Cartographie génomique par analyse de signature ADN sur molécule unique issue de molécules en épingle à cheveux micro-manipulées par pinces magnétiques

François-Xavier Lyonnet Culinas Du Moutier

PSL UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'École Normale Supérieure de Paris

Cartographie génomique par analyse de signature ADN sur molécule unique issue de molécules en épingle à cheveux micro-manipulées par pinces magnétiques

Genomic mapping by DNA fingerprinting analysis using single molecule from hairpin shaped molecule and magnetic tweezers micromanipulation

Soutenue par
François-Xavier Lyonnet Culinas du Moutier
Le 18 décembre 2018

Ecole doctorale n° 564
Physique en ile-de-France

Spécialité
Bio-informatique

Composition du jury :

M. Claude THERMES Directeur de recherche émérite, Institut de Biologie Intégrative de la Cellule CNRS	*Président* *Rapporteur*
Mme Valentina BOEVA Directrice de recherche, Institut Cochin INSERM	*Rapporteur*
M. Thomas BRÜLS Chercheur, Institut de Biologie François Jacob CEA	*Examinateur*
M. Cédric VAILLANT Chercheur, Laboratoire de Physique ENS Lyon	*Examinateur*
M. Vincent CROQUETTE Directeur de recherche, Laboratoire de physique statistique, ENS - CNRS	*Directeur de thèse*
M. Hugues ROEST CROLLIUS Directeur de recherche, Institut de Biologie de l'ENS CNRS	*Directeur de thèse*

ENS | PSL

Sommaire

Financement et Collaboration

Cette thèse a été financée dans le cadre d'une Convention industrielle de Formation par la Recherche (CIFRE) par l'ANRT Et par l'entreprise Depixus (anciennement Picoseq).

La thèse a été réalisée à mi-temps au sein de Depixus, sous la tutelle générale de Chas André (CTO), ainsi que Jimmy Ouellet (Biologie) et Pol d'Avezac (Bio-Informatique). L'investissement humain tant que matériel de l'entreprise a été très importante dans la réussite de la Thèse, l'ensemble des constructions de molécules en épingle à cheveux ont été effectué à Depixus. Soit les constructions ont été effectuées par moi-même, avec l'aide de l'équipe, soit effectuées directement par l'équipe de biologie. Une partie des manipulations sur pince magnétique a également été faite au sein de Depixus. De même en informatique, les logiciels développés par Depixus ont été beaucoup utilisés, principalement à la fin de la thèse.

Département de Physique

École normale supérieure

Le reste du temps s'est déroulé principalement au LPS sous la direction de Vincent Croquette, avec une période dans le laboratoire de Hugues Roest Crollius.

La plupart des manipulations par pinces magnétiques ont été effectuées au sein du Laboratoire de Physique Statistique.

Remerciements

Essayer d'exprimer sa gratitude est un exercice difficile. Penser à chacun et à tous, être juste et authentique est un sacré défi. J'essaye de m'y atteler dans les prochains paragraphes. Quoiqu'il en soit, à tous, cité explicitement ou non ici, merci du fond du cœur.

Tout d'abord, je tiens à remercier Vincent Croquette, mon directeur de thèse. Alors que je sortais d'école d'ingénieur tu m'as tout d'abord accepté en tant que stagiaire, puis as accepté de devenir mon directeur de thèse. Tu as été très présent tout au long de ma thèse, et m'as fait confiance. Par tes talents de pédagogue et de scientifique, tu m'as permis de grandir dans la compréhension de la méthode scientifique et de ses tenants techniques. Tu m'as permis de m'affranchir de limite que je me fixais moi-même, et de m'ouvrir à de nombreux domaines. Tu as su me laisser l'espace pour faire des erreurs, tout en étant là pour m'accompagner et me soutenir. Merci pour ta patience et ton investissement qui sont source d'inspiration.

Je souhaite tout autant remercier Hugues Roest Crollius, merci d'avoir accepté de faire partie de cette aventure, et de t'être engagé dans la codirection de cette thèse alors que ce n'était pas le cœur de ta recherche. Mon travail n'aurait pu aboutir sans nos conversations longues et fructueuses. En plus de ton aide scientifique précieuse, Merci pour ton soutien humain qui m'a permis de dépasser mes appréhensions, et me sortir des moments difficiles.

Gordon Hamilton, c'est toi qui m'as fait entrer dans l'aventure qu'est Depixus. Alors que tu devais jongler entre toutes les tâches qui incombent à un dirigeant d'une jeune structure, tu as réussi à être présent pour chacun des salariés et notamment pour moi. Merci pour ta bienveillance, tes talents de dirigeant et ta profonde humanité. Le temps que tu m'as accordé m'a permis de faire grandir ma confiance en moi et ton amitié m'a été précieuse.

Jimmy Ouellet, merci pour ta patience. Quand j'ai commencé cette thèse, je n'avais presque aucune connaissance en biologie moléculaire. Grâce à toi tout d'abord puis l'aide de l'équipe, j'ai pu progresser énormément. Tu as été présent dès le départ et m'as offert ton aide et ton temps tout au long de la thèse. Tu as eu la patience de supporter toutes mes maladresses. Sans toi, je ne serais pas allé bien loin ! Pour tout ce que tu m'as apporté, j'éprouve beaucoup de gratitude.

Pol d'Avezac, arrivée lors de la 3e année de ma thèse, tu as été pour moi une personne structurante. Tu m'as permis de prendre une direction qui m'a mené vers le bout. Merci de m'avoir permis de me recentrer vers l'essentiel quand c'était nécessaire, et pour toute ton aide, notamment en informatique et traitement de données. Merci aussi pour ton humour décapant et ces discussions qui m'ont sorti de mon confort.

Chas André, je souhaite te remercier particulièrement pour la confiance que tu as su m'inspirer. Discret, mais toujours présent, tu as su me rassurer, et me rappeler que je n'étais pas seul.

Je souhaite ensuite remercier Valentina Boeva et Claude Thermes, mes rapporteur · e · s de thèse.

Merci pour le temps et l'énergie que vous avez mise dans la lecture attentive de mon manuscrit. Merci pour les remarques très pertinentes que vous avez pu me faire parvenir, et merci pour votre patience face à la difficulté que j'ai eue à rédiger.

Et bien sûr, je remercie également Thomas Brüls et Cédric Vaillant d'avoir accepté de faire partie de mon jury de thèse, pour leur lecture et leur question lors de ma soutenance, ainsi que leur patience.

Au sein de l'équipe ABCDLab, du Laboratoire de Physique Statistique, je tiens à remercie l'équipe permanente. Jean-François Allemand, qui avec son humour, son dynamisme, et son attention, a su me porter conseil et me secouer quand c'était nécessaire. David Bensimon pour ses apports scientifique, et Nicolas Desprat pour ses conseils avisés qui vont droit au but. Bertrand Ducos et les discussions que nous avons pu avoir. Je remercie également Marie-Cécilia Duvernoy, qui m'a apporté son soutien à tout au long de la thèse, qu'elle soit à Paris, à Grenoble, ou aux États-Unis, par nos longues discussions, tes conseils m'ont été précieux, et les bons moments passés avec toi aussi. Saurabh Raj, merci pour toute ton aide qui m'a permis de maitriser les pinces magnétiques et pour ton amitié. Caroline Peron Cane, merci pour ta capacité à voir les choses positivement et ta grande sensibilité. Ce sont des qualités exceptionnelles. Fatima, tu as été là dès le départ, nous nous sommes soutenus dans nos différents problèmes et j'espère que ça continuera. Samar Hobeib, tu m'a impressionné par ta capacité à gérer ton rôle de jeune parent et ta thèse en parallèle. Merci pour ces longues discussions. Tal Marcus, pour l'ambiance et la réflexion que tu as amené au labo, et ces parties de poker ;). Maxime Ardre, tu ne m'as pas seulement appris l'art de l'impression 3D, mais tu m'as aussi montré l'impératif de trouver l'équilibre entre perfectionnisme et pragmatisme. Magherita Peliti, pour ces folles discussions italiennes. Martin Rieu, tu m'as redonné de l'énergie à la fin de ma thèse, et j'ai tant apprécié nos défis, nos expérimentations, ta volonté d'être un bon enseignant. Jessica Valle Orero, pour ton yoga et ton énergie. Elena Khomyakova et Thao Tran, avec qui j'ai longuement échangé. Et pour finir, car sans eux la recherche n'avancerait pas, je souhaiterais remercier avec chaleur les agents administratifs. Annie Ribaudeau, qui m'a soutenu à chacune de mes venues. Benoit Paulet, Nora Sadaoui, Marie Gefflot, Fabienne Renia. Mais également le service informatique. Frederic Deprez, Frederic Ayrault, Zaire Dissi, Rémy Portier, Yann Colin. Et toutes les personnes que j'ai pu omettre, et qui pourtant ont participé à la réussite de ma thèse.

Au sein du Laboratoire Dyogen, je tiens tout particulièrement à remercier Alexandra Louis, qui en plus de m'avoir apporté beaucoup de soutien, et m'avoir redonné confiance à plusieurs reprises, m'a permis de découvrir le laboratoire alors que je cherchais mon chemin (métaphoriquement, le laboratoire est physiquement facile à trouver ;)). Je souhaite ensuite remercier toute l'équipe. Camille Berthelot, Yves Clément, Lambert Moyon, Guillaume Louvel, Élise Parey. Vous m'avez permis de découvrir une autre facette du métier de chercheur, et votre bienveillance ainsi que votre esprit de groupe ont été source d'inspiration pour moi.

Ma présence à Depixus a aussi été l'occasion de très belles rencontres. Nathalie Bouchard, merci pour ton attention toute particulière à mon égard, pour ton écoute, tes shots de bonne humeur et ton authenticité. Anais Leseur, merci de m'avoir supporté pendant ces longues heures de préparation d'échantillon où tu m'as expliqué avec patience tout ce que j'avais besoin de savoir. Sylwia Gorlach, merci pour tes préparations et les expériences que tu as effectuées pour moi, sans quoi je n'aurais pas pu présenter une partie de mes données. Merci également pour la profondeur de nos discussions, sur le travail, mais aussi la philosophie. Thibault Vieille, grâce à toi, il m'était impossible de m'endormir face à mon ordinateur :) Plus sérieusement, merci pour ta franchise, tes bons conseils, ton soutien, et tous les efforts que tu as faits pour essayer de m'apprendre l'optique et la physique ! Gael Radou, ta bonne humeur et tes conseils m'ont donné de force. David Salthouse, Jérome Maluenda, Laurène Giraut, Andreas Lefevre, merci pour votre humour, vos conseils, et votre âme d'enfant ! Didier Beaullit pour nos débats, merci également à Rémi Moulinas, Zhen Wang, Hua Bai, et à toutes les personnes que

j'aurais pu omettre.

Enfin, je voudrais remercier mes proches. Tout d'abord ma famille. Merci d'avoir été là pour moi, dans les bons moments comme les moins bons, merci à mon père et à Marie-Aude pour leurs corrections attentives. Et merci à Isabelle, Marie-Émilie et Raphaël. Une attention toute particulière pour Marie-Caroline, qui m'a aidé à faire des choix à des moments difficiles. Ensuite, je voudrais parler de mes colocataires et amis qui ont eu à me supporter pendant ces 4 longues années, avec mon humeur changeant, et toutes les choses du quotidien. Il serait trop long de tous vous citer, mais vous avez tout mon respect et ma gratitude. Un merci tout particulier à Delphine Mahieu pour ses corrections et son soutien, ainsi que Baptiste Galerne et Julia Golher pour leurs conseils avisés. Un merci et pardon à Sasha qui a m'a particulièrement supporté. Enfin, merci à tous mes amis qui de près ou de loin m'ont aidé durant la thèse. Je citerais des noms sans chercher à être exhaustif : Guillaume Cassier, Nicolas Jacques, Ariane Moreau, Simon Guillot, François de Chateleux, Cyprien Demaegdt, Noémie Fayol, et tous les autres. Merci aussi aux Scouts et Guides de France, et toute particulière aux membres des équipes territoriales du Badéo et de Paris Celeste, les membres du groupe d'Antony, et toutes les personnes rencontrés lors de mon cham. Merci également à l'association Starting-Block et toutes les personnes exceptionnelles que j'ai pu y rencontrer.

Un dernier merci à Abdenour Bouzane qui m'a permis d'effectuer ma première expérience de recherche.

Et pour terminer, j'ai une pensée toute particulière pour ma sœur, Anne-Claire, que j'aurais aimé avoir près de moi durant cette aventure.

Chapitre 1

Introduction

Sommaire du chapitre

1.1 Méthodes de séquençages

Dans ce chapitre, nous proposons d'expliquer le contexte de ce travail. La nouvelle approche en molécule unique que nous développons présente des avantages particuliers pour déterminer les marqueurs épigénétiques de l'ADN ou de l'ARN. Mais avant de parler de marqueurs épigénétiques, il faut d'abord déterminer la séquence de la molécule. Dans notre approche, le séquençage de novo complet est plus délicat que la caractérisation des marqueurs épigénétiques, le travail de cette thèse est donc de proposer une alternative à ce problème en réalisant un séquençage très partiel qui doit permettre d'identifier la molécule étudiée puis de déterminer les marques épigénétiques en les positionnant sur la séquence connue. Nous présentons d'abord l'évolution des méthodes de séquençages, leurs possibilités de détermination de marqueurs épigénétiques. Puis nous détaillons comment nous réalisons notre séquençage partiel et nous le comparerons aux techniques existantes.

1.1.1 Première génération

1.1.1.1 Principe de base

Séquencer l'ADN n'est pas une chose facile, les quatre bases ont une taille de $0.3nm$ et se ressemblent beaucoup à notre échelle. Dans la cellule le mécanisme de recopie de l'ADN agit avec une excellente

précision recopiant le génome quasiment sans erreur . L'enzyme qui accomplit ce petit miracle est appelée la polymérase, celle-ci recopie un brin d'ADN pour synthétiser son complémentaire avec une cadence de quelques centaines de bases par seconde. Pour séquencer l'ADN, il suffit donc de faire "parler" la polymérase. Nous allons voir qu'au cours du temps, les chercheurs ont réussi de mieux en mieux dans cette aventure. Revenons quelques instants sur le fonctionnement de la polymérase pour comprendre les méthodes de séquençage qui l'utilisent.

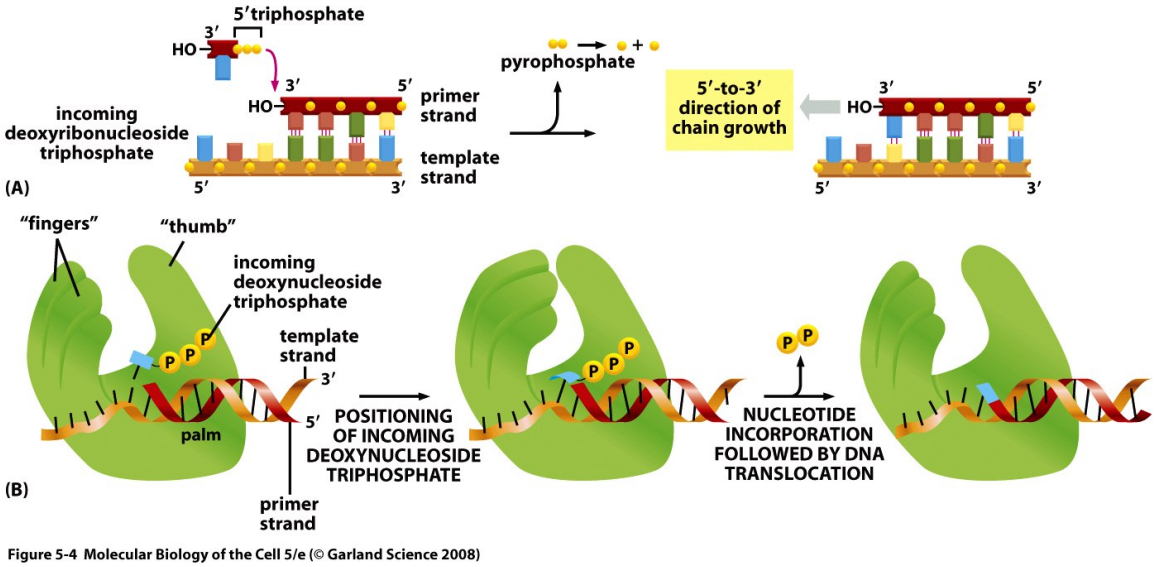


Figure 1.1: Incorporation d'un desoxyribonucléotide triphosphate par une polymérase (Alberts et al. 2002)

La polymérase (souvent schématisé comme une main droite) étend un brin incomplet en se servant du brin complémentaire comme modèle. Le site de la réaction (au niveau de la paume de la main et sous le pouce) stabilise transitoirement une base triphosphate en face de sa base complémentaire, puis si cette stabilisation dure assez longtemps, l'enzyme coupe les deux derniers phosphates et lie la nouvelle base au groupe OH libre à l'extrémité 3' du brin polymérisé. La réaction se répète, à chaque fois le fait que la base qui est complémentaire à l'autre brin conduit à une immobilisation transitoire plus longue que si la base n'est pas complémentaire. Cette sélectivité assure la première étape de la fidélité de la polymérase.

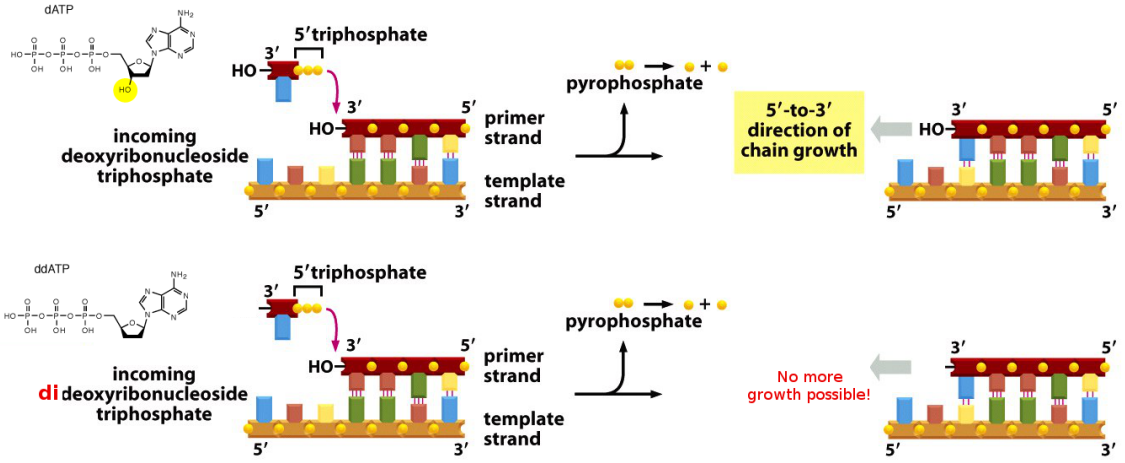


Figure 1.2: Incorporation d'un didesoxyribonucléotide triphosphate par une polymérase (Alberts et al. 2002)

Les chimistes savent fabriquer des analogues des briques élémentaires permettant de synthétiser l'ADN. Les ddNTP ressemblent aux dNTP, mais ils leur manquent le groupement OH auquel est attaché la base suivante, de ce fait leur incorporation est possible, mais elle bloque complètement la synthèse. Ces ddNTP sont ainsi des terminateurs.

1.1.1.2 Séquençage de Sanger

En 1977 (Sanger, Nicklen, et Coulson 1977) propose le séquençage de l'ADN, première génération.

Pour faire parler la polymérase, Sanger propose d'utiliser les ddNTs et les possibilités expérimentales de l'époque: On sait mesurer par électrophorèse la longueur d'un ADN simple brin avec la précision d'une base, on utilise une molécule simple brin correspondant à la molécule que l'on veut séquencer. On en fabrique au préalable un très grand nombre de copies. On hybride à son extrémité 3' un primer[1] contenant un marqueur fluorescent. Celui-ci assurera la détection de la molécule finale et le primeur permet à la polymérase de démarrer la copie. On rajoute des polymérases des dNTPs et on ajoute un des ddNTPs dans le volume de réaction, par exemple le ddATP. Les polymérases recopient l'ADN insérant majoritairement les dNTPs, mais de temps en temps elles se trompent et incorporent un ddATP (en face d'une base T). Ceci conduit à l'arrêt de la copie du brin en question. Une fois la réaction finie, on dénature toutes les molécules qui se sont répliquées et ont finies en incorporant en ddATP (un terminateur). On recommence la même opération pour chacun des autres types de ddNTPs (ddCTP, ddGTP, ddTTP) puis on fait un gel d'électrophorèse permettant de mesurer la longueur des quatre groupes de molécules. Le gel permet de lire directement la séquence.

[1] petite molécule d'ADN en simple brin (oligonucléotide) complémentaire au début du fragment d'ADN, près de son extrémité 3′ qui permet d'amorcer la réaction de polymérisation par la polymérase

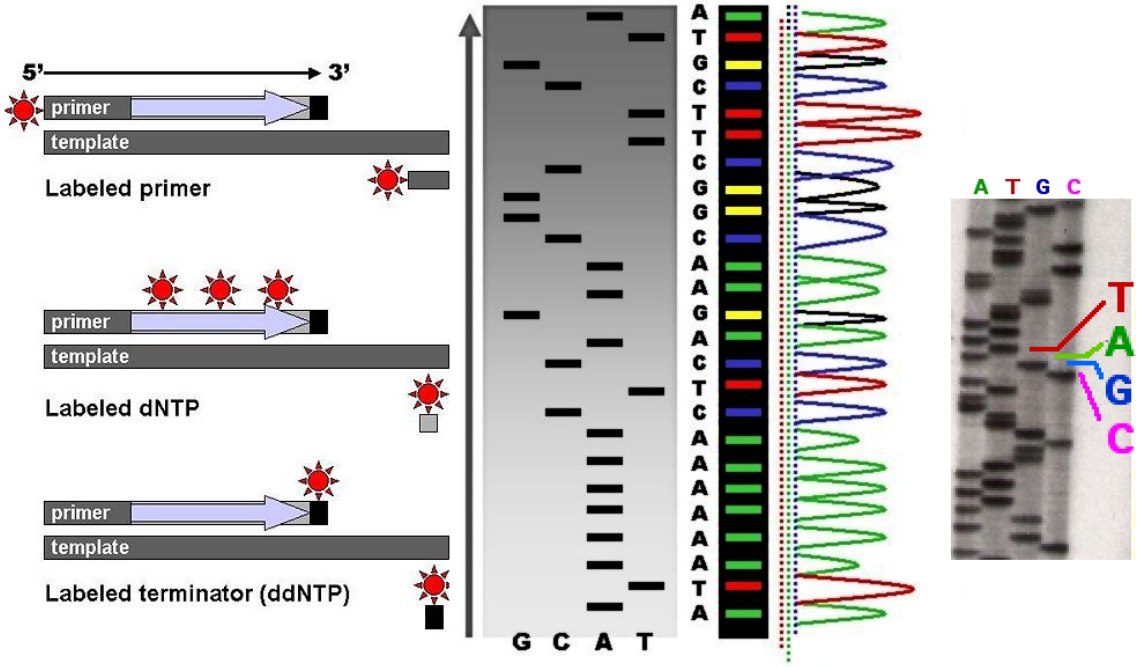


Figure 1.3: Séquençage par Sanger: À gauche le principe de l'élongation des copies de la molécule à séquencer (template). Un primer permet de démarrer la réaction de la polymérase à un point précis de la séquence. Un ddNTP va bloquer l'élongation lors de l'incorporation de la base complémentaire. Plusieurs stratégies de détection sont possibles : en utilisant des marqueurs fluorescents soit sur le primer, sur les nucléotides ajoutés ou sur le terminateur. Au centre, principe de l'image des quatre lignes d'électrophorèses correspondant aux quatre groupes de molécules associées chacune à un ddNTP. À droite image d'un gel.

Mais cette méthode possède également des défauts: on séquence une molécule à la fois, le gel d'électrophorèse permet de déterminer des séquences d'environ mille bases, le séquençage complet d'un génome est donc long, coûteux, et fastidieux. L'information épigénétique des molécules est perdue lors de l'amplification.

Malgré ces défauts, le premier génome humain a été séquencé en utilisant la méthode de Sanger en 2001 (Consortium 2001), mais la méthode, bien qu'exploitable, reste mal adaptée au séquençage des génomes. Elle reste utilisée pour des besoins de séquençage simple de nos jours, notamment grâce à sa longueur de lecture (readlength) relativement grande et à son faible taux d'erreur.

1.1.2 Seconde génération - Next generation sequencing

Le principe qui a amené au séquençage de masse s'est invité dans la compétition farouche qui a eu lieu lors du séquençage du génome humain: comme le génome humain contient 3 milliards de bases et que le séquençage de Sanger ne peut en lire que mille bases à la fois est apparue la nécessité de développer une stratégie d'assemblage des séquences obtenues pour couvrir au mieux le génome humain. La première stratégie, que nous appellerons raisonnable, consiste à séparer le génome en morceaux en suivant un découpage hiérarchique permettant un réassemblage facile. Mais, le fait que certains morceaux d'ADN soient difficiles à séquencer et que le beau découpage comportait quelques erreurs inévitables a fait apparaître les difficultés de cette stratégie. En 1995 Greg Venter a lui proposé une approche radicalement différente: le "shot gun" ((Fleischmann et al. 1995), au lieu de rigoureusement

couper le génome de façon très ordonnée, il propose de tout couper de façon aléatoire, de séquencer les morceaux et de les réassembler comme un puzzle en utilisant les recouvrements de séquences se produisant entre les différents morceaux. Cette proposition qui a rencontré une forte résistance à ses débuts est vite devenue la méthode la plus efficace, et ceci grâce à la montée en puissance des ordinateurs qui ont permis l'assemblage du génome à bas coût.

1.1.2.1 Basée sur le principe du *shotgun sequencing*

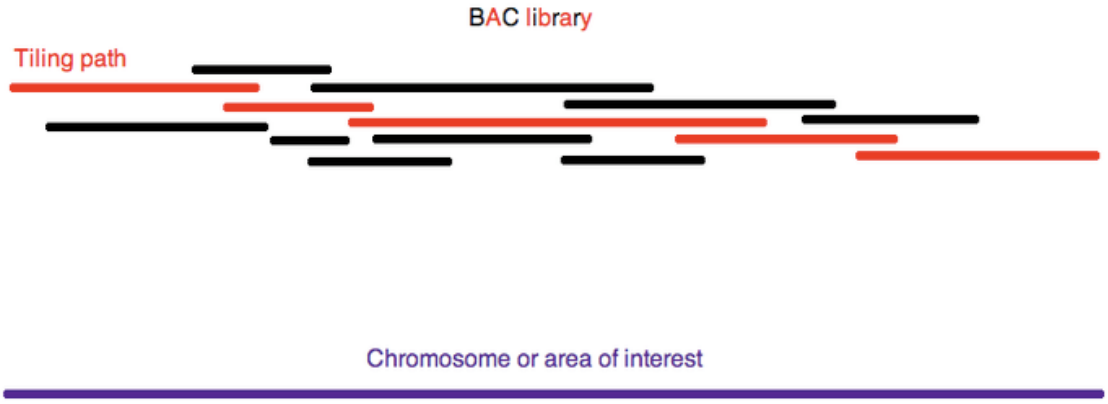


Figure 1.4: Shotgun sequencing: cette technique basée sur l'assemblage de petits fragments d'ADN, utilise le recouvrement partiel de ces fragments pour reconstruire une grande séquence continue (en rouge).

On coupe l'ADN d'un génome de façon aléatoire pour produire un grand nombre de petites molécules qui présentent des recouvrements de séquences. En utilisant ces recouvrements, on reconstruit la molécule originale. Cette technique fait l'hypothèse que la séquence de l'ADN est pseudo aléatoire.

1.1.2.2 Le séquençage à grande échelle et très grand parallélisme

Le ressemelage du génome est possible, car sa séquence est presque aléatoire, seul un petit nombre de régions du génome présentent des motifs répétés qui posent de sérieux problèmes (et n'ont pas été séquencées dans les premiers génomes). Par ailleurs, si le génome était parfaitement aléatoire, la connaissance de 16 bases consécutives permet de façon statistique d'avoir une seule occurrence de cette séquence dans l'ensemble du génome. En pratique une vingtaine de bases sont suffisantes.

La seconde génération de séquenceur s'est faite avec ces concepts : au lieu de séquencer très bien une molécule, on a cherché à séquencer un grand nombre de molécules simultanément quitte à avoir une longueur de lecture réduite.

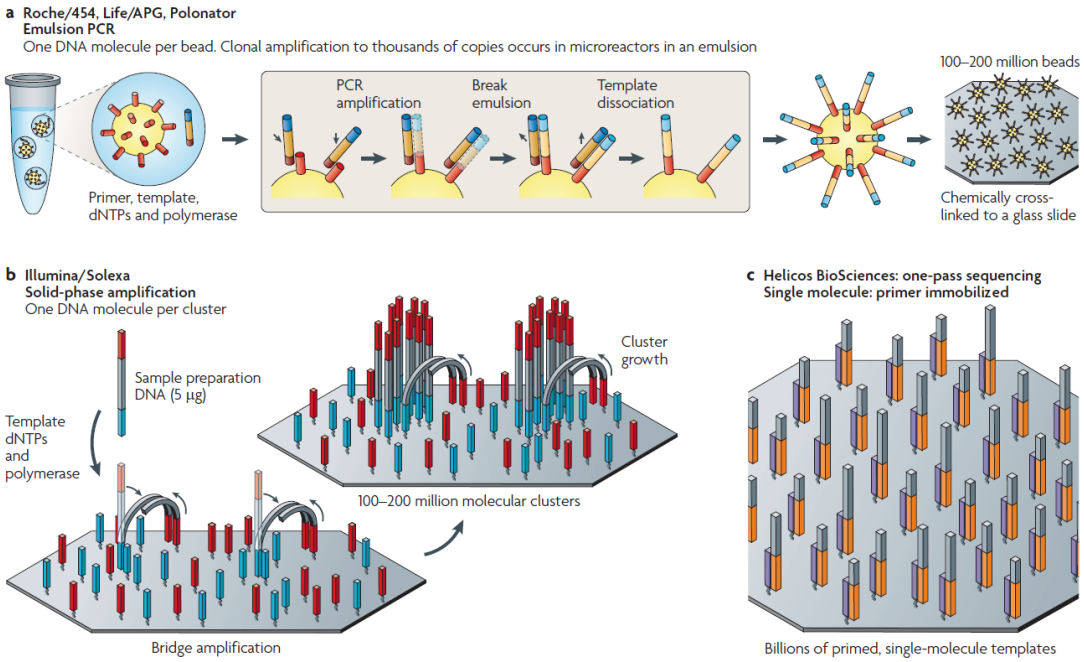


Figure 1.5: Présentation des méthodes de préparation des échantillons en vue du séquençage haut débit: Les méthodes de PCR en émulsion (en haut), de “bridge amplification” (en bas à gauche) et sans amplification (en bas à droite) sont illustrées. La PCR en émulsion est utilisée par Roche/454 et IonTorrent, celle de de “bridge amplification” par Solexa et maintenant Illumina. Helicos est une stratégie molécule unique qui utilise une seule molécule pour la détection et ne nécessite pas d’amplification. (Khan 2014)

On coupe l’ADN génomique en petites molécules, on s’arrange pour fabriquer un grand nombre de colonies différentes contenant chacune 10 à 100 000 copies d’une seule de ces molécules. On réalise cette opération, par exemple en faisant une réaction de PCR sur une émulsion de petites gouttes de réactifs contenant chacune au plus une copie de l’ADN génomique, une bille magnétique recouverte de streptavidine, un primer universel avec une biotine, de la polymérase et des dNTPs. On réalise des cycles de PCR et à la fin chaque bille est recouverte de 10 à 100 000 copies de la molécule d’ADN génomique.

