ pour $m = 5 \times 10^3$. Les fluctuations de $\hat{\theta}_m(x_i)$ en fonction de i ne sont pratiquement pas visibles ($(\sum_{i=0}^{m}(\hat{\theta}_m(x_i) - \langle\hat{\theta}_m\rangle)^2/m)^{1/2} = 2.8 \times 10^{-6}$ où $\langle\hat{\theta}_m\rangle = \sum_{i=0}^{m-1}\hat{\theta}_m(x_i)/m$) du fait d'une grande valeur de $\sqrt{\epsilon_\theta} = 10$ par rapport à la taille du puits.

Ce premier exemple permet de montrer que même lorsque m est grand, il reste pour la valeur θ, une erreur relativement importante quelque soit le choix de la valeur de ϵ. Il en est de même, bien que cela soit moins visible ici, dans l'estimation de μ.

Pour comprendre ce phénomène, nous considérons ici le cas particulièrement explicite d'un processus d'Ornstein-Uhlenbeck dans $\mathbb{R}^d$ satisfaisant à l'EDS pour un choix de $a, \theta > 0$:

$$dX_i = -aX_t dt + \sqrt{2\theta}\, dX_t \tag{5.6}$$

Une solution explicite est donnée par : $X_h = e^{-ah}X_0 + \sqrt{2\theta}\int_0^h e^{-a(h-s)}dW_s$ pour laquelle on a $\mathbb{E}[X_h - X_0 \mid X_0] = (e^{-ah} - 1)X_0$ et $\mathbb{E}[\ \|X_h - X_0\|^2 \mid X_0] = (e^{-ah} - 1)^2\|X_0\|^2 + \frac{\theta d}{a}(1 - e^{-2ah})$. Dans ce cas, la mesure invariante est $\mu \sim \mathcal{N}(0, \theta/a)$ si bien que l'on a $\mathbb{E}_\mu[\ \|X_h - X_0\|^2\ ] = \frac{2\theta d}{a}(1 - e^{-ah})$ et donc :

$$\mathbb{E}_\mu[\|X_h - X_0\|^2/h] = D(1 - \frac{a}{2}h + O(h^2)) \tag{5.7}$$

où $D = 2\theta d$ est le coefficient de diffusion. En particulier, on remarque ainsi une sous-estimation de D pour $h > 0$, phénomène que nous retrouvons dans notre situation sur le modèle des trois puits.

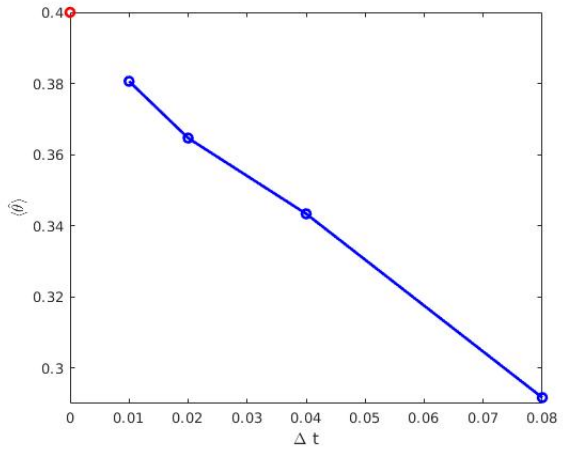

FIGURE 5.6 – Affichage de $\langle\hat{\theta}_m\rangle = \frac{1}{m}\sum_{i=0}^{m_1}\hat{\theta}_m(x_i)$ en fonction de $\Delta t_m = T/m$ pour $m = 1250, 2500, 5000, 10000$.

On peut donc essayer de gagner un ordre dans l'estimation en considérant les deux nouveaux estimateurs (déjà introduits dans [74]) :

$$\begin{cases} \hat{\hat{\mu}}_m(x) \triangleq 2\hat{\mu}_m(x) - \hat{\mu}_{m/2}(x) \\ \\ \hat{\hat{\theta}}_m(x) \triangleq 2\hat{\theta}_m(x) - \hat{\theta}_{m/2}(x) \end{cases} \tag{5.8}$$

La comparaison des estimateurs d'ordre 1 et d'ordre 2 est évaluée à la Fig. 5.7. On constate un effet marqué sur l'estimation de la température locale avec une réduction d'un ordre de grandeur de l'erreur d'estimation pour les grandes valeurs de ϵ. En particulier, on trouve $\langle\hat{\hat{\theta}}\rangle = 3.96 \times 10^{-1}$ nm^2·ns^{-1} pour $\Delta t = 0.01$ ns à comparer à $\hat{\theta} = 3.80 \times 10^{-1}$ nm^2·ns^{-1} obtenu pour l'estimateur direct d'ordre 1 (on rappelle que $\theta = 0.4$ nm^2·ns^{-1}). Elle ne semble pas très sensible dans le cas de l'estimation de μ sans toutefois dégrader les performances.

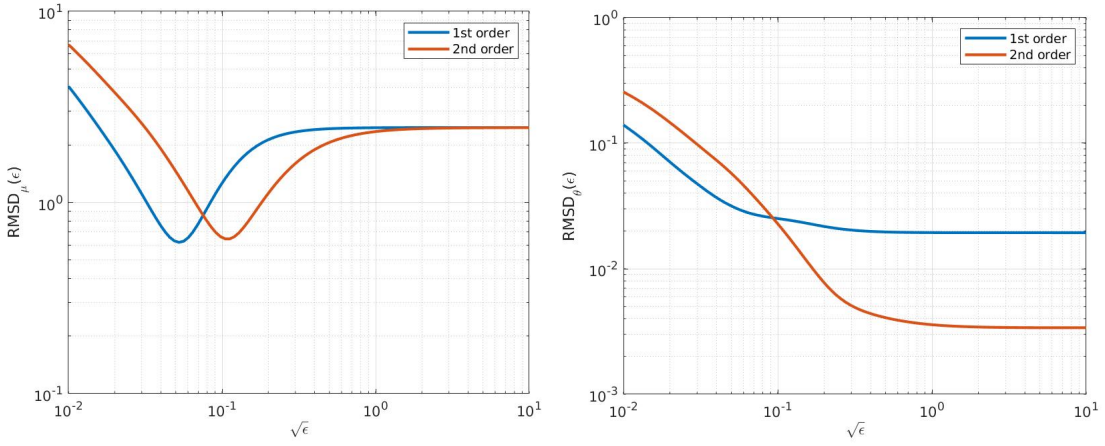

FIGURE 5.7 – Comparaison des performances entre les estimations au premier et au second ordre pour l'estimation de μ (gauche) et pour l'estimation de θ (droite).

Remarque 35. *Dans le but de gagner en précision, l'emploi des estimateurs d'ordre 2 constitue donc une alternative à l'augmentation du nombre de points considérés.*

5.2.2 Cas réel : application à VKORC1 et KIT

Dans un cas réel, une étude des variations quadratiques peut comme nous l'avons vu permettre de modéliser la trajectoire par une EDS (voir Chapitre 2 §2.2.3) et d'obtenir un coefficient de diffusion estimé. Par ailleurs, l'algorithme de κ-segmentation permet l'identification d'un puits de référence, et des quantités associées κ_0 et $\mathbf{R}_0$ dont la connaissance est indispensable à la modélisation par un modèle de trois puits. Dans le but d'appliquer cette méthode aux trajectoires de VKORC1, KIT+KID, KID seul, KD+KID, et KD seul, voici donc les valeurs dont nous aurons besoin ici :

	$\mathbf{D}_0$ ($\mathrm{nm}^2{\cdot}\mathrm{ns}^{-1}$)	κ_0	$\mathbf{R}_0$ (nm)
VKORC1 (Traj. 1)	1000	3419	11
VKORC1 (Traj. 2)	1000	4745	11
KIT+KID	900	643	14
KID	1000	2658	12
KD+KID	750	389	34
KD	900	3007	16

TABLEAU 5.1 – Valeurs de $\mathbf{D}_0$, κ_0 et $\mathbf{R}_0$ pour VKORC1, KIT+KID, KID, KD+KID et KD.

En revanche, disposer d'un pas de temps d'observation assez faible est crucial dans le but d'appliquer les estimateurs définis précédemment : en effet, l'écriture d'une discrétisation suppose que le pas de temps soit assez faible de sorte à ce que $\mu(x_i)$ suffise à expliquer l'itération vers x_{i+1}. Or, la validité d'un régime EDS, indispensable à l'application de nos estimateurs, exige de ne pas descendre en deçà de la valeur pivot $(\Delta t)^*$, lequel n'est potentiellement pas assez fin pour permettre une bonne estimation. Pour lever cette incertitude, on se propose de comparer la distance R^* parcourue par le processus observé pendant un temps $(\Delta t)^*$ correspondant au pas de temps limite permettant une modélisation par EDS, au rayon moyen des puits observés sur une projection en dimension 2 par ACP, i.e $\overline{R}$. Pour chaque système d'intérêt, cela revient à s'assurer que :

$$R^* \triangleq \sqrt{\mathbf{D} \times (\Delta t)^*} \ll \overline{R}$$

Nous avons alors les valeurs suivantes :

	R^* (nm)	$\overline{R}$ (nm)
VKORC1	1	10
KIT+KID	0.94	30
KID	1	18
KD+KID	0.86	20
KD	0.94	15

TABLEAU 5.2 – Comparaison de R^* et $\overline{R}$ pour VKORC1, KIT+KID, KID, KD+KID et KD.

On peut alors raisonnablement envisager d'écrire une discrétisation de l'EDS pour les systèmes

ci-dessus. Il n'en demeure pas moins que l'identification des paramètres optimaux (m, ϵ) soit impossible à effectuer. En effet, ceux-ci étant issus du calcul des écarts types empiriques associés aux estimateurs, il est indispensable de connaître les véritables valeurs de μ et θ en tout point de l'espace conformationnel visité.

Pour contourner cet obstacle, l'idée consiste à créer un équivalent de la trajectoire étudiée dans le paysage des trois puits. En effet, dans ce dernier cadre, les termes de dérives et de diffusion réels sont connus, les écarts types empiriques calculables, et donc l'identification des paramètres optimaux rendue possible. En estimant alors la dérive et la diffusion pour ces paramètres pour un pas de temps équivalent à $(\Delta t)^*$, nous serions capables en nous basant sur la qualité de cette estimation dans le paysage des trois puits, de savoir si le pas de temps pivot serait suffisamment fin pour calculer directement ces estimées pour la trajectoire réelle, et si tel était le cas, de connaître la fiabilité de l'estimation.

Avant de passer aux applications, nous revenons ici sur un dernier point. Dans l'exemple que nous avons considéré pour le modèle des trois puits, la température équivalente est assez élevée de l'ordre de 0.5 $\mathrm{nm}^2 \cdot \mathrm{ns}^{-1}$ pour laquelle la mesure d'équilibre $\nu_{\theta_{eq}}$ est relativement peu creusée au voisinage du centre du puits. Ce ne sera plus le cas pour les modèles équivalents sur des trajectoires réelles pour lesquelles nous trouverons des températures deux fois plus petites créant des écarts de densité beaucoup plus importants entre le centre du puits et les bords. Or, si $\mathrm{err}(x)$ est l'erreur d'estimation (pour la dérive ou la température) en x, l'ergodicité locale du processus nous donne l'approximation :

$$\frac{1}{m}\sum_{k=1}^{m} \mathrm{err}(x_i)^2 \simeq \int \mathrm{err}(x)^2 \nu_{\theta_{eq}}(dx)$$

ce qui quantifie la représentation différentielle des différents points d'un puits dans la prise en compte de l'erreur. A basse température, l'accumulation des points au voisinage du fond du puits amène à privilégier un voisinage restreint autour du puits dans le calcul de l'erreur et un choix de ϵ petit provoquant des erreurs importantes d'estimation en bord de puits (mais qui ne sont très peu prise en compte dans l'erreur quadratique totale). Pour pallier ce défaut, nous introduisons une pondération $w(x) = \frac{d\nu_{\theta_{ref}}}{d\nu_{\theta_{eq}}}(x)$ (calculable sur le modèle des trois puits) pour laquelle on a l'approximation :

$$\frac{1}{m}\sum_{k=1}^{m} w(x_i)\mathrm{err}(x_i)^2 \simeq \int \mathrm{err}(x)^2 \nu_{\theta_{ref}}(dx)$$

ce qui permet de tempérer la température apparente pour la pondération sur les erreurs sur une valeur de référence que nous fixerons dans la suite à $\theta_{ref} = 1$.

Application à VKORC1

Appliquons cette méthode à la Trajectoire 1 de VKORC1, avec $\kappa_0 = 3419$, $\mathbf{D}_0 = 1000\ \mathrm{nm}^2{\cdot}\mathrm{ns}^{-1}$ et $\mathbf{R}_0 = 11$ nm. Nous calculons les écarts types empiriques associés aux estimateurs du premier ordre sur l'équivalent trois puits correspondant. Puis, pour $m = 200000$, nous calculons alors les écarts types associés aux estimateurs du second ordre.

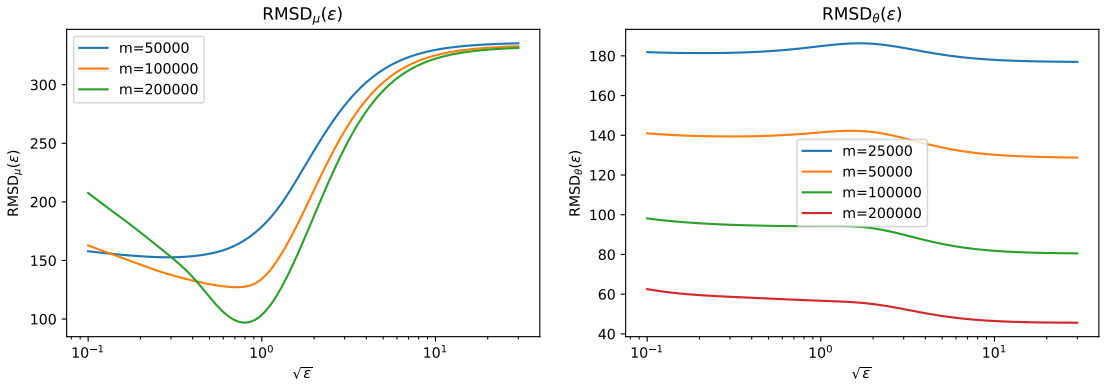

FIGURE 5.8 – Écarts types empiriques associés à l'estimateur $\hat{\mu}$ (gauche) et à l'estimateur $\hat{\theta}$ (droite) en fonction de $\sqrt{\epsilon}$ pour l'équivalent trois puits de la Trajectoire 1.

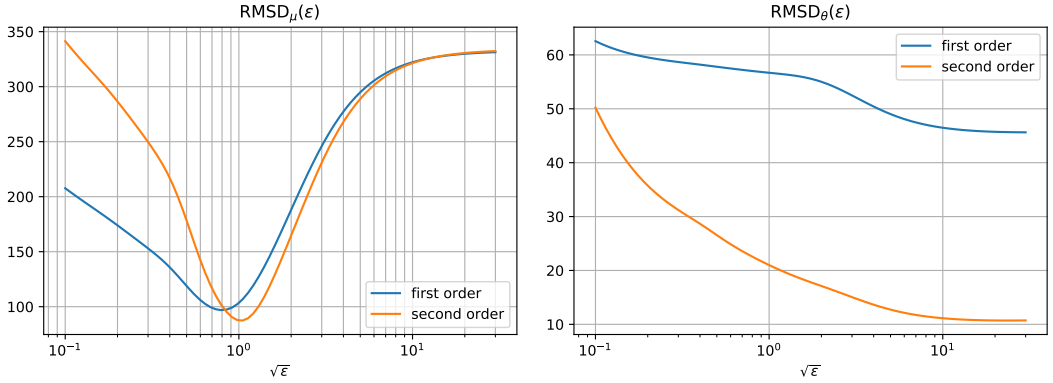

FIGURE 5.9 – Comparaison des performances entre les estimations au premier et au second ordre avec $m = 200000$, pour l'estimation de μ (gauche) et de θ (droite) sur l'équivalent trois puits de la Trajectoire 1.