1.1.2.3 La méthode Illumina

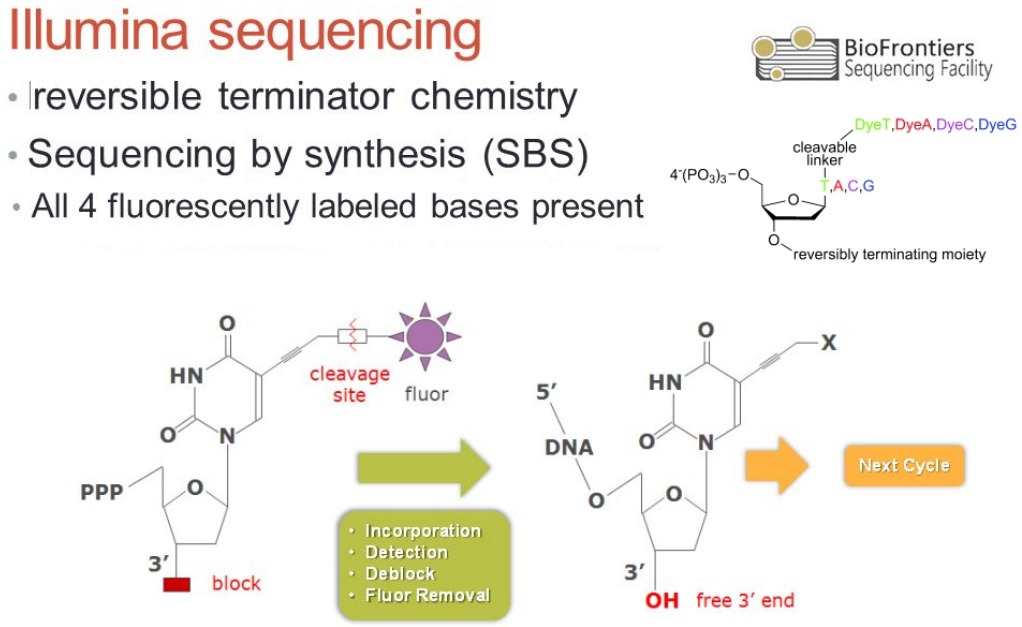


Figure 1.6: dNTP réversible: La stratégie d'Illumina consiste à modifier les bases de l'ADN pour leur associer à chacune un fluorophore clivable de couleur différente ainsi qu'un groupe bloquant la polymérisation.Ainsi, lorsque l'une de ces bases est ajoutée par la polymérase, celle-ci est détectable par la couleur de la fluorescence, tant que le groupement terminateur est présent. Mais ces deux groupements chimiques peuvent être clivés, ce qui rend à la fois la base invisible, et permet l'ajout d'une nouvelle base à la chaîne d'ADN en croissance sur le 3'OH du ribose. Le clivage restaure la base modifiée en la rendant exactement équivalente à une base d'ADN normale.

Le séquençage en temps réel par la méthode Illumina s'appuie sur une technologie clé: la préparation de dideoxynucléotides "terminateurs" réversibles, qui comprend trois principes.

- Fabrication de dNTPs fluorescents avec une couleur différente pour chaque base.
- Insertion d'un groupe bloquant la réplication.
- Ces deux modifications sont clivables par un composé chimique.

La technologie Illumina inclut une phase de préparation des échantillons qui consiste à polymériser un très grand nombre de copies du fragment de séquence inconnue, tout immobilisée sur une surface de verre par une extrémité, et regroupé sous forme de "clusters" ou colonies. Une plaque de verre comprend des centaines de millions de colonies, composées de séquences inconnues différentes d'une colonie à l'autre. Ce stratagème est important afin d'obtenir un signal facilement mesurable. En effet, toutes les molécules d'une colonie ont la même séquence et sont utilisées de manière synchrone comme matrice (brin complémentaire) pour la polymérisation à l'aide des bases modifiées. À l'étape de polymérisation n la même base fluorescente est incorporée dans toutes les molécules d'une colonie au même endroit de la séquence, et le signal de fluorescence est donc la somme de la fluorescence de 10 000 à 100 000 molécules identiques. En lisant la couleur émise par chaque colonie, on détermine ainsi la nature de la base n de chacune des molécules. Après cette lecture, on rince l'échantillon avec un produit qui clive le groupe fluorescent et le groupe bloquant la polymérisation de la base suivante. On peut alors procéder au séquençage de la base $n+1$ de la même façon (Figure 1.7).

1.1.2.3.1 La méthode Illumina génère des séquences de longueur limitée: pourquoi?

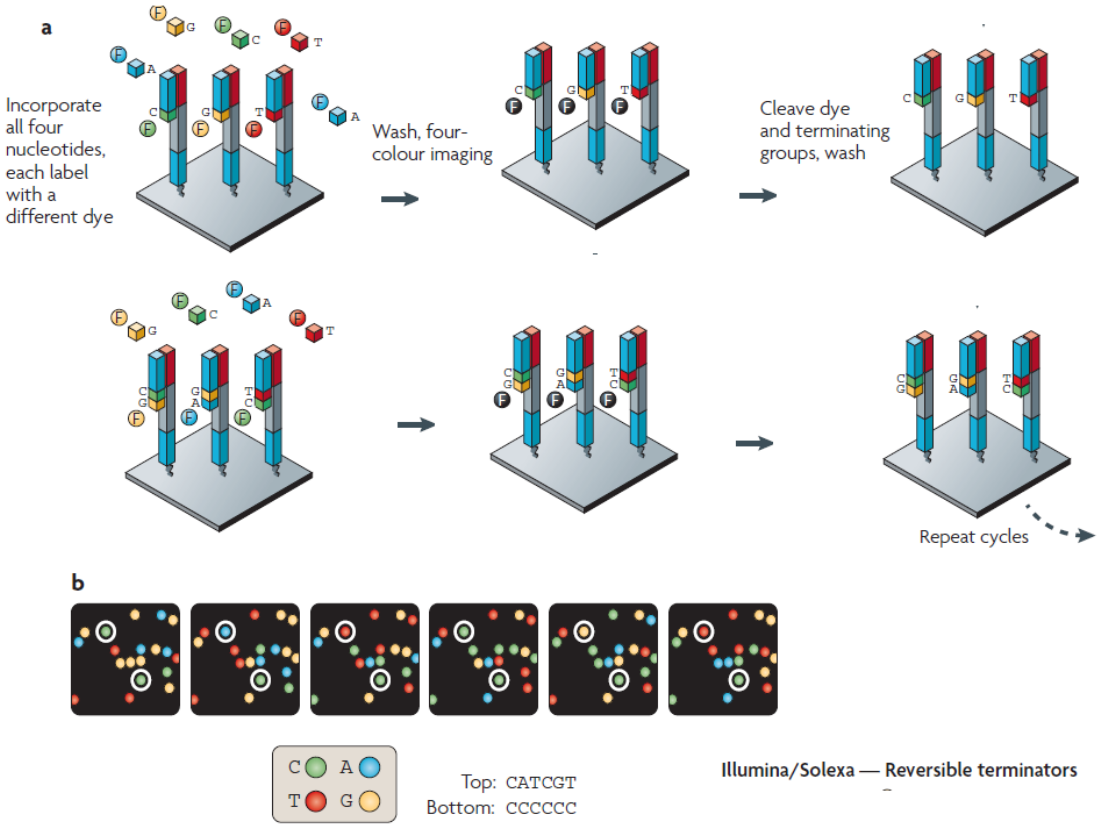


Figure 1.7: Les étapes du séquençage par fluorescence d'Illumina : À chaque étape les colonies de molécules (ici représentées par une seule molécule) présentent une couleur de fluorescence qui permet d'identifier la base incorporée. À la première étape, les bases fluorescentes sont incorporées par la polymérase, puis un rinçage laisse uniquement les molécules fluorescentes qui sont lues par une caméra. Un agent chimique est introduit afin de cliver les groupes fluorescents et de restaurer le groupe 3'OH, et le cycle peut être répété. La caméra observe des taches colorées (en bas) correspondant à chaque colonie qui sont lues à chaque cycle.

La méthode Illumina présente un avantage très important par rapport à la méthode de Sanger basée sur des gels d'électrophorèse: le parallélisme. Une seule plaque de verre de la taille d'une lame de microscope suffit à produire des centaines de millions de séquences! La limitation de cette méthode vient des petites erreurs qui se produisent à chaque étape de polymérisation: les réactions chimiques de clivage et déblocage et surtout la réaction de la polymérase ont une chance de succès d'environ 99%. Cela signifie qu'au cycle 1, environ 1% des molécules n'auront incorporé la base n et le signal fluorescent aura 99% de sa valeur théorique totale. Cela ne pose pas de problème de détection, mais au cycle 2, ces 1% de molécules incorporent la base fluorescente 1 tandis que la majorité des molécules restantes incorpore la base 2. Le signal de fluorescence n'est donc plus parfaitement pur. Au cycle 3, à nouveau 1% des molécules ne réagissent pas et vont incorporer la base $n-1$. Au fur et à mesure que le nombre de cycles augmente, on observe ainsi une désynchronisation de la colonie et le signal de fluorescence est de moins en moins pur. Après une centaine de cycles, sa couleur est le résultat d'un mélange et il n'est plus possible de déterminer la nature de la base. La longueur de lecture de ces méthodes est donc limitée à environ 150 paires de bases.

Un procédé déjà éprouvé et exploité par la méthode de Sanger, appelé “paired-end sequencing” permet de pallier au moins partiellement cette limite. Ainsi, grâce à l’ajout d’une petite séquence connue (appelée “adaptateur”) aux extrémités de la séquence inconnue, il est possible d’identifier les deux lectures qui proviennent chacune d’un côté de cette séquence. Si le fragment de séquence inconnue mesure 250 bp par exemple, il est alors théoriquement possible de séquencer les deux extrémités sur 150 bp qui se chevaucheront légèrement, pour obtenir finalement la séquence complète de 250 bp. Une variante de cette approche, appelée “mate pair sequencing” consiste à séquencer les extrémités de fragments beaucoup plus longs (2-5 kb), non plus pour reconstituer une séquence plus longue, mais pour faciliter l’assemblage des lectures. Cette approche nécessite cependant des étapes supplémentaires, et ne permet pas de pallier directement à la limite de longueur de lecture de la méthode Illumina.

1.1.2.4 NGS: une révolution dans le séquençage

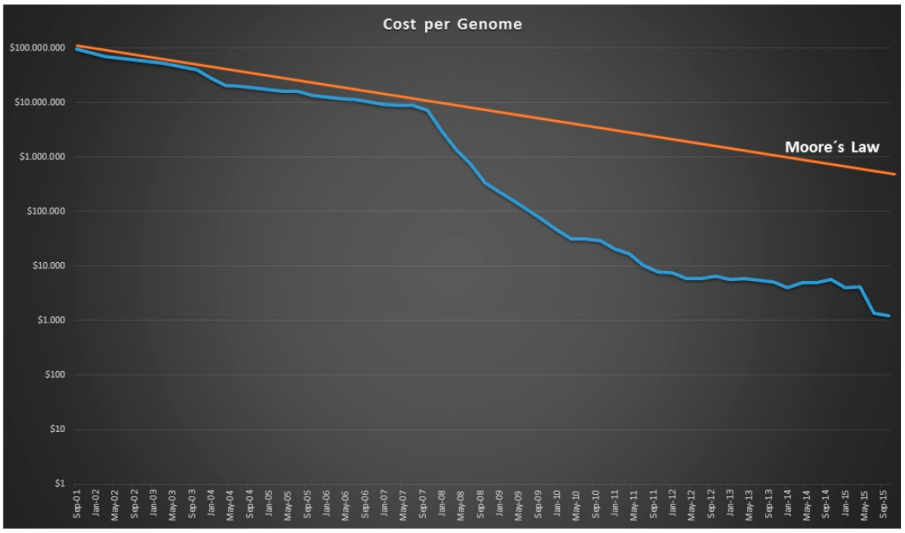


Figure 1.8: Le coût du séquençage a diminué de façon très rapide avec le temps. Ce graphique montre que cette diminution a été plus rapide que la loi de Moore utilisée en électronique.

En conclusion, l’émergence de la technologie Illumina a représenté une véritable révolution dans le domaine du séquençage. Son parallélisme impressionnant a eu pour conséquence une diminution très importante du coût du séquençage par base, en même temps qu’une diminution du temps de séquençage requis. Ainsi, avec les appareils “X-ten” vendus par l’entreprise aujourd’hui, il est possible de reséquencer pour environ 1000 € un génome humain complet en 2 jours. Cette application en génomique humaine est possible, car une version complète du génome humain existe déjà. Dite “de référence”, celle-ci avait été obtenue après de longs et fastidieux efforts de cartographie et de séquençage par la méthode de Sanger (Consortium 2001). Le reséquençage du génome humain est particulièrement efficace pour identifier des mutations d’une seule base (single nucleotide polymorphism, SNP) ainsi que les petites insertions ou délétions d’ADN. D’autres applications en génomique fonctionnelle ont profité de cette technologie, comme le séquençage des transcrits (RNA-seq), ou le séquençage de régions d’ADN reconnues et fixées par une protéine (ChIP-seq).

En revanche, la petite taille des lectures est une limite importante pour le séquençage dit “de novo”, qui consiste à obtenir la séquence d’un organisme sans version initiale de référence. L’assemblage des lectures courtes est en effet rendu difficile dans les régions répétées du génome (généralement des éléments transposables) dont la longueur est plus élevée que la longueur des lectures Illumina.

Beaucoup de génomes ont ainsi été séquencés sous une version très approximative par la méthode Illumina. Par opposition au génome humain, au génome de la souris, au génome de l'arabette ou de la drosophile qui ont tous bénéficié d'un séquençage "complet", avec une reconstitution de la séquence des chromosomes d'un télomère à l'autre, ces génomes partiels sont reconstitués sous forme de fragments, généralement sans position ni orientation précise les uns par rapport aux autres: on parle de "draft genome sequence", ou séquence "brouillon". Les "trous" de séquence sont principalement dus aux régions très biaisées en composition (ex: riches en nucléotides G et C) difficiles à séquencer, ou aux séquences très répétitives. Les répétitions posent surtout un problème d'assemblage: une lecture couvrant une séquence répétée, que celle-ci soit issue d'un centromère, d'un télomère ou d'un élément transposable, ne peut être placée dans le génome de manière fiable. Finalement, les modifications biochimiques de l'ADN, comme la méthylation, sont des informations importantes qui sont perdues lors du séquençage classique par la méthode de Sanger ou d'Illumina, à moins de réaliser des traitements préalables.

De nouvelles méthodes commencent à voir le jour. Dites de "troisième génération", elles résolvent le problème principal de la longueur des lectures et sont susceptibles de permettre un nouveau bond en avant dans le volume et la qualité des génomes séquencés.

1.1.3 Les méthodes de troisième génération

Les méthodes de troisième génération empruntent certains concepts aux méthodes vues précédemment, mais partagent un point commun nouveau: le séquençage est réalisé et observé sur des molécules uniques. Tandis que les précédentes réalisent le séquençage en solution (méthode de Sanger) ou sur surface solide (méthode Illumina) sur des milliers ou millions de molécules identiques à la fois, ces nouvelles stratégies génèrent un signal sur une seule molécule.

1.1.3.1 PacBio et Helicos.

Dans les années 1990 ,(Moerner et Kador 1989) à montrer que l'on pouvait observer la fluorescence d'une seule molécule avec un rapport signal sur bruit suffisant, il a obtenu le prix Nobel en 2014. Comme il est possible de fabriquer des nucléotides fluorescents avec une couleur différente pour chaque base, il est apparu naturel d'utiliser une polymérase pour les incorporer, puis de suivre la fluorescence associée à chaque base pour déterminer la séquence.

Deux stratégies sont apparues pour atteindre ce but. La première appelée Helicos a été proposée par S. Quake (Harris et al. 2008). Elle reprend la méthode d'Illumina en la transposant à l'échelle de la molécule unique: une base est incorporée à la fois, et la couleur de la fluorescence est observée. Mais la qualité du signal de fluorescence a limité la longueur de lecture à une trentaine de bases, ce qui n'a pas permis un développement important. La deuxième stratégie a été proposée par W. Web (Levene et al. 2003), et consiste à laisser la polymérase incorporer les bases sans interrompre la synthèse du brin d'ADN, et à détecter la fluorescence de chaque base pour obtenir la séquence. Ici, le groupe fluorescent permettant l'identification des bases est lié après les trois groupements phosphates du nucléotide libre. Ainsi, il est clivé lors de l'incorporation de la base ce qui conduit à un brin d'ADN synthétisé natif, sans modifications.

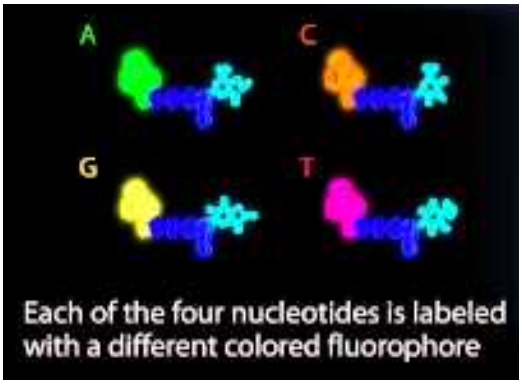


Figure 1.9: Séquençage par Pacific Bioscience : Les quatre bases ont été modifiées pour leur ajouter un fluorophore qui sera clivé par la polymérase au moment ou celle-ci incorpore la base dans l'ADN.

Mais la DNA polymérase ne fonctionne bien que quand la concentration des dNTPs est assez grande. Dans ce cas, tout le volume réactionnel est fluorescent et il est impossible de voir le nucléotide incorporé. W. Web a proposé une solution élégante pour contourner ce problème:

1.1.3.1.1 Utiliser un guide d'onde

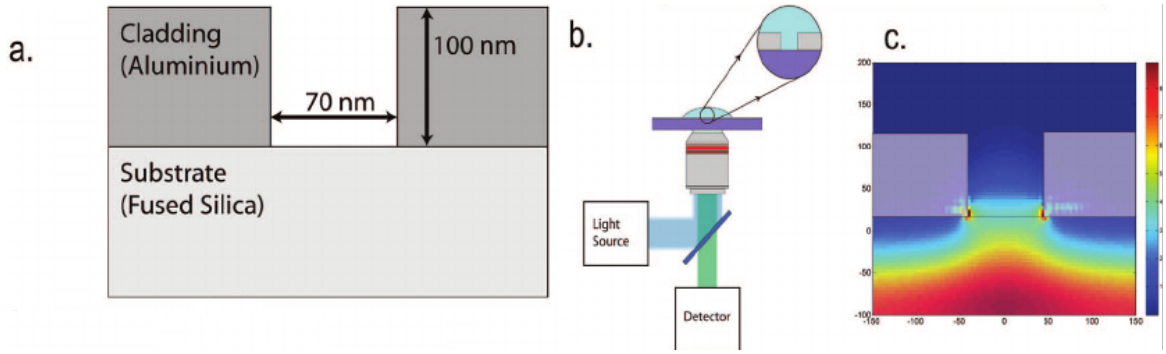


Figure 1.10: Le guide d'onde à zéro mode de Pacific Bioscience : Afin de pouvoir observer la fluorescence des bases incorporées par la polymérase qui a besoin d'une concentration substantielle de nucléotides, Pac-Bio utilise une lamelle couverte d'aluminium et percée de minuscule trou de 70 nm de diamètre. Ceux-ci sont si petits qu'ils ne laissent pas passer la lumière ou plus exactement il ne laisse passer qu'une onde évanescente centrée sur ce petit trou où est attachée la polymérase. De cette façon, le volume éclairé est extrêmement petit et la concentration élevée des nucléotides n'est pas un problème.

Un trou de 70 nm de diamètre dans un film d'aluminium constitue un guide d'onde qui ne laisse passer qu'une onde évanescente dans les premiers 50 nm du trou. Si une ADN-polymérase est attachée par une biotine et une streptavidine au fond du trou, elle est juste au milieu de cette onde évanescente. Comme le volume éclairé est petit, la concentration en dNTPs peut être assez grande pour assurer le bon fonctionnement de la polymérase, mais avec un petit nombre de nucléotides éclairé. Les quelques molécules en solution dans la zone éclairée bougent trop vite pour donner du signal, mais elles deviennent visibles lorsqu'elles s'immobilisent lors de l'incorporation par la polymérase. Le séquençage est réalisé en observant en temps réel les flashs lumineux produits par l'incorporation des bases consécutives comme sur la figure :

1.1.3.1.2 Pacific Bioscience sequencing

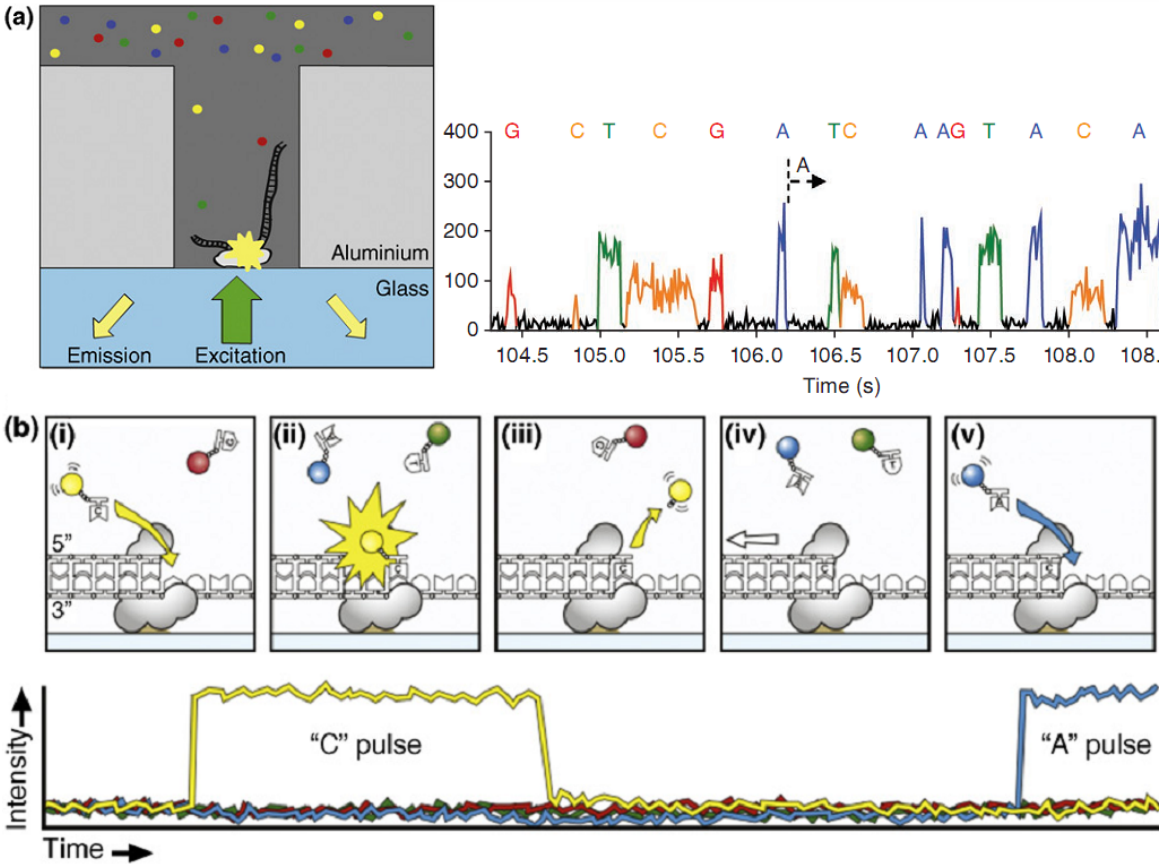


Figure 1.11: Principe du séquençage en temps réel de Pac-Bio : une caméra observe en temps réel la polymérase dans son puits. Les bases individuelles fluorescentes diffusantes en solution bougent trop vite pour donner un signal de fluorescence détectable, par contre lorsqu'elles entrent dans la polymérase elles s'immobilisent quelques millisecondes donnant alors un signal visible (a). En observant les impulsions de lumières (b) de différentes couleurs, on retrouve la séquence. L'incorporation de chaque base est stochastique et donc de durée très variable, parfois très courte (t = 107.25s)

Un des problèmes dans le séquençage de Pac-Bio est que le temps d'incorporation d'une base est stochastique. La distribution de ces temps serait une exponentielle si le processus suivait une loi de Poisson. C'est presque le cas comme on peut le voir dans la figure fig. 1.12. Cette distribution contient beaucoup d'évènements avec des temps d'incorporation très courts qui dans certains cas ne seront pas détectés, conduisant à une erreur.

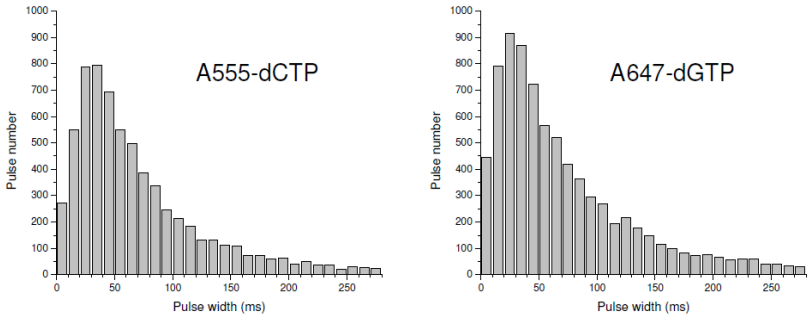


Figure 1.12: Temps d'incorporation des bases CTP et GTP dans le séquençage de Pac-Bio: dans les expériences de molécule unique, la cinétique des réactions est donnée par une loi de Poisson qui correspond ici à une exponentielle. C'est presque le cas ici, sauf au temps très court, car une fois que la polymérase a incorporé une base il lui faut un peu de temps avant de pouvoir en incorporer une autre. Cette très grande variabilité du temps d'incorporation pose des problèmes: si le temps d'incorporation est très court (ce qui arrive assez souvent) il est assez facile de ne pas voir le pulse fluorescent correspondant.

La préparation des molécules pour le séquençage de type PacBio se fait en ajoutant une boucle à chaque extrémité de la molécule double brins comme sur la Fig. fig. 1.13.

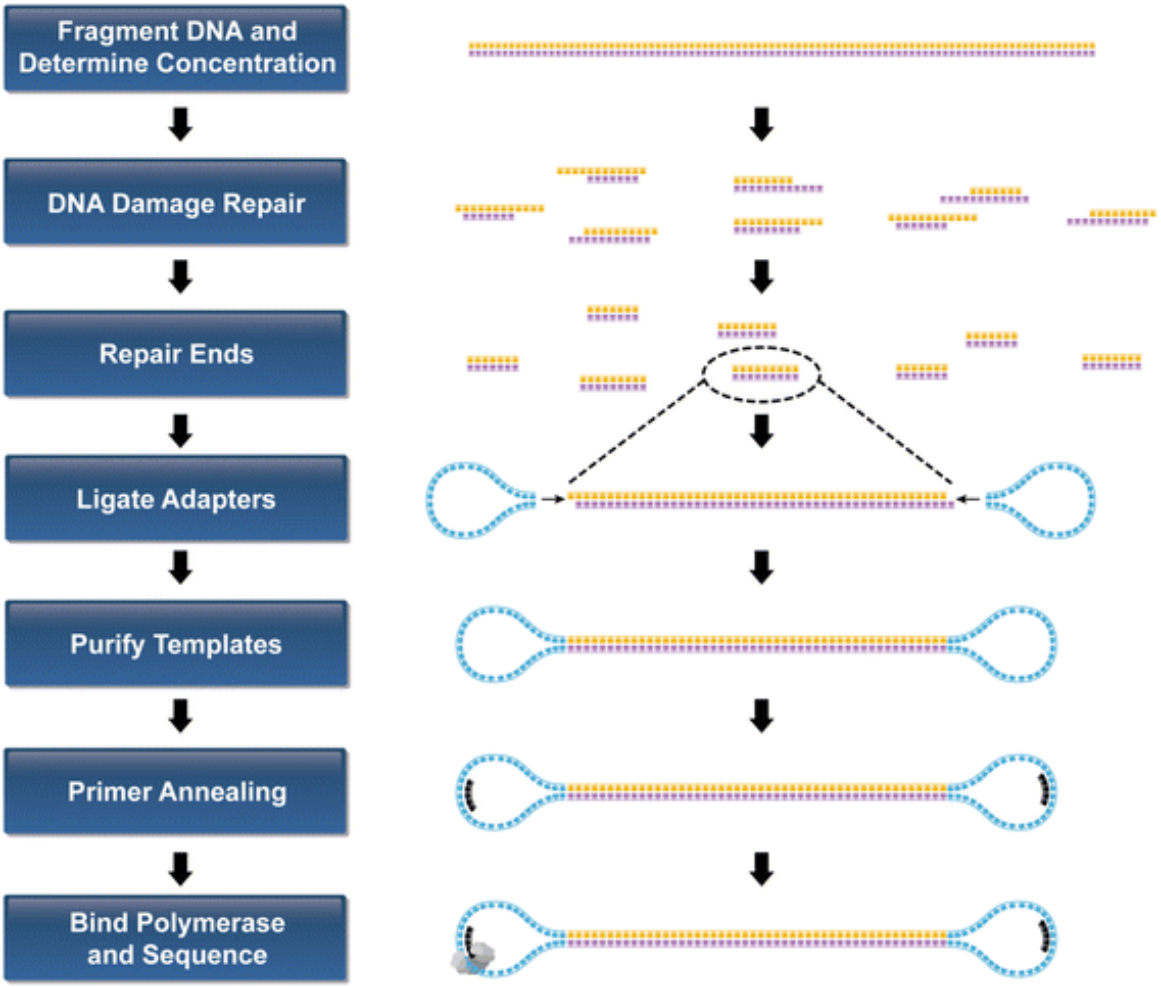


Figure 1.13: Préparation de la librairie d'ADN utilisée dans Pac-Bio: le morceau d'ADN double brin (en marron) est ligué à deux boucles d'ADN simple brin (en bleu). Un oligonucléotide (en noir) est hybridé à l'intérieur de la boucle pour servir d'amorce à la polymérase qui vient s'associer à cette amorce et sera attachée au milieu du trou dans la plaque d'aluminium grâce à une molécule de biotine.

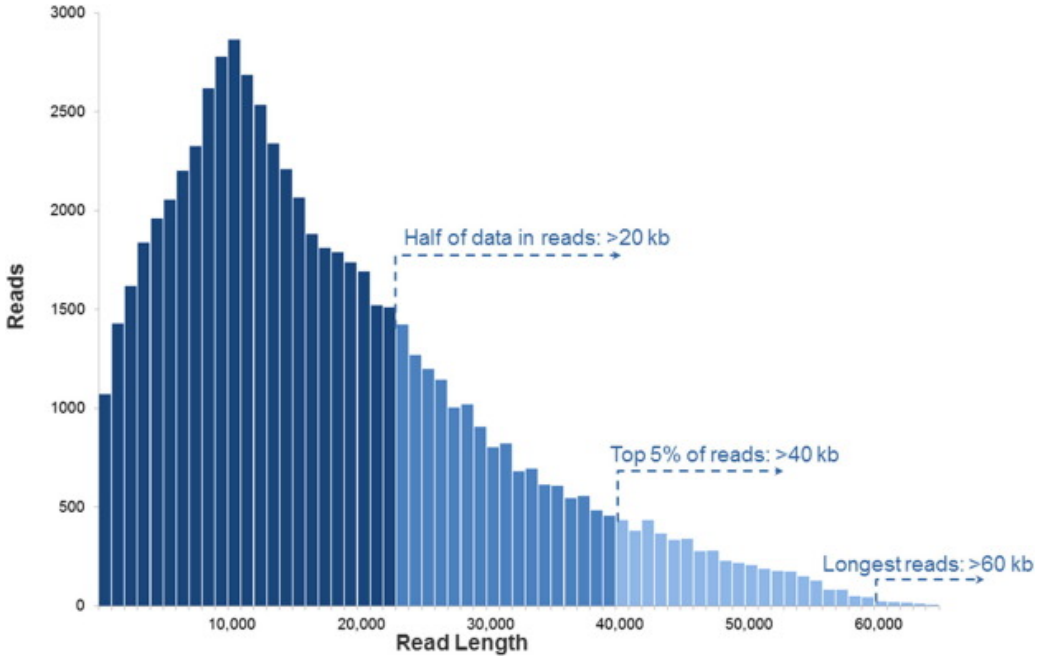


Figure 1.14: Longueur de lecture dans le séquençage de Pac-Bio : De façon intéressante, la polymérase incorpore des bases fluorescentes sur de grandes longueurs comme on peut le voir ici. La forme exponentielle de cette courbe à grande distance indique qu'une seule réaction chimique est impliquée dans ce phénomène : la polymérase est endommagée peu à peu lors des réactions de fluorescence . Par contre la longueur de lecture de quelques dizaines de kB est un atout pour la méthode.

Comme le montre la distribution des longueurs de séquences obtenues, cette stratégie est très efficace, et peut générer des lectures de plus de 50 kb. Cependant, le taux d'erreur de 13% est élevé. En effet, comme le temps d'incorporation des nucléotides est stochastique, il peut être trop court pour permettre la détection d'une incorporation. Par ailleurs un mauvais nucléotide peut s'attarder dans la polymérase avant de s'en aller, conduisant à une fausse détection.

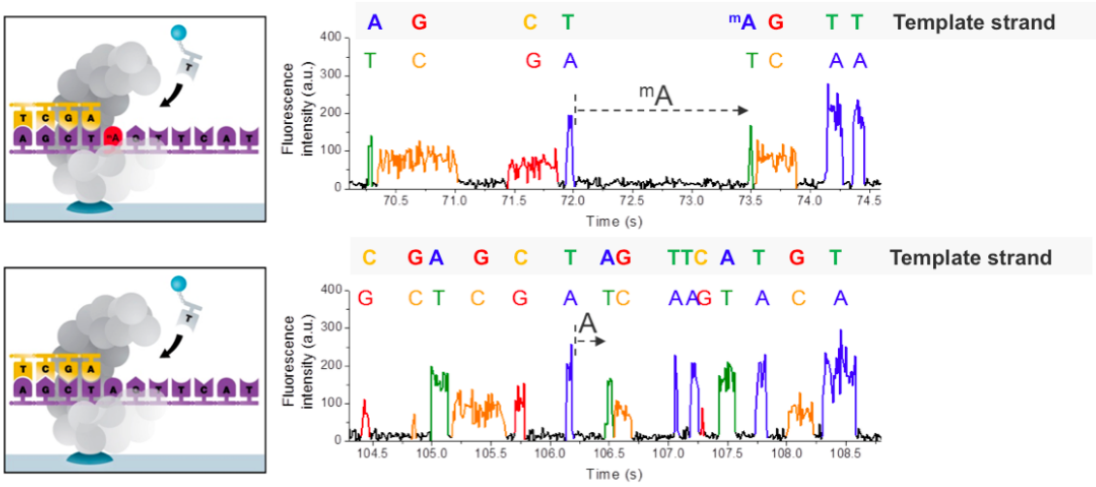


Figure 1.15: Détection des marqueurs épigénétique chez Pac-Bio. Un deuxième temps caractérise la cinétique de la polymérase : le temps séparant deux incorporations successives. Si l'ADN template comporte une base méthylée (ici une 6mA), la polymérase met plus temps à incorporer la base suivante ce qui permet de détecter la présence de cette marque épigénétique. Ce temps entre incorporations suit aussi une loi de Poisson qui présente de grandes fluctuations, pour être sûr que nous avons bien affaire à une base méthylée il faut disposer d'une statistique suffisante.

1.1.3.1.3 les avantages et inconvénients de Pacific Bioscience

- Comme la molécule d'ADN forme un anneau, il est possible de séquencer plusieurs fois la molécule et son brin complémentaire en laissant la polymérase faire plusieurs tours si l'anneau est suffisamment petit.
- Les erreurs sont aléatoires et indépendantes. Ainsi, le fait de réaliser plusieurs tours sur une molécule circularisée réduit le taux d'erreur. À l'heure actuelle, c'est le séquençage de troisième génération le plus précis.
- Les modifications épigénétiques changent la cinétique d'incorporation: la méthode détecte bien les méthylations des Adénosines (6mA), mais moins bien les 5mC.
- Mais cette précision a un prix, suivant les cas il faut faire 10 à 30 tours et la longueur de lecture est réduite très fortement.
- Le coût de l'instrument (350000$) et son encombrement sont importants au regard du débit faible, relativement aux autres méthodes de troisième génération.

1.1.3.1.4 Le séquençage par nanopore.

La nature a fabriqué des protéines qui s'insèrent dans une membrane phospholipidique et établissent une communication entre les deux côtés avec un trou nanométrique. Depuis 1990 les chercheurs ont réalisé qu'ils pouvaient faire passer des molécules d'ADN simple brin par électrophorèse dans des pores et détecter leur passage par le changement de conductivité provoqué par ce passage. Très vite, ils se sont rendu compte qu'un simple brin poly-A n'avait pas la même signature qu'un poly-T etc. Ils ont donc proposé que l'on puisse séquencer l'ADN par cette méthode. Mais il a fallu beaucoup de temps et d'énergie pour passer du concept à un une réalité.

La grande difficulté qui a longtemps bloqué le processus c'est que l'ADN simple brin passe beaucoup trop vite dans le nanopore, une base passe en 10 microsecondes typiquement tandis que le courant mesuré lors du passage des ions dans le nanopore est très petits 50 pA et le changement de ce courant

qu'il faut détecter est encore plus faible quelques pA. Dans ces conditions le rapport signal sur bruit était trop défavorable. La grande avancée a consisté à ralentir le déplacement de l'ADN en le couplant d'abord à une polymérase puis maintenant à une hélicase, ce qui conduit à typiquement 100 bases par secondes augmentant le rapport signal sur bruit.

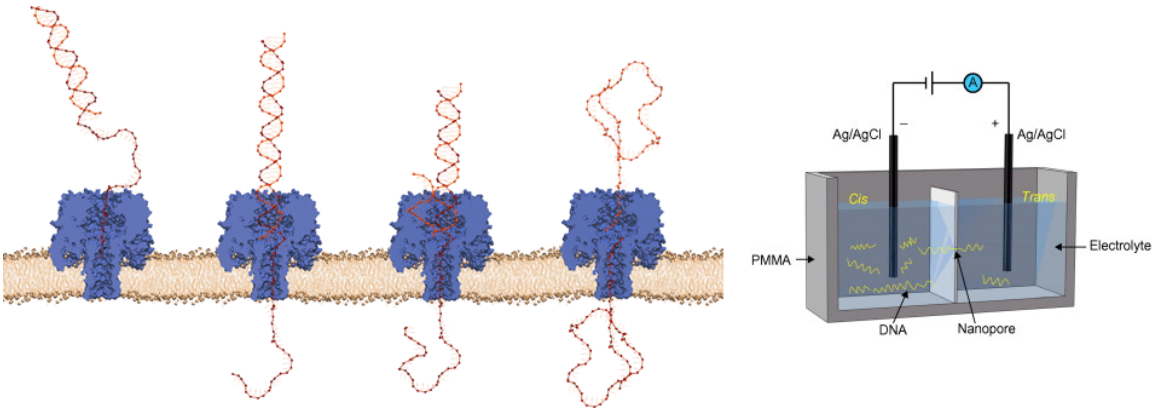


Figure 1.16: Principe de fonctionnement des Nanopores : des protéines s'insèrent dans une membrane phospholipidique en format un tout petit trou dans celle-ci. La présence de ce trou est détectée en plaçant la membrane entre deux compartiments entre lesquels on applique une tension électrique. Si un nanopore s'est incorporé dans la membrane, un courant que quelques dizaines de pA apparaissent. Si une molécule d'ADN simple brin passe dans le nanopore attirée par le champ électrique, elle va réduire le courant durant son passage.

La nature a fabriqué des protéines qui s'insèrent dans une membrane phospholipidique et établissent une communication entre les deux côtés avec un trou nanométrique. Si on applique une différence de potentiel entre les deux côtés de la membrane, un courant va y circuler si le nanopre est incorporé. Pour 100 mV on observe un courant de 50 pA si la solution est riche en sel.

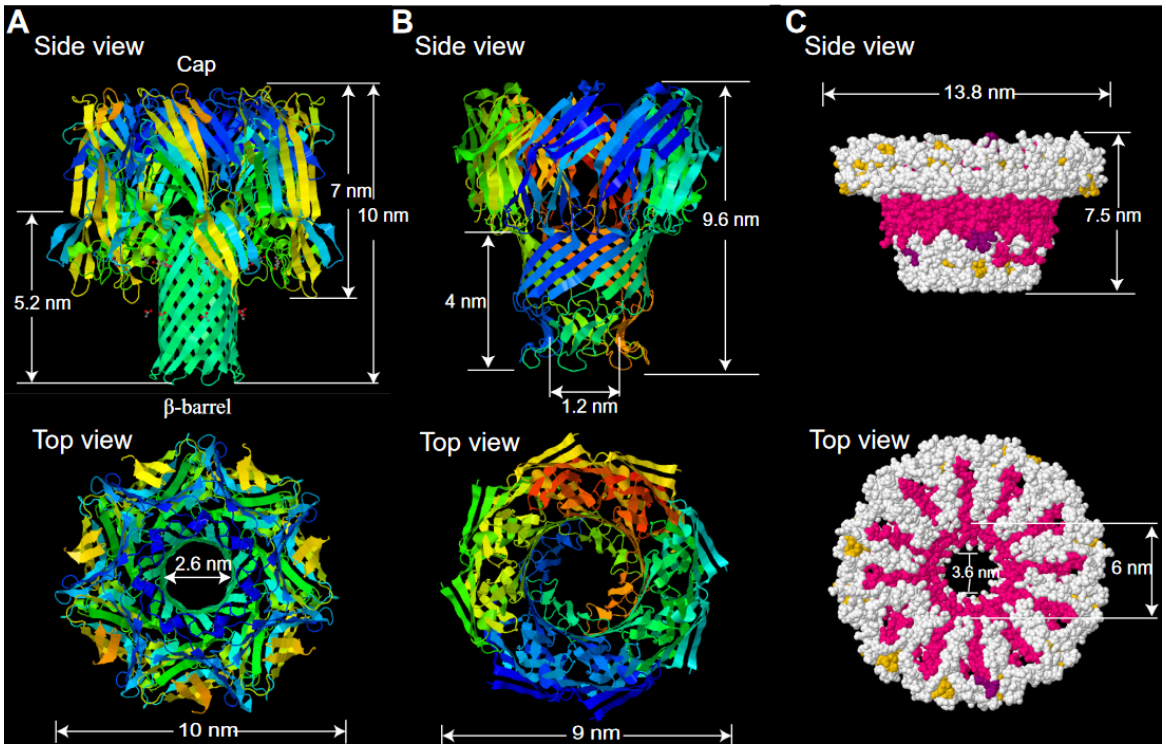


Figure 1.17: Il existe plusieurs protéines pouvant jouer le rôle de Nanopore: les bactéries ont développé ces protéines typiquement pour lyser d'autres cellules.

Il existe différents types de pores pouvant servir à séquencer l'ADN. Le plus classique est l'alpha-emolysin (à gauche) dont la partie du haut est assez large pour laisser entrer une molécule d'ADN double brin tandis que la partie inférieure qui traverse la membrane est plus étroite et ne laisse passer qu'un simple brin. Par ailleurs, le col de cette protéine est plus grand qu'une base d'ADN et il y a en fait 4 à 5 bases simple brin dans le pore en même temps.

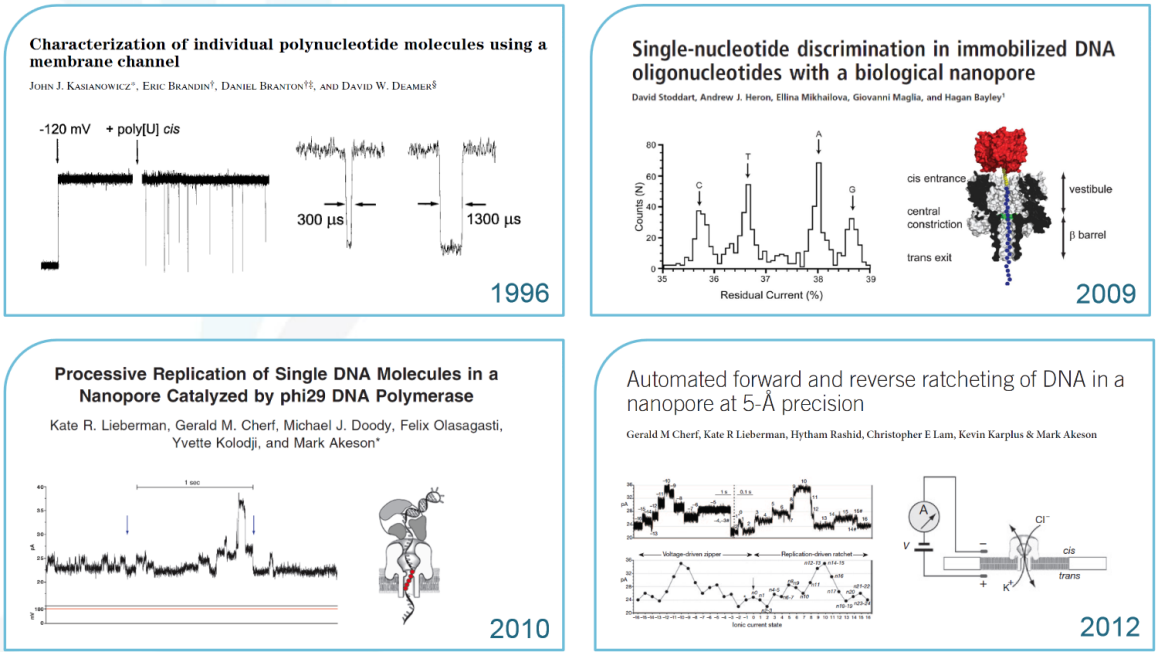


Figure 1.18: Le développement du séquençage par Nanopore a mis du temps à devenir une réalité, plusieurs étapes ont été nécessaires: en 1996 est apparue l'idée générale. (Kasianowicz et al. 1996) En 2009 une étape fut franchie en montrant que des séquences simple brin constitué d'une seule base produisait des blocages de courant différent permettant de discerner chacune des bases, mais celles-ci passent encore beaucoup trop vite dans le nanopore (Stoddart et al. 2009). En 2010, on réalise que l'on peut ralentir le passage de l'ADN en utilisant une polymérase (Lieberman et al. 2010). En 2012 en améliorant la sélectivité du nanopore les premiers signaux de séquençage sont obtenus (Cherf et al. 2012).

Le fait que 4 ou 5 bases (correspondant à un k-mer) soient simultanément dans le pore complique et aide le séquençage: cela complique les choses, car la réduction du courant provoqué par l'ADN résulte ainsi de la convolution de 4 ou 5 bases consécutives. Il n'est donc pas évident de remonter à la séquence à partir de ce courant de blocage, en fait il existe plusieurs k-mers de séquences différentes produisant presque exactement le même courant. D'autre part, le k-mer qui passe est très fortement corrélé au k-mer suivant attendu : il ne diffère que d'une base. Cette corrélation est précieuse pour retrouver la séquence, elle impose une contrainte forte sur le signal qui aide grandement à déterminer la bonne séquence. D'autre part, s'il est très probable que pour certaines bases aléatoirement le temps de passage de celle-ci dans le pore soit très court (ce qui complique sa détection), il y a peu de chance que cela soit le cas pour 4 ou 5 bases qui se suivent. C'est un avantage sur PacBio qui lui est complètement dépendant du temps de passage d'une base pour sa détection. De façon analogue cela va aider pour la détection des bases modifiées (les marqueurs épigénétiques), car celles-ci vont rester plus longtemps dans le pore et donc seront plus faciles à détecter.

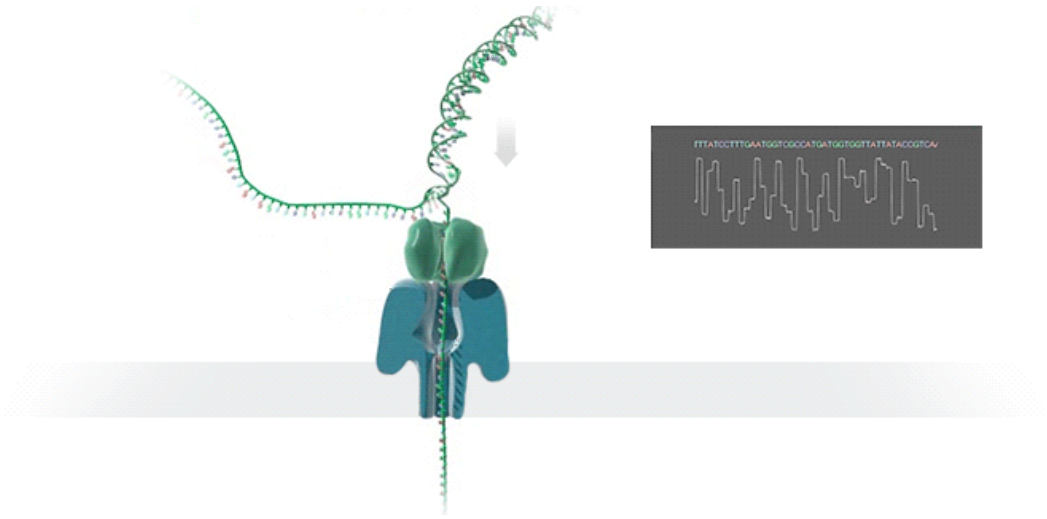


Figure 1.19: Illustration du fonctionnement d'un Nanopore associé à une hélicase. En dégrafant un ADN double brin avec une vitesse adaptée, l'hélicase fait passer un brin d'ADN dans le nanopore dont le signal est modulé par sa séquence. La différence de taille des bases permet de différencier les différents niveaux de courant.

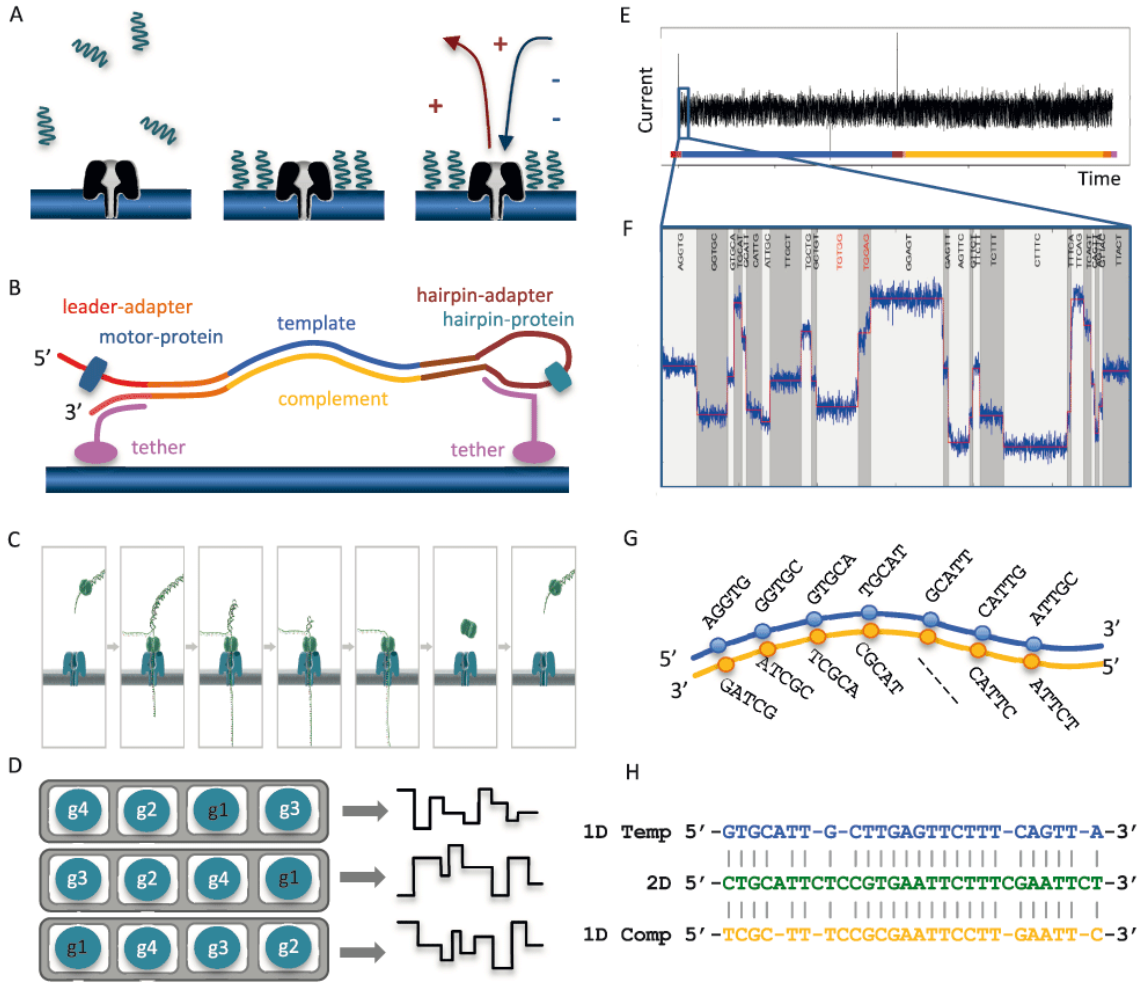


Figure 1.20: Principe de la fabrication d'une librairie 2D et nature du signal électrique observé avec le Nanopore. Comme le signal est assez bruyant, le taux d'erreur est assez important, une façon de le réduire est de lire plusieurs fois la même molécule. Ici, l'ADN double brin B) est lié à une fourche à gauche et une boucle à droite. Quand l'extrémité 5' entre dans le nanopore, l'hélicase qui y est attachée régule la vitesse de translation du template, à la fin, la boucle conduit le brin complémentaire a passé également dans le nanopore offrant deux lectures de la séquence. Il faut noter que le nanopore n'est pas aussi court qu'une seule base, mais qu'il y a en moyenne 5 bases dans son canal. Le signal observé est donc la convolution de 5-mers G) qui sont corrélés par 4 bases ce qui aide à la reconstruction de la séquence. Le fait de lire deux fois la séquence réduit les erreurs H).

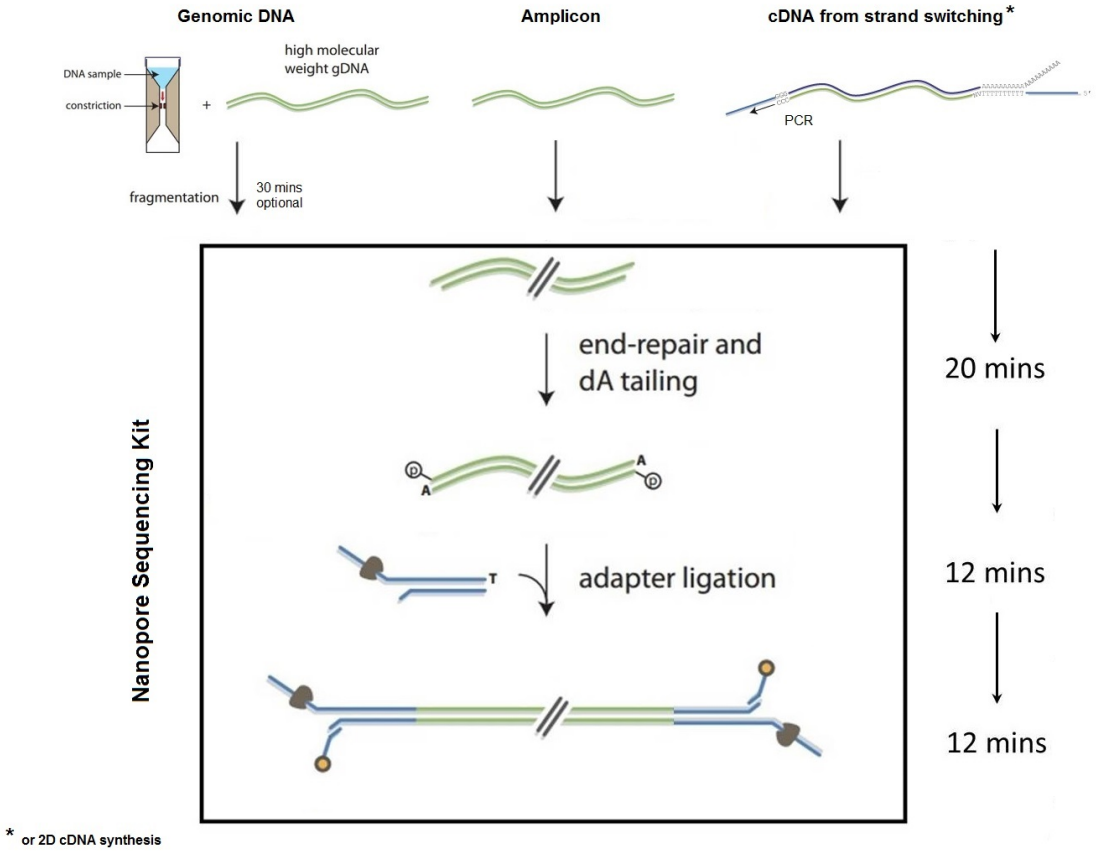


Figure 1.21: Préparation de la librairie 1D2 pour le séquençage Nanopore. L'idée est également de faire passer un brin et son complémentaire dans le nanopore, mais cette fois les deux brins ne sont pas attachés par une boucle, ils sont ancrés dans la membrane de telle façon que quand le premier brin est passé le second se trouve très près du naopore et a une grande chance (60%) de passer également.

L'autre avantage indiscutable de la technologie nanopore est le fait que la détection se fasse par la mesure d'un courant électrique: le développement formidable de l'électronique et sa miniaturisation font que détecter un courant électrique de 50 pA se fait presque très facilement à un coût extrêmement bas. Les caméras qui mesurent la lumière des molécules fluorescentes ont aussi fait des progrès impressionnants, mais leur prix et leur encombrement n'a rien à voir.

Figure 1.22: Un des très gros avantages du système de séquençage de type nanopore est qu'il utilise un dispositif très petit et peu onéreux, on voit ici le Mi-ion d'Oxford Nanopore qui est juste une grosse clé USB connectée directement à un PC (non représenté).

Fort de cet avantage, Oxford Nanopore a développé un instrument très petit se branchant au port

USB 3.0 d'un ordinateur pour un coût d'environ 1000 $. Les débuts ont été un peu difficiles, mais l'amélioration des kits de préparation et de séquençage ont maintenant amené Oxford Nanopore dans la cour des grands et ne nombreux articles de séquençage sont apparus grâce à cet instrument.

Pour effectuer un séquençage, il faut d'abord préparer une librairie à partir de l'ADN à séquencer. Le protocole ajoute à chaque extrémité de la molécule double brin une fourche avec des adaptateurs. Cette fourche contient sur un bras une hélicase dont la translocation est bloquée et sur l'autre un groupement ayant une grande affinité pour la membrane. Lorsque l'extrémité du bras portant l'hélicase entre dans le nanopore, celle-ci se débloque et commence à avancer en séparant les deux brins de l'ADN, régulant la vitesse de passage celui qui est engagé dans le nanopore. Lorsque le processus arrive au bout de la molécule, le brin complémentaire qui est resté accroché à la membrane a une assez forte chance de se faire capturer par le nanopore ce qui permet de séquencer le brin complémentaire dans 60% des cas.

La précision du séquençage est comparable à celle de PacBio (environ 10 à 15% d'erreurs), mais lorsque le brin complémentaire accompagne le premier brin la qualité du séquençage augmente de façon notoire.

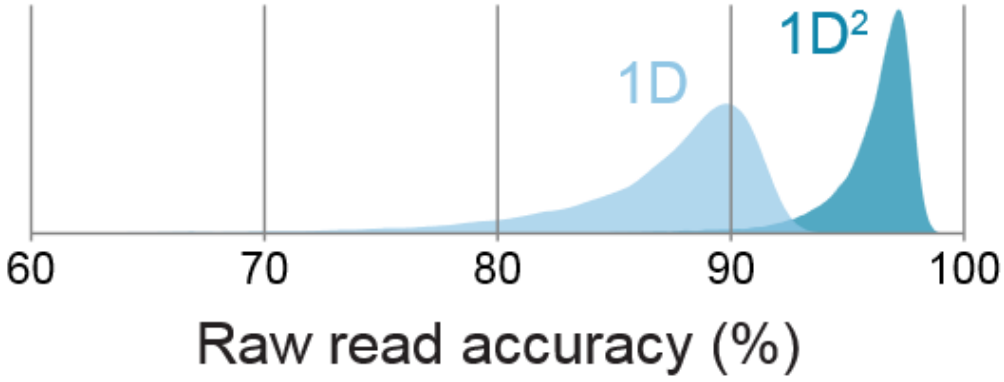


Figure 1.23: La précision du séquençage par nanopore est limitée par la qualité du signal et le nombre de passages d'une séquence particulière. Pour un seul passage (1D) cette précision est en moyenne de 90% elle est nettement améliorée par le second passage du brin complémentaire (1D2)

La modulation du courant ionique faite par chaque base est assez faible, et ce courant s'accompagne d'un bruit important qui est la cause de ce niveau d'erreur. Par contre la processivité de l'hélicase dans cette configuration est extraordinaire et la longueur de lecture atteint des dizaines et des centaines de kilobases !

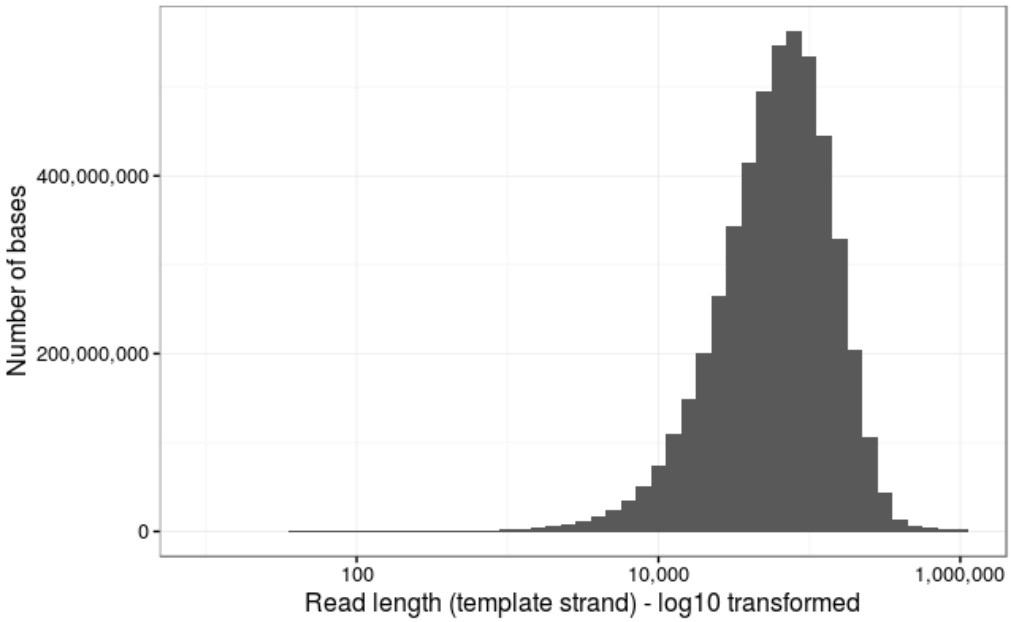


Figure 1.24: La longueur des fragments lus et séquencés est exceptionnelle, elle atteint 50 à 100 kB.

- Évolution très impressionnante de Oxford Nanopore, longueur de lecture atteignant quelques centaines de kb !
- En quelques heures, on peut obtenir 5 à 10 Gb de séquences
- Un taux d'erreur de l'ordre de 10 % en lecture simple.
- Améliorée en 1D2.
- Permets de séquencer des régions très répétées comme le centrosome. (« The Long View on Sequencing » 2018)
- Les bases modifiées sont également détectables, le fait qu'elles restent durant 4 ou 5 étapes dans le canal aide à leur détection.