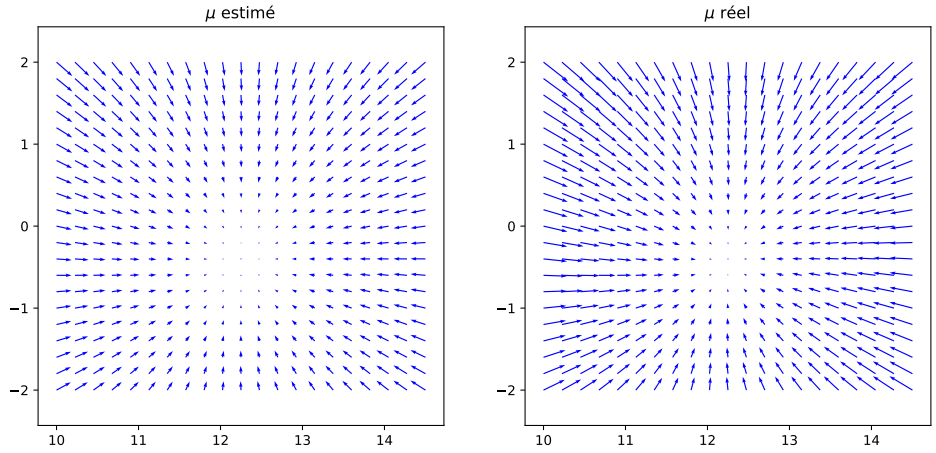

FIGURE 5.10 – Comparaison de l'estimation de μ (second ordre) avec μ réel pour l'équivalent trois puits de la Trajectoire 1, sur une grille localisée au voisinage du deuxième puits.

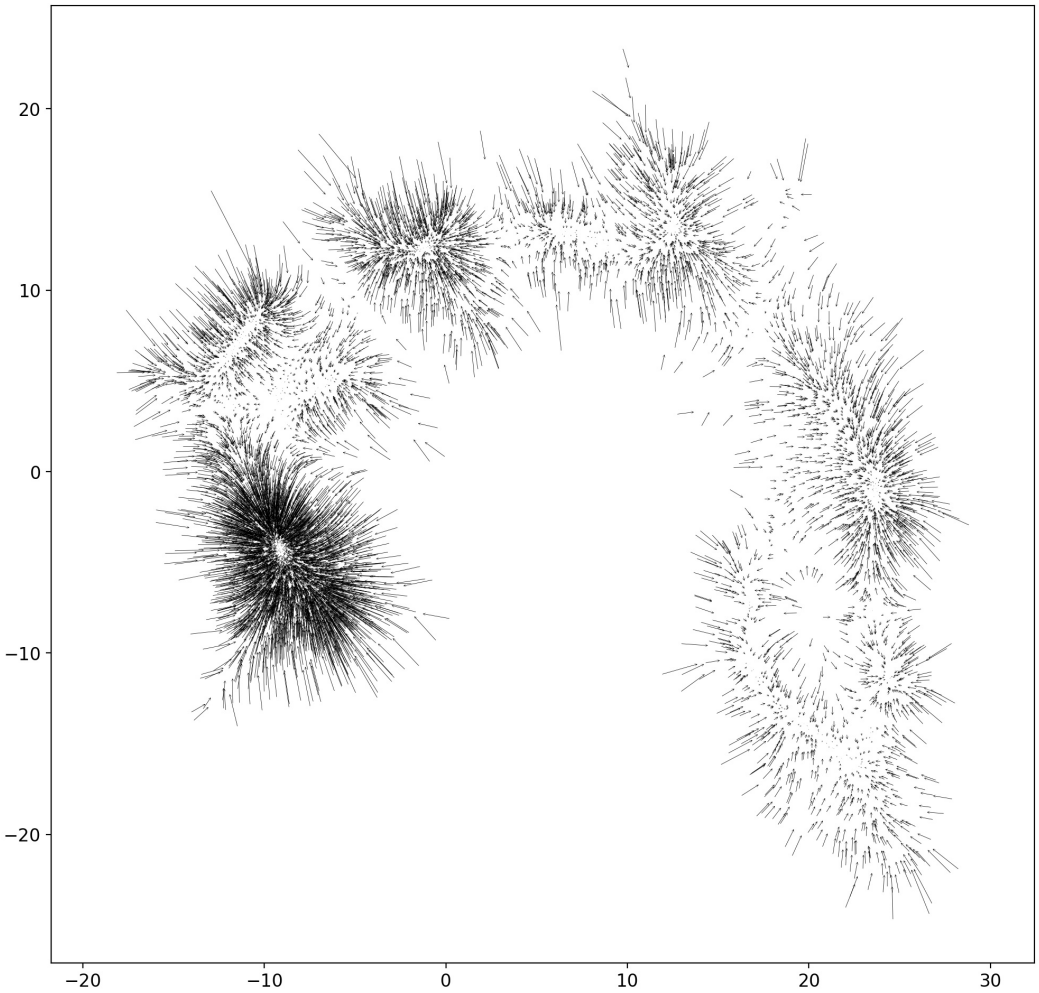

FIGURE 5.11 – Estimation de la dérive avec les paramètres optimaux $(m_\mu, \sqrt{\epsilon_\mu}) = (10^5, 1)$ pour la Trajectoire 1 (pour plus de lisibilité, la représentation est sous-échantillonnée $(1/15)$).

Remarque 36. *La trajectoire que nous utilisons ici n'est pas observée au pas* $(\Delta t)^*_{\text{VKOR}} = 1$ ps *mais au pas* $\Delta t = 5$ ps. *Nous constatons néanmoins d'une part que la discrétisation est toujours raisonnable du fait que* $\sqrt{\mathbf{D}_{\text{VKOR}} \times \Delta t} = 2.23 \text{ nm} \ll \overline{R}_{\text{VKORC1}} = 10 \text{ nm}$, *et d'autre part nous supposons que le pas d'échantillonnage n'influe guère sur les paramètres optimaux. Il en sera de même pour les trajectoires suivantes.*

Appliquons cette méthode à la Trajectoire 2 de VKORC1, avec $\kappa_0 = 4745$, $\mathbf{D}_0 = 1000 \text{ nm}^2 \cdot \text{ns}^{-1}$ et $\mathbf{R}_0 = 11$ nm. Nous calculons les écarts types empiriques associés aux estimateurs du premier ordre sur l'équivalent trois puits correspondant. Puis, pour $m = 200000$, nous calculons alors les écarts types associés aux estimateurs du second ordre :

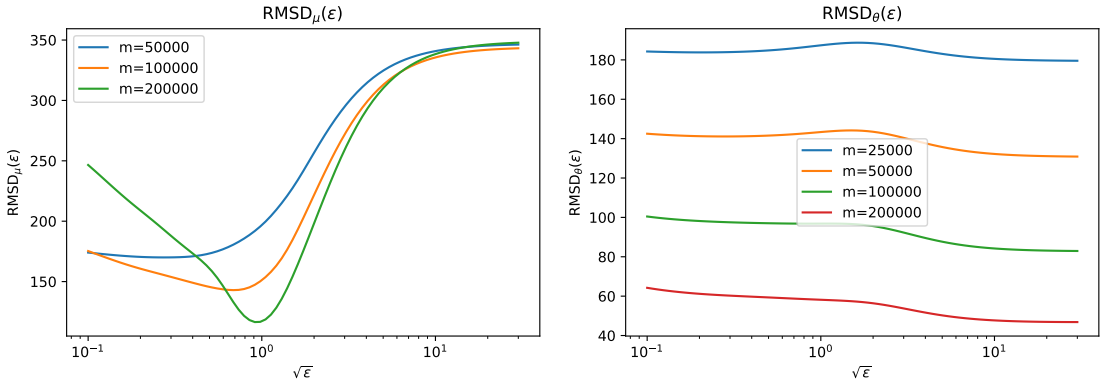

FIGURE 5.12 – Écarts types empiriques associés à l'estimateur $\hat{\mu}$ (gauche) et à l'estimateur $\hat{\theta}$ (droite) en fonction de $\sqrt{\epsilon}$ pour l'équivalent trois puits de la Trajectoire 2.

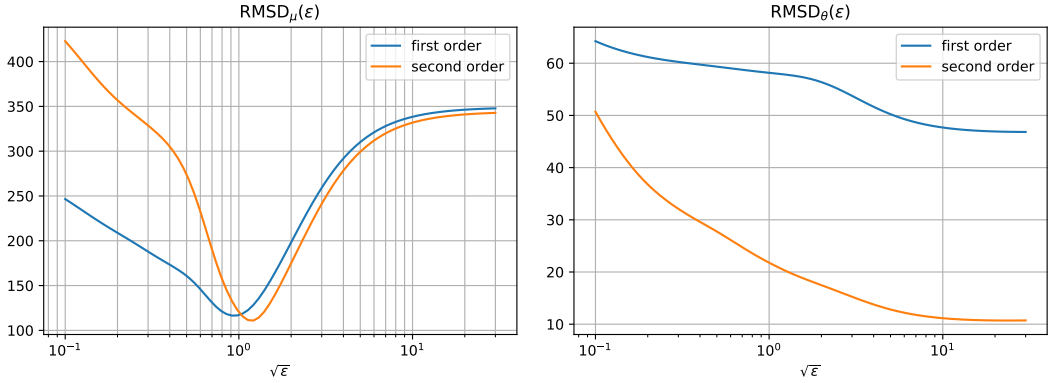

FIGURE 5.13 – Comparaison des performances entre les estimations au premier et au second ordre pour l'estimation de μ (gauche) de θ (droite), avec $m = 200000$, sur l'équivalent trois puits de la Trajectoire 2.

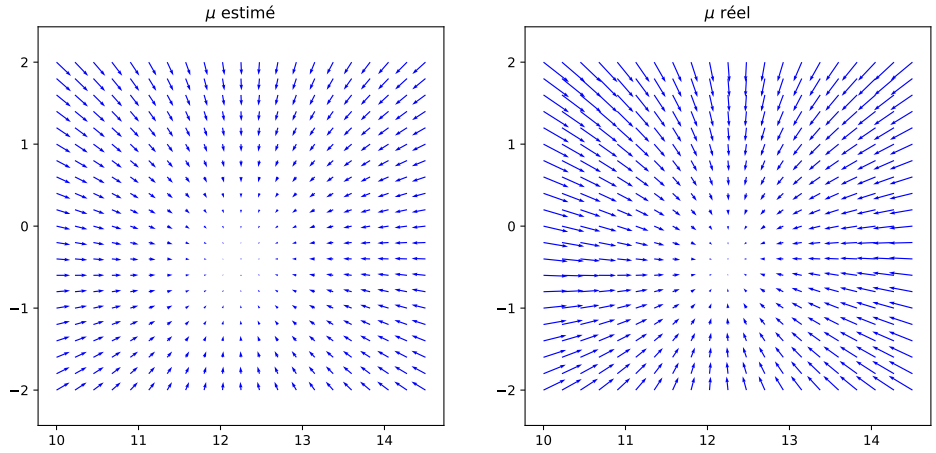

FIGURE 5.14 – Comparaison de l'estimation de μ (second ordre) avec μ réel pour l'équivalent trois puits de la Trajectoire 2, sur une grille localisée au voisinage du deuxième puits.

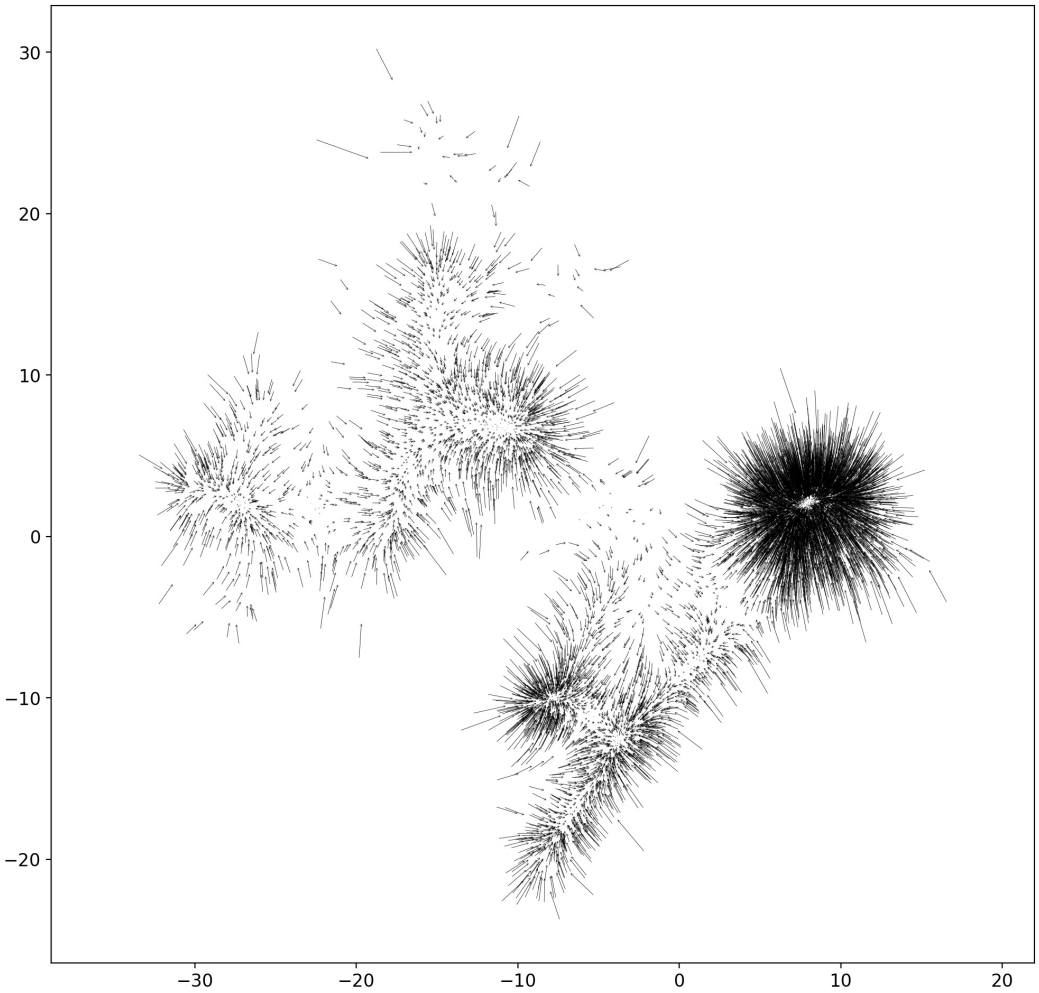

FIGURE 5.15 – Estimation de la dérive avec les paramètres optimaux $(m_\mu, \sqrt{\epsilon_\mu}) = (10^5, 1.2)$ pour la Trajectoire 2 (pour plus de lisibilité, la représentation est sous-échantillonnée (1/15)).

Application à KIT+KID

Construisons alors un équivalent de la trajectoire KIT+KID de 2 µs. Avec les notations précédentes, nous avons $\kappa_0 = 643$, $\mathbf{D}_0 = 900$ nm^2·ns^{-1} et $\mathbf{R}_0 = 14$ nm. Nous calculons les écarts types empiriques associés aux estimateurs du premier ordre sur l'équivalent trois puits correspondant. Puis, pour $m = 200000$, nous calculons alors les écarts types associés aux estimateurs du second ordre :

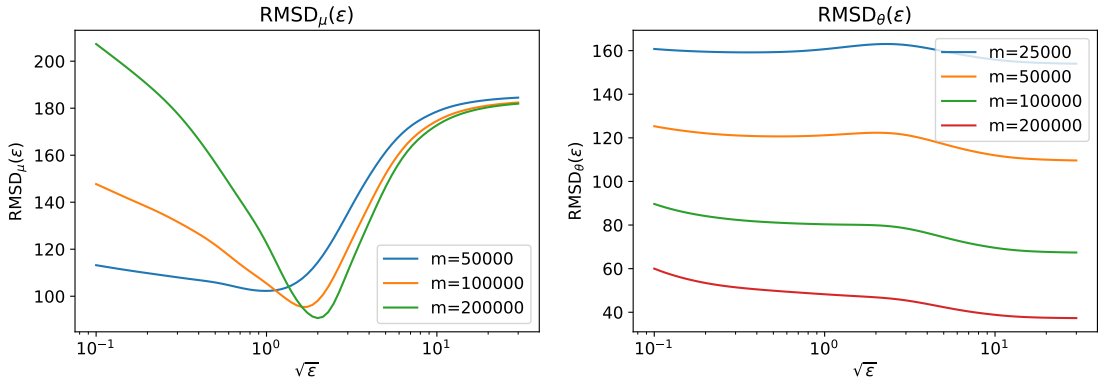

FIGURE 5.16 – Écarts types empiriques associés à l'estimateur $\hat{\mu}$ (gauche) et à l'estimateur $\hat{\theta}$ (droite) en fonction de$\sqrt{\epsilon}$ pour l'équivalent trois puits de la Trajectoire KIT+KID.