1.1.3.1.5 Détection des marqueurs épigénétiques

Notre génome contient toutes les informations pour construire un être humain, mais il manque les informations de régulation : au cours du développement une série de gènes sont activés pour développer une partie spécifique du corps par exemple, mais une fois que celle-ci est différenciée il ne faut plus que ces gènes s'expriment afin d'éviter d'en développer une seconde. Ces informations de régulation sont écrites sur notre ADN comme des marques particulières qui consistent typiquement à modifier les cytosines pour leur ajouter un groupe méthyle. On peut comparer ces marques aux accents ajoutés aux lettres de notre alphabet pour en modifier le ton. Ces méthylations tout au long de l'existence et leurs apparitions ou disparitions subites peuvent parfois être liées à des maladies, ou des changements environnementaux importants. Il existe également d'autres modifications épigénétiques. Chez les bactéries, la méthylation de l'Adénosine est la plus fréquente, mais le nombre de types de modifications est encore plus grand puisqu'on en compte en tout une vingtaine. Pour la plupart ces modifications ne sont pas détectables directement avec le séquençage de seconde génération.

Comme nous l'avons dit, dès qu'une molécule d'ADN est copiée par une polymérase, tous les marqueurs épigénétiques sont effacés. C'est le cas de la PCR ou de toute amplification faite lors de la construction d'une librairie pour réaliser un séquençage (indispensable lors de l'utilisation de la première et la seconde génération). Pour accéder à ces informations, il existe plusieurs possibilités: Il est possible de traiter l'ADN génomique avec du bisulfite qui modifie les bases Cytosine qui ne sont pas méthylées, puis de séquencer ces molécules modifiées dont on a connaissance préalable de la séquence brute, et détection la mutation d'une base cytosine en une base adénine. Cela signifie alors que cette cytosine n'était pas méthylée. Cette méthode présente des inconvénients. Le bisulfite détruit une partie de l'ADN dégradant ainsi la qualité de la librairie, par ailleurs comme toute réaction chimique elle n'a pas un rendement parfait et entraine à la fois des faux positifs et faux négatifs. L'information sur la méthylation est donc plutôt statistique.

Les méthodes molécules uniques apparaissent mieux adaptées pour la détection de la méthylation dans la mesure où la construction d'une librairie ne nécessite pas d'amplification de l'ADN. Avec PacBio, la polymérase qui incorpore les nucléotides est un peu retardée lorsqu'une base modifiée est présente, il suffit donc, en principe, de détecter les bases qui produisent ce retard. Cependant cette détection est loin d'être parfaite: la polymérase incorpore chaque base avec un temps caractéristique qui présente une distribution exponentielle, ce qui veut dire que pour une base ce temps d'incorporation varie énormément, si la base est modifiée le temps moyen de cette incorporation est allongé, mais il est fréquent d'observer qu'une base non modifiée met beaucoup de temps pour être incorporée et l'inverse qu'une base présentant une modification épigénétique s'incorpore très rapidement. Cette détection est donc statistique et son taux d'erreur intrinsèque est grand. Ceci impose de faire passer plusieurs fois la molécule à étudier dans la polymérase ce qui est possible grâce à la forme en boucle de l'ADN, mais ce qui conduit à une longueur de lecture diminuée. Par ailleurs, si une méthylation sur l'adénosine produit un retard important, celle sur la cytosine est beaucoup plus faible. Ceci complique fortement la détection de la méthylation chez les eucaryotes avec PacBio.

Pour Oxford Nanopore, la base modifiée provoque une diminution du courant passant dans le nanopore comparativement à la base qui ne l'est pas, au lieu de devoir déterminer 4 bases pour séquencer il faut en ajouter une cinquième correspondante, par exemple, à la cytosine méthylée. Après avoir observé les plateaux en courant correspondant à l'avancée de chaque base dans le pore, il faut maintenant analyser la valeur de ces courants pour lui attribuer un 4-mer ou un 5-mer contenant 5 bases. Oxford Nanopore a un avantage ici sur PacBio, car une base modifiée met 4 ou 5 itérations à transiter dans le pore, le critère statistique de détection est donc meilleur. De fait, les résultats sur la détection des 5mC par Oxford Nanopore sont assez bons. Néanmoins on est en droit de se demander combien de modifications épigénétiques cette technique va pouvoir détecter : si nous introduisons les 20 modifications épigénétiques connues dans l'alphabet des bases à détecter, il est assez probable que le logiciel identifiant celles-ci voit son taux d'erreur augmenter.

1.1.3.2 Bilan des méthodes molécules uniques

À l'heure actuelle, le séquençage de troisième génération a fait des progrès considérables, mais il n'a pas encore détrôné le séquençage de seconde génération type Illumina. Les raisons sont multiples, mais plusieurs facteurs font la force d'Illumina: - son parallélisme massif produisant une centaine de gigas bases - le coût très bas du séquençage d'un génome - et son taux d'erreur relativement bas. Le taux d'erreur est un aspect fondamental pour les applications de diagnostic, jusqu'ici seul le séquençage de première génération est complètement accepté pour déterminer l'existence d'une mutation génétique. Le taux d'erreur d'Illumina, son parallélisme et son coût

permet de séquencer un génome avec un taux de recouvrement variant de *10X* à *30X* à un prix abordable, ceci permet de réduire le taux d'erreur et donc d'avoir une grande utilité en médecine. L'exemple de la détection de la trisomie par le séquençage a démontré son potentiel de façon éclatante.

La troisième génération en molécules uniques possède un atout indéniable: sa longueur de lecture inespérée, celle-ci a permis le réassemblage du génome humain en corrigeant des erreurs impossibles à détecter avec le séquençage de seconde génération. Le taux d'erreur est un problème pour le diagnostic, mais l'association d'un séquençage PacBio ou Oxford Nanopore avec un séquençage Illumina permet un assemblage facile avec les grandes longueurs de lecture du séquençage molécules unique et une correction des erreurs avec le séquençage Illumina. L'autre solution pour améliorer la précision est de faire plusieurs fois le séquençage molécules unique : par exemple en imposant une librairie de molécules courtes dans PacBio, on relit ainsi plusieurs fois la même séquence et on diminue rapidement le taux d'erreur. Comme dans PacBio les erreurs ne sont pas corrélées, on obtient ainsi le séquençage avec le plus bas taux d'erreur, le prix à payer est la réduction concomitante de la longueur de lecture. Si on veut lire 30 fois la même séquence, il faut réduire la longueur de la molécule du même facteur ce qui conduit à une longueur de lecture effective de quelques centaines de bases, comparables à ce que fait Illumina. Oxford Nanopore peut relire 2 fois la séquence à analyser dans la stratégie 1D2, mais les erreurs présentent une certaine corrélation ce qui limite les possibilités. Il est aussi possible de préparer la librairie d'Oxford Nanopore à partir de molécules obtenues par une amplification "rolling circle", ce qui permet de convertir la grande longueur de lecture en une lecture multiple de la même séquence et donc de réduire les erreurs.

Il est intéressant de remarquer que la troisième génération de séquençage n'exploite pas encore tout le potentiel que l'on est en droit d'attendre des molécules uniques : imaginons que nous n'ayons aucune limitation, l'intérêt de séquencer à l'échelle de la molécule unique est de pouvoir séquencer directement l'ADN du patient sans utiliser d'étape d'amplification et de pouvoir, par exemple, 1) retrouver dans un échantillon raisonnable les anomalies d'un gène particulier 2) de retrouver quelques molécules anormales d'un gène particulier au sein d'une majorité de molécules normales (ceci correspond à la détection de cellules cancéreuses). Un millilitre de sang contient environ 5.10^6 génomes, en imaginant un rendement d'extraction de 100% et le fait qu'il faut séquencer entre 10 et 100 molécules pour avoir un résultat fiable, nous voyons qu'il faut séquencer 1 molécule sur 5.10^4 pour atteindre le but 1). C'est presque le cas de PacBio qui séquence 1 molécule sur 70 000 introduites dans le séquenceur, chez Oxford Nanopore c'est 1 sur 2 000 000. Malheureusement la préparation de la librairie pour le séquençage par ces méthodes est loin d'avoir un rendement de 100% et en pratique si le but 1) est presque atteignable, on est loin d'atteindre 2). Il y a encore beaucoup d'améliorations à attendre pour optimiser les méthodes molécules uniques. Les défauts de Pacbio et d'Oxford Nanopore sont au moins de deux types: le séquençage d'une base se fait vite et pratiquement une seule fois, ce qui conduit au niveau d'erreur élevée, le séquençage se fait, soit au fond d'un puits très étroit de 70 nm de large, soit dans un seul nanopore très petit au sein d'une membrane de quelques microns. L'accessibilité de ce site de séquençage est donc faible et pour compenser ce défaut il faut ajouter beaucoup d'ADN dont la majorité ne sera pas séquencée.

Dans la méthode que nous proposons, nous essayons d'adopter une stratégie où il est possible : d'interroger plusieurs fois la même molécule et où le taux de molécule séquencé n'est pas réduit par la taille du site de séquençage. La détection d'un type de marque épigénétique se fait en introduisant un anticorps s'accrochant à ce type de base modifiée par exemple 5mC, on détecte la position d'accrochage de l'anticorps au cours de cycles de force où les anticorps s'accrochent

de façon transitoire. À chaque cycle, la présence de l'anticorps peut donner ou non un blocage, la répétition de plusieurs cycles permet de garantir la bonne détection malgré le caractère aléatoire de cette détection molécule unique. Une fois les positions des 5mC enregistrées, on peut rincer l'échantillon et introduire un second anticorps, par exemple le 6mA, et recommencer une nouvelle détection. On peut répéter cette opération pour toutes les modifications épigénétiques que l'on souhaite détecter et pour lesquels on possède un anticorps. Mais il faut également connaitre la séquence de la molécule d'ADN étudiée, avec notre méthode, le séquençage de novo est encore délicat. Mais, et c'est le sujet de cette thèse nous proposons de faire un séquençage partiel simple qui permet d'identifier la molécule étudiée pourvu que celle-ci appartienne à un génome connu.

1.1.3.3 Séquençage mécanique réalisé à l'aide de pinces optiques.

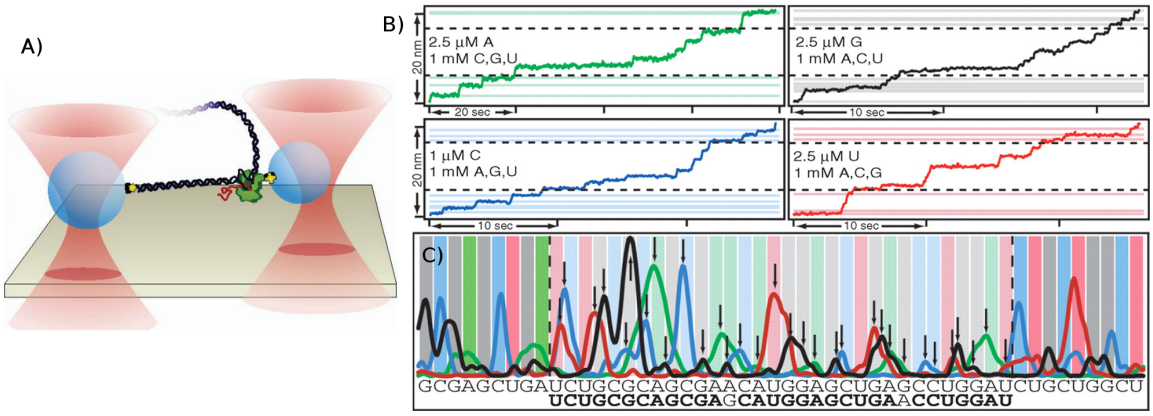


Figure 1.25: Schéma de principe et signaux correspondants au séquençage réalisé par pinces optiques. A) Une RNA polymérase est attachée à une bille bloquée dans une pince optique, elle est engagée dans une molécule d'ADN double brin dont l'extrémité est attachée à une seconde bille maintenue dans une deuxième pince optique. Lorsque la RNA polymérase transcrit l'ADN en ARN, elle avance dans l'ADN et la longueur de la molécule diminue de 0.34 nm par base transcrite. B) En effectuant cette expérience avec une quantité réduite d'un des nucléotides, la trajectoire présente des pauses à chaque occurrence de la base en question. En répétant cette opération pour chacun des quatre nucléotides on obtient un signal C) qui permet de séquencer la molécule d'ADN.

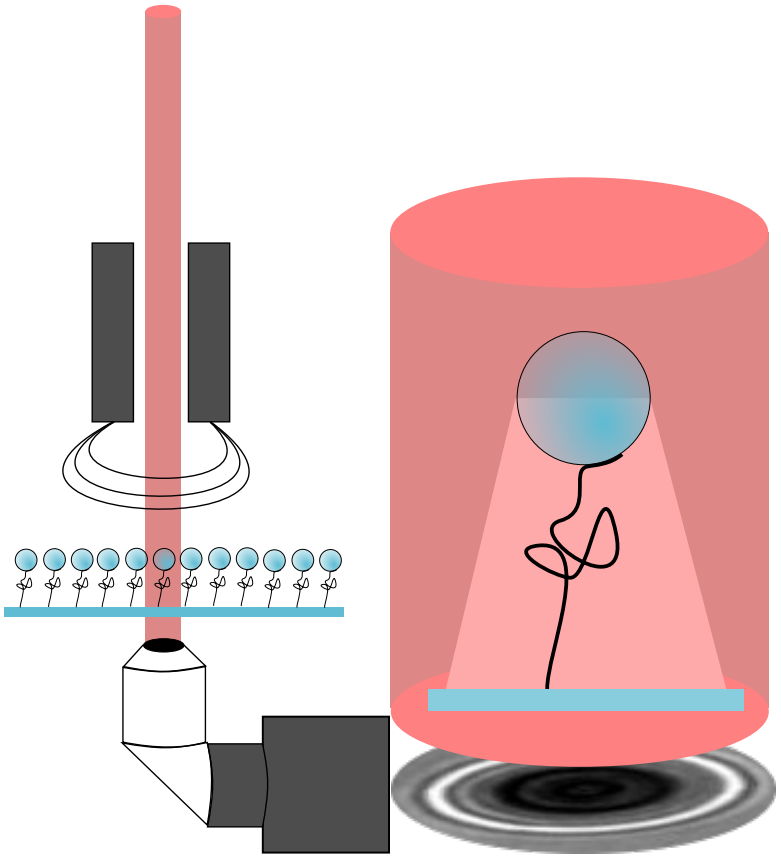
En 2006, S.M. Block (Greenleaf et Block 2006) a optimisé ses pinces optiques et a obtenu une résolution de l'ordre de l'Angstroem tandis qu'il observé une RNA polymérase en train de transcrire l'ADN. Dans son expérience, il tire sur une molécule d'ADN par une extrémité accrochée à une bille et par une RNA-polymérase engagée dans l'ADN est accroché à une seconde bille Figure 1.25. Chaque fois que la polymérase transcrit une base, la molécule se raccourcit de 0.34 nm, ce que S.M. Block détecte. En faisant une expérience de type Sanger où un des nucléotides est introduit à basse concentration, la progression de la polymérase est ralentie chaque fois que cette base doit être incorporée, ce qui permet d'identifier la position de cette base le long de la séquence Figure 1.25. En répétant quatre fois cette opération pour chacune des bases, il est possible de reconstruire la séquence de l'ADN Figure 1.25.

Cette méthode impressionnante par la résolution de l'instrument souffre malheureusement d'un problème : une fois que la RNA polymérase a transcrit l'ADN avec le premier type de nucléotide à basse concentration, il n'y a pas de moyen de faire reculer la polymérase pour réaliser les trois autres mesures avec les nucléotides différents. Les auteurs ont fait la même opération sur d'autres molécules ayant la même séquence. La méthode ne permet donc pas de séquencer sous cette forme une molécule unique. Une deuxième limitation provient du fait qu'il est difficile de multiplexer les pinces optiques surtout

avec cette très grande résolution, donc il n'est possible de séquencer (partiellement) qu'une molécule à la fois.

1.1.3.4 Séquençage par pinces magnétiques

L'équipe de biophysique des molécules uniques travaille depuis plus de dix ans sur un appareil de mesure spécifique à pinces magnétiques. Les pinces magnétiques permettent de manipuler des molécules d'ADN individuelles en les plaçant entre une bille paramagnétique et un support en verre. Ces billes, d'un diamètre d'un micron, sont visualisées par un microscope optique associé à une caméra. Des aimants placés au-dessus de l'échantillon permettent de tirer sur les molécules. Le microscope optique assure la mesure de la position en 3 dimensions de plusieurs billes à la cadence vidéo. Cet appareil fonctionne en combinant différents principes de physique.



(a) Microscope à pinces magnétiques (b) Diffraction induite par la lumière sur les billes paramagnétiques

Figure 1.26: Illustration d'un microscope à pince magnétique. Dans la figure (a) nous observons le schéma général d'un microscope (les éléments ne sont pas à l'échelle). Un faisceau lumineux non cohérent issu de LED traverse l'ensemble du système. Passant à travers des aimants, l'échantillon, puis la lumière viennent se projeter sur le capteur d'une caméra. Les aimants peuvent à la fois translater de haut en bas, changeant le champ magnétique à la surface de l'échantillon, et effectuer une rotation, sans bloquer le passage du faisceau lumineux. L'échantillon est une cellule microfluidique d'une hauteur d'environ $50\mu m$ pouvant contenir des molécules d'ADN au sens large, ainsi que protéines. Ces molécules sont attachées à des billes superparamagnétiques qui sont influencées par les aimants, qui pourront leur appliquer un mouvement de rotation. Mais aussi, en faisant varier la distance des aimants à la surface. Augmenter ou diminuer l'attraction des billes aux aimants, appliquant une force plus ou moins grande sur les protéines ou l'ADN. Dans la figure (b), on s'intéresse plus particulièrement à la diffraction de la lumière au voisinage d'une bille paramagnétique. La lumière directe et celle déviée par la bille forme des anneaux d'interférence de taille variable à mesure que la bille s'approche au s'éloigne du point focal de la ligne optique.

En projetant un faisceau de lumière dont les rayons sont parallèles sur une bille microscopique, ceux-ci sont diffractés par les billes et ils interfèrent avec les rayons qui ne sont pas déviés et forment des anneaux d'interférence concentriques de tailles croissantes. On peut les visualiser en plaçant une caméra dans le plan image de l'objectif du microscope. L'écartement de ces anneaux de diffraction varie de façon univoque avec la distance entre la bille et le plan de mise au point du microscope. Après calibration, on peut déduire la position verticale de la bille en observant simplement l'image de ses anneaux. Cette méthode de mesure est un peu moins précise que la méthode des pinces optiques, la

résolution est de quelques nanomètres, mais par contre, elle permet de suivre quelques centaines de billes en simultanée.

En attachant une molécule d'ADN entre la surface de verre de l'échantillon et une bille magnétique, on peut déterminer la longueur de celle-ci. En plaçant des aimants au-dessus des billes magnétiques, on peut lui appliquer une force variable en ajustant la hauteur des aimants par rapport aux billes. La force est d'autant plus grande que les aimants sont proches de celles-ci. Il est possible de calibrer cette force de façon absolue à partir de l'amplitude du mouvement brownien de la bille perpendiculaire à la direction de la force, et de la longueur de la molécule.[2] Comme l'aimantation des billes est assez homogène, après avoir calibré les aimants, il est possible de prédire la force appliquée aux billes à 30%près simplement en connaissant la distance des aimants à l'échantillon. Les pinces magnétiques permettent également d'appliquer une contrainte de torsion sur des molécules adaptées en tournant les aimants. Le domaine d'application des pinces magnétiques est large (Gosse et Croquette 2002), et a permis des avancées dans la compréhension d'abord des propriétés physiques de l'ADN et notamment des ces différentes conformations, ainsi que la dynamique régissant la formation d'ADN double brin. Puis, dans la compréhension de différentes protéines interagissant avec l'ADN, particulièrement dans le cadre du réplisome d'organisme (Strick et al. 1996), (Hodeib et al. 2017), (Manosas et al. 2012), (Neuman et al. 2009), [Koster et al. (2005)

1.1.3.4.1 Le séquençage de novo par pinces magnétiques utilisant le séquençage par hybridation.

La force appliquée par la bille permet, si elle est supérieure à 15pN, de déhybrider les deux brins d'une molécule d'ADN en épingle à cheveux. Pour séquencer l'ADN avec les pinces magnétiques, nous utilisons l'ouverture et la fermeture d'une molécule en épingle à cheveux qui conduit à un changement de longueur de ~1 nm par base ouverte. Ceci est trois fois mieux que le facteur de conversion observé par S.M. Block, mais cette longueur reste plus petite que la résolution des pinces magnétiques. De ce fait, séquencer l'ADN est délicat. Comme nous allons l'expliquer dans la suite, il est possible de détecter l'hybridation d'un petit oligonucléotide qui peut s'hybrider à l'ADN lorsque la molécule et ouverte et provoquant un blocage lors de la fermeture de cette molécule. Nous pouvons donc utiliser la méthode du séquençage par hybridation. Dans cette méthode, à l'origine on a juste besoin de savoir quels oligonucléotides s'hybrident avec la molécule à séquencer pour retrouver la séquence en utilisant le recouvrement entre les oligonucléotides pour assembler la séquence (Dramanac et al. 1989). Cependant, cette méthode est conçue pour des oligonucléotides k-mer pas trop court pour que les hybridations soient rares et elle demande de tester tous les k-mers qui sont très nombreux. Un exemple a été réalisé avec des 6-mers fluorescent hybridés sur des molécules d'ADN dans l'article suivant (Pihlak et al. 2008). Avec les pinces magnétiques, on dispose d'une information supplémentaire, non seulement nous savons si l'oligonucléotide s'hybride, mais nous connaissons sa position d'hybridation à quelques bases prés. Le séquençage de novo est donc possible, mais il réclame beaucoup de mesures pour fonctionner correctement. L'équipe de Depixus développe des stratégies pour rendre ce séquençage plus efficace, mais pour le moment il reste délicat. Dans ce travail, nous proposons d'utiliser une information partielle de séquençage pour retrouver des molécules dans des génomes connus. Cette opération est relativement facile à réaliser et constitue une étape vers le séquençage de novo.

1.1.4 Fingerprinting - Signature

Nous avons vu dans la section précédente que le séquençage par pince magnétique reste en développement actif. Cependant, une voie intermédiaire existe pour à la fois posséder le locus d'un fragment

[2] La relation entre la force F, l'amplitude des fluctuations browniennes dx, la longueur de la molécule L est obtenue à partir de la formule d'Einstein : $\frac{1}{2}k_BT = \frac{1}{2}\frac{F}{L} \times dx^2$ où k_B est la constante de Boltzmann et T est la température.

d'ADN et, ce qui nous intéresse particulièrement, son état de méthylation. Alors que le séquençage consiste à extraire la succession des nucléotides d'un fragment d'ADN, on peut imaginer de récupérer cette information de manière partielle. Le séquençage partiel par pince magnétique est alors effectué en un temps plus court et est moins cher. En effet, dans de nombreux cas, obtenir la séquence d'un fragment d'ADN n'est pas nécessaire. L'immense base de données de séquences de référence de plus 73 000 Organismes réunis par les chercheurs depuis le début du séquençage au milieu des années 1970 permet un large panel d'applications basées sur le reséquençage de l'ADN (NCBI 2017).

1.2 Problématique

Les pinces magnétiques semblent, avec la conception des molécules en épingle à cheveux, et les premiers travaux sur l'utilisation d'oligonucléotides interrompant transitoirement la fermeture des molécules (Ding et al. 2012), ouvrir de larges possibilités dans le cadre la détection de marqueurs épigénétiques (méthylations), à la fois sur l'ADN génomique, mais également sur les ARN.

Cependant, avant d'être capable d'utiliser cette stratégie, un travail de démonstration reste à faire. Alors que le séquençage par cette méthode semble une voie plus complexe, l'utilisation de signatures pourrait permettre d'identifier une région génétique et d'accéder à son état de méthylation plus simplement.

Mon travail de thèse a pour objectif de démontrer la faisabilité à la fois théorique et pratique de l'utilisation de signature pour retrouver le locus d'origine d'une molécule d'ADN dans un génome de référence connu. Ce travail, qui peut s'apparenter à du reséquençage partiel de génome nécessite d'explorer différents éléments pour parvenir au succès de cette stratégie.

Tout d'abord, il s'agit de caractériser le ou les oligonucléotides, modifiés ou non, qui permettront le séquençage partiel d'une molécule, permettant le retrouver le locus d'origine d'une molécule. Pour cela, il faudra à la fois simuler le comportement des oligonucléotides et s'accommoder de leurs particularités théoriques, pour ensuite les tester expérimentalement, sur des molécules en épingle à cheveux, dont certaines auront été conçues et construites pour l'occasion.

Une fois la démonstration fait de l'utilisabilité des oligonucléotides pour le séquençage partiel et la création de *signatures* unique sur un génome,[3] nous pourrons nous concentrer sur une part importante du travail : concevoir des algorithmes, qui pourront efficacement, à partir d'un ensemble d'hybridation le long d'un fragment d'ADN d'un génome connu, déterminer le locus d'origine de ce fragment. Deux approches ont été retenues lors de la rédaction de ce manuscrit. Nous étudierons donc les caractéristiques et les performances de ces méthodes dans la réalisation de leur tâche de cartographie à l'aide de simulation.

Enfin, en utilisant de l'ADN d'E.coli, nous tenterons de démontrer l'utilisabilité en condition réelle d'une des deux approches développées et testées avec des données simulées. Le manuscrit expliquera également la démarche de construction et d'acquisition des données en aveugle, afin de parvenir à une démonstration convaincante.

1.3 Signature par pince magnétique

1.3.1 Introduction

Dans la partie précédente nous avons décrit comment les pinces magnétiques permettent la micromanipulation de molécules d'ADN en leur appliquant une force de l'ordre de quelques pN et le suivi de leur

[3] nous reviendrons sur cette terminologie dans le prochain chapitre.

élongation en temps réel avec une résolution de l'ordre du nanomètre. À présent, nous allons présenter leur utilisation pour l'acquisition des données de séquençage par hybridation d'oligonucléotides.

Également, dans le cadre de ce travail, nous allons identifier et même d'authentifier une molécule d'ADN provenant d'un ensemble de molécules connues. Pour cela, un objectif important est de déterminer la *signature* d'une molécule, terme qui pourrait également être remplacé par *empreinte digitale ADN* (*fingerprint* en anglais). Comme dans le cas d'une empreinte digitale réelle, une molécule doit posséder une signature qui lui est propre (unique) et qui permet de l'identifier.

Ainsi, connaissant l'ensemble des molécules possibles[4] la signature d'une molécule d'ADN correspond à un jeu d'informations spécifiques suffisantes pour identifier cette molécule de manière unique dans cet ensemble. Cette signature repose sur une *fonction de signature*, c'est-à-dire un processus reproductible fournissant ces données sur une molécule à la fois *in silico* et *in vitro*. L'ensemble des fonctions de signatures que nous traiterons seront basées sur l'hybridation de manière spécifique d'une ou plusieurs séquences ADN par une sonde. Concrètement, une signature consistera donc en une liste de coordonnées le long de la séquence correspondant aux positions reconnues de manière spécifique par la ou les sondes. Cependant, comme nous avons vu que les processus d'hybridation sont stochastiques, les mesures réalisées sur une molécule sont variables. Une bonne fonction de signature doit néanmoins permettre de discriminer deux molécules de séquence différente malgré ces variations.

Cette fonction de signature devra ainsi minimiser la différence entre deux molécules de séquence identique, et maximiser la différence entre deux molécules de séquence différente, tout en respectant l'ensemble des contraintes induites par les conditions expérimentales, à la fois liées aux techniques de microscopie et à la physique de l'ADN.

Ce chapitre consistera à expliquer la méthode permettant l'acquisition de signature à partir de pinces magnétiques, puis se concentrera sur les limites auxquelles nous sommes confrontés lors de l'acquisition des données.

1.3.2 Molécules en épingle à cheveux

Outre les informations sur la séquence, les pinces magnétiques permettent également de déterminer les modifications épigénétiques d'une molécule d'ADN. Celle-ci doit former une épingle à cheveux où la molécule d'ADN double-brins est refermée à une extrémité par une boucle, tandis qu' à l'autre extrémité, une fourche en forme d'Y est attachée, chaque bras de l'Y se terminant par une molécule d'attachement spécifique. Une molécule de biotine permet de coupler à une bille paramagnétique, et plusieurs molécules de digoxigénine permettent de fixer l'ADN à la surface de verre.[5]

[4] par exemple, l'ensemble des molécules du génome d'un organisme connu, celui-ci étant entièrement séquencé, un génome de référence.

[5] dig-dUTP

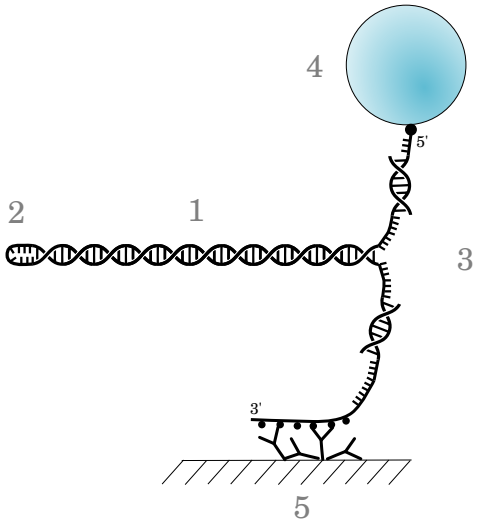


Figure 1.27: Molécule en épingle à cheveux: 1. Fragment d'ADN d'intérêt dont les extrémités simples brins contiennent un site spécifique 2. Petit fragment d'ADN simple brin se refermant sur lui-même formant une boucle dont l'extrémité est complémentaire d'une extrémité du fragment d'intérêt. 3. Construction d'ADN formant la *fourche* de la molécule en épingle à cheveux. Constitué d'une molécule simple brin à l'extrémité de laquelle nous avons inséré une molécule de biotine, hybridée partiellement à un second brin se terminant par des molécules de digoxigénine. 4. Bille paramagnétique recouverte de streptavidine, formant avec la biotine une liaison par un complexe de type clef-serrure. 5. Lamelle de verre recouverte des anticorps anti-digoxigénine, se liant spécifique aux molécules dig-dUTP de la *fourche*

Cette construction d'ADN couplée aux billes magnétiques est introduite dans une cellule microfluidique, dont la surface inférieure est une fine lamelle de verre recouverte d'anticorps anti-digoxigénine. La surface supérieure est une feuille de Mylar[6], tandis que la cellule est un simple scotch double face d'une épaisseur de $50\mu m$. La lamelle de verre mesure 24x60 mm, tandis que la cellule mesure 5x45 mm.

Les billes associées à la construction en épingle à cheveux sont injectées dans la cellule, où elles s'y accrochent grâce aux anticorps anti-dig recouvrant la surface.

1.3.3 Description d'une expérience

Comme expliqué dans la sec. 1.1.3.4 en approchant les aimants de la surface de l'échantillon, le champ magnétique induit une aimantation des billes paramagnétiques, dirigées selon la direction du champ produit par les aimants. Les aimants génèrent un gradient de champ qui produit ainsi une force sur les billes paramagnétiques. Cette force varie avec la distance séparant les aimants des billes et avec le volume de ces dernières[7]

Typiquement, pour des billes paramagnétiques MyOne™ de $1\mu m$ de diamètre, on peut exercer une force entre 0.1pN et 25pN en faisant varier la distance de 2 mm à 0.2 mm. La force F peut être approximée par $F = F_0.e^{-a.Z+b.Z^2}$, avec $F_0 = 35$pN, $a \simeq 3\text{mm}^{-1}$et $b \simeq 0.15\text{mm}^{-2}$.

Cette gamme de force permet de manipuler de manière non destructive les molécules en épingle à

[6]film de polytéréphtalate d'éthylène d'une épaisseur comprise entre 25 et $50\mu m$
[7]La quantité de matériaux ferromagnétique est proportionnelle au cube du rayon des billes

cheveux. En effet, les liaisons covalentes entre les nucléotides peuvent supporter une force moyenne de 500pN avant de se briser. Les liaisons streptavidine-biotine supportent une force moyenne de 100pN. La liaison dig / anti-dig est la plus faible, en effet elle peut résister à une force moyenne de quelques dizaines de pN. La faiblesse de cette liaison est compensée par le fait que la molécule possède plusieurs dig-dUTP qui peuvent se lier simultanément à la surface, augmentant la fixation globale de la molécule sur la surface.