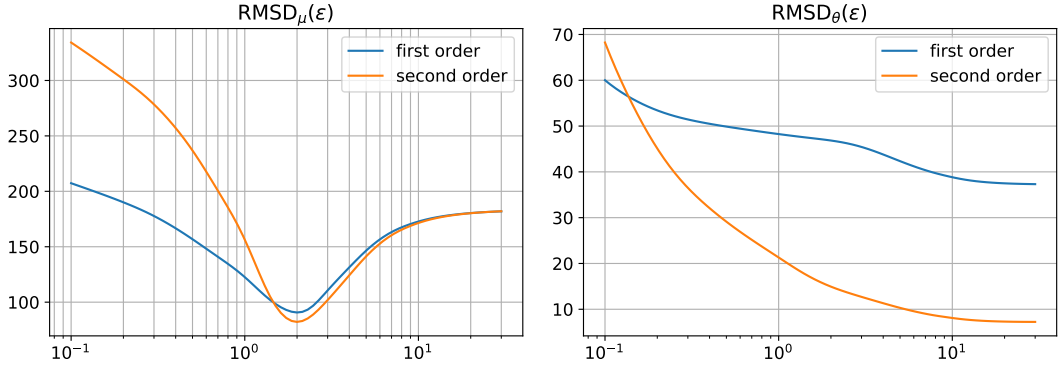

FIGURE 5.17 – Comparaison des performances entre les estimations au premier et au second ordre pour l'estimation de μ (gauche) et de θ (droite), avec $m = 200000$ sur l'équivalent trois puits de la Trajectoire KIT+KID.

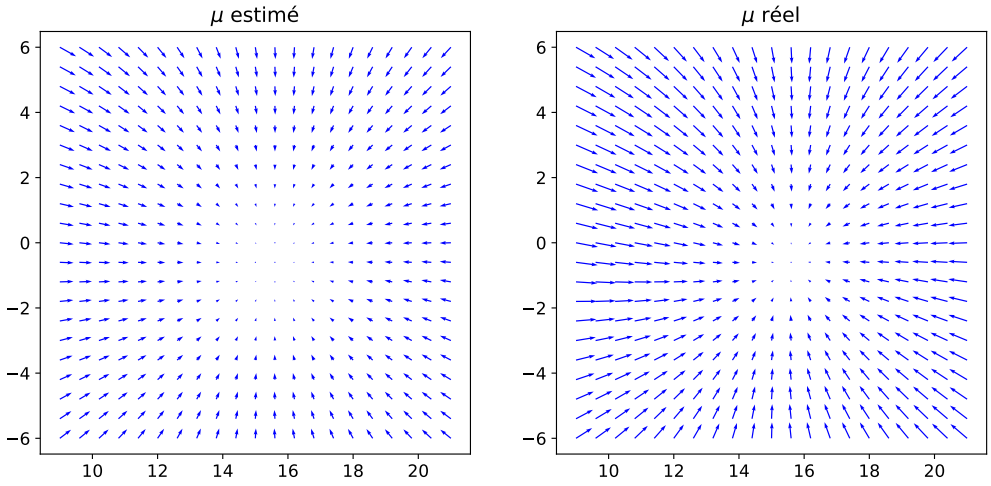

FIGURE 5.18 – Comparaison de l'estimation de μ (second ordre, à gauche) avec μ réel sur une grille localisée au voisinage du deuxième puits pour l'équivalent trois puits de la Trajectoire KIT+KID.

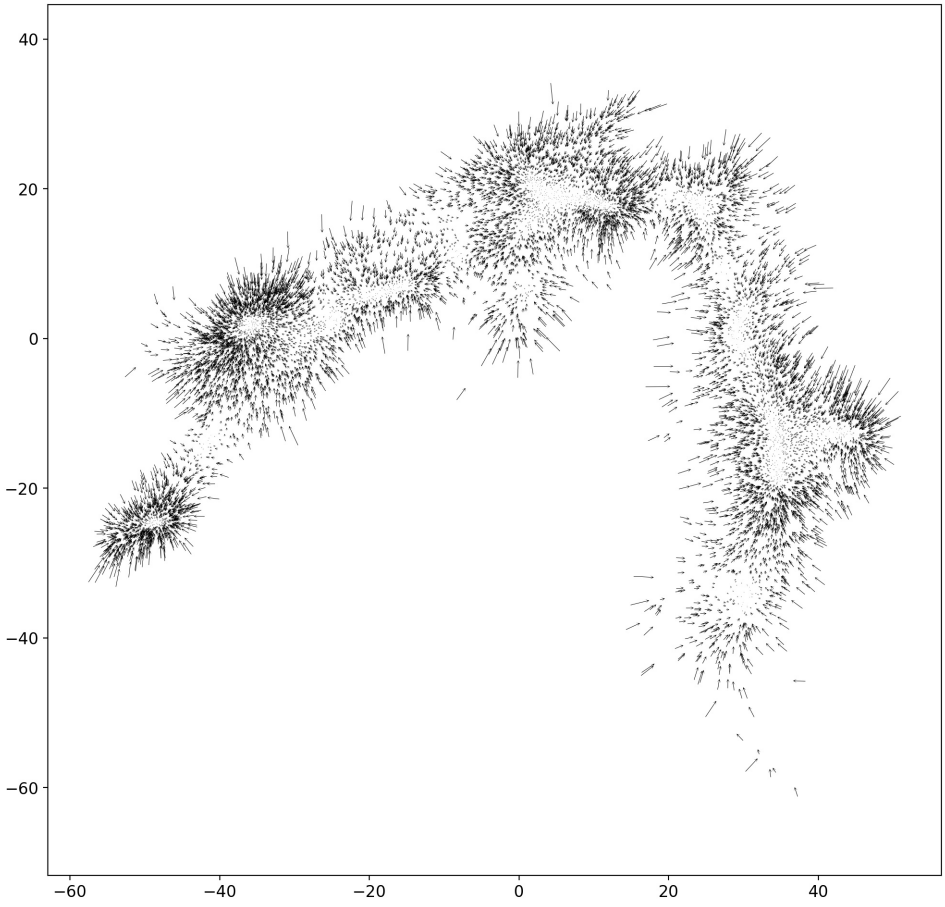

FIGURE 5.19 – Estimation de la dérive avec les paramètres optimaux $(m_\mu, \sqrt{\epsilon_\mu}) = (10^5, 2)$ pour la Trajectoire KIT+KID (pour plus de lisibilité, la représentation est sous-échantillonnée sur 5000 points.

Application à KID

Appliquons cette même méthode à KID seul avec $\kappa_0 = 2658$, $\mathbf{D}_0 = 1000$ nm^2·ns^{-1} et $\mathbf{R}_0 = 12$ nm. Nous calculons les écarts types empiriques associés aux estimateurs du premier ordre sur l'équivalent trois puits correspondant. Pour $m = 200000$, nous calculons alors les écarts types associés aux estimateurs du second ordre :

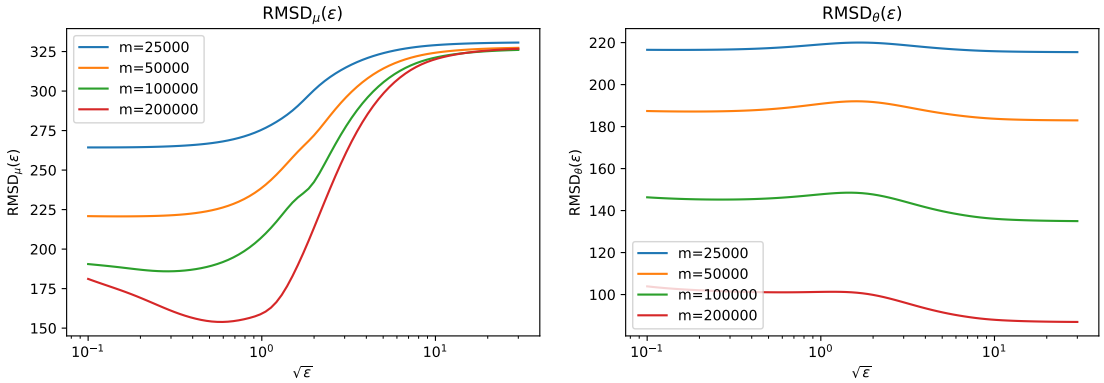

FIGURE 5.20 – Écarts types empiriques associés à l'estimateur $\hat{\mu}$ (gauche) et à l'estimateur $\hat{\theta}$ (droite) en fonction de $\sqrt{\epsilon}$ pour l'équivalent trois puits de la Trajectoire KID.

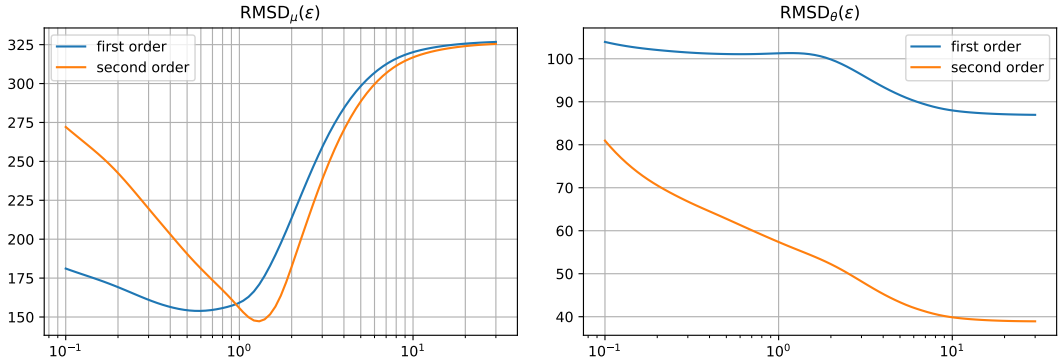

FIGURE 5.21 – Comparaison des performances entre les estimations au premier et au second ordre : pour l'estimation de μ (gauche) et pour l'estimation de θ (droite), avec $m = 200000$ sur l'équivalent trois puits de la Trajectoire KID.

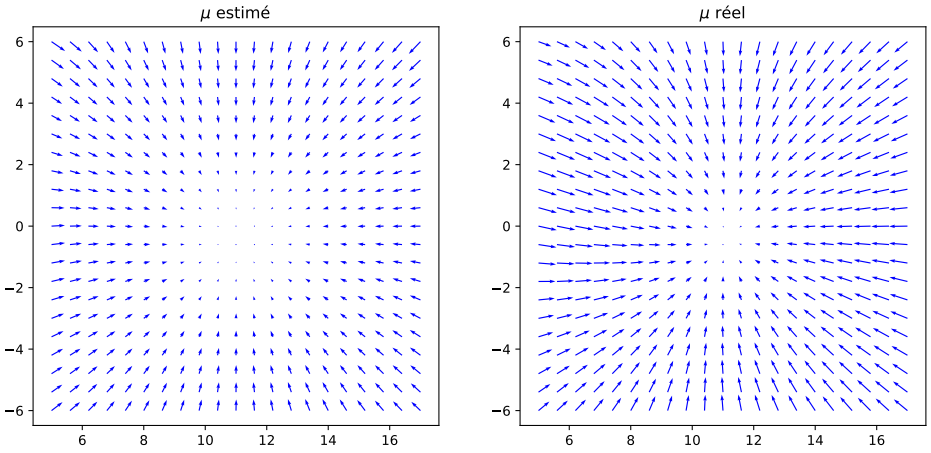

FIGURE 5.22 – Comparaison de l'estimation de μ (second ordre) avec μ réel pour l'équivalent trois puits de la Trajectoire KID, sur une grille localisée au voisinage du deuxième puits.

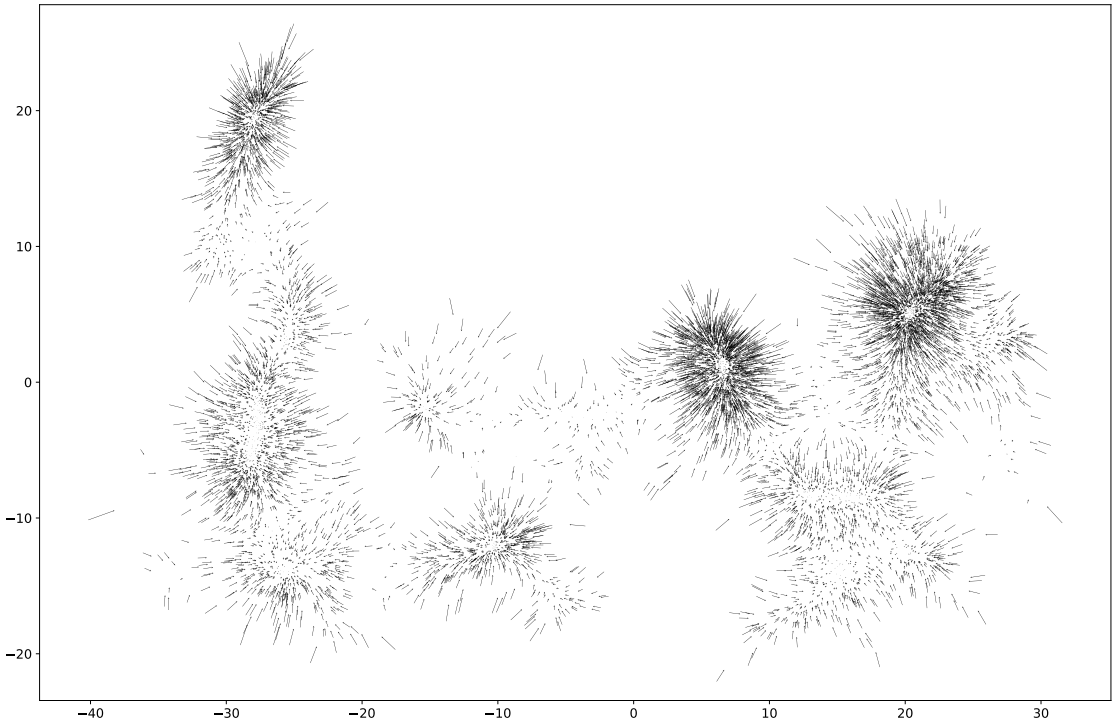

FIGURE 5.23 – Estimation de la dérive avec les paramètres optimaux $(m_\mu, \sqrt{\epsilon_\mu}) = (10^5, 1.4)$ pour la Trajectoire KID (pour plus de lisibilité, la représentation est sous-échantillonnée (1/15)).

Application à KD+KID

Appliquons cette même méthode à KD+KID avec $\kappa_0 = 2658$, $\mathbf{D}_0 = 1000$ nm^2·ns^{-1} et $\mathbf{R}_0 = 34$ nm. Nous calculons les écarts types empiriques associés aux estimateurs du premier ordre sur l'équivalent trois puits correspondant. Pour $m = 200000$, nous calculons alors les écarts types associés aux estimateurs du second ordre :

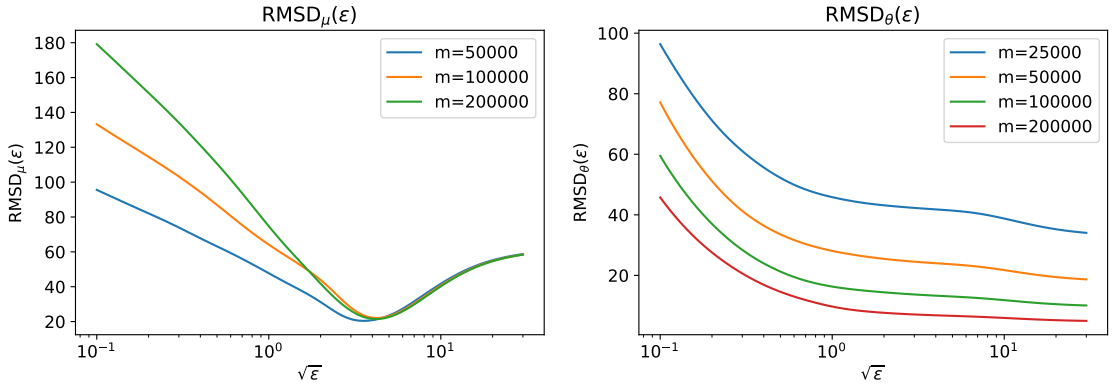

FIGURE 5.24 – Écarts types empiriques associés à l'estimateur $\hat{\mu}$ (gauche) et à l'estimateur $\hat{\theta}$ (droite) en fonction de$\sqrt{\epsilon}$ pour l'équivalent trois puits de la Trajectoire KD+KID.