1.3.3.1 De l'épingle à cheveux à la molécule simple brin

En appliquant une force supérieure à 15pN sur une structure en épingle à cheveux, les bases de la partie double brin se *dégrafent* une à une, déroulant la molécule, jusqu'à ce qu'elle soit entièrement ouverte, en simple brin. La vitesse de ce phénomène est, dans les conditions expérimentales normales, plus grande que 1000 bases ouvertes en 30 ms, soit la durée typique d'une image de la caméra.

En diminuant la force en dessous de 12pN, les liaisons hydrogène impliquées dans l'appariement des bases complémentaires deviennent plus fortes que celle exercée par les pinces magnétiques : la molécule se *referme* sur elle-même, reformant sa structure en épingle à cheveux.

La possibilité de répéter cette opération d'ouverture / fermeture de la molécule n'est limitée que par la probabilité de rupture des liens reliant la molécule à la surface, à la bille magnétique, ainsi que les liaisons entre les nucléotides. On peut raisonnablement effectuer plusieurs centaines de cycles d'ouverture et fermeture successifs d'une molécule, sans observer de dégradation et dans certains cas, plusieurs milliers de cycles.

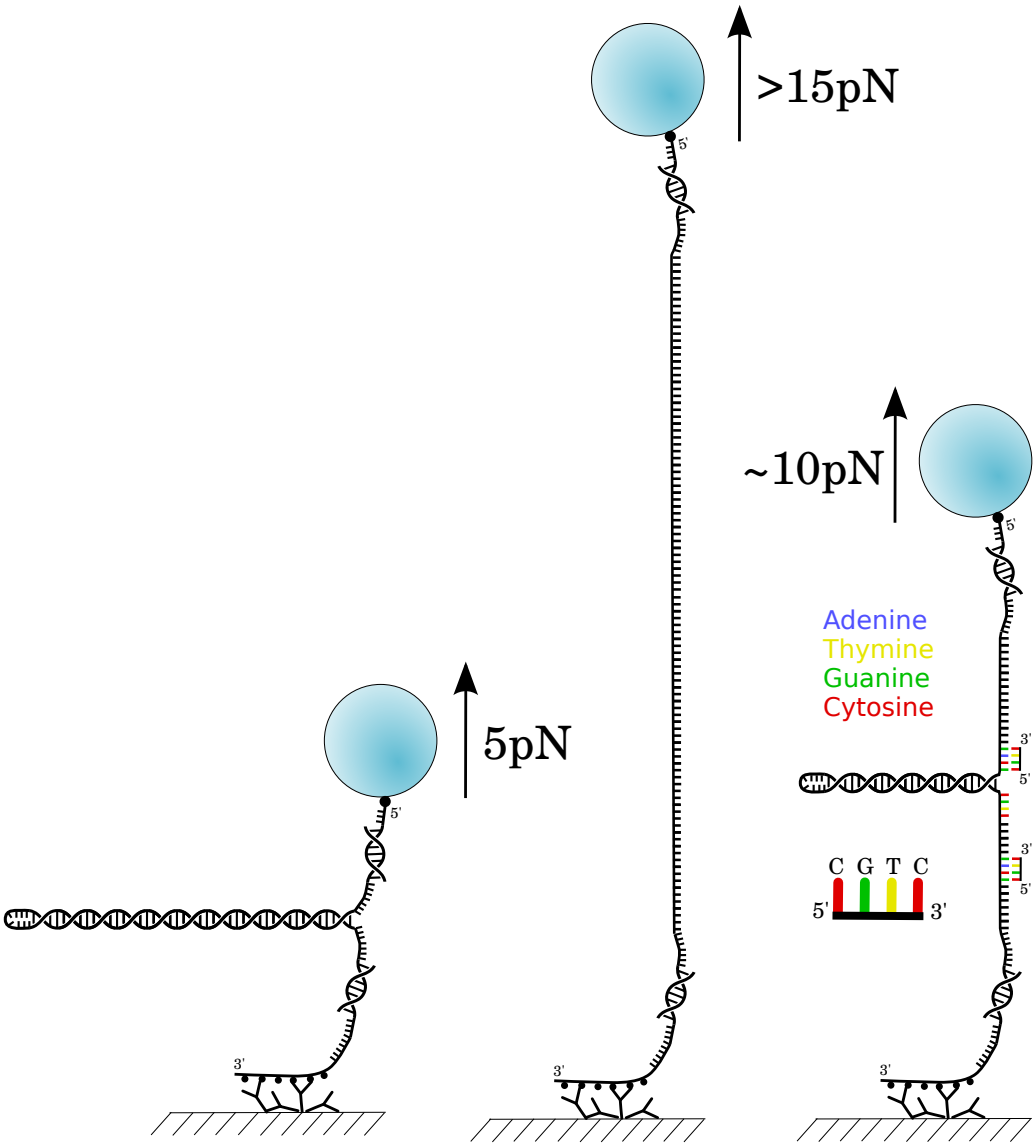


Figure 1.28: Comportement schématique d'une molécule en épingle à cheveux soumise à différentes forces durant une expérience. La molécule la plus à gauche est soumise à une force de 5 pN, qui sépare la bille paramagnétique de la surface. Au centre, la molécule s'est entièrement déroulée, exposant le simple brin de la séquence d'intérêt. À droite, une molécule en train de se fermer sur laquelle s'hybride en plusieurs positions un oligonucléotide sur sa séquence complémentaire. Les annotations 5'et 3' indiquent le sens de l'ADN et permettent notamment de se rendre compte qu'un oligonucléotide ne peut théoriquement pas s'hybrider à la même position sur les deux brins, à moins que sa séquence soit palindromique.

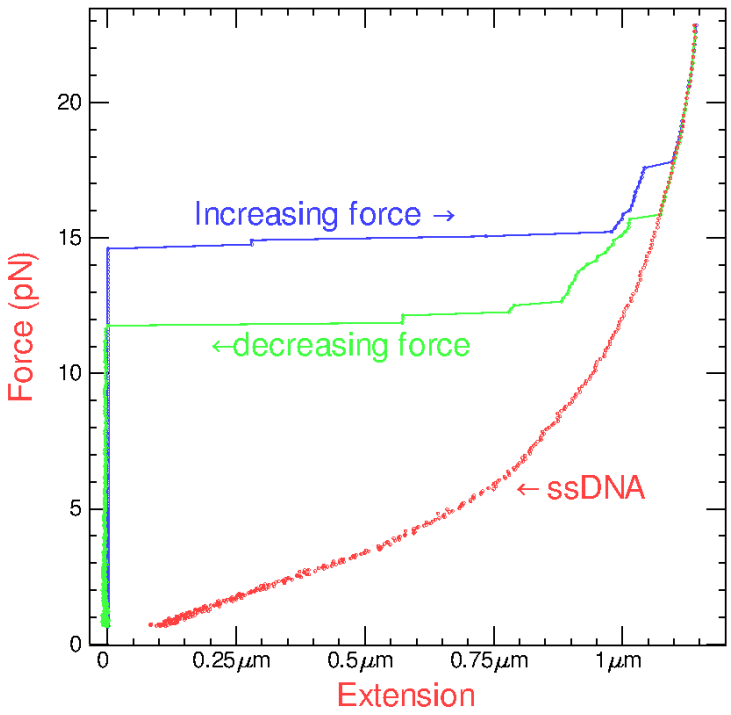


Figure 1.29: Courbe force extension illustrant l'ouverture et la fermeture d'une molécule en épingle à cheveux en fonction de la force appliquée. La courbe bleue est mesurée en augmentant la force à partir de 0, la molécule s'ouvre typiquement à 15 pN. La courbe verte correspond à la même mesure en diminuant la force, la molécule se referme vers 12 pN. Si nous introduisons un oligonucléotide complémentaire de la séquence de l'apex de notre molécule, celui-ci va s'hybrider une fois la molécule ouverte et il va empêcher la nucléation de la formation de la double hélice comme on peut le voir avec la courbe rouge qui correspond à l'élasticité du simple brin. Cet oligonucléotide est éjecté à très basse force de l'ordre de 1 pN permettant la refermeture de la molécule.

1.3.3.2 Hybridation d'oligonucléotides sur une molécule en épingle à cheveux

Lorsque la molécule est en position ouverte, l'ADN est exposé sous forme simple brin. L'ADN est alors disponible pour interagir avec d'autres molécules. Si nous injectons de telles sondes dans la cellule au début de l'expérience, celles-ci peuvent se fixer sur le simple brin de la molécule exposée, à l'endroit où il y a une affinité spécifique. Lorsque l'on referme la molécule, en réduisant la force en dessous de $\approx 12pN$, on observe un blocage transitoire à une position intermédiaire, la fourche d'ADN est arrêtée par la sonde fixée. Notre dispositif mesure l'élongation de la molécule d'ADN avec une précision d'environ 3nm, ce qui nous permet de détecter la position des différents points de blocage spécifiques. La durée moyenne des blocages est dictée par l'énergie libre de fixation de la sonde et la force appliquée lors de la fermeture. La sonde finit par se détacher, permettant la fermeture totale de la molécule.

Nous allons nous concentrer principalement sur une classe particulière de sondes, les oligonucléotides. Ce sont de courtes chaines de nucléotides ayant une forte affinité avec tout fragment d'ADN simple brin de séquence complémentaire. La dynamique de ces appariements spécifiques, nommées hybridations, est décrite par (SantaLucia 1998)

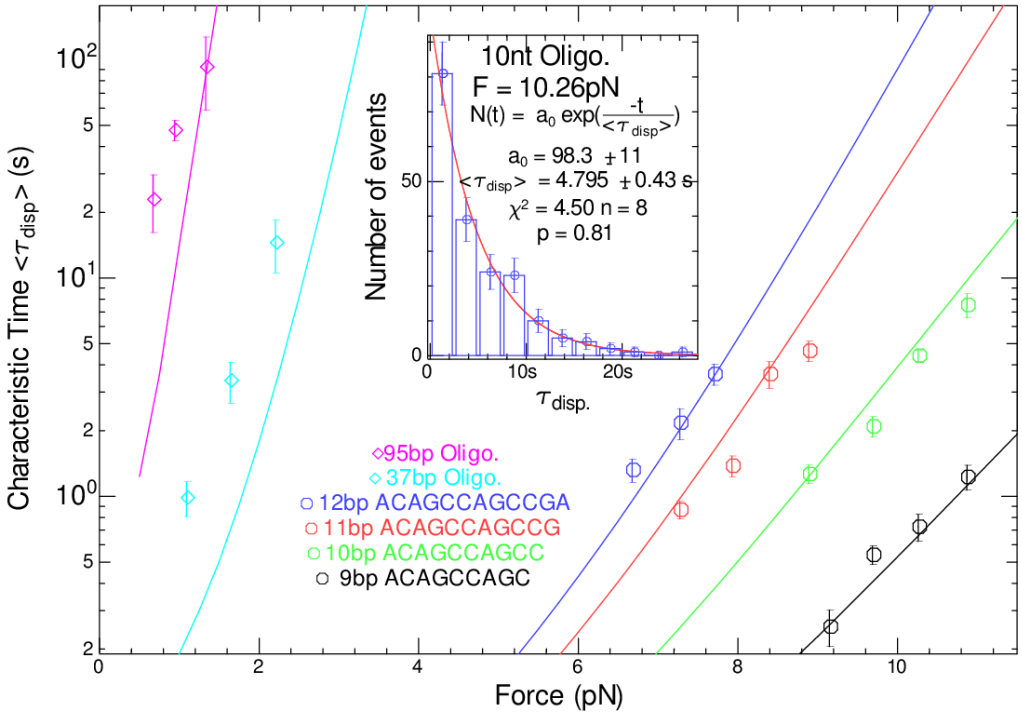


Figure 1.30: Le temps de blocage de la fourche se refermant sur la molécule en épingle à cheveux due à un oligonucléotide dépend fortement de sa longueur, de son énergie libre d'hybridation et de la force appliquée à la molécule durant la fermeture. Comme on peut le voir sur l'insert de cette figure, le temps de déplacement présente une distribution de probabilité qui suit une loi exponentielle avec un temps moyen T_{disp}. Celui-ci augmente très vite avec la longueur de l'oligonucléotide et avec la force appliquée. Pour des oligonucléotides d'ADN et des forces autour de 10 pN un oligonucléotide de 10 bases bloque la fourche pendant 5 secondes. Plus sa composition est riche en GC plus il est stable.

1.3.3.2.1 Thermodynamique d'une hybridation

L'hybridation d'un oligonucléotide à un substrat d'ADN qui n'est pas soumis à la force d'une fourche pour l'expulser, est un problème classique en biophysique dont la thermodynamique est caractérisée par la variation d'énergie libre ΔG que l'on peut calculer en sommant les énergies des différents dinucléotides qui le constituent (SantaLucia 1998). L'oligonucléotide est en équilibre entre une phase hybridée et une phase détachée que l'on peut caractériser par la variable K_d qui définit sa constante affinité. Avec $K_d = \frac{K_{on} C_{oligo}}{K_{off}}$ où C_{oligo} est la concentration d'oligonucléotide, K_{off} est le taux de décrochage avec $K_{off} = \frac{1}{T_{disp}}$ et K_{on} est le taux d'hybridation. On a $\Delta G = RTlog_e(K_d)$, donc l'affinité $K_d = e^{\frac{\Delta G}{RT}}$ avec ΔG négatif. Ces données thermodynamiques ne précisent pas la valeur de T_{disp}, mais le rapport des temps d'accrochage et de décrochage. Il faut donc mesurer l'une de ces constantes pour pouvoir prédire la cinétique.

Dans le cas où la fourche est soumise à une force inférieure à 15 pN et donc où elle se referme, le temps de déplacement T_{disp} est beaucoup plus court, car la fourche tend à éjecter l'oligonucléotide. *Simona Cocco* a proposé un modèle qui permet de calculer T_{disp} en fonction du ΔG de l'oligonucléotide, des énergies élastiques de l'ADN simple brin et double brin et de la force appliquée. En pratique, nous pouvons utiliser cette approche pour prédire T_{disp}, et nous observons que K_{on} dépend peu de la longueur de l'oligonucléotide on a $T_{on} = \frac{1}{K_{on} C_{oligo}}$, donc la probabilité qu'un oligonucléotide s'hybride

dépend principalement de sa concentration et de la température.

Outre la nécessité d'adapter la composition des oligonucléotides pour améliorer la qualité de la signature résultante, ce que nous verrons dans le chapitre suivant, il est important de comprendre la dynamique des hybridations.

Les oligonucléotides étant des molécules d'ADN simple brin, le T_{disp} de l'hybridation est maximale lorsque l'hybridation s'effectue avec de l'ADN strictement complémentaire, cependant, si l'affinité de l'oligonucléotide est suffisante, nous pouvons constater que le T_{disp} peut rester suffisant pour être mesurable avec les pinces magnétiques dans le cas d'hybridation partiellement complémentaire (Turner 1996) . Nous parlerons dans ce cas de mésappariement. D'autres situations peuvent provoquer des situations d'hybridation mesurable inattendue, comme l'hybridation triple brins.

A contrario, certaines conditions peuvent augmenter le K_{on} des oligonucléotides. Si les oligonucléotides ont une affinité trop forte avec eux-mêmes, cela peut réduire leur disponibilité pour s'hybrider avec des molécules en épingle à cheveux, ce qui peut se produire sur des oligonucléotides ayant une séquence complémentaire avec eux-mêmes, et pouvant donc s'autohybrider. De la même manière, un oligonucléotide possédant des extrémités complémentaires peut éventuellement se replier sur lui-même réduisant sa disponibilité.

Nous allons voir dans les chapitres suivants qu'il est souvent intéressant de diminuer la taille des oligonucléotides de manière à augmenter leur fréquence d'apparition le long d'une séquence. Cette réduction de la taille a un effet important sur le T_{disp} fig. 1.30, à partir d'une certaine taille (pour l'ADN, $8Bp$, selon la composition de l'oligonucléotide), le T_{disp} n'est plus suffisant pour être détecté par les pinces magnétiques. Pour pallier ces problèmes, nous utilisons des modifications de l'ADN, qui augmentent fortement le T_{disp} des oligonucléotides. Alors que les propriétés thermodynamiques des bases ADN présentes dans la nature sont bien connues, la qualité des données thermodynamiques dans le cas des bases modifiées est moins bonne, comme nous le verrons dans la suite du chapitre.

Dans le cas de ces oligonucléotides, nous ne pouvons utiliser le modèle présenté ci-dessous pour calculer leur dynamique que de manière partielle. Mais, en utilisant un sous-ensemble réduit d'oligonucléotides, nous pouvons déterminer, et optimisons empiriquement les paramètres expérimentaux de manière à obtenir un T_{disp}, et k_{on} compatibles avec l'acquisition de signatures.

1.3.3.2.2 Utilisation de cycles d'ouverture et fermeture de molécule

Le processus d'ouverture et fermeture tel que nous venons de le présenter n'altère pas la molécule en épingle à cheveux, tant qu'elle ne casse pas. Il est donc possible d'effectuer de nombreux cycles d'ouverture et de fermeture successive de molécules.

Ce procédé permet à chaque cycle d'observer une ou plusieurs hybridations sur la molécule, aux positions pour lesquelles les oligonucléotides sont spécifiques. La répétition des cycles permet de détecter le plus grand nombre de positions d'hybridations spécifiques avec une statistique satisfaisante. Pour éviter les problématiques de masquage[8], c'est-à-dire qu'une hybridation proche de l'apex (centre de nucléation) de la molécule en épingle à cheveux ne dure jusqu'à la fin du cycle, masquant les positions auxquelles sont hybridés d'autres oligonucléotides situés plus loin, on limite la concentration des oligonucléotides pour avoir un taux d'hybridation à une position donnée de 10% en moyenne, et on ajuste le temps d'expérience pour accommoder le temps caractéristique d'hybridations de l'oligonucléotide.

[8] cf. le 3ème cycle de la Figure 1.31

La dynamique d'hybridation d'anticorps spécifiques aux groupes méthyle est similaire à celle d'oligonucléotides. Nous pouvons donc les utiliser de la même manière et sur les mêmes molécules pour déterminer l'état de méthylations de celles-ci.

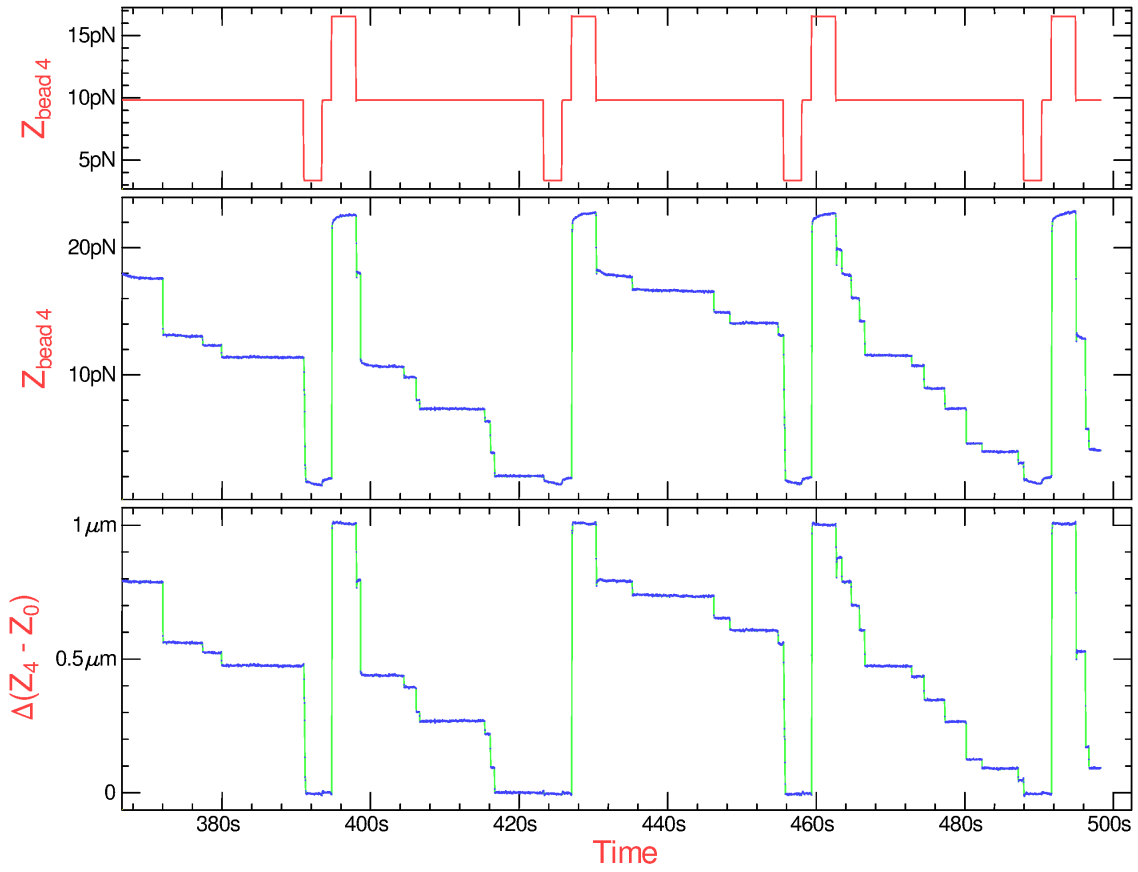


Figure 1.31: Signal d'hybridations multiples observé lors de cycles d'ouverture et de fermeture d'une molécule en épingle à cheveux et illustration de petites dérives thermiques. En haut, la modulation de force imposée : une phase à 17 pN permet d'ouvrir les molécules, une seconde phase à un peu moins de 10 pN permet de la refermer et d'observer les blocages. Une troisième phase à 3 pN permet d'enlever tous les oligonucléotides et de nettoyer la molécule pour le cycle suivant. Juste après cette phase, la force est augmentée à 10 pN avant de monter à 17 pN. Cette petite phase à 10 pN permet de vérifier que la molécule est bien fermée et elle nous sert de point de référence pour l'extension de la molécule. Au milieu, le signal expérimental brut observé montre des blocages dus à un oligonucléotide de 9 bases qui peut s'hybrider 18 fois sur la molécule en épingle à cheveux. Chaque fois que la force est modifiée, une petite relaxation est induite par la différence de température entre les aimants et l'échantillon. En bas, on a retranché à ce signal celui d'une bille collée à la surface qui permet d'enlever la quasi-totalité des signaux parasites.

L'utilisation de ces cycles constitue une faiblesse, mais à la fois la force du système. En effet, la multiplication des cycles augmente la durée totale d'une acquisition de données. Mais la possibilité d'utiliser un grand nombre de cycles sans altérer la molécule permet d'utiliser autant de sondes que nécessaire pour récupérer toutes les informations qui pourront paraître pertinentes sur la molécule.

1.3.3.2.2.1 Dérives de la mesure

Comme nous l'avons décrit dans la sec. 1.1.3.4, le système optique du microscope à pinces magnétiques est très sensible à la température. Pour assurer une précision de l'ordre du nanomètre, il est nécessaire de mettre en place une régulation fine de la température.[9] Dans la version de l'instrument utilisé, les aimants ne sont pas asservis en température. À chaque changement de force une légère relaxation de la température peut-être observée au cours du temps. Cela modifie la hauteur mesurée de molécule, sans changement physique de la molécule elle-même.

À court terme, d'autres modifications de mesure peuvent être observées. Le lien entre les anticorps qui recouvre la lamelle de verre et la queue dioxigénine des molécules en épingle à cheveux étant faible, nous pouvons observer des phénomènes de décrochages des dioxigénines provoquant des sauts de hauteurs, notamment pour les billes en positions ouvertes, toute la mesure de la molécule est alors décalé. Les dioxigénines ne se replaçant pas toujours aux mêmes positions sur la surface. Plusieurs décalages et recalages à des hauteurs différentes peuvent être observés au cours d'une expérience.

Certaines autres conditions expérimentales peuvent avoir une influence sur la stabilité à long terme de la position en hauteur. Le Mylar étant souple, une modification de la pression peut entraîner sa déformation. Par exemple, la présence et le déplacement de bulles d'air à l'entrée ou à l'intérieur de la cellule, ou encore la présence d'un flux du buffer dans la cellule. Mais aussi tout simplement de lentes dérives thermiques, car certaines parties de l'instrument de mesure ne sont pas thermalisées.

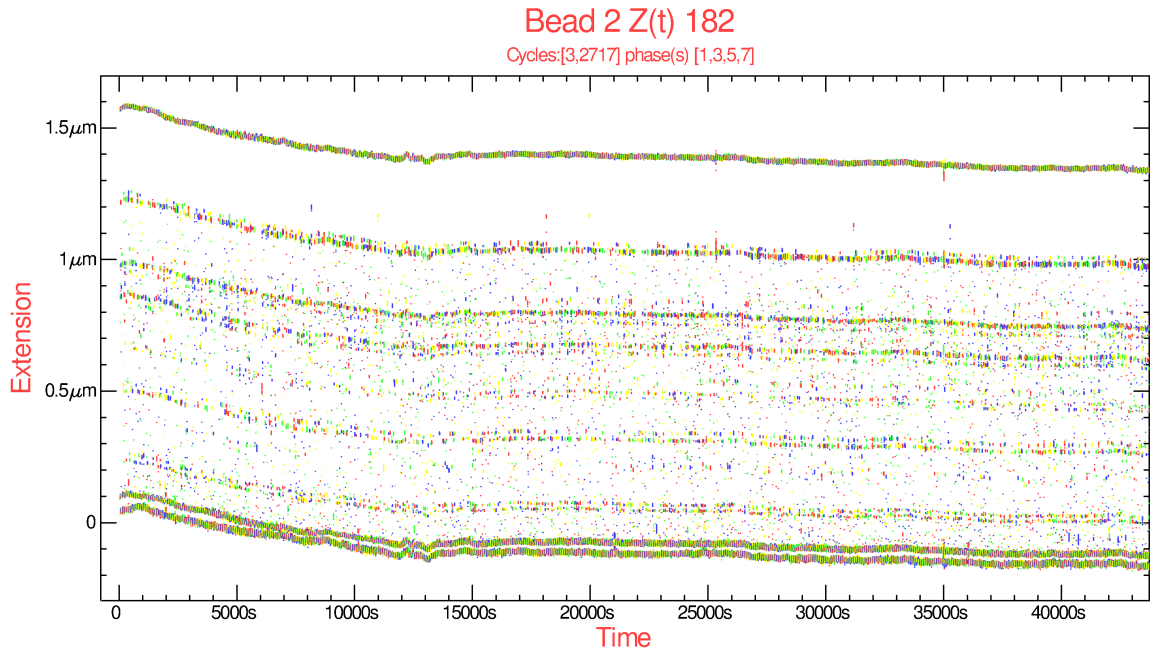


Figure 1.32: Dérives des positions de blocage au cours du temps. Sur cet enregistrement de plus de 10 heures, on voit que les positions de blocages et de références dérivent doucement dans le temps. Le fait de recaler chaque cycle à sa phase de référence permet de supprimer la plus grande partie de ces dérives.

[9] Dans le cas des systèmes actuels, au centième de kelvin

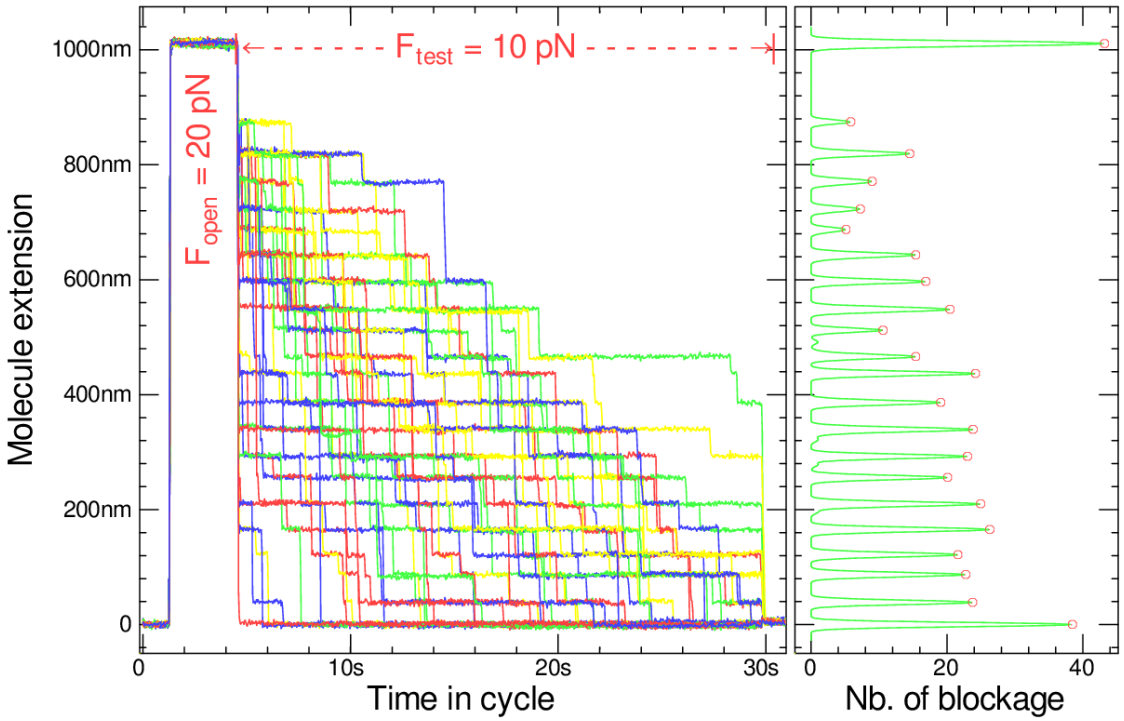


Figure 1.33: Superposition de 42 cycles d'ouverture et de fermeture illustrant le principe de la mesure. À gauche, les cycles auxquels on a enlevé les dérives thermiques sont superposés, les blocages présentent des durées aléatoires, mais leurs positions se superposent avec une assez bonne précision. À droite, l'histogramme des positions de blocage permet de mesurer précisément leurs positions. L'intensité des pics représente le nombre de blocages observés à chaque position. On peut également mesurer pour chaque pic la durée du blocage T_{disp} (qui n'est pas représentée ici). Plusieurs points sont importants pour pouvoir corréler ces mesures à la séquence de la molécule d'ADN. Le plus important est la précision de la mesure : une base correspond typiquement à 0.9 nm tandis que le bruit dans chaque blocage est de l'ordre de 2 nm. L'expérience représentée ici est de très bonne qualité, les différents blocages se superposent de manière remarquable, il y a peu de dérives, mais surtout les positions d'hybridation sont réparties avec une très grande régularité, les pics de l'histogramme sont très bien différenciés. Ce n'est malheureusement pas le cas généralement.

1.3.3.3 Extraction des positions d'hybridation

Pour extraire les positions de blocage, il faut superposer le résultat de plusieurs cycles d'ouverture et de fermeture. Nous réalisons cette opération en alignant les cycles en phase et en ajustant leur position en z pour qu'elle soit nulle durant la phase de référence (entre 0 et 1,5S dans la figure fig. 1.33). Avant cette opération, on soustrait au signal des billes intéressantes celui de billes fixes. Ceci permet d'enlever une partie des dérives. Pour trouver les positions de blocage dans la phase de mesure [4.5S,30S], nous utilisons un algorithme qui fait l'histogramme des points de mesures en z de tous les cycles. Lorsque la fourche se déplace entre deux blocages, un ou deux points sont associés, par contre chaque blocage implique un grand nombre de points à la même position z. L'histogramme ainsi construit présente des pics aux positions de blocage, chacun d'eux possède une largeur de quelques nanomètres qui représente le bruit de mesure expérimental. Une fois les maximums de cet histogramme trouvés, on dispose de la position de blocage et il est alors facile de mesurer la durée des blocages et le nombre de fois que chacun est visité.