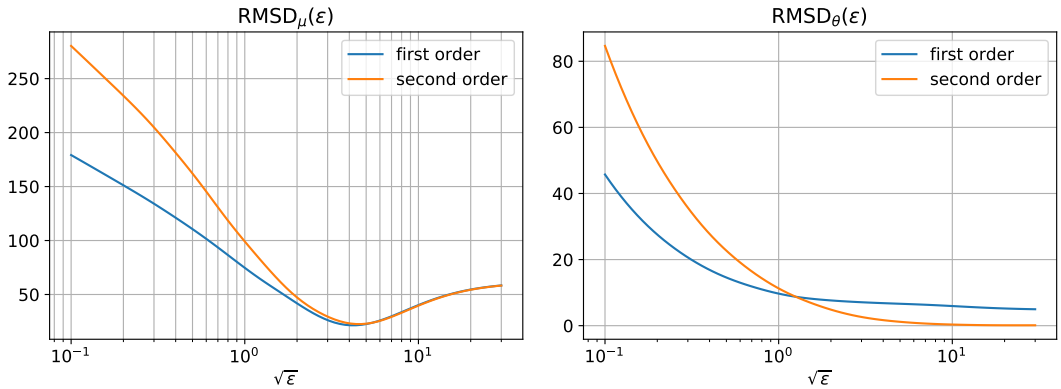

FIGURE 5.25 – Comparaison des performances entre les estimations au premier et au second ordre : pour l'estimation de μ (gauche) et pour l'estimation de θ (droite), avec $m = 200000$ sur l'équivalent trois puits de la Trajectoire KD+KID.

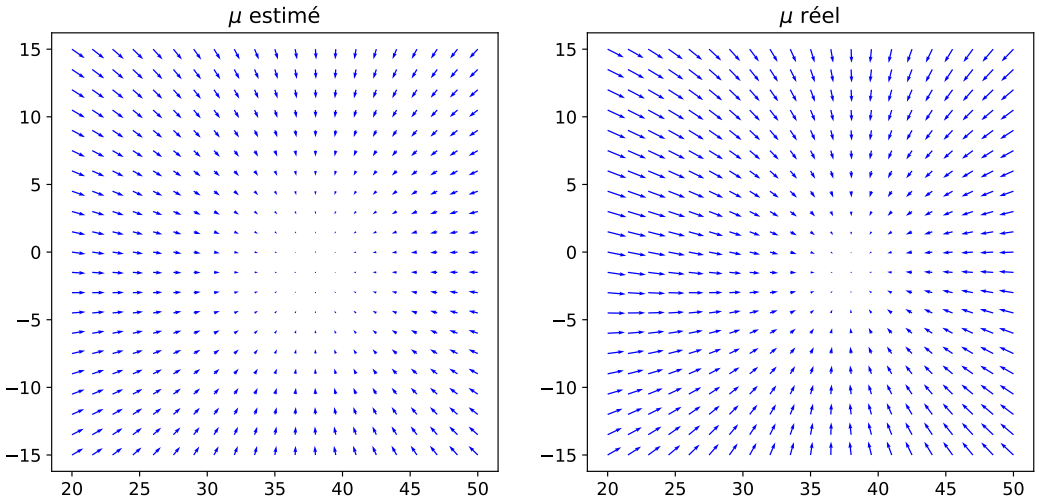

FIGURE 5.26 – Comparaison de l'estimation de μ (second ordre) avec μ réel pour l'équivalent trois puits de la Trajectoire KD+KID, sur une grille localisée au voisinage du deuxième puits.

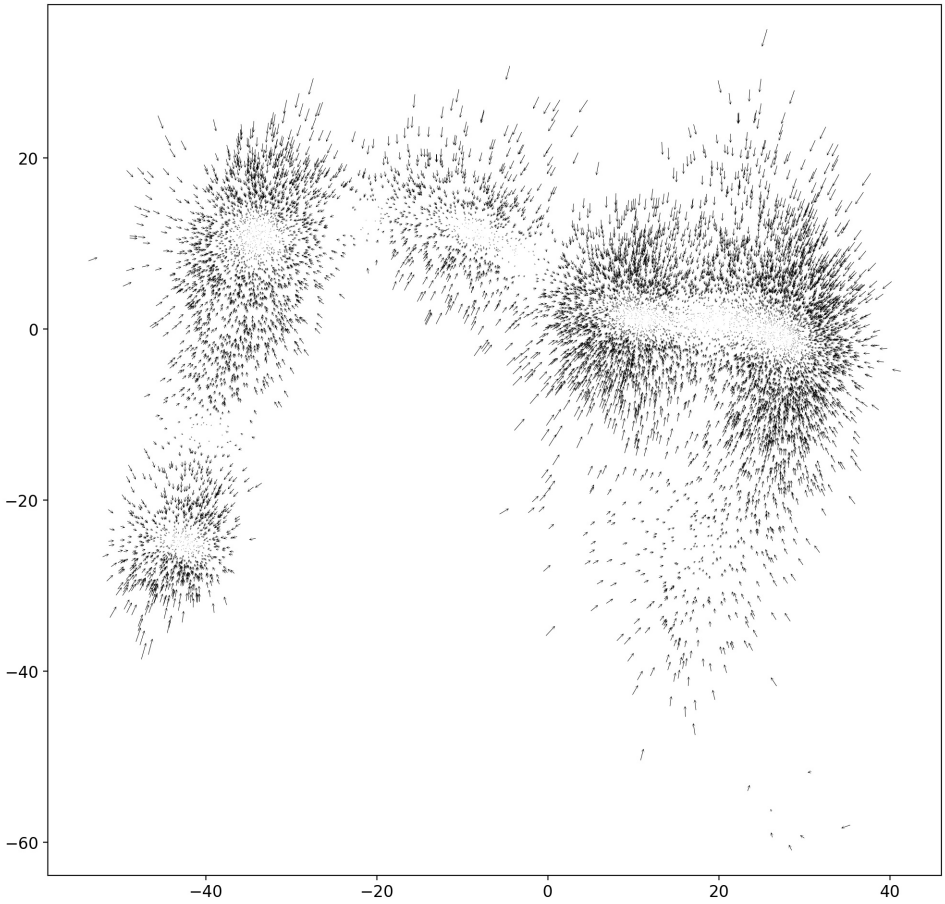

FIGURE 5.27 – Estimation de la dérive avec les paramètres optimaux $(m_\mu, \sqrt{\epsilon_\mu}) = (10^5, 4.5)$ pour la Trajectoire KD+KID (pour plus de lisibilité, la représentation est sous-échantillonnée $(1/15)$).

Application à KD seul

Appliquons cette même méthode à KD seul $\kappa_0 = 2658$, $\mathbf{D}_0 = 1000$ nm^2·ns^{-1} et $\mathbf{R}_0 = 18$ nm. Nous calculons les écarts types empiriques associés aux estimateurs du premier ordre sur l'équivalent trois puits correspondant. Puis, pour $m = 200000$, nous calculons les écarts types associés aux estimateurs du second ordre :

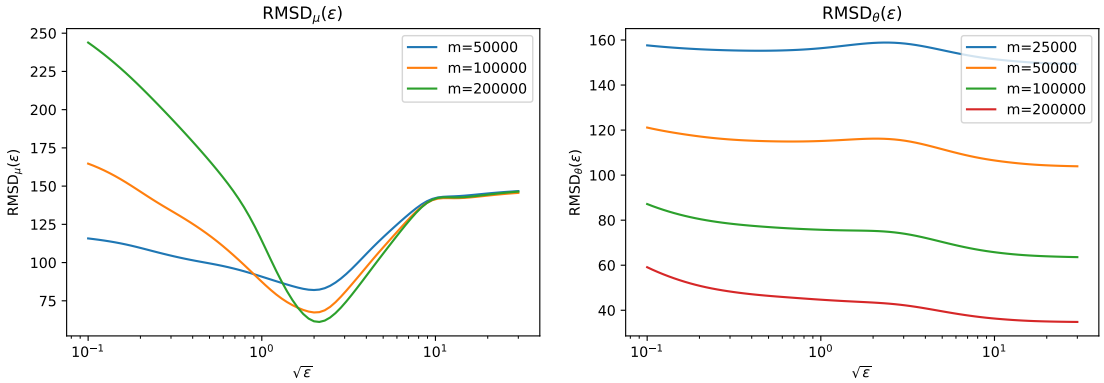

FIGURE 5.28 – Écarts types empiriques associés à l'estimateur $\hat{\mu}$ (gauche) et à l'estimateur $\hat{\theta}$ (droite) en fonction de $\sqrt{\epsilon}$ pour l'équivalent trois puits de la Trajectoire KD.

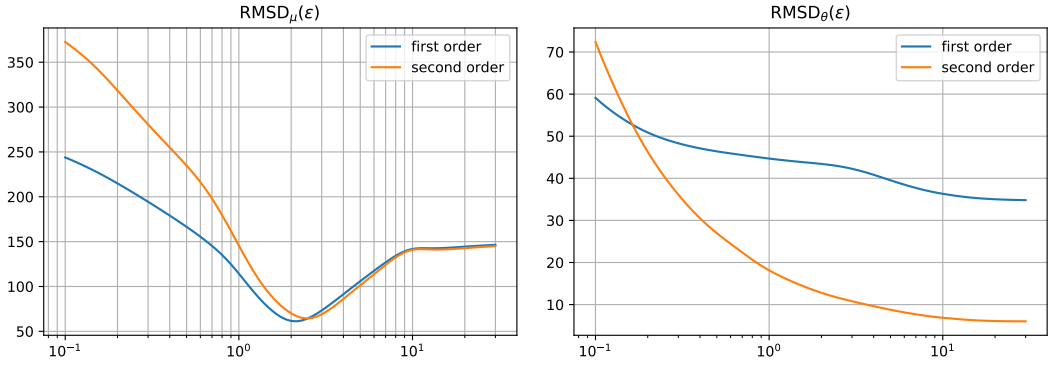

FIGURE 5.29 – Comparaison des performances entre les estimations au premier et au second ordre : pour l'estimation de μ (gauche) et pour l'estimation de θ (droite), avec $m = 200000$ sur l'équivalent trois puits de la Trajectoire KD.

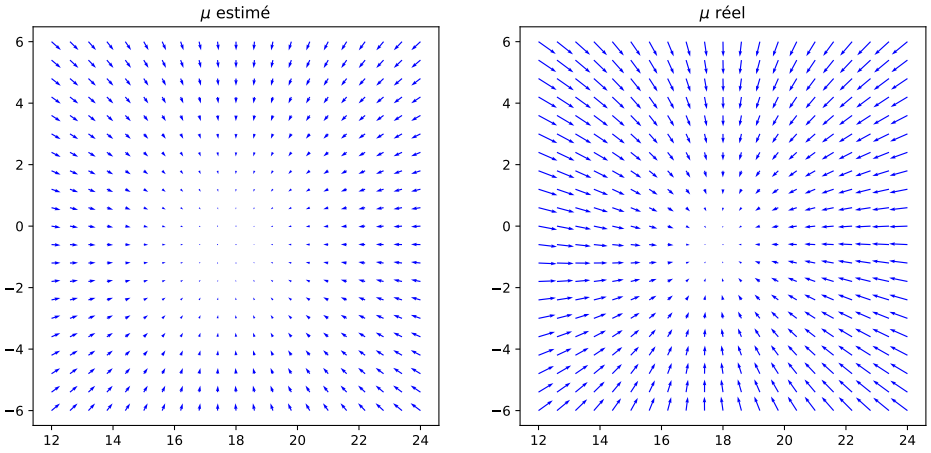

FIGURE 5.30 – Comparaison de l'estimation de μ (second ordre) avec μ réel pour l'équivalent trois puits de la Trajectoire KD, sur une grille localisée au voisinage du deuxième puits.

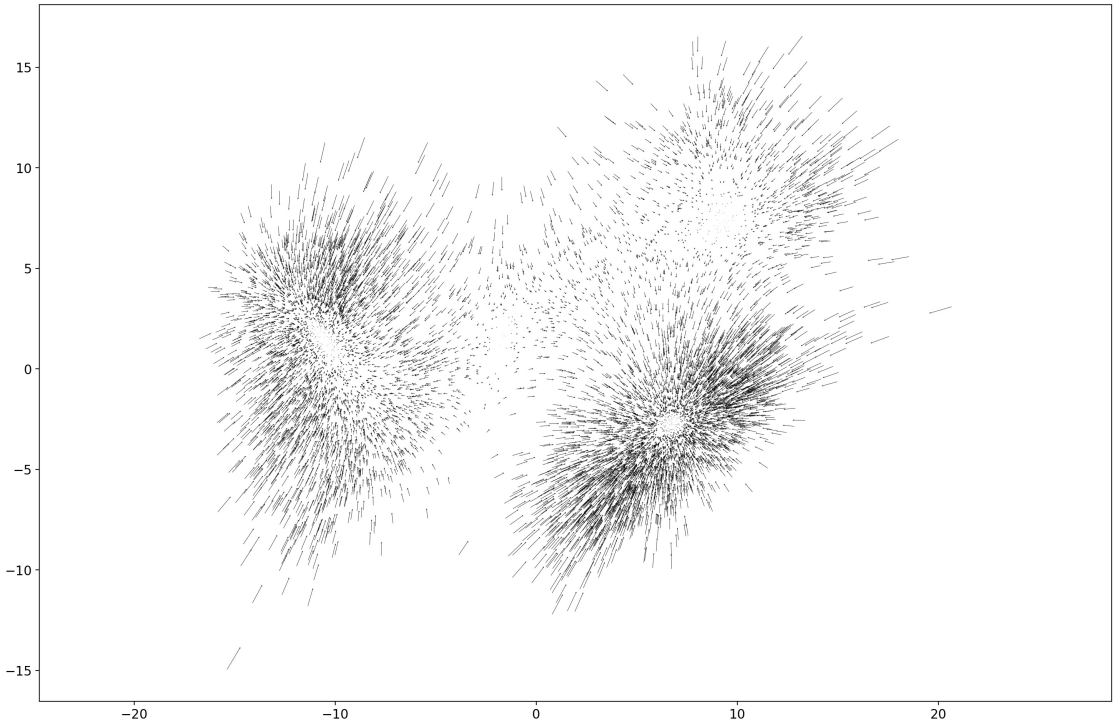

FIGURE 5.31 – Estimation de la dérive avec les paramètres optimaux $(m_\mu, \sqrt{\epsilon_\mu}) = (10^5, 1.6)$ pour la Trajectoire KD (pour plus de lisibilité, la représentation est sous-échantillonnée $(1/15)$).

5.3 Comparaison avec le gradient de l'énergie estimée

Nous souhaitons confronter ici les résultats précédemment obtenus de l'estimation du terme de dérive $\mu(x)$ dans le cadre d'une modélisation par EDS, avec le gradient d'une fonction d'énergie estimée à partir des données.

5.3.1 Principe

Rappelons que la trajectoire étudiée est notée $\mathbf{X} = (x_i)_{1\leq i\leq N}$, avec $x_i \in \mathbb{R}^d$. En s'inspirant de la notion de similarité telle qu'elle a été décrite précédemment (voir Chapitre 2 §2.3.1), on commence par définir l'estimateur empirique de la densité, défini sur tout l'espace conformationnel $\mathcal{X}$ par :

$$\hat{\pi}(x) = \frac{1}{N}\sum_{i=1}^{N} \exp(-\|x - x_i\|/2\epsilon)$$

Supposons alors que cette densité de probabilité soit reliée à une fonction d'énergie notée U selon une distribution de Maxwell-Boltzmann : $\pi \propto \exp(-\beta U)$, avec $\beta = 1/k_B\theta$ où θ est la température du système et k_B la constante de Boltzmann. On a ainsi que :

$$\widehat{U} = -\frac{1}{\beta}\log(\hat{\pi}) + \text{Cte} \tag{5.9}$$

5.3.2 Application aux données de DM

Pour une trajectoire de simulation de DM donnée, nous proposons alors de calculer les valeurs de $-\nabla\widehat{U}$ en les x_i, et de les comparer aux valeurs de $\hat{\mu}(x_i)$ calculées précédemment. Dans ce qui suit, nous présentons pour chaque système étudié le paysage d'énergie estimé selon (5.9), le gradient correspondant calculé sur une fenêtre du plan, et les $\hat{\mu}$ calculés sur cette même fenêtre.

Application à VKORC1

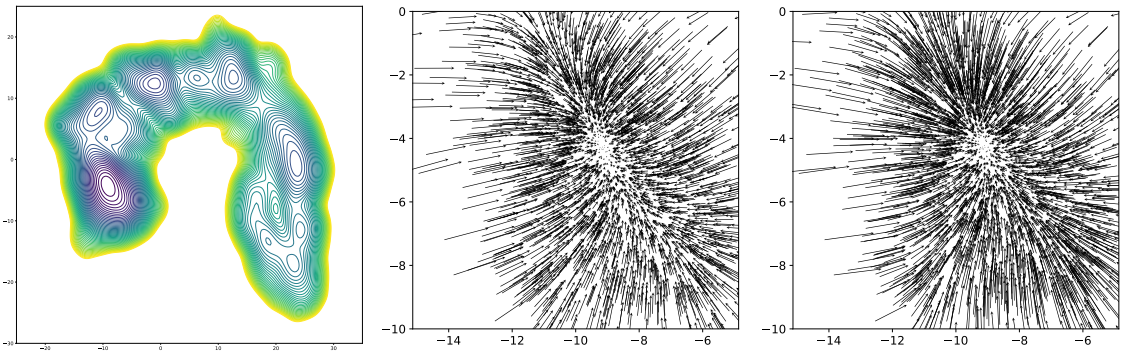

FIGURE 5.32 – Résultats pour la Trajectoire 1 de VKORC1 : lignes de niveaux de l'énergie estimée $\widehat{U}$ (gauche) et comparaison du gradient de l'énergie estimée $-\nabla\widehat{U}$ (centre) avec la dérive estimée $\hat{\mu}$ sur la restriction du plan à une fenêtre localisée au niveau du dernier puits (voir Fig. 4.3, Well no. 1, $\kappa = 3419$).

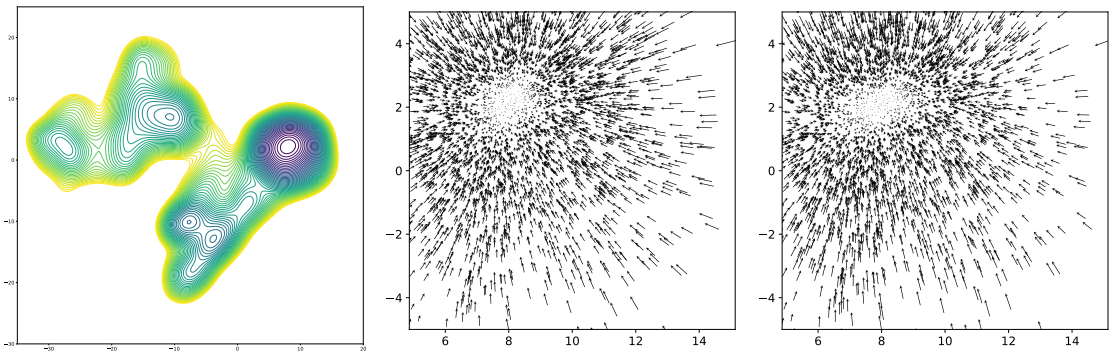

FIGURE 5.33 – Résultats pour la Trajectoire 2 de VKORC1 : lignes de niveaux de l'énergie estimée $\widehat{U}$ (gauche) et comparaison du gradient de l'énergie estimée $-\nabla\widehat{U}$ (centre) avec la dérive estimée $\hat{\mu}$ sur la restriction du plan à une fenêtre localisée au niveau du dernier puits (voir Fig. 4.7, Well no. 4, $\kappa = 4745$).

Application à KIT avec KID

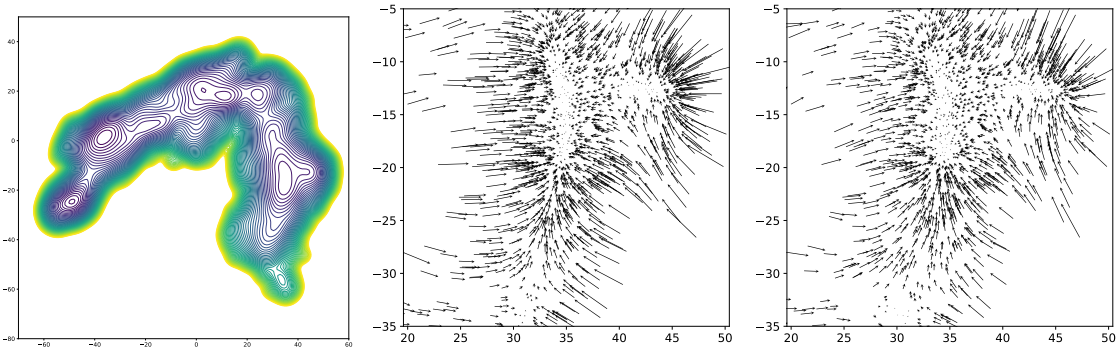

FIGURE 5.34 – Résultats pour la Trajectoire KIT+KID : lignes de niveaux de l'énergie estimée $\widehat{U}$ (gauche) et comparaison du gradient de l'énergie estimée $-\nabla\widehat{U}$ (centre) avec la dérive estimée $\hat{\mu}$ sur la restriction du plan à une fenêtre localisée au niveau du premier puits (voir Fig. 4.13, Well no. 1, $\kappa = 376$).

Application à KID seul

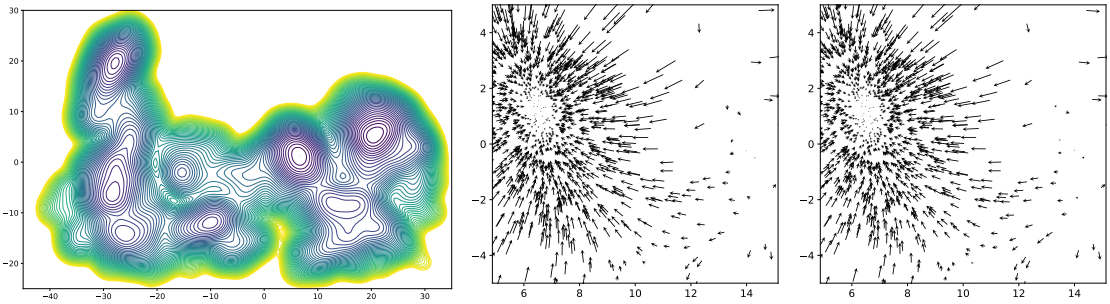

FIGURE 5.35 – Résultats pour la Trajectoire KID : lignes de niveaux de l'énergie estimée $\widehat{U}$ (haut, gauche) et comparaison du gradient de l'énergie estimée $-\nabla\widehat{U}$ (haut, droite) avec la dérive estimée $\hat{\mu}$ (bas) sur la restriction du plan à une fenêtre localisée au niveau du troisième puits (voir Fig. 4.17, Well no. 3, $\kappa = 2658$).

Application à KD avec KID

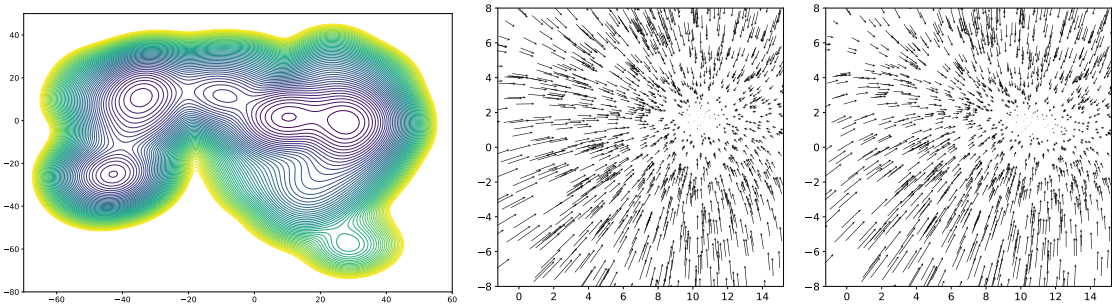

FIGURE 5.36 – Résultats pour la Trajectoire KD+KID : lignes de niveaux de l'énergie estimée $\widehat{U}$ (gauche) et comparaison du gradient de l'énergie estimée $-\nabla\widehat{U}$ (centre) avec la dérive estimée $\hat{\mu}$ sur la restriction du plan à une fenêtre localisée au niveau du second puits (voir Fig. 4.21, Well no. 2, $\kappa = 389$).

Application à KD seul

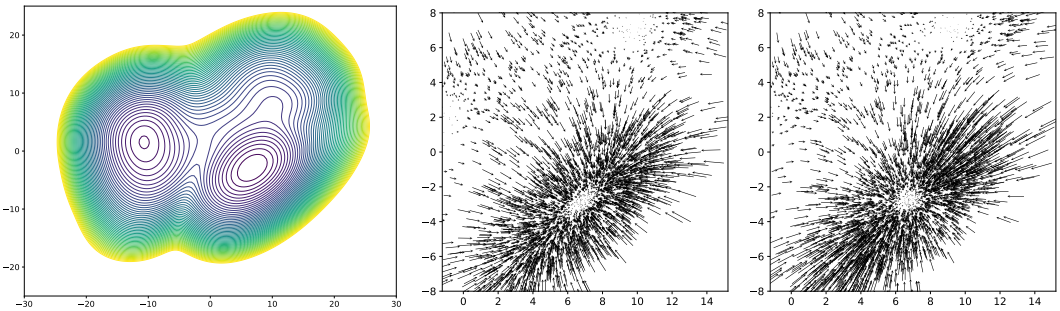

FIGURE 5.37 – Résultats pour la Trajectoire KD : lignes de niveaux de l'énergie estimée $\widehat{U}$ (gauche) et comparaison du gradient de l'énergie estimée $-\nabla\widehat{U}$ (centre) avec la dérive estimée $\hat{\mu}$ sur la restriction du plan à une fenêtre localisée au niveau du premier puits (voir Fig. 4.25, Well no. 1, $\kappa = 3007$).

5.3.3 Bilan

On constate alors la très forte adéquation entre d'une part le gradient de l'énergie estimée $-\nabla\widehat{U}$ et d'autre part la dérive estimée $\hat{\mu}$. Cela nous invite à penser que l'on pourrait écrire $\mu = -\nabla U$, autrement dit, que les trajectoires observées pourraient raisonnablement être modélisées par des équation de type Langevin suramorties, faisant ainsi correspondre le terme d'énergie U à une notion d'énergie libre. Il s'agit d'un constat remarquable en ce sens qu'en général, à partir d'une trajectoire quelconque, il se peut tout à fait que ces deux quantités diffèrent (voir Chapitre 2 §2.4).

On constate par ailleurs que l'on retrouve raisonnablement les équilibres locaux mis en évidence par la κ-segmentation au chapitre précédent. En effet, les centres des puits mis en évidence par la segmentation comme des conformations à l'origine d'une maximisation locale de κ en sortie de puits, correspondent aux minimums locaux de l'énergie estimée empiriquement. A ce stade, les trois approches évoquées (κ-segmentation, estimation de la dérive, et calcul de l'énergie estimée) constituent trois points de vues complémentaires permettant d'acquérir des informations sur la trajectoire quant à la nature de sa dynamique et aux équilibrations locales qu'elle présente.

Chapitre 6

Conclusions et perspectives

6.1 Conclusions

6.1.1 Modèle de Langevin

Nous avons abordé dans un premier temps la question de la modélisation des trajectoires de simulations de DM. L'étude du cadre conceptuel de notre travail nous a naturellement mené, moyennant quelques hypothèses, à la considération d'un modèle de Langevin suramorti, cas particulier d'une modélisation stochastique, dont le terme de dérive s'exprime comme le gradient d'un potentiel d'énergie libre (Chapitre 2 §2.1). En cherchant à retrouver les traces de ce comportement à partir des données en elles-mêmes, et plus précisément via le calcul de la variation quadratique associée à chaque trajectoire de simulation de DM, nous avons mis en évidence l'existence d'un pas d'échantillonnage pivot $(\Delta t)^*$ au delà duquel nous pouvions considérer les données comme étant des observations discrètes d'une EDS (Chapitre 2 §2.2).

En nous penchant sur la considération d'un modèle markovien sur nos données, nous avons vu que celui-ci ne pouvait avoir de sens que si le processus sous-jacent répondait à une équation de type Langevin en ce sens qu'à la limite, la chaîne de markov construite sur les données reproduisait, à un facteur de température près, le générateur infinitésimal d'une équation de Langevin (Chapitre 2 §2.3) : c'est dans ce cas bien précis qu'une méthode de re-simulation des trajectoires observées à pu être envisagée.

Finalement, bien qu'une modélisation par une équation de Langevin apparaisse comme un cadre naturel pour notre étude, il est indispensable de passer par une étude de la nature du terme de dérive μ afin de prétendre à un tel modèle. C'est ainsi que, par le truchement d'un modèle équivalent des trajectoires de simulation de DM au sein du paysage des trois puits, nous avons pu établir une méthode d'estimation des paramètres de dérive et de diffusion dans le cadre d'un modèle d'EDS (Chapitre 5 §5.2), puis comparer d'une part le terme de dérive estimé $\hat{\mu}$ partant d'un modèle général d'EDS, avec d'autre part le gradient de l'énergie estimée $\nabla\widehat{U}$ à partir d'un estimateur empirique de la densité (Chapitre 5 §5.3). La forte adéquation entre les profils observés, qui rappelons-le ne peut être anodine du fait que ces deux quantités ne coïncident pas dans tous les cas, constitue un premier pas vers la confirmation d'un modèle de Langevin.

6.1.2 Algorithme de κ-segmentation

Nous avons construit un nouveau critère permettant de quantifier la stabilité des puits, le *nombre de tours* κ (Chapitre 3 §3.1). Sans dimension et invariante par changements d'échelles de temps et d'espace, cette quantité effectue le quotient de la distance parcourue par le processus alors modélisé par une EDS entre des instants s et t dans le cas d'une pure diffusion (sans dérive), par le rayon maximum effectivement parcouru par la trajectoire (avec dérive) entre ces mêmes instants. Outre des propriétés asymptotiques remarquables, cette variable met à disposition des ordres de grandeurs qui permettent non seulement de comparer les puits entre eux, mais également de juger de leur pertinence d'une manière absolue.

La construction d'un algorithme de segmentation des trajectoires de simulation de DM permet alors, à partir de la lecture automatique d'une matrice du nombre de tours régulée par différents critères de rejets, d'exhiber différents puits et d'en fournir certaines caractéristiques (valeur de κ associée, rayon, *frames* de début et de fin, temps de sortie)(Chapitre 3 §3.2).

L'application de cet algorithme aux données réelles de VKORC1 et KIT offre ainsi une lecture des trajectoires beaucoup plus fournie que ne le permettraient les méthodes classiques, en prenant en compte de surcroît l'information dynamique que contiennent les trajectoires (Chapitre 4). En effet, en plus d'enrichir considérablement le niveau de compréhension des phénomènes de stabilisations locales par rapport à une approche de type RMSD, cette méthode évite l'écueil des algorithmes de clustering, consistant à faire émerger de faux puits.

Qui plus est, on constate que par de simples considérations dynamiques, la segmentation parvient à retrouver des centres de puits qui correspondent à des minimums locaux de l'énergie estimée empiriquement. La κ-segmentation enrichit néanmoins considérablement cette approche en ce sens qu'elle ne repose pas sur l'hypothèse d'un système à l'équilibre et qu'elle tient compte de la dynamique qui régit les données.

6.2 Perspectives

6.2.1 Aspect multi-échelle

Un développement complémentaire consisterait à offrir à l'utilisateur de la κ-segmentation un point de vue multi-échelle. Plus précisément, les puits exhibés par la segmentation peuvent eux-mêmes faire l'objet d'une nouvelle segmentation invoquant de nouveaux paramètres : de cette façon, l'utilisateur aura une lecture hiérarchique des différents sous-puits de chaque puits. En effet, la réalité des puits est en pratique certainement très éloignée d'un puits "lisse" comme on pourrait se le figurer. Cet aspect multi-échelle est palpable au réglage des paramètres de l'algorithme. Bien souvent, il faut considérer des plages de conformations initiales assez grandes pour observer une sortie de puits : il faut laisser au processus le temps de sortir des différents sous-puits présents. Par ailleurs, nous voyons que différentes valeurs de K (taille de la zone post-maximum local de κ à étudier pour détecter une éventuelle fausse sortie) peuvent faire apparaître des sous-puits d'un même puits. Il serait ainsi bon de permettre à l'utilisateur, après une première segmentation effectuée, de sélectionner certains segments qui lui apparaîtront comme dignes d'intérêt, et que ceux-ci fassent l'objet d'une segmentation, et éventuellement d'une ACP locale avant d'appliquer l'algorithme.

Cependant, la légitimité d'une telle approche doit être questionnée au regard du caractère potentiellement très local de la validité d'un régime EDS sur nos données.

6.2.2 Re-simulation non-paramétrique

Comme nous l'avons vu précédemment (Chapitre 2 §2.3), la re-simulation rapide de trajectoires créées à partir des données observées et visitant ces dernières dans un ordre différent, ne peut être cohérente que dans le cas d'un processus sous-jacent de type Langevin. Cette dernière condition représente ainsi un obstacle à une méthode générale qui pourrait s'avérer très utile afin d'effectuer des tests statistiques, en mimant de nouvelles acquisitions de données, comblant ainsi le faible nombre de réalisations du processus sous-jacent dont nous disposons. Dans le but de se délester des paramètres dynamiques de dérive et de diffusion inhérents au modèle d'EDS, il serait bon de prolonger cet aspect de re-simulation en se basant sur des méthodes de *local bootstrap* [62], qui permettent de considérer de nouvelles trajectoires par des procédures de rééchantillonnage.