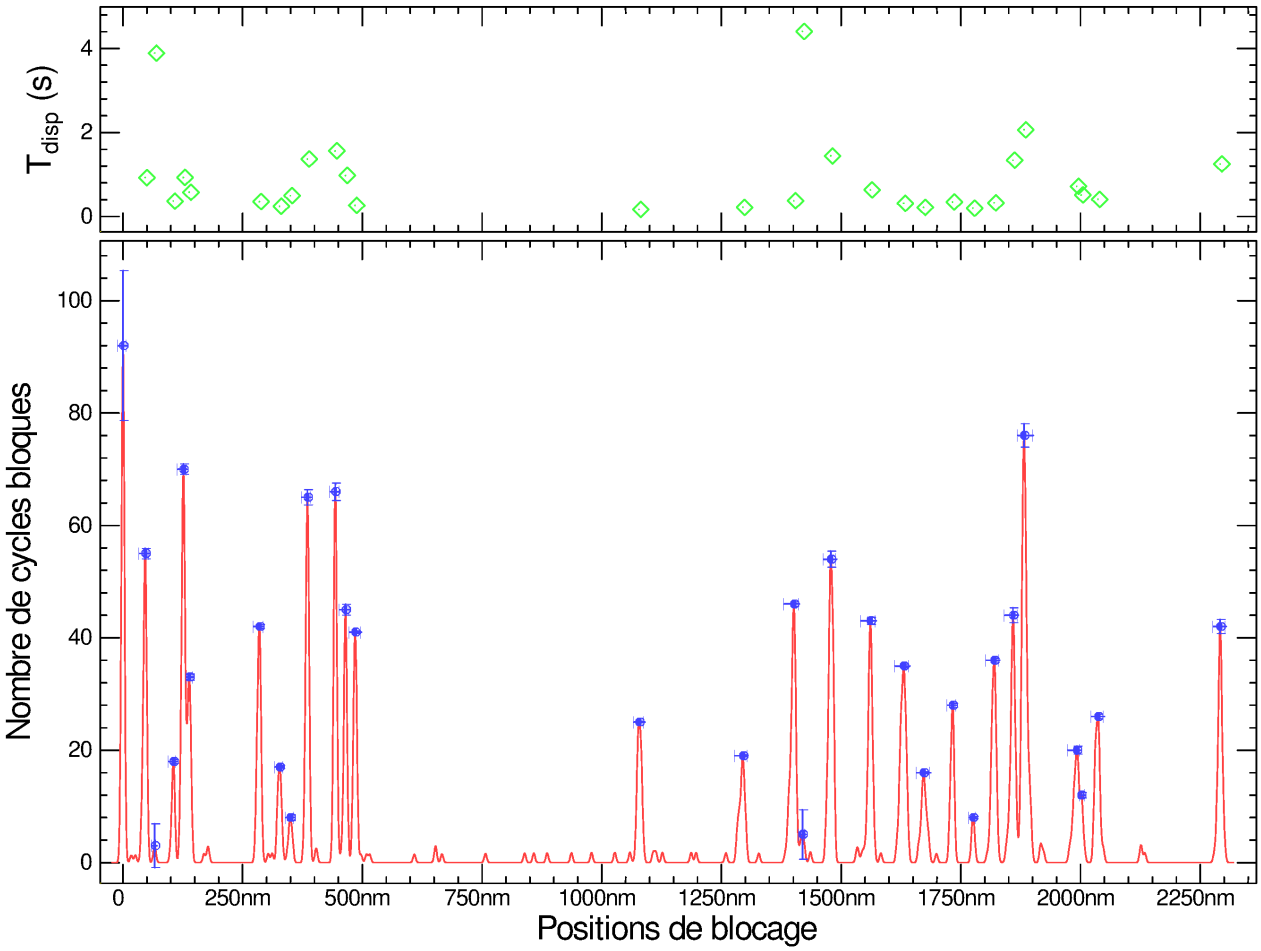


Figure 1.34: Exemple d'une empreinte digitale obtenue sur une molécule en épingle à cheveux de 2.5 kb avec l'oligonucléotide CGTC. En bas, nous observons la position des blocages et leur fréquence : le pic en 0 correspond à la molécule fermée, il ne fait pas partie de la signature. Il nous indique le nombre de cycles total. Les blocages présentent des fréquences assez variables: tandis qu'une proportion de pics a une fréquence supérieure à 20, il existe des blocages nettement moins fréquents. La courbe du haut indique le temps de blocage T_{disp} mesuré pour les pics les plus importants. Il existe également un certain nombre de pics très petits qui est probablement non significatif.

Ce prétraitement des données est une étape décisive qui détermine la qualité du résultat final. Si ce prétraitement donne souvent de bons résultats, comme on peut le voir ici, il existe souvent des cas où le prétraitement ne fonctionne pas aussi bien. Sans rentrer dans tous les détails, voici quelques raisons typiques : dans certaines expériences, la ou les billes fixes ne le sont pas vraiment ou cessent de l'être après un certain nombre de cycles. Par ailleurs, dans l'opération de soustraction d'une bille fixe, on enlève bien les dérives, mais on rajoute le bruit de la bille fixe au signal. Si la bille fixe ne l'est pas, ou si son bruit est trop important, on ne peut pas utiliser cette procédure, on peut alors utiliser la répétabilité des dérives à court terme pour les soustraire, mais cette procédure est moins bonne et les pics des histogrammes s'élargissent dégradant la précision. L'étirement de la molécule dérive également un peu avec le temps ce qui conduit aussi à un élargissement des pics de l'histogramme. En pratique si deux positions de blocage sont trop proches, les pics correspondant dans l'histogramme vont se chevaucher et compliquer leur mesure.

1.3.3.4 Cartographie d'une molécule sur un génome de référence

Nous avons vu lors de sec. 1.3.3.2 que nous pouvons détecter l'hybridation d'oligonucléotides le long d'une molécule d'ADN au moyen de pinces magnétiques. L'ensemble des positions de ces hybridations constitue la signature de nos molécules.

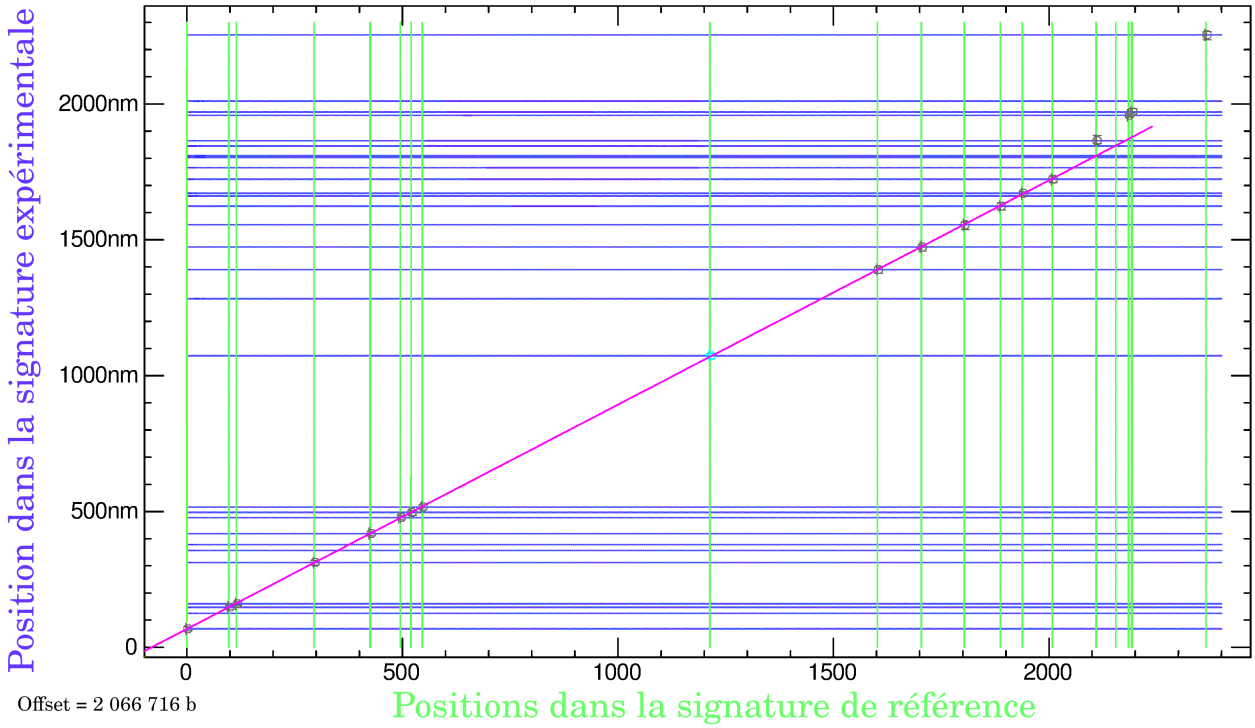


Figure 1.35: Comparaison d'un locus particulier du génome de référence d'E. coli avec une signature expérimentale, en utilisant la méthode par régression Voir 3. En l'occurrence, on constate une très bonne adéquation entre les deux signatures

Comme nous allons le voir par la suite, l'identification d'une signature est d'autant plus efficace qu'un grand nombre d'hybridations est observé. Par ailleurs, ce nombre d'hybridations dépend fortement de la séquence de l'oligonucléotide. Utiliser deux oligonucléotides différents permet d'augmenter le nombre d'hybridations. On peut alors soit les faire passer l'un après l'autre dans la cellule, soit les mélanger. Une signature est *homogène*, lorsque l'on peut associer de manière univoque chaque position d'hybridation à un oligonucléotide. Une signature peut-être considérée comme homogène lorsqu'elle n'est composée que des hybridations d'un unique oligonucléotide, ou lorsque l'on a mesuré les hybridations de chaque oligonucléotide séparément. Une signature est *hétérogène* lorsqu'il n'est pas possible d'associer chaque position d'hybridation à un oligonucléotide précis, par exemple, lorsque plusieurs oligonucléotides sont testés simultanément.

1.3.4 Contraintes liées à l'acquisition

1.3.4.1 Bruit de mesure

Lors du repliement d'une molécule, la détermination de la position de blocage par une hybridation est observée grâce aux interférences optiques qui décorent l'image de la bille. Cette mesure souffre de plusieurs causes de bruit qui en perturbent la précision. Le bruit de photons ajoute des fluctuations

de longueur de l'ordre de 1 nm entre deux images consécutives, le mouvement brownien de la bille attachée à sa molécule apporte une seconde contribution qui dépend de la longueur de la molécule et de sa raideur donc de la force F appliqué. La contribution de ce mouvement brownien est de l'ordre de 1 à 2 nm. À ces sources intrinsèques de bruit, viennent s'ajouter des dérives lentes du bruit en $\frac{1}{f}$) et quelques perturbations extérieures. Dans les cas favorables, le bruit de mesure est de 2 nm, mais il peut atteindre facilement 5 nm dans les cas défavorables. Dans nos expériences, une paire de bases correspond à peu près à 1 nm.

Ce bruit sur la position des hybridations engendre divers effets sur la qualité des signatures produites. Il va créer des ambiguïtés lors de la reconnaissance des certaines hybridations trop proches les unes des autres. Celles-ci sont alors fusionnées dans la signature expérimentale. Le bruit de mesure réduit également la diversité des signatures. La probabilité que deux signatures correspondant à des positions différentes sur le génome de référence soient considérées comme identiques augmente.

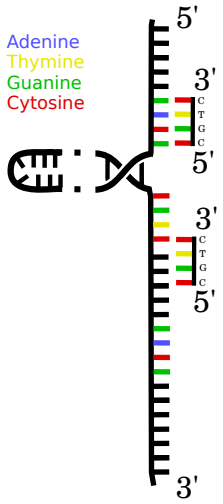


Figure 1.36: Deux hybridations d'oligonucléotides suffisamment proches pour qu'elle ne puisse être distinguée avec la résolution des pinces magnétiques

1.3.4.1.1 Étirement variable d'une molécule.

La force appliquée à chaque molécule varie légèrement, car toutes les billes n'ont pas exactement la même quantité de matière magnétique. Celle-ci varie de 10 à 20%. Les variations de la force de bille à bille produisent des étirements variables d'une molécule à l'autre. La distribution de l'étirement observée lors des expériences est de

$$X \sim \mathcal{N}(\mu \simeq 0.88; \sigma \simeq 0.04)$$

Ne pouvant pas déterminer au préalable l'étirement qu'aura chaque molécule observée, il faut pouvoir considérer deux signatures identiques à un étirement près comme étant identiques strictement.

L'étirement d'une molécule est considéré principalement comme homogène. La relation entre les positions de blocage et les l'index de la base dans la séquence est donc linéaire. On constate cependant qu'il existe de petites non-linéarités dans cet étirement dont l'origine n'est pas entièrement comprise à l'heure actuelle. Les signatures devraient également admettre une certaine tolérance à ce type de non-linéarité.

1.3.4.1.2 Hybridations manquantes

Le nombre d'hybridations détectées à une position donnée est régi par une loi de poisson dépendant des paramètres k_{on} , [Oli]. k_{on} varie avec les conditions expérimentales: température T, teneur en sels du tampon [Na+] et [Mg2+], force appliquée F. L'expérience permet de déterminer le nombre de cycles où l'hybridation a été observée N_h et le temps d'accrochage T_{off}. On a $T_{off} = \frac{1}{k_{off}}$ et $N_h = k_{on}[Oli]T_{unzip}$. Où T_{unzip} est le temps pendant lequel la molécule d'ADN est ouverte. En ajustant la concentration de l'oligonucléotide [Oli], dans une moindre mesure T_{unzip} et surtout en choisissant le nombre de cycles d'ouverture / fermeture de la molécule Voir Signature par pince magnétique 1.3, on peut obtenir une valeur de N_h qui garantisse que cette hybridation est pertinente et ainsi limiter le nombre d'hybridations manquantes.

$$O + T \underset{k_{off}}{\overset{k_{on}}{\rightleftarrows}} OT$$

. La concentration en template T est négligeable dans cette réaction, au regard de la concentration en oligonucléotides. Le k_{on} est affecté par la concentration en oligonucléotides, mais le k_{off} est indépendant de la concentration en Template et est uniquement fonction de la température, de la teneur en sels et de la force appliquée.

1.4 Conclusion

À l'issue de cette partie, nous avons pu décrire de manière générale le système des pinces magnétiques, l'acquisition de données sur ce système, la méthode de construction des molécules en épingle à cheveux et la manière d'en tirer un ensemble de position d'hybridations. Nous avons pu voir également les limites connues relatives aux données acquises, et un ensemble de procédures permettant de représenter les données informatiquement en vue de les exploiter plus tard.

Chapitre 2

Sélection d'oligonucléotides pour signature

Sommaire du chapitre

2.1 Introduction

Lors du précédent chapitre, nous avons pu discuter à la fois du système de pinces magnétiques, ainsi que de la manière de les utiliser pour créer des signatures de molécules. Nous avons également présenté un ensemble d'hypothèses de simulation, pour reproduire *in silico* des molécules semblables aux données expérimentales.

Cependant, pour réaliser des expériences qui apportent une bonne spécificité de signatures, il est nécessaire de maîtriser un des paramètres essentiels de l'expérience: les oligonucléotides. Après avoir discuté de la nécessité de cette sélection, je reviendrai sur les critères que nous pouvons définir pour les choisir de manière efficace. Ces contraintes me servent ensuite à développer un outil permettant de définir un ensemble d'oligonucléotides répondant à ces contraintes. Ce jeu d'oligonucléotides sera validé expérimentalement. Cependant, cette sélection est devenue moins utile avec la découverte par Depixus au cours de ce travail de la possibilité d'utiliser des oligonucléotides de 4 bases qui présentent un grand nombre d'hybridations. La démarche de sélection reste pour autant intéressante, et soulève de nouvelles interrogations par rapport à la spécificité des hybridations d'oligonucléotides.

2.1.1 Nécessité de la sélection des oligonucléotides

2.1.1.1 Génome non aléatoire et fréquence des positions d'hybridation

En observant les génomes de différents organismes vivants, il n'est pas surprenant de constater que les schémas de nucléotides de petite taille ne sont pas uniformément distribués (Holste, Grosse, et Herzel 2001) : le génome d'un organisme vivant n'est pas aléatoire et dispose d'une organisation interne. Par exemple, en considérant les 3-mers, nous pourrions observer un biais dans les parties codantes de l'ADN, correspondant à la succession de codons. Par conséquent, le modèle aléatoire considérant qu'un k-mer devrait en moyenne apparaître toutes les 4^k bases n'est pas suffisant pour expliquer les fréquences d'hybridation sur le génome d'un organisme donné.

De même, nous pouvons constater en observant la Figure 2.1 qu'un oligonucléotide s'hybridant très fréquemment sur un génome donné pourra s'hybrider de manière dramatiquement plus faible dans un autre génome. La distribution des fréquences est différente.

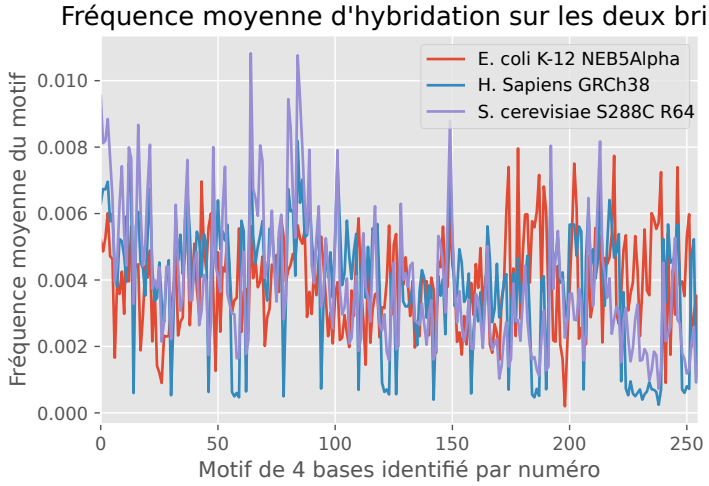


Figure 2.1: Nombre d'occurrences de chaque k-mer de taille k sur le génome d'E. coli et sur le génome humain. On remarque des variations assez grandes justifiant le choix de certains oligonucléotides.

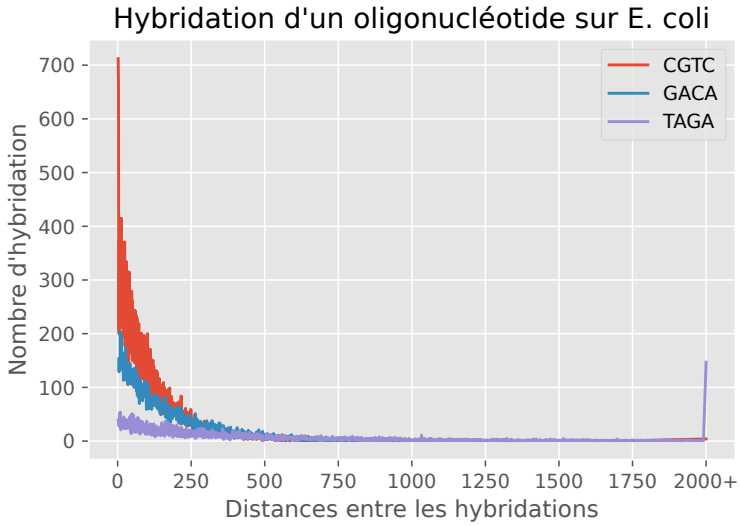


Figure 2.2: Distribution des distances entre deux hybridations pour trois k-mers.

Nous allons voir dans la partie sec. 4 que la spécificité d'une signature dépend du nombre d'hybridations qui la compose, et de la complexité du génome. Par conséquent, nous devons définir une stratégie visant à optimiser le choix des oligonucléotides utilisés pour définir cette signature. De plus, étant donné la variabilité des fréquences des k-mers entre les génomes, il sera nécessaire de réaliser cette optimisation pour chaque génome étudié.

Dans un génome aléatoire et pour un oligonucléotide donné, nous nous attendons à ce que la distribution des distances entre deux hybridations répondent à une loi des événements rare ou loi de Poisson. Bien que les génomes ne soient pas aléatoires, la figure 2.2 semble montrer que cette caractéristique reste conservée. Cette distribution sera essentielle par la suite à la fois pour la sélection des oligonucléotides, mais également dans le processus de cartographie en tant que tel.

2.1.1.2 Compromis de taille / fréquence pour le choix des oligonucléotides

Des oligonucléotides longs offriront une forte spécificité sur le génome, et seront extrêmement discriminants, cependant, pour obtenir une signature spécifique pour chaque partie du génome, il en faudrait un très grand nombre. Un oligo de 11 bases s'hybriderait environ tous les 4^{11} pairs de bases, soit en moyenne, deux fois sur le génome d'*E.coli* (une fois sur chaque brin), cependant, pour que chaque signature soit représentée, il faudrait hybrider un à un chaque oligonucléotide de 11 bases sur le génome. Même en groupant un très grand nombre d'oligonucléotides ensemble, le coût et le temps d'une expérience seraient très importants. Il est donc nécessaire de réduire la taille des oligonucléotides. En réduisant la taille, nous obtenons un plus grand nombre d'hybridations par signature. Mais, si ce nombre devient trop important, les hybridations sont trop proches passent sous la résolution de notre système. Nous ne sommes plus capables de distinguer deux hybridations. Les signatures deviennent donc moins spécifiques.

Un second problème s'ajoute. Le temps d'hybridation des oligonuclétotides dépend notamment de leur taille. un oligo long (ex: 11 bases), aura un temps d'hybridation moyen très long ce qui oblige à augmenter la durée des expériences. À contrario, avec des oligonucléotides courts (ex: 6 bases) Le temps d'hybridation devient trop court pour la résolution temporelle du système. Pour avoir des expériences de durée raisonnable, nous souhaitons un temps d'hybridations moyen supérieur à 1 seconde et inférieur à 10 seconde.

Il s'agit donc de trouver un compromis acceptable entre quantité d'hybridations, temps d'hybridations et spécificité des signatures.

Nous venons de voir les contraintes liées aux fréquences d'occurrences des oligonucléotides dans les génomes des organismes. D'autres contraintes liées directement au comportement des oligonucléotides dans une expérience d'hybridation viennent s'y ajouter. L'ensemble de ces contraintes délimite l'espace des oligonucléotides utilisable pour la création de signatures.

Nous avons vu dans la Section 1.3.3.2.1 que le temps et le taux d'hybridation dans les cycles d'une expérience dépendaient de la taille et de la composition des oligonucléotides. Différentes modifications permettent de réduire la taille des oligonucléotides tout en maintenant un niveau de signal correct. Nous avons décidé dans un premier temps de nous limiter aux modifications les plus simples, en utilisant des oligonucléotides uniquement composés de base LNA[1](Koshkin et al. 1998). Nous avons choisi de travailler avec des oligonucléotides de taille minimum 6 bases avec 4 bases GC (bases fortes) pour nous assurer d'une hybridation suffisamment spécifique. Le problème principal avec l'utilisation des bases LNA est la difficulté de prédire le temps de blocage avec précisions. Simona Cocco et V. Croquette ont écrit un logiciel calculant le temps de déplacement d'un oligonucléotide sous la pression d'une fourche d'ADN en se basant sur les énergies libres des proches voisins [@{SantaLuciaunified-viewpolymer1998]. Ce logiciel évalue le temps pour décrocher chaque paire de bases en comparant ces énergies libres d'appariement aux énergies mécaniques correspondant aux changements de longueur de la molécule. Pour des oligonucléotides d'ADN, l'accord entre cette modélisation et les données expérimentales est assez bon comme on peut le voir sur la figure fig. 1.30. Avec des oligonucléotides d'ADN, il faut au minimum 9 bases pour obtenir un blocage mesurable. Nous avons remplacé les informations des énergies de proches voisins de l'ADN par les valeurs publiées pour des nucléotides LNA (You et al. 2006) et appliqué la même méthode pour prédire leur temps de blocage. Comme on peut le voir dans le tableau tbl. **??**, les temps de blocage prédit sont beaucoup trop grands par rapport à ceux que nous avons mesuré fig. 2.6, ceci est dû au fait qu'une petite erreur sur les énergies libres est amplifiée exponentiellement par la méthode utilisée. Néanmoins, les variations relatives d'un oligonucléotide à l'autre sont cohérentes avec les résultats expérimentaux et conduise déjà à des conditions d'utilisations très différentes. Les oligonucléotides ayant les temps les plus courts ont des chances de ne pas provoquer de blocages assez nets pour être détectés.

Cependant, nous observons que l'hybridation des oligonucléotides de 6 bases dont le motif est le plus présent sur E. coli donne lieu en moyenne à un nombre d'hybridations très faible pour un fragment donné du génome, d'environ 3 hybridations par fragments pour un fragment de 2000Bp. Et surtout, ce nombre est inférieur au nombre minimal permettant d'atteindre une signature unique Voir section 4. Cette stratégie n'est donc pas raisonnable pour générer une signature.

[1]Locked Nucleic Acid : Nucléotide RNA modifié ayant une affinité très importante à l'ADN

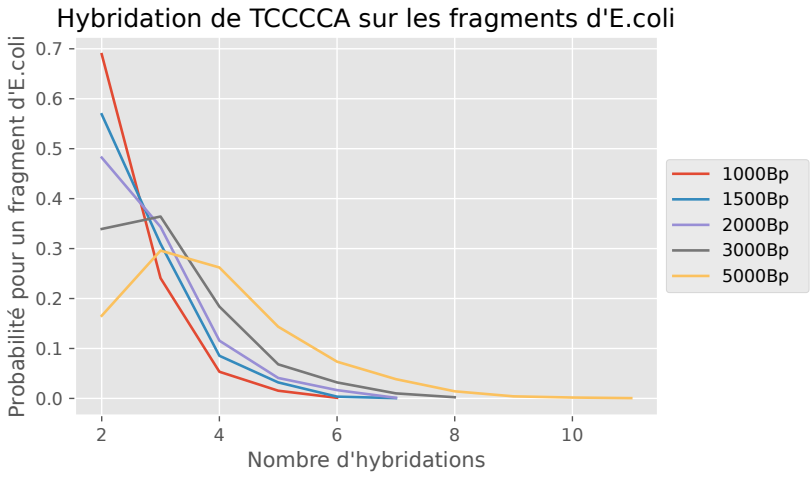


Figure 2.3: Distribution du nombre d'hybridations de l'oligo TCCCCA dans un fragment d'ADN de taille variable le long d'E.coli. On observe que même avec un fragment de 5000Bp, le maximum la distribution est à ~3 Hybridations, ce qui est insuffisant pour une cartographie.

Puisqu'un oligonucléotide ne peut pas, à lui seul générer une signature spécifique, nous devons envisager d'utiliser plusieurs oligonucléotides sur une même molécule.

Pour cela deux manières de faire sont possibles. Les oligonucléotides peuvent être injectés successivement pour former une signature où chaque position d'hybridation est libellée avec un oligonucléotide différent ou alors, on peut mélanger plusieurs oligonucléotides et générer la signature à partir de ce mélange d'oligonucléotides.

Dans les deux cas, les positions d'hybridation de la signature seront très similaires, dans la mesure où les positions d'hybridations ont une distance supérieure à la résolution du système. Cependant, en injectant successivement les oligonucléotides nous disposons d'une information sur les positions respectives de chaque oligonucléotide le long du fragment d'ADN, et donc une signature plus spécifique.

L'acquisition de ces informations supplémentaire a néanmoins un coût. En effet, la durée moyenne d'une acquisition est multipliée par le nombre d'oligonucléotides testés. Le traitement et l'alignement des données s'en trouvent complexifié, et la probabilité d'avoir un problème sur une molécule est augmentée.

En utilisant la méthode du mélange d'oligonucléotides, nous rencontrons d'autres problèmes, notamment celui d'interactions particulières entre les oligonucléotides. Mais en mettant face à face le coûts supplémentaires inhérents à l'utilisation d'oligonucléotides séparés, et la spécificité estimé, d'un mélange d'oligonucléotides, le choix la été fait d'utiliser la méthode du mélange dans un premier temps.

2.1.2 Utilisation simultanée d'oligonucléotides

L'utilisation simultanée de plusieurs oligonucléotides induit des contraintes supplémentaires par rapport leur à utilisation séparée :

2.1.2.1 Contraintes sur la séquence

Tout d'abord, nous devons nous assurer que les hybridations de deux oligonucléotides sur le fragment d'intérêt ne pourront pas être confondues. Si les extrémités des oligonucléotides partagent une partie

de leur séquence, leur probabilité de s'hybrider de manière proche augmente. Soit k correspondant à la taille des oligonucléotides et s_d, s_g le nombre de bases partagées par les deux oligonucléotides aux extrémités droites et gauches, la distribution des hybridations se chevauchant sur un brin correspond à la probabilité de l'hybridation des deux oligonucléotides successifs duquel on retire la séquence partagée : $X \sim B(N, \frac{4^{s_g}+4^{s_d}}{4^{2k}})$. Nous devons maintenant considérer les deux brins de la molécule d'ADN d'intérêt. En effet, un oligonucléotide peut s'hybrider sur un brin et sur son complémentaire. La séquence partagée dans notre cas nécessite la comparaison de la séquence des deux oligonucléotides, mais également la comparaison du premier oligonucléotide avec le complément renversé du second. Soit s'_d et s'_g les bases partagées avec le complément inverse : $X \sim B(N, \frac{2(4^{s_g}+4^{s_d}+4^{s'_g}+4^{s'_d})}{4^{2k}})$, Ce problème de séquence chevauchante se pose uniquement sur les oligos de petite taille, quand le chevauchement provoquerait l'impossibilité de résoudre deux hybridations avec la résolution du système.

2.1.2.2 Compatibilité : dynamique d'hybridation

Pour qu'un oligonucléotide soit utilisable, nous devons également éviter les risques d'autohybridation entre deux copies de l'oligonucléotide, ce qui se produirait si sa séquence comporte une région palindromique. Par exemple, TACGTA est un palindrome qui s'hybriderait sur toute sa longueur avec lui même, diminuant fortement sa disponibilité pour s'hybrider avec la séquence cible. Le problème se pose également dans le cas ou plusieurs oligonucléotides sont utilisés en mélange.

Par ailleurs, nous devons nous assurer que le taux et le temps d'hybridation des différents oligonucléotides soient équivalents, de manière à obtenir des données avec des amplitudes du même ordre de grandeur. Ne possédant pas d'outils *in silico* suffisant lors du développement de cette méthode, nous nous sommes basés sur des données empiriques.

2.2 Méthode de sélection d'oligonucléotides

L'objectif de la méthode est, pour un ensemble d'oligonucléotides et un génome de référence donné, de sélectionner le sous-ensemble d'oligonucléotides qui minimise la probabilité d'obtenir deux signatures identiques, en respectant les contraintes expérimentales présentées dans la partie sec. 1.3.4 et dans la section précédente.

Sont exclus d'office un ensemble d'oligonucléotides sur des critères liés à leur dynamique d'hybridation. Ceux-ci comprennent les critères empiriques connus sur la modification des bases : la taille des oligonucléotides doit être suffisante (5 à 6 bases minimum pour des oligonucléotides LNA). Les bases fortes doivent être majoritaires. Les bases Adénines doivent être en faible quantité, car ce sont les bases les moins fortes en LNA. Il ne doit pas être possible pour l'oligonucléotide de s'hybrider avec lui-même, et la possibilité de l'oligonucléotide de se replier sur lui-même est interdite.

Les oligonucléotides sont ensuite comparés deux à deux, de manière à constituer une matrice de compatibilité entre oligonucléotides qui vérifie le chevauchement des oligonucléotides sur la séquence et l'autohybridation des oligonucléotides entre eux.

Tous les oligonucléotides n'étant pas compatibles les uns avec les autres selon les critères susmentionnés, il est nécessaire de déterminer les ensembles d'oligonucléotides compatible entre eux, et permettant le meilleurs taux d'hybridation. Pour cela j'ai développé un algorithme dit *« greedy »* qui part des N meilleurs oligonucléotides basés sur la quantité d'hybridations sur le génome de référence, et à chaque itération, le meilleur oligonucléotide compatible avec les l'ensembles des oligonucléotides déjà sélectionnés par l'algorithme sont ajouté à la liste.

Nous obtenons ensuite N sous-ensembles d'oligonucléotides compatible qu'il est possible de classer en fonction de la quantité d'hybridations.