6.2.3 Interface graphique

Un point crucial est que l'algorithme de κ-segmentation ne saurait prendre toute sa valeur que par la confrontation de ses sorties avec une visualisation concrète des entités biologiques étudiées. En effet, l'utilisateur doit être en mesure d'effectuer des aller-retours, avec facilité et rapidité, entre des *frames* exhibées par la segmentation afin de rester au plus proche des problématiques liées à la biologie. Il pourra par exemple visualiser de manière statique les différentes conformations correspondant aux centres des puits exhibés, mais aussi disposer de *vidéos* de la trajectoire pour les plages de temps correspondant à des segments transitoires. Il est en effet indispensable de ne pas perdre de vue la réalité biologique des problèmes étudiés, et de pouvoir donner un sens concret aux résultats de la κ-segmentation, laquelle demeure un outil exploratoire permettant d'extraire de l'information à partir des trajectoires de simulation de DM, et de revenir aux problématiques initiales qui relèvent de la biologie, avec un point de vue enrichi par la méthode.

6.2.4 Application à NMDA et à d'autres molécules

L'exemple du récepteur NMDA présenté au Chapitre 1 §1.2.3 n'a pas pu faire l'objet d'un traitement par les méthodes développées, en revanche, il est à prévoir qu'une analyse par ces nouvelles approches des données de ce récepteur, ainsi que d'autres molécules, soit en mesure d'apporter un éclairage nouveau sur les enjeux biologiques étudiés.

Bibliographie

[1] M. Abraham et al. « GROMACS User Manual, version 5.0.7. » *Department of Biophysical Chemistry, University of Groningen* (2015).

[2] I. Ahmed et al. « Glutamate NMDA receptor dysregulation in Parkinson's disease with dyskinesias ». *Brain* 134.4 (mar. 2011), p. 979-986.

[3] S.-H. Ahn et J. Birgmeier. « Using Spectral Clustering to Sample Molecular States and Pathways » (2015).

[4] B. J. Alder et T. E. Wainwright. « Phase Transition for a Hard Sphere System ». *The Journal of Chemical Physics* 27.5 (1957), p. 1208-1209.

[5] V. Alexandrov et al. « Normal modes for predicting protein motions : A comprehensive database assessment and associated Web tool ». *Protein science : a publication of the Protein Society* 14 (avr. 2005), p. 633-43.

[6] K. S. Arun, T. S. Huang et S. D. Blostein. « Least-Squares Fitting of Two 3-D Point Sets ». *IEEE Trans. Pattern Anal. Mach. Intell.* 9.5 (1987), p. 698-700.

[7] M. Baaden. « Dynamique moléculaire in silico ». *Institut de Biologie Physico-Chimique, Paris* (mai 2003).

[8] « Baccalauréat Scientifique Antilles/Guyane » (2013).

[9] I. C. de Beauchêne et L. Tchertanov. « How missense mutations in receptors tyrosine kinases impact constitutive activity and alternate drug sensitivity : insights from molecular dynamics simulations ». *Receptors and Clinical Investigation* 3.3 (2016).

[10] D. Beeman. « Some multistep methods for use in molecular dynamics calculations ». 1976.

[11] R. Bell et J. Matschiner. « Vitamin K Activity of Phylloquinone Oxide ». *Arch. Biochem. Biophys* 2.141 (1970), p. 473-476.

[12] A. Bennasroune et al. « Tyrosine kinase receptors as attractive targets of cancer therapy ». *Critical reviews in oncology/hematology* 50 (mai 2004), p. 23-38.

[13] P. Bougerol. « Calcul stochastique des martingales continues ». *UPMC, Cours Master 2 Probabilités et Finance* (déc. 2015), p. 57-59.

[14] R Broglia et G. Tiana. « The Physics of Protein Folding and of Drug Design » (mar. 2005).

[15] B. Brooks et al. « CHARMM : A Program for Macromolecular Energy, Minimization, and Dynamics Calculations ». *Journal of Computational Chemistry* 4 (sept. 2004), p. 187 -217.

[16] J. C. Butcher. « The Numerical Analysis of Ordinary Differential Equations : Runge-Kutta and General Linear Methods ». (1987).

[17] D. Case et al. « The AMBER biomolecular simulation programs ». *Journal of computational chemistry* 26 (déc. 2005), p. 1668-88.

[18] F. Cazals et al. « Conformational Ensembles and Sampled Energy Landscapes : Analysis and Comparison ». *Journal of Computational Chemistry* 36 (juin 2015).

[19] N. Chatron. « VKORC1 and vitamin K agonists resistance : a molecular modelling study » (mar. 2017).

[20] N. Chatron et al. « Identification of the functional states of human Vitamin K epoxide reductase from molecular dynamics simulations ». *RSC Advances* 7 (nov. 2017), p. 52071-52090.

[21] C. Chipot. « Les méthodes numériques de la dynamique moléculaire ». *Université Henri Poincaré, Nancy* (juin 2002).

[22] N. T. Christoph Göbl. « Application of Solution NMR Spectroscopy to Study Protein Dynamics ». *Entropy* 14.3 (fév. 2012), p. 581-598.

[23] R. Coifman et al. « Diffusion Maps, Reduction Coordinates, and Low Dimensional Representation of Stochastic Systems ». *Multiscale Modeling and Simulation* 7 (jan. 2008).

[24] P. E. W. David D. Boehr Ruth Nussinov. « The role of dynamic conformational ensembles in biomolecular recognition ». *Nature Chemical Biology* 232 (2009), p. 789-796.

[25] G. G. Dodson, D. P. Lane et C. S. Verma. « Molecular simulations of protein dynamics : new windows on mechanisms in biology. » *EMBO reports* 9 2 (2008), p. 144-50.

[26] H. B. Fernandes et L. A. Raymond. « Animal Models of Cognitive Impairment : Chapter 2, NMDA Receptors and Huntington's Disease ». *Frontiers in Neuroscience* (2009).

[27] J.-F. L. Gall. « Calcul stochastique et processus de Markov (Cours de Master 2) » (2010).

[28] E. García-Portugués et al. « Toroidal diffusions and protein structure evolution ». *Applied Directional Statistics : Modern Methods and Case Studies* (avr. 2018), p. 77-110.

[29] M. M. Ghahremanpour et al. « MemBuilder : A Web-Based Graphical Interface to Build Heterogeneously Mixed Membrane Bilayers for the GROMACS Biomolecular Simulation Program ». *Bioinformatics (Oxford, England)* 30 (nov. 2013).

[30] J. Gilreath, L. Tchertanov et M. Deininger. « Novel approaches to treating advanced systemic mastocytosis » (2019).

[31] L. Goodstadt et C. Ponting. « Vitamin K epoxide reductase : Homology, active site and catalytic mechanism ». *Trends in biochemical sciences* 29 (juil. 2004), p. 289-92.

[32] K. B. Hansen et al. « Structure, function, and allosteric modulation of NMDA receptors ». *The Journal of General Physiology* 150.8 (2018), p. 1081-1105.

[33] R. Hockney, S. Goel et J. Eastwood. « Quiet high resolution computer models of a plasma ». *J. Comput. Phys.* 14.2 (fév. 1974), p. 148-158.

[34] S. Hubbard et W Todd Miller. « Hubbard SR, Miller WT.. Receptor tyrosine kinases : mechanisms of activation and signaling. Curr Opin Cell Biol 19 : 117-123 ». *Current opinion in cell biology* 19 (mai 2007), p. 117-23.

[35] F. Inizan et al. *The First 3D Model of the KIT Full-Length Cytoplasmic Domain Reveals a New Look for an Old Receptor.* (2019).

[36] R. A. Jarvis et E. A. Patrick. « Clustering Using a Similarity Measure Based on Shared Near Neighbors ». *IEEE Transactions on Computers* C-22.11 (1973), p. 1025-1034.

[37] I. Jolliffe et J. Cadima. « Principal component analysis : A review and recent developments ». *Philosophical Transactions of the Royal Society A : Mathematical, Physical and Engineering Sciences* 374 (avr. 2016).

[38] W. Jorgensen et al. « Comparison of Simple Potential Functions for Simulating Liquid Water ». *J. Chem. Phys.* 79 (juil. 1983), p. 926-935.

[39] P. Junhui et al. « Clustering algorithms to analyze molecular dynamics simulation trajectories for complex chemical and biological systems ». *Chinese Journal of Chemical Physics* 31 (août 2018), p. 404-420.

[40] D. K. Bhattacharya, E. Clementi et W. Xue. « Stochastic dynamic simulation of a protein ». *International Journal of Quantum Chemistry* 42 (juin 1992), p. 1397 -1408.

[41] J. Klepeis et al. « Long-timescale molecular dynamics simulations of protein structure and function ». *Current opinion in structural biology* 19 (mai 2009), p. 120-127.

[42] S. Kmiecik et al. « Coarse-Grained Protein Models and Their Applications ». *Chemical Reviews* 116 (juin 2016).

[43] T. Lelièvre et G. Stoltz. « Partial differential equations and stochastic methods in molecular dynamics ». *Acta Numerica* 25 (mai 2016), p. 681-880.

[44] M. Levitt. « A simplified representation of protein conformations for rapid simulation of protein folding. » *Journal of molecular biology* 104 1 (1976), p. 59-107.

[45] L. Li et D. Hanahan. « Hijacking the Neuronal NMDAR Signaling Circuit to Promote Tumor Growth and Invasion ». *Cell* 153 (mar. 2013), p. 86-100.

[46] T. Li. « Identification of the gene for vitamin K epoxide reductase ». *Nature* 427.6974 (2004), p. 541-544.

[47] W. Li et al. « Structure of a bacterial homologue of vitamin K epoxide reductase ». *Nature* 463.7280 (2010), p. 507.

[48] U. von Luxburg. « A tutorial on spectral clustering ». *Statistics and Computing* 17.4 (2007), p. 395-416.

[49] U. von Luxburg, O. Bousquet et M. Belkin. « On the Convergence of Spectral Clustering on Random Samples : The Normalized Case ». *Proceedings of the 17th Annual Conference on Learning Theory* 3120 (juil. 2004), p. 457-471.

[50] O. M. Becker, Y. Levy et O. Ravitz. « Flexibility, Conformation Spaces, and Bioactivity ». *Journal of Physical Chemistry* 104 (fév. 2000).

[51] A. D. MacKerell et al. « All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins ». *The Journal of Physical Chemistry* 102.18 (1998), p. 3586-3616.

[52] C. Maffeo et al. « Modeling and Simulation of Ion Channels ». *Chemical Reviews* 112.12 (2012), p. 6250-6284.

[53] J. A. McCammon, B. R. Gelin et M. Karplus. « Dynamics of folded proteins ». *Nature* 267 (1977), p. 585-590.

[54] E. A. Nadaraya. « On estimating regression ». *Theory of Probability & Its Applications* 9.1 (1964), p. 141-142.

[55] B. Nadler et al. « Diffusion Maps, Spectral Clustering and Reaction Coordinates of Dynamical Systems » (2006).

[56] B. Nadler et al. « Difusion maps, spectral clustering and reaction coordinates of dynamical systems ». *Applied and Computational Harmonic Analysis* 21 (juil. 2006), p. 113-127.

[57] W. Nadler et al. « Molecular and stochastic dynamics of proteins ». *Proceedings of the National Academy of Sciences of the United States of America* 84 (déc. 1987), p. 7933-7.

[58] K. R. Naqvi. « The origin of the Langevin equation and the calculation of the mean squared displacement : Let's set the record straight » (mar. 2005).

[59] Oliet, Stéphane H.R. et Papouin, Thomas. « De l´importance de la localisation des récepteurs du glutamate NMDA ». *Med Sci (Paris)* 29.3 (2013), p. 260-262.

[60] Z. Palmai et al. « How does binding of agonist ligands control intrinsic molecular dynamics in human NMDA receptors ? » *PLoS ONE* 13.8 (juil. 2018).

[61] V. S. Pande, K. A. Beauchamp et G. R. Bowman. « Everything you wanted to know about Markov State Models but were afraid to ask. » *Methods* 52 1 (2010), p. 99-105.

[62] E. Paparoditis et D. Politis. « The Local Bootstrap for Markov Processes ». *Journal of Statistical Planning and Inference* 108 (nov. 2002), p. 301-328.

[63] J. C. Phillips et al. « Scalable molecular dynamics with NAMD ». *Journal of computational chemistry* 26 16 (2005), p. 1781-802.

[64] A. Quintas-Cardama et al. « Novel approaches in the treatment of systemic mastocytosis ». *Cancer* 107 (oct. 2006), p. 1429-39.

[65] G. R. Bowman, X. Huang et V. Pande. « Using generalized ensemble simulations and Markov state models to identify conformational states ». *Methods (San Diego, Calif.)* 49 (juin 2009), p. 197-201.

[66] A. Rahman. « Correlations in the Motion of Atoms in Liquid Argon ». *Phys. Rev.* 136 (1964), p. 405-411.

[67] A. H. Rezvani. « Animal Models of Cognitive Impairment : Chapter 4, Involvement of the NMDA System in Learning and Memory ». *Frontiers in Neuroscience* (2009).

[68] S. Rost. « Mutations in VKORC1 cause warfarin resistance and multiple coagulation factor deficiency type 2 ». *Nature* 427.6974 (2004), p. 537-541.

[69] J. Shao et al. « Clustering Molecular Dynamics Trajectories : 1. Characterizing the Performance of Different Clustering Algorithms ». *Journal of Chemical Theory and Computation* 3.6 (2007), p. 2312-2334.

[70] D. Shaw et al. « Anton, a special-purpose machine for molecular dynamics simulation ». *Communications of The ACM - CACM* 51 (jan. 2007), p. 1-12.

[71] A. Sirur, D. De Sancho et R. B. Best. « Markov state models of protein misfolding ». *The Journal of Chemical Physics* 144.7 (2016), p. 75101.

[72] C. Sollier et L. Drouet. « Vitamine K ». *Oléagineux, Corps gras, Lipides* 18 (mar. 2011), p. 94-98.

[73] D. van der Spoel et al. « GROMACS : fast, flexible, and free ». *Journal of computational chemistry* 26 (déc. 2005), p. 1701-18.

[74] R. Stanton. « A nonparametric model of term structure dynamics and the market price of interest rate risk ». *The Journal of Finance* 52.5 (1997), p. 1973-2002.

[75] K. M. Thayer, B. Lakhani et D. L. Beveridge. « Molecular Dynamics–Markov State Model of Protein Ligand Binding and Allostery in CRIB-PDZ : Conformational Selection and Induced Fit ». *The Journal of Physical Chemistry B* 121.22 (2017), p. 5509-5514.

[76] e. a. Tie J.K. « Membrane topology mapping of vitamin K epoxide reductase by in vitro translation/cotranslocation ». *J. Biol. Chem* 280.16 (2005), p. 16410-16416.

[77] J.-K. Tie et D. Stafford. « Vitamin K – Structure and Function of Vitamin K Epoxide Reductase ». *Vitamins and hormones* 78 (fév. 2008), p. 103-30.

[78] A. E. Torda et W. F. van Gunsteren. « Algorithms for clustering molecular dynamics configurations ». *Journal of Computational Chemistry* 15.12 (1994), p. 1331-1340.

[79] L. Verlet. « Computer "'Experiments"' on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules ». *Physical Review* 159 (juil. 1967), p. 98.

[80] G. S. Watson. « Smooth regression analysis ». *Sankhyā : The Indian Journal of Statistics, Series A* (1964), p. 359-372.

[81] W. Yang. « Master Course ». *Molecular Dynamics, Chap. 3, Non-Bonded Interactions and Boundary Conditions* (2013).

[82] Y. Zhang et al. « Dysfunction of NMDA receptors in Alzheimer's disease ». *Neurological sciences : official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology* 37.7 (2016), 1039–1047.

[83] S. Zhu et al. « Mechanism of NMDA Receptor Inhibition and Activation ». *Cell* 165 (2016), p. 704-714.

Annexe A

Annexes

A.1 Données simulées

Nous faisons dans cette brève section l'inventaire des données qui ont été traitées au cours de ce travail.

A.1.1 Le modèle des trois puits

Nous considérons dans ce travail un modèle jouet tiré de [55], que nous désignons par *modèle des trois puits*. On se place dans un paysage énergétique défini dans le plan par une fonction d'énergie notée $U : \mathbb{R}^2 \to \mathbb{R}$, et qui fait distinctement apparaître trois puits énergétiques : deux puits prononcés en $(-1, 0)$ et $(1, 0)$, ainsi qu'un troisième bien moins prononcé en $(0, 5/3)$ (Voir Fig. A.1). Le paysage énergétique étant plat à l'extérieur de ces bassins, on décide d'ajouter un terme quadratique à la fonction de manière à contraindre le processus à se réorienter vers les puits lorsqu'il s'en éloigne. Pour tout $(x, y) \in \mathbb{R}^2$:

$$U_c(x, y) = 3e^{-x^2}[e^{-(y-1/3)^2} - e^{-(y-5/3)^2}] - 5e^{-y^2}[e^{-(x-1)^2} + e^{-(x+1)^2}] + c(x^2 + y^2)$$

où $c > 0$. On définit une densité de probabilité $f(x) = \frac{1}{Z_\theta} \exp(-\frac{U_c(x)}{\theta})$, où θ est un paramètre de température fixé arbitrairement, et Z_θ la constante de normalisation permettant d'obtenir une densité.

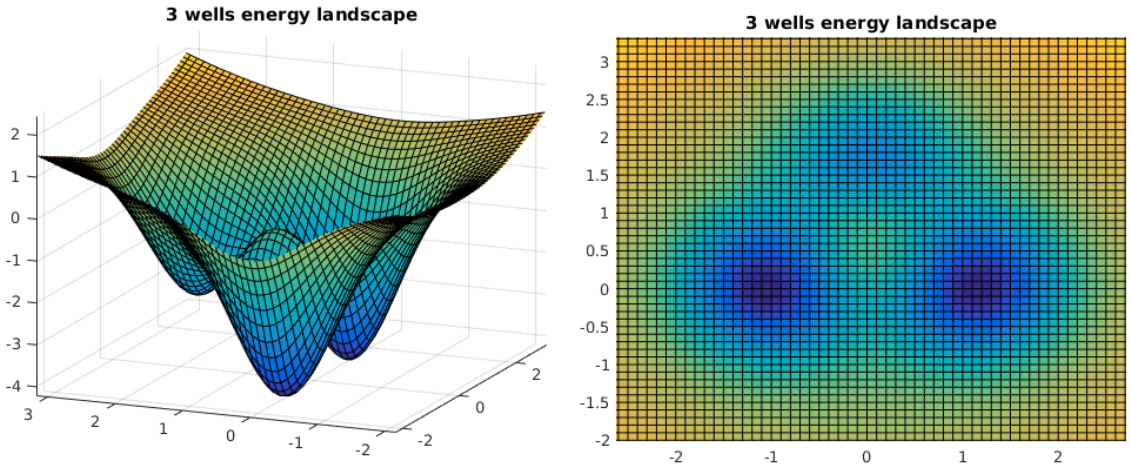

FIGURE A.1 – Énergie des 3 puits pour $c = 0.1$

A.1.2 Macromolécules biologiques et données de simulation de leur DM

Pour VKORC1, ont été traitées deux trajectoires de 1 µs échantillonnées toutes les 5 ps, correspondant à une version *hyperflexible* de la protéine. Ces deux trajectoires sont nommées Trajectoire 1 et Trajectoire 2.

Pour KIT, a été traitée une trajectoire de 2 µs échantillonnée toutes les 10 ps générée pour la protéine intégrale, KIT avec le domaine KID : on la nomme Trajectoire KIT+KID. Les diverses trajectoires concernant des sous-systèmes de la protéine intégrale sont obtenus par des restrictions de la Trajectoire KIT+KID aux résidus correspondant. Ces trajectoires seront nommées : Trajectoire KID, Trajectoire KD et Trajectoire KD+KID.

	Trajectoires de DM des protéines simulées
VKORC1	2×1 µs / 5 ps
KIT+KID	1×2 µs / 10 ps
KID	1×2 µs / 10 ps
KD	1×2 µs / 10 ps
KD+KID	1×2 µs / 10 ps

TABLEAU A.1 – Trajectoires simulées de VKORC1, KIT+KIT et KID seul

A.2 Pré-traitements d'une trajectoire de DM avant analyse

A.2.1 Recalage en translation et rotation (recalage rigide)

Soit $x = (x_a)_{a \in A}$ une conformation à recaler sur $x^{\text{réf}}$. En écrivant $\nabla_\tau J(\Theta, \tau) = 0$ on obtient : $\sum_{a \in A} m_a(\Theta x_a + \tau - x_a^{\text{réf}}) = 0$ et donc :

$$\tau = -\Theta\ \overline{x} + \overline{x^{\text{réf}}} \tag{A.1}$$

où $\overline{x} \triangleq \frac{1}{M} \sum_{a \in A} m_a x_a \in \mathbb{R}^3$ avec $M \triangleq \sum_{a \in A} m_a$ est le barycentre de la conformation x. En substituant τ par son expression (A.1) en point critique on peut écrire J comme une fonction de

Θ :

$$\begin{aligned} J(\Theta) &\triangleq \inf_{\tau\in\mathbb{R}^3} J(\Theta,\tau) = \sum_{a\in A} m_a \|\Theta(x_a-\overline{x}) - (x_a^{\text{réf}} - \overline{x^{\text{réf}}})\|^2 \\ &= \sum_{a\in A} m_a \|\Theta(x_a-\overline{x})\|^2 + \sum_{a\in A} m_a \|x^{\text{réf}} - \overline{x^{\text{réf}}}\|^2 \\ &\underbrace{- 2 \sum_{a\in A} m_a \langle \Theta(x_a-\overline{x}), x_a^{\text{réf}} - \overline{x^{\text{réf}}}\rangle}_{= \langle \Theta, \sum_{a\in A} m_a (x_a^{\text{réf}} - \overline{x^{\text{réf}}})(x_a-\overline{x})^T\rangle} \end{aligned}$$

En posant $S \triangleq \sum_{a\in A} m_a (x_a^{\text{réf}} - \overline{x^{\text{réf}}})(x_a-\overline{x})^T \in \mathcal{M}_3(\mathbb{R})$, le problème revient à calculer :

$$\Theta^* = \underset{\Theta\in SO_3(\mathbb{R})}{\text{argmax}} \{\langle \Theta, S\rangle\}.$$

Or, on peut écrire (SVD) que $S = U\Lambda V^T$, avec U et V matrices unitaires de $\mathcal{M}_3(\mathbb{R})$, et $\Lambda = \text{diag}(\lambda_1,\lambda_2,\lambda_3)$, avec $\lambda_i \geq 0$. En notant alors $(u_i)_i$ et $(v_i)_i$ les bases orthonormées décrites par U et V respectivement, $(w_i)_i$ celle induite par l'action de la rotation Θ sur la base $(v_i)_i$ (i.e. on pose $w_i = \Theta v_i$) et $(e_j)_j$ la base canonique de $\mathbb{R}^3$, on a que :

$$\begin{aligned} \langle \Theta, S\rangle &= \langle \Theta, U\Lambda V^T\rangle = \langle U^T\Theta V, \Lambda\rangle \\ &= \langle \sum_{i=1}^3 u_i^T w_i, \sum_{j=1}^3 \lambda_j e_j e_j^T\rangle = \sum_{i,j=1}^3 \lambda_j \langle u_i, e_j\rangle\langle w_i, e_j\rangle \\ &\leq \sum_{j=1}^3 \lambda_j \sqrt{\underbrace{\sum_{i=1}^3 \langle u_i, e_j\rangle^2}_{=1} \underbrace{\sum_{i=1}^3 \langle w_i, e_j\rangle^2}_{=1}} \text{ (Cauchy-Schwarz)} \\ &\leq \sum_{j=1}^3 \lambda_j . \end{aligned}$$

où l'on remarque que ce dernier maximum est atteint lorsque $U^T\Theta V = \text{I}_3$. Au final, on a donc que :

$$\begin{cases} \Theta^* = V^T U \\ \tau* = -\Theta^*\overline{x} + \overline{x^{\text{réf}}} \end{cases}$$

Pour recaler la trajectoire toute entière sur $x^{\text{réf}}$, on devra ainsi calculer le couple (Θ^*, τ^*) pour toute conformation x de la trajectoire à recaler, et effectuer les opérations suivantes pour chacun des atomes $a \in A$:

$$x_a \longleftarrow \Theta^* x_a + \tau^*$$

A.2.2 Analyse en Composantes Principales (ACP)

Pour une trajectoire $\mathbf{X} \in \mathcal{M}_{d,N}(\mathbb{R})$, on commence par centrer la matrice en posant $\widetilde{\mathbf{X}} = \mathbf{X} - \overline{\mathbf{X}}$ où $\overline{\mathbf{X}}$ est une moyenne de $\mathbf{X}$ effectuée sur chaque ligne (i.e. sur chaque composante de $\mathbb{R}^d$). L'idée consiste alors à identifier un premier vecteur $u_1^* \in \mathbb{R}^d$ tel que la variance du nuage de points projeté sur celui-ci $\widetilde{\mathbf{X}}^T u_1^*$ soit maximale. Pour $u \in \mathbb{R}^d$, cette variance s'écrit :

$$\frac{(\widetilde{\mathbf{X}}^T u)^T(\widetilde{\mathbf{X}}^T u)}{N} = \frac{u^T \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T u}{N} = u^T C u$$

où $C = \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T/N \in \mathcal{M}_d(\mathbb{R})$ est la matrice de covariance de $\widetilde{\mathbf{X}}$. On cherche alors :

$$u_1^* = \underset{u\in\mathbb{R}^d}{\operatorname{argmax}}\{u^T C u\}$$

La matrice C étant symétrique réelle définie-positive, elle est diagonalisable dans une base orthonormée, que nous résumons au sein d'une matrice unitaire P de sorte que pour $u \in \mathbb{R}^d$:

$$u^T C u = u^T P^T D P u = v^T D v$$

où $v = Pu \in \mathbb{R}^d$ et $D = \operatorname{diag}(\lambda_k)_{1\leq k\leq d}$. Par conséquent :

$$\begin{aligned} u^T C u &= \sum_{k=1}^{d} \lambda_k v_k^2 \\ &\leq \lambda_{\max}(C) \times \|v\|^2 \end{aligned}$$

où $\lambda_{\max}(C) = \max\limits_{1\leq k\leq d}\{\lambda_k\}$. Mais alors, P étant unitaire, on a que $\|v\|^2 = \|Pu\|^2 = \|u\|^2$. En maximisant sous la contrainte $\|u\|^2 = 1$, on voit donc que :

$$\max_{u\in\mathbb{R}^d:\|u\|^2=1}\{u^T C u\} = \lambda_{\max}(C)$$

Ce maximum étant atteint pour u_1^* correspondant au vecteur propre normalisé associé à la plus grande valeur propre de C :

$${u_1^*}^T C u_1^* = \lambda_{\max}(C)\|u_1^*\|^2 = \lambda_{\max}(C)$$

La recherche d'un second axe de projection correspond au calcul de :

$$u_2^* = \underset{u\in \operatorname{Vect}(u_1^*)\bigoplus {u_1^*}^{\perp}}{\operatorname{argmax}}\{u^T C u\}$$

En maximisant à nouveau sous la contrainte $\|u\|^2 = 1$, on voit que le maximum recherché correspond à la seconde plus grande valeur propre de C, et qu'il est atteint pour le vecteur propre normalisé qui lui est associé, lequel est déjà décrit dans la même matrice P que précédemment. Ainsi, en procédant de manière itérative, on définit une base optimale de $\mathbb{R}^d$ maximisant la variance des données et constituée des vecteurs propres normalisés $(u_k^*)_{1\leq k\leq d}$ de C. Quitte à les réordonner au sein de la matrice P suivant l'ordre décroissant des valeurs propres, une ACP de dimension $d_0 < d$ consiste alors à ne retenir que ces d_0 premiers vecteurs, et donc à projeter les données sur le sous-espace $\bigoplus_{k=1}^{d_0} \operatorname{Vect}(u_k^*)$. Pour ce faire, on calcule les coefficients de projection Γ et les données projetées $\widetilde{\mathbf{X}}_{\text{proj}}$ de la façon suivante :

$$\begin{cases} \Gamma = P^T \widetilde{\mathbf{X}} \\ \widetilde{\mathbf{X}}_{\text{proj}} = \overline{\mathbf{X}} + P\Gamma \end{cases}$$

puis on conserve les d_0 première lignes de Γ et de $\widetilde{\mathbf{X}}_{\text{proj}}$.

A.3 Preuves

A.3.1 Proposition 2 §2.3.1

Soit $x \in \mathcal{X}$. Calculons :

$$P_\epsilon q(x) - q(x) = \int_{y\in\mathcal{X}} P_\epsilon(x,y)\mu(y)[q(y)-q(x)]dy = \frac{\int_{y\in\mathcal{X}} W_\epsilon(x,y)\mu(y)[q(y)-q(x)]dy}{\int_{y\in\mathcal{X}} W_\epsilon(x,y)\mu(y)dy}$$
$$= \frac{\int_{h\in\mathcal{X}} \exp(-\frac{\|h\|^2}{2\epsilon})\mu(x+h)[q(x+h)-q(x)]dh}{\int_{h\in\mathcal{X}} \exp(-\frac{\|h\|^2}{2\epsilon})\mu(x+h)dh}$$

En multipliant en haut et en bas par $(1/2\pi\epsilon)^{d/2}$, on fait apparaître la loi $\sqrt{\epsilon}\times\mathcal{N}(0_{\mathbb{R}^d}, \mathrm{I}_d)$. Pour $\gamma \sim \mathcal{N}(0_{\mathbb{R}^d}, \mathrm{I}_d)$, on peut alors écrire que :

$$P_\epsilon q(x) - q(x) = \frac{\mathbb{E}\Big[\mu(x+\sqrt{\epsilon}\gamma)\big[q(x+\sqrt{\epsilon}\gamma)-q(x)\big]\Big]}{\mathbb{E}\Big[\mu(x+\sqrt{\epsilon}\gamma)\Big]}$$

Par ailleurs, on a que :

$$\mu(x+\sqrt{\epsilon}\gamma) \underset{\epsilon\to 0}{=} \mu(x) + \sqrt{\epsilon}\langle\nabla\mu(x),\gamma\rangle + R_{\gamma,\epsilon}$$

où $R_{\gamma,\epsilon} \underset{\epsilon\to 0}{=} O(\gamma^2\epsilon)$. Et donc :

$$\begin{aligned}\mathbb{E}\Big[\mu(x+\sqrt{\epsilon}\gamma)\Big] &\underset{\epsilon\to 0}{=} \mathbb{E}\Big[\mu(x)+\sqrt{\epsilon}\langle\nabla\mu(x),\gamma\rangle + o(\sqrt{\epsilon})\Big]\\ &\underset{\epsilon\to 0}{=} \mu(x) + \sqrt{\epsilon}\times\underbrace{\mathbb{E}[\langle\nabla\mu(x),\gamma\rangle]}_{=0} + o(\sqrt{\epsilon})\\ &\underset{\epsilon\to 0}{=} \mu(x) + \mathbb{E}[R_{\gamma,\epsilon}]\end{aligned}$$

Mais alors, il existe $M > 0$, tel que :

$$|R_{\gamma,\epsilon}| \underset{\epsilon\to 0}{\leq} M\gamma^2\epsilon$$

Et donc :

$$\begin{aligned}\frac{|\mathbb{E}[R_{\gamma,\epsilon}]|}{\sqrt{\epsilon}} &\underset{\epsilon\to 0}{\leq} \frac{\mathbb{E}[|R_{\gamma,\epsilon}|]}{\sqrt{\epsilon}}\\ &\underset{\epsilon\to 0}{\leq} M\sqrt{\epsilon}\end{aligned}$$

De sorte que : $\mathbb{E}[R_{\gamma,\epsilon}] \underset{\epsilon\to 0}{=} o(\sqrt{\epsilon})$. D'où :

$$\begin{aligned}\mathbb{E}\Big[\mu(x+\sqrt{\epsilon}\gamma)\Big] &\underset{\epsilon\to 0}{=} \mu(x) + o(\sqrt{\epsilon})\\ &\underset{\epsilon\to 0}{\sim} \mu(x)\end{aligned}$$