2.2.1 Sélection d'un jeu d'oligonucléotides adapté au génome d'E. coli

2.2.1.1 Choix des paramètres

Pour sélectionner un ensemble d'oligonucléotides utilisables pour un système de signature dans le génome d'E. coli, les paramètres suivants sont utilisés :

Paramètre	Valeur
Nombre de bases LNA	6
Taille des oligonucléotides	6
Nombre minimum de Cytosines / Guanines	4
Nombre maximum de bases pouvant s'hybrider aux extrémités entre les oligonucléotides	2
Nombre maximum de bases se chevauchant lors de l'hybridation sur une séquence	0

Les paramètres ont été choisis pour minimiser la taille des oligonucléotides tout en ayant un T_{off} suffisamment grand pour permettre d'être détecté. Nous avons choisi une taille de 6 bases avec au moins 4 bases fortes (GC), en nous basant sur les connaissances empiriques du laboratoire. Pour éviter que ces oligonucléotides ne s'hybrident trop fortement entre eux nous avons limité leur recouvrement à 2 bases.

2.2.1.2 Sélection in silico

À l'issue du processus de sélection, nous avons trouvé 78 Jeux d'oligonucléotides candidats.

Nous sélectionnons celui maximisant la spécificité des signatures

Table 2.2: Jeu d'oligonucléotide sélectionné et simulation du temps d'hybridation des oligonucléotides LNA lors de la fermeture de molécules en épingle à cheveux pour les paramètres expérimentaux typiques, en utilisant des données non publiées, ainsi que les adaptations des énergies LNA présentés dans (You et al. 2006)

Id	Séquence	Temps de blocage prédit
0	CCGCAA	118
1	CCGTCA	366
2	TCCCCA	4325
3	TCGCCA	169
4	CTGGCA	687
5	TCGGCA	140
6	TCCCGA	1300
7	CTGCGA	257
8	CCGCGA	644
9	CTGGGA	907
10	CCGTAG	235
11	CCGCAG	453

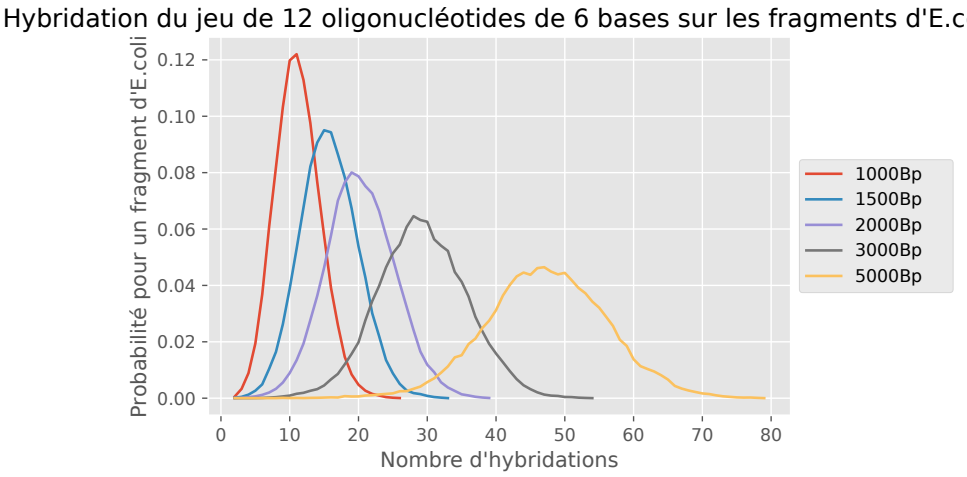


Figure 2.4: Représentation de la distribution de signature générée avec le jeu d'oligonucléotides tbl. 2.2 pour différentes tailles du fragment génomique.

Avec le jeu d'oligonucléotide sélectionné, 98% des fragments d'ADN de 2000 pb échantillonnés sur le génome d'*E. coli* disposent de plus de 10 hybridations le long de leur séquence. Les fragments de 2000 Bp sont construits en découpant des fragments de 2000Bp tous les 1500Bp le long du génome d'E. coli

Nous aimerions disposer d'un nombre d'hybridations supérieur de manière à nous assurer de la spécificité des signatures.[2] Cependant, pour des oligonucléotides de 6 bases et en tenant compte de nos critères de sélection, ce jeu de 12 oligonucléotides est celui qui offre le plus grand nombre d'hybridations moyen par fragment. Il semble suffisant pour continuer l'expérimentation.

À partir de ce jeu d'oligonucléotides théorique, il convient de vérifier expérimentalement les hypothèses que nous avons établies dans cette section.

2.2.2 Validation expérimentale des oligonucléotides

Les données thermodynamiques concernant les oligonucléotides d'ADN dans la littérature sont assez précises, mais ont été obtenues sur des oligonucléotides de taille minimale 9 à 10 bases. Nous savons relativement bien prédire leur temps de blocage sur une molécule lors d'une hybridation. Pour les oligos cependant, comme indiqué dans le paragraphe précédent, les contraintes d'hybridation nous imposent des oligonucléotides LNA de 6 bases, dont l'affinité est renforcée par rapport à une structure ADN. La thermodynamique pour les oligonucléotides LNA sont également relativement connues. Cependant, avec des limitations similaires que pour l'ADN. Pour ces raisons, les données thermodynamiques sont nettement moins bien connues pour les oligonucléotides LNA de 6 bases, ce qui pose des difficultés pour prédire le temps de blocage de nos oligonucléotides avec précision.

Pour éviter de sélectionner un oligonucléotide qui aurait un temps de blocage très court et afin de déterminer les oligonucléotides ayant un temps et une fréquence d'hybridation suffisante pour notre projet de signature[3] nous devons tester expérimentalement ces oligonucléotides. Cela peut se faire soit

[2]Nous verrons plus en détail le nombre d'hybridations nécessaires pour assurer la spécificité d'une signature pour E. coli dans la partie sec. 4

[3]Il est nécessaire de que le temps et le taux d'hybridations de nos oligonucléotides ne soient ni trop courtes, car elle ne serait pas visible par notre caméra cadencée entre 30Fps et 70Fps selon les configurations, ni trop longue, provoquant un phénomène de masquage. De même pour la fréquence d'hybridations, qui même si elle peut-être mitigé par l'adaptation de la concentration en oligonucléotide, il est souhaitable, d'avoir un taux d'hybridation similaire pour tous les oligonu-

en les testant sur des molécules en épingle à cheveux déjà construite et disponible au laboratoire dont la séquence présente une complémentarité exacte, soit en fabriquant une molécule en épingle à cheveux contenant cette séquence complémentaire afin de mesurer le temps de blocage expérimentalement. Une hybridation peut être considérée comme ayant une durée suffisante, quand elle dure plus d'une seconde en moyenne, ce qui permet une marge d'erreur largement suffisante pour différencier une hybridation du bruit de fond de l'expérience.

À partir des 12 oligonucléotides issus de la construction par l'algorithme glouton présenté en sec. 2.2, nous avons mis en place le protocole suivant afin de tester la dynamique de chacun des oligonucléotides.

Nous construisons *in silico* une séquence d'ADN d'intérêt contenant les 12 oligonucléotides d'intérêt, pour ensuite la faire synthétiser et l'incorporer à une molécule en épingle à cheveux. Le principe est de construire une séquence en concaténant les 12 oligonucléotides, chacun séparé par 4 bases d'espacement dont la distribution de nucléotides suit une loi binomiale ayant une probabilité de 70% de bases faible (Adénine et Thymine), de manière à équilibrer la présence de base forte (Cytosine et Guanine) dans la séquence des 12 oligonucléotides, et ainsi à limiter la quantité possible de structures secondaires pouvant provoquer des blocages structurels de la molécule[4]

La séquence d'intérêt sélectionnée est la suivante :

> Molécule FLY0:
>
> 5' - TGGGGAATTATCGGGACTCTCTACGGGTGTTGGCGATGAATGACGGTT
> AATGCCGAATCCCTGCGGCTTGTGCCAGACTCTCCCAGAATGTCGCGG
> ACTTTCGCAGAGCATTGCGGACTT - 3'

Nous avons ensuite incorporé la séquence d'intérêt dans une molécule en épingle à cheveux en suivant le protocole présenté en annexe. Le protocole a été conçu par Jimmy Ouellet, j'ai effectué la construction, avec la précieuse aide d'Anaïs Leseur, Laurène Giraut, Jérôme Maluenda, Sylwia Gorlach Andréas Lefevre. L'ensemble des manipulations sur pince magnétique et les analyses subséquentes mentionnées dans ce chapitre ont été effectués par moi-même.

En attendant la fin de la synthèse et de la fabrication de la molécule en épingle à cheveux constituée de notre séquence d'intérêt. Nous avons cherché au laboratoire et à Depixus, l'ensemble des molécules en notre possession, déjà préparé et permettant l'hybridation des 12 oligonucléotides. Leurs provenances importent peu dans le cadre de notre travail, c'est uniquement leur séquences qui est importantes. La séquence d'intérêt de chaque molécule est présente en annexe de cette thèse.

---

cléotides lors d'une expérience hétérogène, de manière à pouvoir déterminer un seuil unique d'acceptation des positions d'hybridation basée sur fréquence d'hybridation, et le temps.

[4]La présence de structures secondaire a été testée à l'aide de (UNAFold Markham et Zuker 2008)

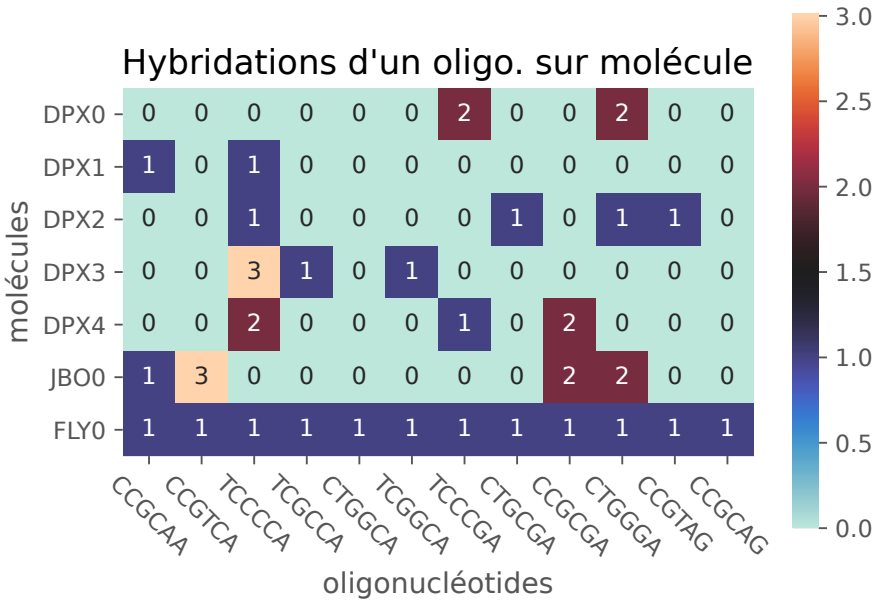


Figure 2.5: Nombre d'hybridation par molécule et oligonucléotide, pour les molécules en épingle à cheveux sélectionnées pour minimiser le nombre d'expériences nécessaires à la caractérisation de la dynamique d'hybridation des oligonucléotides. La molécule FLY0 est la molécule construite sur mesure.

Pour 10 des 12 oligonucléotides,[5] nous effectuons une expérience en répétant des cycles d'ouverture et de fermeture de manière à tester leur hybridations sur différentes molécules en épingle à cheveux, selon la possibilité d'hybridations des oligonucléotides sur ces molécules cf 2.5, Les expériences ont utilisé dans un premier temps sur les molécules DPX0 à DPX4 et JBO0, déjà présentent au laboratoire, puis sur notre molécule contenant tout les oligonucléotides: FLY0.

En attendant la molécule FLY0, un mélange des molécules DPX0 à DPX4 réalisé pour d'autres projets de notre laboratoire est utilisé pour la caractérisation des oligonucléotides. Cela nous permet de paralléliser les tests d'un oligonucléotide sur plusieurs séquences . Pour cela nous procédons en deux temps. Après avoir injecté dans une cellule microfluidique un mélange des cinq molécules, nous utilisons un système d'identification spécifique pour différencier chaque molécule du mix (à la manière des labels dans les systèmes Illumina). Pour cela, nous utilisons avant l'injection des 12 oligonucléotides à tester, un jeu d'oligonucléotides de 8 à 10 bases sélectionnées spécialement pour s'hybrider de manière différenciée molécules DPX0 à DPX4. Ces oligonucléotides forment une signature unique pour chaque molécules, qui permet *a posteriori* de les identifier. Cette étape permet de différencier les molécules en fonction de leur motif. Ensuite, sont injecté un à un les oligonucléotides à tester.

Les oligonucléotides possédant des hybridations théoriques sur la molécule JBO0 sont testés sur celles-ci.

Une fois la construction de la molécule FLY0 effectuée, les oligonucléotides restants y sont testés.

À l'issue de ces expériences, nous possédons des données sur l'ensemble des oligonucléotides, qu'ils

[5]Dans un premier temps, seuls 10 oligonucléotides ont été commandés et testés. À l'issue des tests, les résultats étant satisfaisants sans nécessiter les deux oligonucléotides restants, ils n'ont finalement pas été utilisés. Il s'agit de CCGTAG et CTGGCA. nous noterons qu'il aurait été pertinent de conserver l'oligo CCGTAG et d'éliminer CCGCAG en tenant compte du nombre de molécules permettant de tester.

soient testés avec le mix de molécule DPX0-DPX4, JBO0, ou notre molécule spécialement conçue: FLY0.

Table 2.3: Experiences menées pour caractériser le temps d'hybridation des molécules.

Molécule	Oligo	T°C
FLY0	TCCCCA	23°C
DPX4	TCCCCA	23°C
DPX4	TCCCCA	27°C
FLY0	CTGGCA	23°C
FLY0	CTGGCA	23°C
FLY0	TCGGCA	23°C
DPX4	TCGGCA	27°C
FLY0	TCCCGA	23°C
DPX4	TCCCGA	27°C
FLY0	CCGTCA	23°C
FLY0	CCGTCA	27°C
JBO0	CCGTCA	23°C
FLY0	TCGCCA	23°C
DPX4	TCGCCA	27°C
FLY0	CCGCAG	23°C
FLY0	CCGCGA	23°C
JBO0	CCGCGA	23°C
FLY0	CCGCAA	27°C
JBO0	CCGCAA	23°C
FLY0	CTGCGA	27°C

Pour des questions de disponibilité des instruments dans le laboratoire, les expériences ont été effectuées sur différents microscopes à pinces magnétiques. La force à appliquer pour ouvrir les molécules varie selon la magnétisation des billes paramagnétiques, soit entre 15 pN et 18 pN. De même, lors de la fermeture, la force à appliquer varie entre 12 pN et 8 pN. Pour récolté le plus de données possible, le palier de positions des aimants est optimisé pour permettre l'ouverture et la fermeture de la plus grande partie des billes paramagnétiques. Le temps d'hybridations de tous les oligonucléotides est obtenu pour une force moyenne de 10 pN. L'information utile présentée ici est l'évolution de ce temps d'hybridation avec la séquence, montrant en particulier que certains oligonucléotides ne sont pas utilisables.

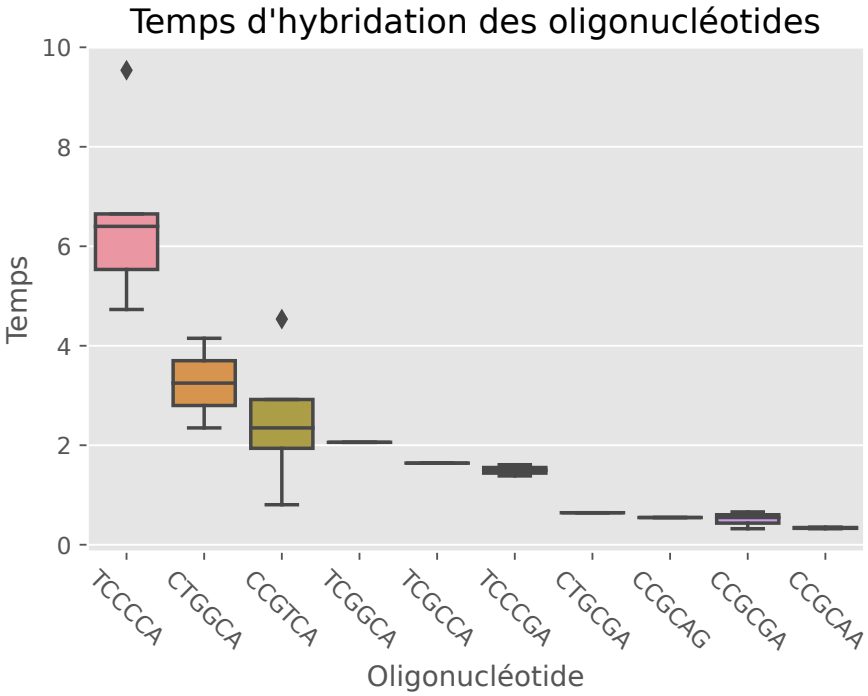


Figure 2.6: Temps médian d'hybridation des oligonucléotides en fonction de leur séquence, à une force d'environ 12 pN, à 23°C. On constate que le temps d'hybridation des oligonucléotides dépend fortement de leur séquence. Alors que `TCCCCA, CTGGCA, CCGTCA, TCGCCA, TCCCGA` possèdent des temps d'hybridations variant entre 6 et 1 seconde. `CTGCGA, CCGCAG CCGCGA CCGCAA` ont des temps d'hybridation très faible (moins de 1 seconde), qui ne permet pas leur utilisation en conditions normales dans le cadre de la création de signatures.

Le temps d'hybridation des oligonucléotides est très variable selon la séquence observée. En effet, selon l'oligonucléotide, le temps médian d'hybridation varie de 6 secondes `TCCCCA` à moins d'une seconde `CTGCGA, CCGCAG, CCGCGA, CCGCAA` fig. 2.6. Il en va de même pour le nombre de cycles d'ouverture / fermeture de molécules possédant une hybridation. Il varie de 40% à 10% des cycles, et ce à conditions expérimentales équivalentes. En effet, dans nos expériences, nous imposons qu'un blocage dure plus longtemps que quelques images de la caméra pour être gardé, il en résulte que nous perdons les blocages courts qui sont plus nombreux pour les oligonucléotides dont le T_{off} est faible.

À cette l'étape, nous pouvons déterminer que les oligonucléotides `CTGCGA, CCGCAG, CCGCGA, CCGCAA` ne seront pas exploitables pour effectuer des signatures de molécules. En effet, leur temps et nombre d'hybridation sont insuffisants, et donc le rapport signal sur bruit trop faible pour être certain de différencier des hybridations réelles du bruit de fond des expériences.

Le jeu d'oligonucléotides est donc réduit, ce qui pourrait mettre en péril sa capacité à permettre l'acquisition de signatures spécifiques. Cependant, un nouvel élément va nous permettre de règler cette probléme.

Dès les premiers tests, nous constatons la présence d'hybridations surnuméraires le long des différentes molécules testées. En comparant les positions d'hybridation avec la séquence des molécules en épingles à cheveux, la majorité des hybridations surnuméraires sont expliquées par un mésappariement en face de la dernière base de l'oligonucléotide.

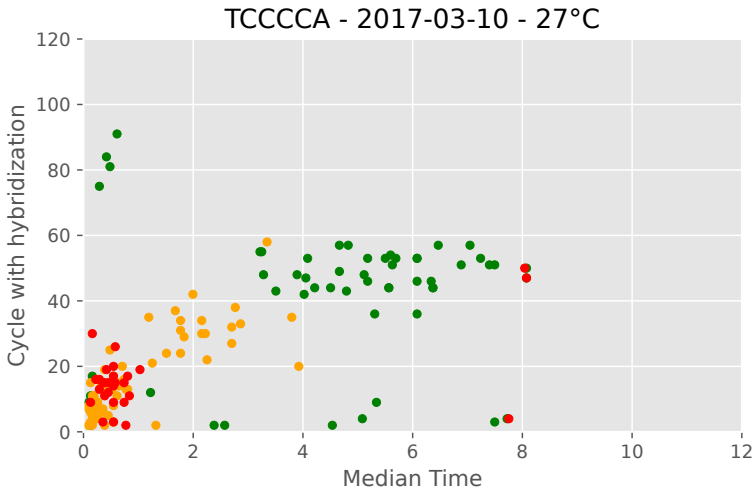


Figure 2.7: Synthèse d'une expérience sur la molécule en épingle à cheveux DPX4, et l'oligonucléotide TCCCCA à 27°C. Chaque point correspond à une position d'hybridation pour une bille donnée. 232 positions d'hybridation au totale sont considérées sur 22 molécules observées. Nous observons ici le temps d'hybridation et le nombre de cycles sur lesquelles on rencontre une hybridation à la position, indépendamment de cette position. Les points verts correspondent aux hybridations à des positions attendues, les points orange à des positions avec une mésappariement aux extrémités, et les points rouges à des positions d'hybridation inexpliquées.

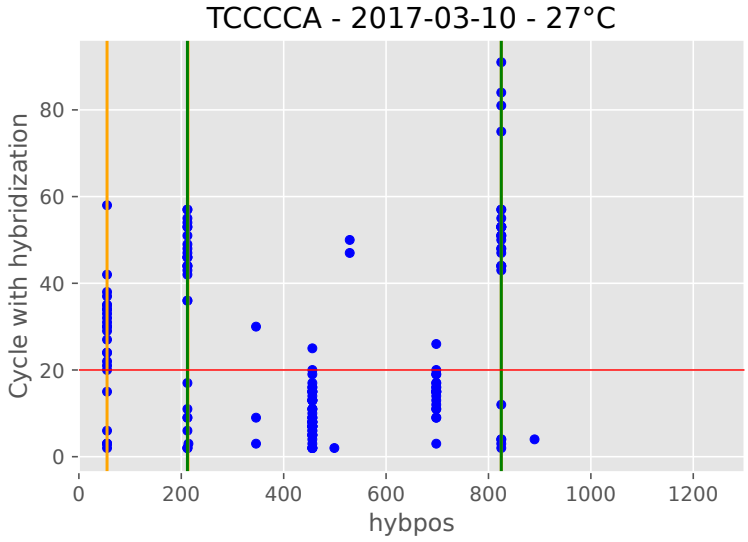


Figure 2.8: À présent, nous observons le temps moyen et le nombre de cycles possédant des hybridations en fonction de la position. Les lignes verticales représentent les positions connues d'hybridation. En vert, les positions correspondant aux hybridations parfaites, en orange, les positions des mésappariements. La ligne rouge horizontale correspond au seuil sélectionné pour la détection automatique des hybridations.

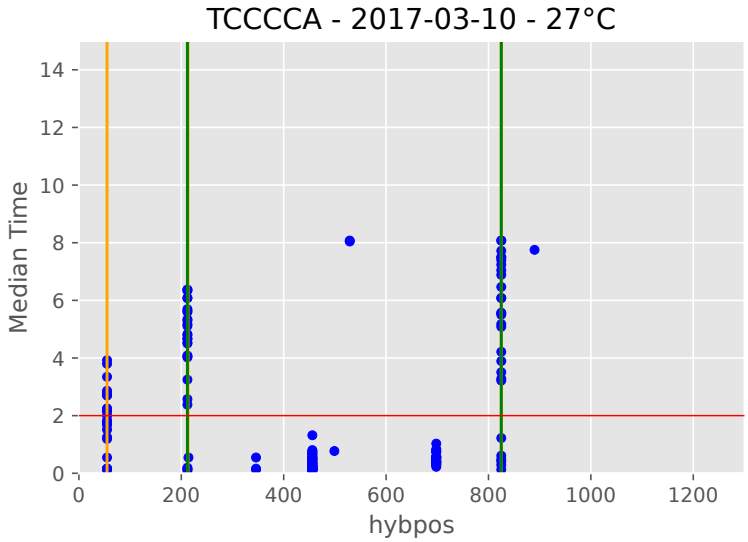


Figure 2.9: De la même manière que pour la figure précédente, nous observons le temps moyen par hybridation pour l'oligonucléotide TCCCCA. Là encore, nous sélectionnons les hybridations au-delà d'un temps moyen suffisamment important qui élimine la plupart des hybridations trop courtes pour être considérées comme réelles.

En observant l'ensemble des hybridations correspondant à des mésappariements, et leur temps d'hybridation, nous constatons que les positions de mésappariements possèdent des temps et un nombre d'hybridations plus importantes quant le mésappariement intervient du côté 3' de l'oligonucléotide, correspondant pour toutes les molécules à un nucléotide Adénine.

![Ensemble des molécules en épingle à cheveux sélectionnées pour minimiser le nombre d'expériences nécessaires à la caractérisation de la dynamique d'hybridation des oligonucléotides. La molécule FLY0 est la molécule construite sur mesure.

Bien que ne possédant pas une information exhaustive sur tous les mésappariements des oligonucléotides, une hypothèse est formulée quant à l'ensemble des hybridations perceptibles. Les oligonucléotides de 6 bases sont à présent considérés comme possédant 5 bases, en ignorant l'adénine présente en 3'. Cela nous permet de former deux groupes de 3 oligonucléotides de 5 bases, minimisant les mésappariements. Considérant ces oligonucléotides avec 5 bases, leurs taux d'hybridation sur le génome d'E. coli sont beaucoup plus importants. Nous pouvons avec ces paramètres prévoir deux jeux de 3 oligonucléotides avec un nombre suffisant d'hybridations pour être utilisées dans le cadre de la signature de molécules.

Table 2.4: Jeux d'oligonucléotides sélectionnés pour la génération de signature sur E. coli

Mix1	Mix2
TCCCC(A)	TCCCG(A)
TCGCC(A)	CCGTC(A)
TCGGC(A)	CTGGC(A)

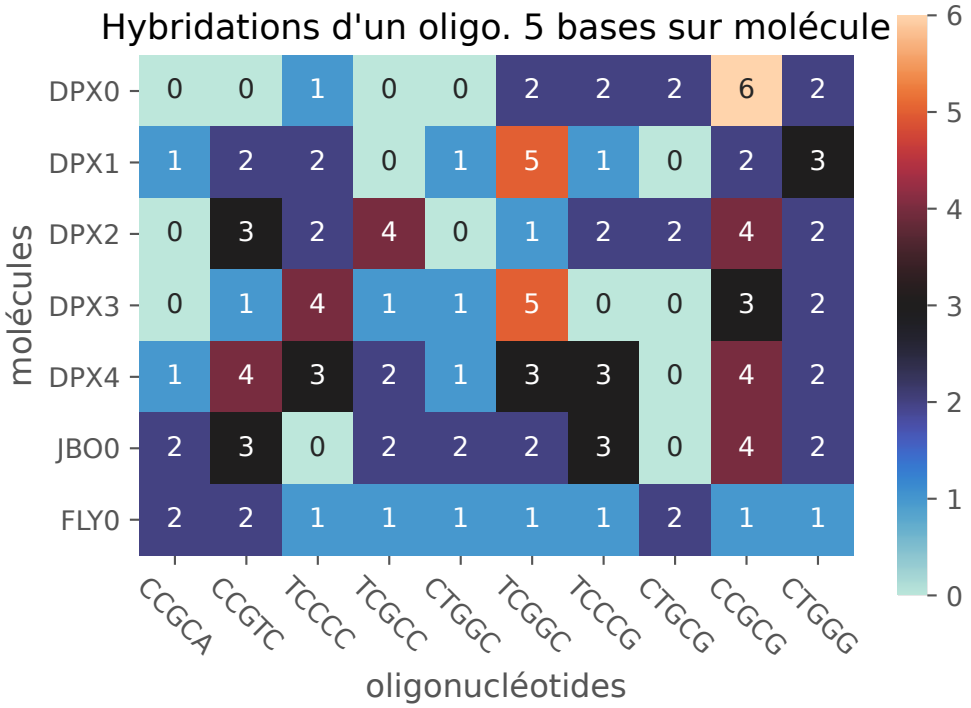


Figure 2.10: Nombre d'hybridation par molécule et oligonucléotide, pour les molécules en épingle à cheveux sélectionnées pour minimiser le nombre d'expériences nécessaires à la caractérisation de la dynamique d'hybridation des oligonucléotides, En tenant compte de la possibilité d'une méhybridation des bases A en 3'

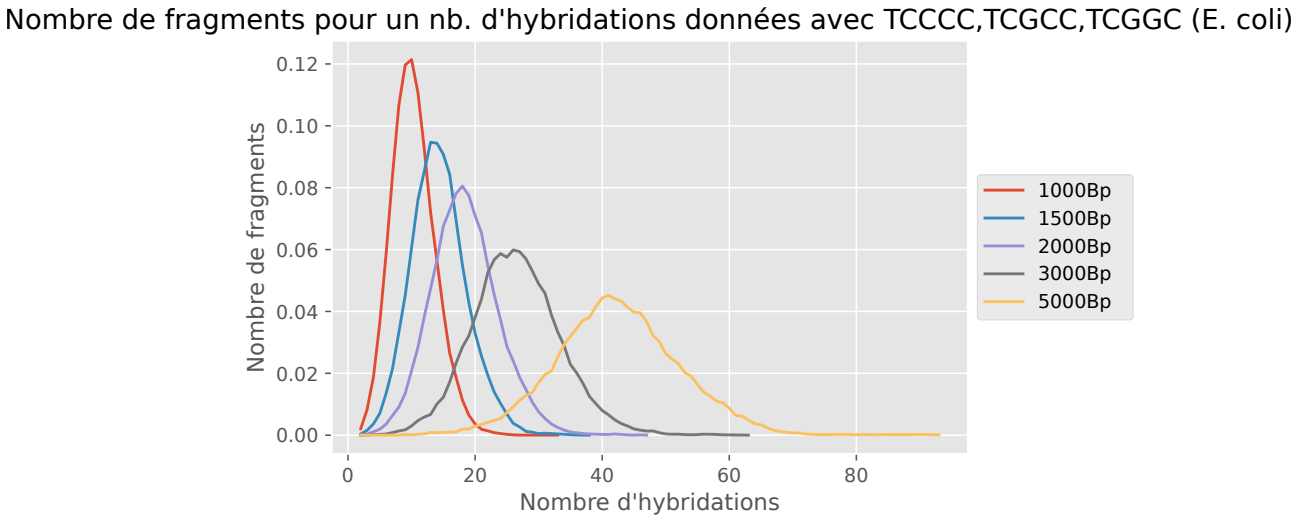


Figure 2.11: Représentation de la distribution de signatures générée avec le jeu d'oligonucléotides Mix 1 2.4 pour différentes tailles du fragment génomique.

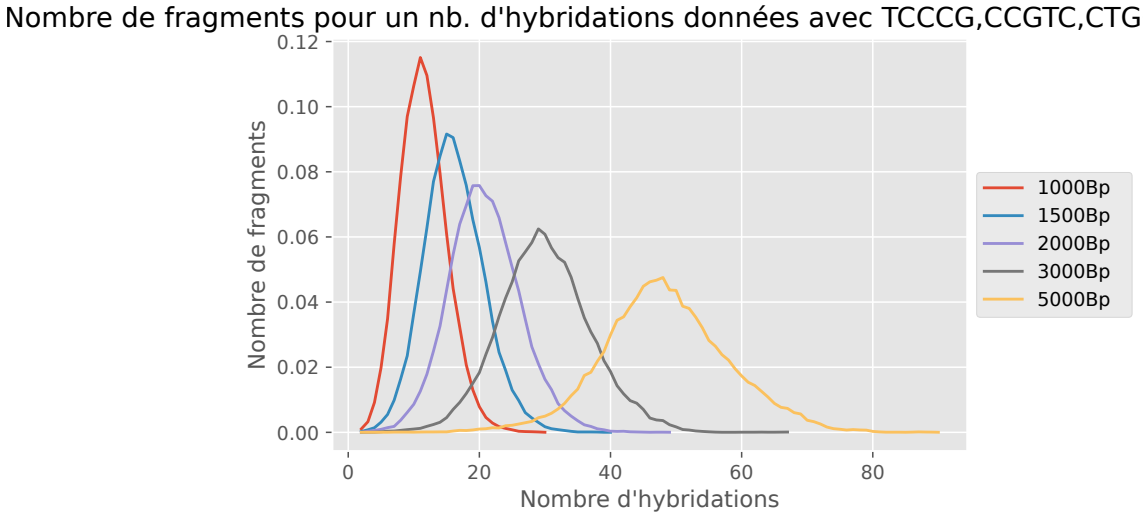


Figure 2.12: Représentation de la distribution de signatures générée avec le jeu d'oligonucléotides Mix 2 2.4 pour différentes tailles du fragment génomique.