D'autre part, nous avons que :

$$\begin{aligned}
&\mu(x+\sqrt{\epsilon}\gamma)[q(x+\sqrt{\epsilon}\gamma)-q(x)]\\
&\underset{\epsilon\to 0}{=}\left(\mu(x)+\sqrt{\epsilon}\langle\gamma,\nabla\mu(x)\rangle+o(\sqrt{\epsilon})\right)\left(\sqrt{\epsilon}\langle\gamma,\nabla q(x)\rangle+\frac{\epsilon}{2}\langle H_q(x)\gamma,\gamma\rangle+o(\epsilon)\right)\\
&\underset{\epsilon\to 0}{=}\sqrt{\epsilon}\mu(x)\langle\gamma,\nabla q(x)\rangle+\frac{\epsilon}{2}\mu(x)\langle H_q(x)\gamma,\gamma\rangle+\epsilon\langle\gamma,\nabla\mu(x)\rangle\langle\gamma,\nabla q(x)\rangle+o(\epsilon)\\
&\underset{\epsilon\to 0}{=}\sqrt{\epsilon}\mu(x)\sum_{j=1}^{d}\partial_j q(x)\gamma_j+\frac{\epsilon}{2}\mu(x)\sum_{i=1}^{d}\sum_{j=1}^{d}\partial_{ij}q(x)\gamma_i\gamma_j+\epsilon\sum_{i=1}^{d}\sum_{j=1}^{d}(\partial_i\mu(x))(\partial_j q(x))\gamma_i\gamma_j+o(\epsilon)
\end{aligned}$$

Par conséquent, en montrant de même que $\mathbb{E}[o(\epsilon)]\underset{\epsilon\to 0}{=}o(\epsilon)$, nous avons que :

$$\begin{aligned}
\mathbb{E}\Big[\mu(x+\sqrt{\epsilon}\gamma)[q(x+\sqrt{\epsilon}\gamma)-q(x)]\Big]\underset{\epsilon\to 0}{=}&\sqrt{\epsilon}\mu(x)\sum_{j=1}^{d}\partial_j q(x)\underbrace{\mathbb{E}[\gamma_j]}_{=0}\\
&+\frac{\epsilon}{2}\mu(x)\sum_{i=1}^{d}\sum_{j=1}^{d}\partial_{ij}q(x)\underbrace{\mathbb{E}[\gamma_i\gamma_j]}_{=\delta_{ij}}\\
&+\epsilon\sum_{i=1}^{d}\sum_{j=1}^{d}(\partial_i\mu(x))(\partial_j q(x))\underbrace{\mathbb{E}[\gamma_i\gamma_j]}_{=\delta_{ij}}\\
&+o(\epsilon)
\end{aligned}$$

D'où :

$$\begin{aligned}
\mathbb{E}\Big[\mu(x+\sqrt{\epsilon}\gamma)[q(x+\sqrt{\epsilon}\gamma)-q(x)]\Big]&\underset{\epsilon\to 0}{=}\frac{\epsilon}{2}\mu(x)\Delta q(x)+\epsilon\langle\nabla q(x),\nabla\mu(x)\rangle+o(\epsilon)\\
&\underset{\epsilon\to 0}{\sim}\frac{\epsilon}{2}\mu(x)\Delta q(x)+\epsilon\langle\nabla q(x),\nabla\mu(x)\rangle
\end{aligned}$$

On obtient que :

$$P_\epsilon q(x)-q(x)\underset{\epsilon\to 0}{\sim}\frac{\frac{\epsilon}{2}\mu(x)\Delta q(x)+\epsilon\langle\nabla q(x),\nabla\mu(x)\rangle}{\mu(x)}$$

Et donc :

$$\frac{P_\epsilon q(x)-q(x)}{\epsilon}\underset{\epsilon\to 0}{\sim}\frac{\Delta q(x)}{2}+\frac{\epsilon\langle\nabla q(x),\nabla\mu(x)\rangle}{\mu(x)}$$

En remarquant finalement que $\nabla\mu(x)=-\mu(x)\frac{\nabla U(x)}{\theta}$, on aboutit à la formule désirée :

$$\frac{P_\epsilon q(x)-q(x)}{\epsilon}\underset{\epsilon\to 0}{\sim}\frac{\Delta q(x)}{2}-\frac{1}{\theta}\langle\nabla q(x),\nabla U(x)\rangle$$

Cherchons maintenant à déterminer l'équation de Fokker-Planck progressive (*forward*) correspondante, agissant cette-fois ci sur les mesures. L'opérateur en question, noté $\mathcal{L}_f$ est défini de manière unique par :

$$\int_{x\in\mathcal{X}}(\mathcal{L}_b q)(x)\mu(x)dx=\int_{x\in\mathcal{X}}q(x)(\mathcal{L}_f\mu)(x)dx$$

Calculons :

$$\int_{x\in\mathcal{X}}(\mathcal{L}_b q)(x)\mu(x)dx=\int_{x\in\mathcal{X}}\frac{\Delta q(x)}{2}\mu(x)dx-\frac{1}{\theta}\int_{x\in\mathcal{X}}\langle\nabla q(x),\nabla U(x)\rangle\mu(x)dx$$

Le premier terme peut s'écrire :

$$\begin{aligned}\int_{x\in\mathcal{X}} \Delta q(x)\mu(x)dx &= \int_{x\in\mathcal{X}} \mathrm{div}(\nabla q(x))\mu(x)dx \\ &= \underbrace{\int_{x\in\mathcal{X}} \mathrm{div}(\mu(x)\nabla q(x))dx}_{=0} - \int_{x\in\mathcal{X}} \langle \nabla\mu(x), \nabla q(x)\rangle dx \\ &= -\int_{x\in\mathcal{X}} \langle \nabla\mu(x), \nabla q(x)\rangle dx \\ &= \int_{x\in\mathcal{X}} \Delta\mu(x) q(x) dx \quad \text{par symétrie du produit scalaire}\end{aligned}$$

Pour le second terme, on a que :

$$\begin{aligned}\int_{x\in\mathcal{X}} \langle \nabla q(x), \nabla U(x)\rangle \mu(x)dx &= \int_{x\in\mathcal{X}} \langle \nabla q(x), \mu(x)\nabla U(x)\rangle dx \\ &= \underbrace{\int_{x\in\mathcal{X}} \mathrm{div}(q(x)\mu(x)\nabla U(x))}_{=0} - \int_{x\in\mathcal{X}} q(x)\mathrm{div}(\mu(x)\nabla U(x)) \\ &= -\int_{x\in\mathcal{X}} q(x)\mathrm{div}(\mu(x)\nabla U(x))\end{aligned}$$

Au total, on a donc que :

$$\int_{x\in\mathcal{X}} q(x)\left[\frac{\Delta\mu(x)}{2} + \frac{1}{\theta}\mathrm{div}(\mu(x)\nabla U(x))\right]dx = \int_{x\in\mathcal{X}} q(x)(\mathcal{L}_f\mu)(x)dx$$

Et par unicité, on a donc que :

$$(\mathcal{L}_f\mu)(x) = \frac{\Delta\mu(x)}{2} + \frac{1}{\theta}\mathrm{div}(\mu(x)\nabla U(x))$$

Or, pour $\mu = \mu^\infty$, on a que :

$$\frac{\Delta\mu^\infty(x)}{2} = \frac{1}{2}\mathrm{div}(\nabla\mu^\infty) = \frac{1}{2}\left(-\frac{2}{\theta}\mathrm{div}(\mu^\infty\nabla U(x))\right) = -\frac{1}{\theta}\mathrm{div}(\mu^\infty\nabla U(x))$$

Et donc :

$$(\mathcal{L}_f\mu^\infty)(x) = 0$$

A.3.2 Proposition 3 §2.3.2

Calculons :

$$\begin{aligned}
P_\epsilon^{(\alpha)}q(x) - q(x) &= \int_{y\in\mathcal{X}} P_\epsilon^{(\alpha)}(x,y)\mu(y)q(y)dy - q(x)\int_{y\in\mathcal{X}} P_\epsilon^{(\alpha)}(x,y)\mu(y)dy \\
&= \int_{y\in\mathcal{X}} P_\epsilon^{(\alpha)}(x,y)\mu(y)[q(y)-q(x)]dy \\
&= \frac{\int_{y\in\mathcal{X}} W_\epsilon^{(\alpha)}(x,y)\mu(y)[q(y)-q(x)]dy}{\int_{y\in\mathcal{X}} W_\epsilon^{(\alpha)}(x,y)\mu(y)dy} \\
&= \frac{\int_{y\in\mathcal{X}} \frac{W_\epsilon(x,y)}{(d_\epsilon(x)d_\epsilon(y))^\alpha}\mu(y)[q(y)-q(x)]dy}{\int_{y\in\mathcal{X}} \frac{W_\epsilon(x,y)}{(d_\epsilon(x)d_\epsilon(y))^\alpha}\mu(y)dy} \\
&= \frac{\int_{y\in\mathcal{X}} \exp(-\frac{\|x-y\|^2}{2\epsilon})\frac{\mu(y)}{(d_\epsilon(x)d_\epsilon(y))^\alpha}[q(y)-q(x)]dy}{\int_{y\in\mathcal{X}} \exp(-\frac{\|x-y\|^2}{2\epsilon})\frac{\mu(y)}{(d_\epsilon(x)d_\epsilon(y))^\alpha}dy} \\
&= \frac{\int_{h\in\mathcal{X}} \exp(-\frac{\|h\|^2}{2\epsilon})\frac{\mu(x+h)}{(d_\epsilon(x+h))^\alpha}[q(x+h)-q(x)]dh}{\int_{h\in\mathcal{X}} \exp(-\frac{\|h\|^2}{2\epsilon})\frac{\mu(x+h)}{(d_\epsilon(x+h))^\alpha}dh}
\end{aligned}$$

Or, pour tout $x \in \mathcal{X}$, on a que :

$$\begin{aligned}
d_\epsilon(x) &= \frac{1}{(2\pi\epsilon)^{d/2}}\int_{y\in\mathcal{X}} \exp\left(-\frac{\|x-y\|^2}{2\epsilon}\right)\mu(y)dy \\
&= \frac{1}{(2\pi\epsilon)^{d/2}}\int_{h\in\mathcal{X}} \exp\left(-\frac{\|h\|^2}{2\epsilon}\right)\mu(x+h)dh \\
&= \mathbb{E}[\mu(x+\sqrt{\epsilon}\gamma)] \\
&\underset{\epsilon\to 0}{\sim} \mu(x)
\end{aligned}$$

avec $\gamma \sim \mathcal{N}(0_{\mathbb{R}^d}, I_d)$. Et donc :

$$P_\epsilon^{(\alpha)}q(x) - q(x) \underset{\epsilon\to 0}{=} \frac{\int_{h\in\mathcal{X}} \exp(-\frac{\|h\|^2}{2\epsilon})\mu^{1-\alpha}(x+h)[q(x+h)-q(x)]dh}{\int_{h\in\mathcal{X}} \exp(-\frac{\|h\|^2}{2\epsilon})\mu^{1-\alpha}(x+h)dh}$$

En faisant intervenir comme précédemment une loi $\gamma \sim \mathcal{N}(0_{\mathbb{R}^d}, I_d)$, on a que :

$$P_\epsilon^{(\alpha)}q(x) - q(x) \underset{\epsilon\to 0}{=} \frac{\mathbb{E}\left[\mu^{1-\alpha}(x+\sqrt{\epsilon}\gamma)[q(x+\sqrt{\epsilon}\gamma)-q(x)]\right]}{\mathbb{E}\left[\mu^{1-\alpha}(x+\sqrt{\epsilon}\gamma)\right]}$$

De même que précédemment, on a d'une part que :

$$\begin{aligned}
\mathbb{E}\left[\mu^{1-\alpha}(x+\sqrt{\epsilon}\gamma)\right] &\underset{\epsilon\to 0}{=} \mu^{1-\alpha}(x) + \sqrt{\epsilon}\times \underbrace{\mathbb{E}\left[\langle\nabla\mu^{1-\alpha}(x),\gamma\rangle\right]}_{=0} + o(\sqrt{\epsilon}) \\
&\underset{\epsilon\to 0}{\sim} \mu^{1-\alpha}(x)
\end{aligned}$$

Et d'autre part, nous avons :

$$\begin{aligned}
&\mu^{1-\alpha}(x+\sqrt{\epsilon}\gamma)[q(x+\sqrt{\epsilon}\gamma)-q(x)] \\
&\underset{\epsilon\to 0}{=} \sqrt{\epsilon}\mu^{1-\alpha}(x)\langle\nabla q(x),\gamma\rangle + \mu^{1-\alpha}(x)\frac{\epsilon}{2}\langle H_q(x)\gamma,\gamma\rangle + \epsilon\langle\nabla\mu^{1-\alpha}(x),\gamma\rangle\langle\nabla q(x),\gamma\rangle + o(\epsilon)
\end{aligned}$$

De sorte que :

$$\mathbb{E}\Big[\mu^{1-\alpha}(x+\sqrt{\epsilon}\gamma)[q(x+\sqrt{\epsilon}\gamma)-q(x)]\Big] \underset{\epsilon\to 0}{\sim} \mu^{1-\alpha}(x)\frac{\epsilon}{2}\Delta q(x)+\epsilon\langle\nabla\mu^{1-\alpha}(x),\nabla q(x)\rangle$$

Puis, en remarquant que $\nabla\mu^{1-\alpha}(x) = -\frac{1-\alpha}{\theta}\mu^{1-\alpha}(x)\nabla U(x)$, on a :

$$\begin{aligned}\frac{\mathbb{E}\Big[\mu^{1-\alpha}(x+\sqrt{\epsilon}\gamma)[q(x+\sqrt{\epsilon}\gamma)-q(x)]\Big]}{\mathbb{E}\Big[\mu^{1-\alpha}(x+\sqrt{\epsilon}\gamma)\Big]} &\underset{\epsilon\to 0}{\sim} \frac{\mu^{1-\alpha}(x)\frac{\epsilon}{2}\Delta q(x)+\epsilon\langle\nabla\mu^{1-\alpha}(x),\nabla q(x)\rangle}{\mu^{1-\alpha}(x)}\\ &\underset{\epsilon\to 0}{\sim} \frac{\epsilon}{2}\Delta q(x)-\epsilon\frac{1-\alpha}{\theta}\langle\nabla U(x),\nabla q(x)\rangle\end{aligned}$$

Et finalement :

$$\lim_{\epsilon\to 0}\Big[\frac{P_\epsilon^{(\alpha)}-I}{\epsilon}\Big]q(x) = \frac{\Delta q(x)}{2}-\frac{1-\alpha}{\theta}\langle\nabla q(x),\nabla U(x)\rangle$$

université PARIS-SACLAY | ÉCOLE DOCTORALE de mathématiques Hadamard (EDMH)

Titre: Développement de méthodes mathématiques pour l'analyse de trajectoires conformationnelles en dynamique moléculaire

Mots clés: dynamique moléculaire, macromolécules, segmentation, équilibres locaux

Résumé: Les ressources informatiques actuelles (CPU/GPU) permettent de générer une grande quantité de trajectoires de simulation de la dynamique moléculaire (DM). Cependant, en plus de ne pas toujours tirer profit du contenu dynamique des trajectoires, les méthodes d'analyse actuelles visant à extraire de l'information pertinente au regard des macromolécules étudiées présentent certaines lacunes. Si la RMSD manque de finesse quant à la description des phénomènes d'équilibrations locaux, les méthodes de clustering sont sujettes à la détection d'artefacts: ces deux méthodes ont par ailleurs en commun une exploitation limitée de l'information dynamique que contiennent les trajectoires. Nous présentons au court de ce travail une façon d'aborder la question de la modélisation des trajectoires observées, et nous verrons dans quelles mesures celles-ci peuvent être modélisées par une EDS, ou mieux, une équation de Langevin; nous étudierons de plus la légitimité de l'utilisation de modèles de Markov. Nous développons également une méthode de segmentation des trajectoires, appelée κ-segmentation, permettant la détection et la quantification d'équilibres locaux tout en évitant la détection d'artefacts, comblant ainsi les déficiences des méthodes citées. Nous considérons enfin une modélisation de toute trajectoire de DM par un modèle jouet, permettant l'accès à une estimation de la dérive qui pourra être confrontée au gradient de l'énergie estimée.

Title: Development of mathematical methods for the analysis of conformational trajectories in molecular dynamics

Keywords: molecular dynamics, macromolecules, segmentation, local equilibrium

Abstract: Molecular dynamics (MD) simulations can produce nowadays huge amount of data using high-throughput CPU/GPU clusters. However, in addition not to consider the kinetic aspect of the data, the routine use of MD simulations for a study of real macromolecules present some drawbacks. The RMSD approach remains too vague for the description of local equilibrium, whereas classical clustering methods can provide artifacts: the both points of view share a limited exploitation of the information held by the MD data regarding the dynamics of the underlying process. We will present here a work towards the dynamics modelling of such trajectories, and we will show to what extend they can me described as resulting from an SDE, or better, a Langevin equation; the legitimacy of Markov models will also me addressed. Furthermore, we present a new method addressing the detection of local equilibrium, the κ-segmentation algorithm, allowing to detect and quantidy local equilibrium, while avoiding the detection of artifacts, and thusly, making up for the downsides of the quoted methods. Finally, we will show how to represent any real MD-data through a toy-model allowing drift estimation, which can be compared to the gradient of the estimated energy.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France