2.3 Conclusion

La méthode bio-informatique présenté dans les premières sections de ce chapitre semblait être une approche satisfaisante pour permettre la génération d'un jeu d'oligonucléotides de k-nucléotides suffisamment spécifique pour l'application sur les signatures pour un génome de la taille d'E. coli, Nous avons en fait découvert que la connaissance insuffisante des caractéristiques thermodynamiques des oligonucléotides limite son applicabilité directe.

Cependant la méthode a permis d'établir une première sélection d'oligonucléotides ne posant pas de problème en termes de superposition et ainsi a permis, après étude empirique des caractéristiques des oligonucléotides, d'établir des sous-jeux d'oligonucléotides satisfaisants.

Quand les connaissances des caractéristiques des oligonucléotides seront meilleures, nous pourrons envisager de nous passer de l'analyse empirique des oligonucléotides individuelle.

La méthode présentée ici permet de générer des jeux d'oligonucléotides pour un génome donné. Malheureusement, la variété de la composition des génomes ne permet pas d'envisager d'utiliser le même jeu pour tous les génomes. La méthode présentée devra être appliquée de nouveau pour chaque organisme étudié.

Par la suite, nous allons voir que la possibilité nouvelle d'utiliser des oligonucléotides de 4 bases rend l'utilisation d'un jeu d'oligonucléotides obsolète : un seul oligonucléotide devient suffisant.

Mais le travail réalisé ici n'en devient pas inutile pour autant. En effet ces bases pourront être utiles pour les problématiques de séquençage par pinces magnétiques.

La méthode naïve de séquençage consiste effectuer des cycles d'ouverture et de fermeture de molécule en épingle à cheveux en injectant successivement chaque oligonucléotide de k bases dans la cellule, puis en assemblant les p hybridations pour chaque expérience, de manière à en extraire la séquence.

En fonction de la valeur inférieure atteinte pour la taille des oligonucléotides de séquençage k,[6] il

[6] Actuellement Depixus tente de séquencer en utilisant des oligonucléotides de 3 paires de bases

faudra adopter ou non une stratégie de mélange (batching) d'oligonucléotides afin de limiter le nombre d'expériences à effectuer en vue de réduire le nombre d'expériences.

Chapitre 3

Méthodes de cartographie

Sommaire du chapitre

Nous souhaitons à présent déterminer si deux molécules identiques peuvent être reconnues sur la base de leurs signatures. Pour cela, deux paramètres doivent être évalués:

- Un critère de similarité - ou distance - entre deux signatures
- Un critère de confiance dans ce critère de similarité.

Tout au long de mon travail de thèse, j'ai envisagé différentes méthodes pour comparer les signatures. Deux approches ont retenu particulièrement mon attention, qui présentent des avantages et des inconvénients qui leur sont propres.

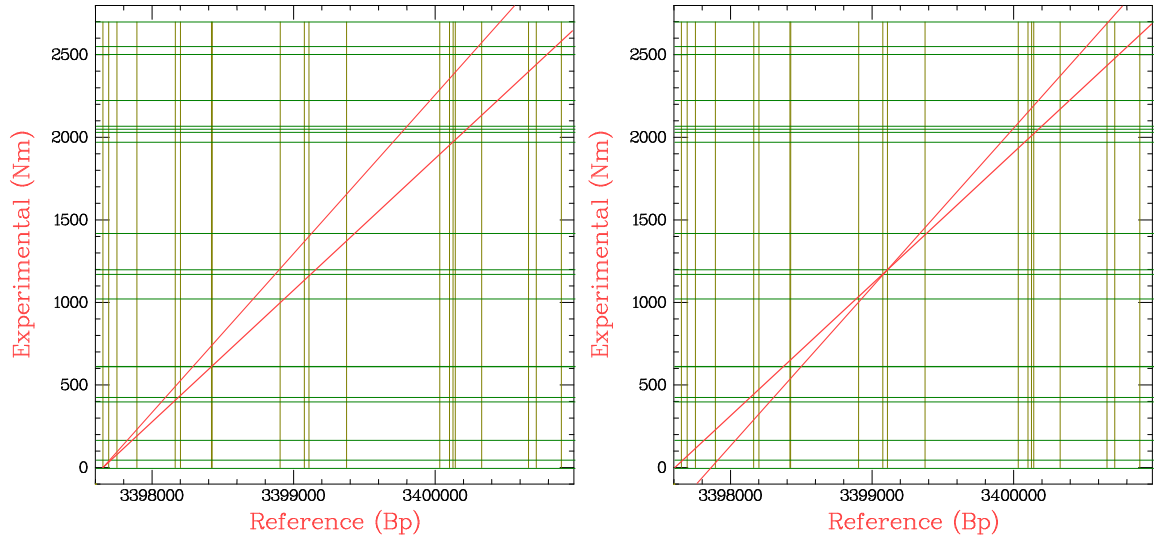
Avec la **méthode par régression**, on cherche à optimiser l'ajustement de deux signatures entre elles. À l'aide d'une régression linéaire, du test du chi-square, et d'une sélection judicieuse des points de correspondances entre deux signatures en s'accommodant du bruit expérimental et de l'étirement variable des signatures, on cherche à trouver l'ensemble des paires de signatures ayant une adéquation suffisante pour être considérées comme possédant la même position d'origine.

Dans la **méthode par classification** de segments, on cherche à simplifier les signatures, en se permettant une perte d'information raisonnable : pour cela on classe les segments qui séparent deux positions d'hybridation en fonction de leur taille, et on utilise la classe à laquelle appartient chaque segment pour l'identifier. Cette perte d'information (de segment à classe de segment), permet de s'affranchir

du bruit de mesure et de l'étirement variable. On utilise ensuite une méthode plus conventionnelle d'alignement de séquence pour déterminer quels sont les paires de signature qui semblent provenir de la même origine.

Dans cette partie, je vais présenter le fonctionnement de ces deux méthodes.

3.1 Méthode par régression



(a) Pivot à l'origine

(b) Pivot centré

Figure 3.1: Représentation du principe de la méthode par régression. On trace une ligne horizontale à chaque position sur l'axe Y correspondant à un blocage observé expérimentalement, et une ligne verticale à chaque position du génome sur l'axe des abscisses où nous attendons une hybridation. On choisit une extension suivant l'abscisse compatibles avec les valeurs d'étirement expérimentales. On fait défiler les positions d'hybridations le long du génome en cherchant une région où les patterns de lignes horizontales et verticales coïncident. Sur cette figure en particulier, nous sélectionnons le pivot, c'est-à-dire l'hybridation de référence, commune entre la signature expérimentale et la signature de référence. Dans la figure (a), le pivot est choisi à l'origine de la correspondance entre les deux signatures; dans la figure (b), le pivot est choisi de manière centrée.

3.1.1 Données d'entrée

Avec cette méthode, l'ensemble des signatures de références sont représentées par une succession de positions d'hybridation dont nous indiquons la position en paires de bases, elles sont créées à partir de la séquence d'un génome de référence. Les données expérimentales sont les positions d'hybridations observées en nanomètres, acquises par microscope en pince magnétique.

Le parcours d'un génome de référence correspond donc au parcours de l'ensemble des hybridations sur le génome.

3.1.2 Contexte

Le principe de la méthode est plus facilement compréhensible à l'aide de la fig. 3.1. On porte en ordonnée les positions des hybridations mesurées en *nm* matérialisés par des droites horizontales.

On porte en abscisse les positions d'hybridations attendues sur le génome en paires de bases. La fenêtre génomique choisie sur l'axe des X est tel que le nombre de bases corresponde à la longueur de la molécule mesurée entre l'ouverture totale et la fermeture de la molécule divisée par le facteur d'étirement moyen. Ceci nous donne la taille de la molécule en paires de bases avec une précision typique de 10%. Le fait que la taille de cette zone soit un peu trop grande ne joue pas de rôle dans la méthode et n'est utilisé qu'à titre d'illustration. La suite de la méthode nous permettra de préciser la correspondance entre les nanomètres (expérimental) et paires de base (génome de référence). On fait ensuite défiler dans cette fenêtre les positions attendues dans le génome, en recherchant la situation représentée sur la figure où les deux faisceaux de droites sont en correspondance. Dans le cas d'une correspondance parfaite, les points d'intersection s'alignent sur une diagonale dont la pente nous donne le facteur d'étirement, et le χ^2 la qualité de la coïncidence. Dans la pratique, il manque souvent des hybridations dans l'expérience et certaines lignes verticales ne trouveront alors pas de ligne horizontale correspondante. Il y a également des hybridations supplémentaires correspondantes à des mésappariements ou à des erreurs qui produisent l'effet inverse.

De façon pratique, pour trouver la meilleure correspondance entre une signature observée et une signature attendue, nous les comparerons en les centrant autour d'une hybridation particulière que nous appellerons le pivot. Nous pouvons avec cette représentation, exprimer la méthode comme l'action de chercher à faire passer une droite par l'ensemble des points d'intersection entre les hybridations des deux signatures permettant le meilleur ajustement. Dans cette représentation les intersections se situant autour de la diagonale constituent l'ensemble des intersections candidates pour être jointes par notre droite.

L'ensemble des points candidats peut être calculé en fonction du facteur d'étirement et de ses variations possibles.

Dans le cas d'un système sans bruit et à étirement fixe, deux signatures identiques correspondraient à un ensemble de points d'intersection dont les positions suivrait fonction linéaire $y = a(x - x_0)$ d'une pente a correspondant à l'étirement du système.

3.1.2.1 Choix du pivot

À présent, nous devons envisager un étirement variable des molécules. Pour cela il suffit de considérer l'ensemble des fonctions affines passant par le pivot, et dont l'étirement a est compris dans des bornes acceptables sec. 1.3.4.1.1 : en prenant comme hypothèse $\mathbb{P}(s_{min} \leq \bar{S} \leq s_{max}) = 99.7\%$, nous pouvons choisir les bornes d'acceptabilité à $0.76 \leq a \leq 1$. Ces valeurs correspondent respectivement à des forces de : 5.8 pN et 13 pN qui sont les valeurs extrêmes utilisables dans l'expérience.

Cela se traduit par la région délimitée par les deux droites rouges dans la fig. 3.1.

Pour réduire l'ensemble des ajustements possibles, il faut minimiser l'espace dans lequel se situent les intersections candidates :

$$A = \int_{s_{min}}^{s_{max}} \left(\int_0^{N_h} ax + b dx \right) da$$

$$\Rightarrow b = \frac{N_h}{2}$$

, N_h la taille du fragment d'ADN étudié

On constate qu'un pivot au centre de la signature permettra de minimiser la quantité d'intersections à considérer. C'est donc la position d'hybridation la plus centrée qui sera sélectionnée comme pivot. D'autres critères pourraient entrer en ligne de compte dans ce choix, par exemple la densité d'hybridations autour de l'hybridation candidate au rôle de pivot.

Une fois le pivot sélectionné, nous pouvons commencer le parcours des signatures possibles.

3.1.3 Comparaisons de signature

Nous comparons successivement l'ensemble des signatures du génome comprises dans la fenêtre au fragment dont nous cherchons à trouver la position d'origine dans le même génome. L'algorithme fonctionne de la manière suivante :

3.1.3.1 1. Sélection des points

Nous calculons l'ensemble des points d'intersection acceptables en fonction du pivot et du bruit de mesure et de l'étirement variable, dans un intervalle de confiance donnée. L'ensemble des points est ajouté à une liste de points .

$$M_{hyb} = u_{x,y} \in A, \forall (x,y) \in (s_1, s_2)$$

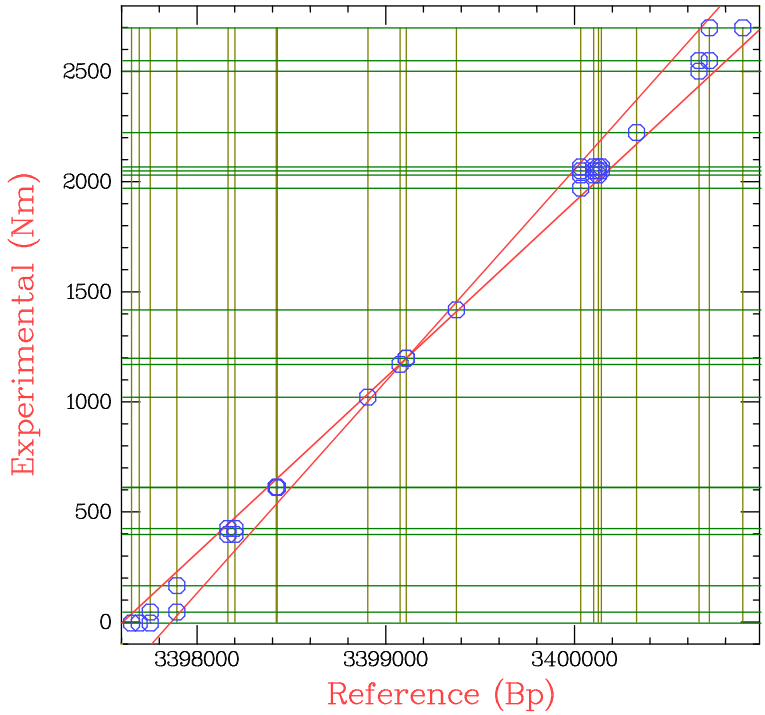


Figure 3.2: Représentation de l'ensemble des intersections acceptables pour les signatures expérimentales attendues et un pivot donné. Ces intersections sont entourées d'un cercle bleu. Nous pouvons constater des zones plus ou moins denses.

Cependant il est clair sur la fig. 3.2 que les régions où les hybridations sont rapprochées produisent trop de points d'intersection potentiels. Par exemple, les deux hybridations ayant lieu autour de la position 3 398 200 bp donnent lieu à quatre intersections, alors qu'il est évident ici que seuls les deux

points sur la diagonale parallèle aux droites rouges sont à conserver. Mais dans beaucoup de cas, il est nettement moins facile de trier les hybridations acceptables par un algorithme adapté. Il nous a donc fallu mettre au point une façon de trier les points possibles d'intersection.

3.1.3.2 2. Sélection des intersections univoques

On recherche au sein de cette liste l'ensemble des intersections pour lesquelles l'hybridation pour les deux signatures n'est mise en jeu que dans une unique intersection. Les intersections sont dites univoques.

$M_{univoque} = (u \in M_{hybs}, count(u_{x,k}, k \in s_2) = 1)$

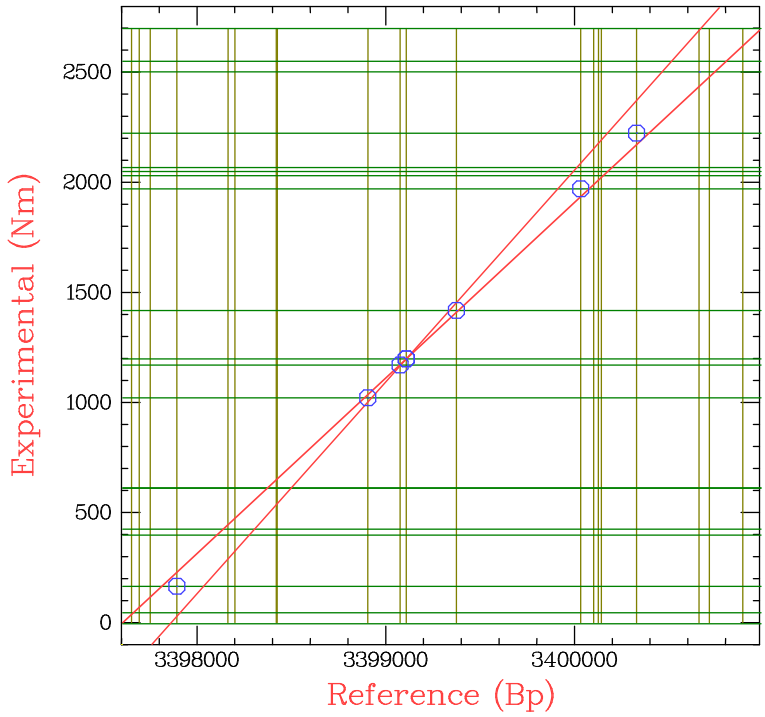


Figure 3.3: Description de la recherche des points univoques. À partir du pivot et des deux droites rouges qui délimitent les points compatibles avec la fourchette d'étirement connue, on ne garde dans un premier temps, que les points d'intersection qui ne présentent aucune ambiguïté.

3.1.3.3 3. Régression sur les intersections univoques

À partir des coordonnées des intersections univoques, nous pouvons effectuer une régression pour obtenir une première approximation du facteur d'étirement entre les deux signatures. Pour cela, on utilise l'estimateur des moindres carrées ordinaire.

$$\begin{cases} \hat{\beta}_{pivot} = \dfrac{\mathrm{cov(X,Y)}}{\mathrm{var(X)}} \\ \hat{\beta}_{etirement} = \bar{y} - \hat{\beta}_{pivot}\bar{x} \end{cases}$$

[1]

[1]Source : https://fr.wikipedia.org/wiki/R%C3%A9gression_lin%C3%A9aire#Estimateur_des_moindres_carr.C3.A9s_ordinaires

La fonction affine issue de la régression servira de base pour l'ensemble des calculs suivants.

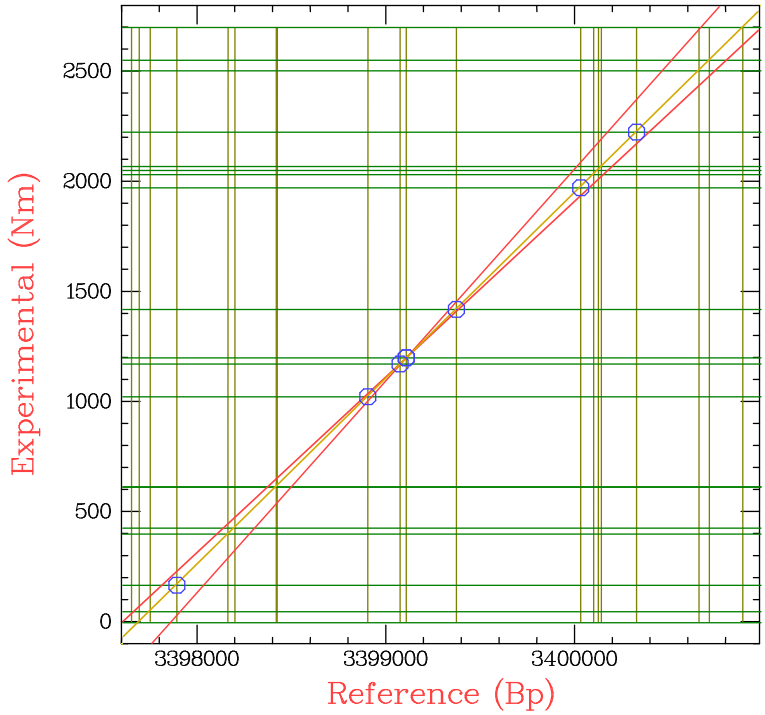


Figure 3.4: L'ajustement des points univoques par une droite permet d'obtenir une première approximation du facteur d'étirement réel.

3.1.3.4 4. Redéfinition de l'ensemble des points acceptables.

Maintenant que nous disposons d'une approximation de l'étirement de la molécule, nous pouvons nous affranchir de ce paramètre pour nous pencher plus spécifiquement sur les contraintes du bruit sur la mesure. De la même manière que pour l'étirement, le choix des paramètres est basé sur la connaissance des paramètres expérimentaux avec un intervalle de confiance donné.

Les points acceptables pour ce nouveau jeu de paramètres sont calculés en prenant l'ensemble des points autour de la droite tel que $y = x\hat{\beta}_{pivot} + \hat{\beta}_{etirement} \pm 3\sigma(Noise)$

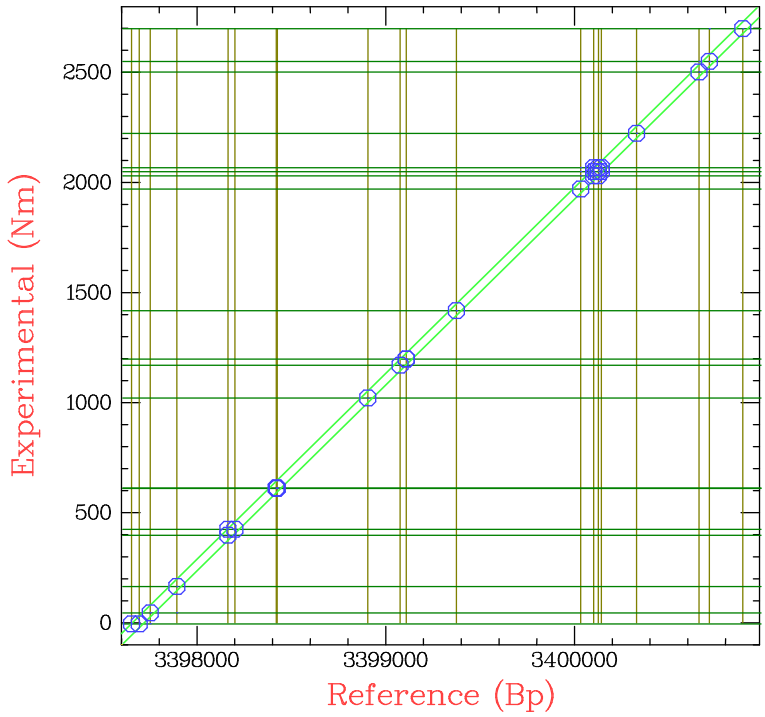


Figure 3.5: En utilisant la proximité (définie à partir du bruit typique de l'expérience) des points avec la première droite trouvée, on augmente le nombre de coïncidences ce qui nous permet de faire un nouvel ajustement linéaire avec plus de points.

3.1.3.5 5. Sélection des intersections équivoques

Les points équivoques correspondent aux intersections n'ayant pas une assignation unique, qui avaient été écartées lors de l'étape "Selection des intersections univoques", pour le calcul de la première approximation de l'étirement.

Pour sélectionner les intersections, nous allons considérer chaque hybridation une à une pour déterminer quel point d'intersection minimise le résidu comparé à la fonction affine de l'étirement précédemment défini. Pour assurer une minimalisation globale et une assignation unique des hybridations, nous utilisons une méthode d'assignation *greedy*.

Cette méthode assure une sélection des intersections optimales. Par contre, il n'y a pas d'assurance que pour toutes les hybridations d'une signature s_1, il existera une intersection sélectionnée : pour une hybridation donnée de s_1, si elle ne possède aucune intersection avec la signature s_2, ou si toutes les hybridations de s_2 possédants une intersection avec s_1 sont assignées à une autre hybridation de s_1, l'hybridation sera écartée.

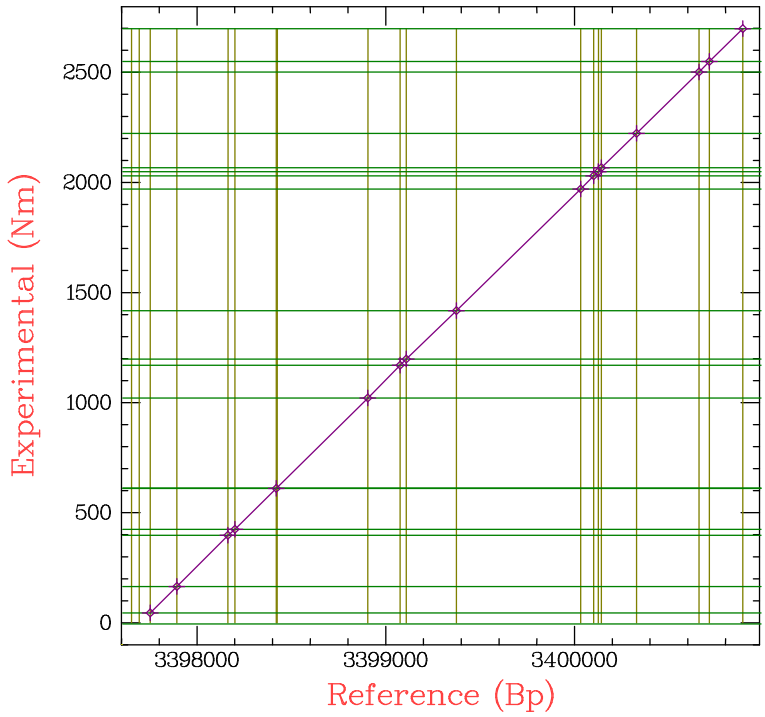


Figure 3.6: Ajustement final des points sélectionnés par la méthode de régression.

3.1.3.6 6. Évaluation de la solution

Nous possédons à présent un ensemble de points qui maximisent le nombre d'hybridations assignées et dont l'ajustement à l'étirement est estimé $\widehat{\beta}_{etirement}$

Pour évaluer la qualité de cet ajustement, et avoir une information sur la vraisemblance que les deux signatures sont identiques, on utilise la méthode du χ^2.

Tout d'abord, nous effectuons de nouveau une régression linéaire, mais à présent sur l'ensemble des points sélectionnés. On effectue ensuite le test de qualité d'ajustement du χ^2.

$$\chi^2(a,b) = \sum_{i=1}^{N} (\frac{y_i - a - bx_i}{\sigma_i})^2$$

Avec a qui correspond à l'abscisse à l'origine, b la pente de la droite qui correspond en fait à notre étirement, N le nombre de points utilisés dans l'ajustement et σ_i l'erreur attendue pour chaque point, déterminé par la résolution de l'experience.

3.1.3.7 7. Rejet et Acceptation d'une solution

Lors d'une des étapes de la méthode par régression, dans le souci d'aller plus vite, on peut décider de rejeter précocement une solution, car elle présente des caractéristiques mauvaises (par exemple elle a peu de points univoques) ce qui peut conduire à éliminer précocement des séquences solutions qui auraient pu finalement être correctes. Ce rejet précoce se fait sur différents critères, notamment en cas d'une valeur statistique du χ^2 trop mauvaise. Typiquement, à partir d'un certain nombre de matchs considérés, si le χ^2 dépassant largement la valeur critique en fonction des degrées de liberté, on

considère, à l'étape de la sélection des points équivoques, que la coïncidence entre les deux séquences ne pourra pas être suffisante. On élimine alors la position de séquence de mauvaise qualité.

À partir du moment où une solution est évaluée jusqu'au bout, il faut déterminer si sa vraisemblance est suffisante pour l'ajouter à la liste des solutions acceptée.

Pour cela on évalue plusieurs critères:

- Le nombre d'hybridations sélectionnées comparées au nombre d'hybridations totales des deux signatures.
- La vraisemblance de la solution :

$$\chi^2_{val} \leq Q_{\chi^2_k}(1 - pvalue), k = N - 1 \text{ (degrés de liberté)}$$

On trie ensuite les solutions par vraisemblance normalisée

$$\chi^2_{val}/Q_{\chi^2_k}(1 - pvalue)$$

3.2 Méthode par classification de segments

3.2.1 Introduction

3.2.1.1 Approche par segments

Plutôt que d'utiliser les positions d'hybridation comme éléments constitutifs d'une signature, nous pouvons décider d'étudier la distance qui sépare deux hybridations d'oligonucléotides. Ce simple changement de perspective ouvre un large champ de méthodes possibles. En effet, une signature devient alors une succession de segments de différentes tailles. Cette transformation permet d'envisager la problématique de cartographie comme une recherche de motif.

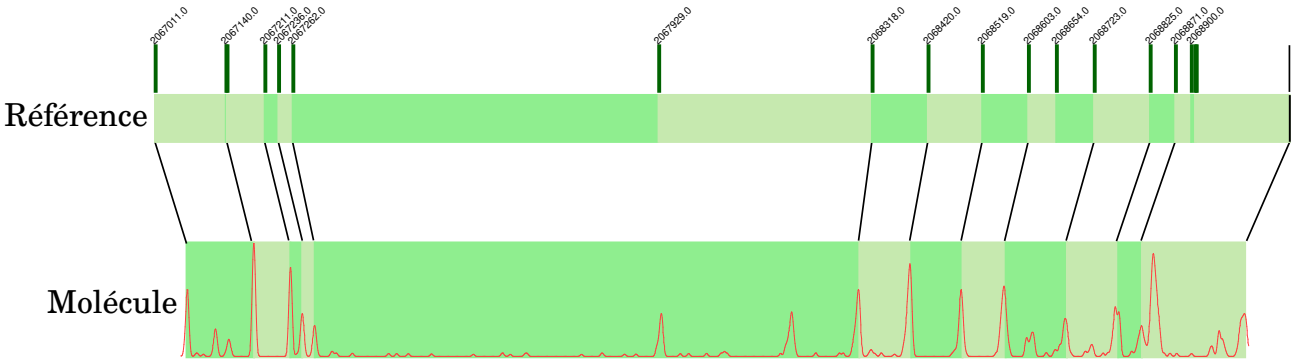


Figure 3.7: Représentation d'un fragment du génome de référence (théorique) et d'une molécule expérimentale sous forme de segment. La première ligne correspond à la position des hybridations sur le génome de référence. La seconde correspond à la traduction en segments de ce fragment. Sur la dernière ligne, on trouve en rouge l'histogramme issu d'une expérience effectuée sur une molécule au locus correspondant à la référence. En vert, les segments sélectionnés à partir de l'histogramme. Certaines hybridations, dont l'amplitude est plus faible, ne sont pas sélectionnées, car le seuil de sélection a été choisi relativement haut, pour éviter la présence d'hybridations surnuméraires. Le locus représenté correspond à celui de la molécule PS825HP, avec l'oligonucléotide CGTC. Les deux teintes de verts utilisés ne sont là qu'à titre de visualisation des différents segments.

Observons comment nous pouvons envisager les signatures sous ce nouvel angle à l'aide de la Figure fig. 3.7. On compare une signature expérimentale sous forme de séquences et de locus équivalent sur le génome de référence. Dans la partie basse de la figure, est présenté (en rouge) l'histogramme qui correspond aux données sur une molécule sur laquelle se sont hybridés des oligonucléotides CGTC,[2] de laquelle ont été extraites des positions d'hybridation représentées sous la forme de la frontière entre deux segments[3] La ligne supérieure correspondent à la représentation du locus d'origine de la molécule sur le génome de référence d'E. coli NEB5α (position: 2066716 bp) (Anton et Raleigh 2016) Les positions des pics de l'histogramme de la molécule expérimentale sont représentées en nanomètres tels que mesurées. Le génome de référence est représenté en paires de bases, selon le facteur de conversion 1bp = 0.88nm avec sa valeur attendue en moyenne pour les conditions expérimentales (Smith, Finzi, et Bustamante 1992), voir sec. 1.3.4

À partir de cet exemple, nous pouvons réfléchir sur la manière de comparer les deux signatures théoriques et expérimentales à partir de la succession de leurs segments :

Dans la signature segmentée de la molécule expérimentale et en tenant compte des critères de présélection d'hybridation, nous constatons que certaines positions d'hybridations, quoique présentes dans le génome de référence ne sont pas présentes dans les données expérimentales. Les raisons probables qui permettent d'expliquer ces contradictions sont soit que ces positions n'ont pas été validées par les critères de sélection des positions d'hybridations, soit qu'elles n'ont pas donné lieu à des blocages expérimentaux. À l'inverse, si nous avions gardé l'ensemble des positions sans sélection liées au nombre d'hybridation à une position donnée, nous aurions pu constater des positions d'hybridation présentes dans la signature expérimentale, mais pas dans la signature du génome de référence. Cela se traduit dans les deux cas par un segment divisé en deux sur l'une des deux signatures. Un système traitant ces signatures devra s'accommoder de ces segments fusionnés / divisés.

Nous constatons aussi la différence d'élongation entre la signature de référence, disposant de l'étirement moyen, et celle de la molécule expérimentale, dont l'étirement est plus faible. Pour traiter ces segments, nous devons trouver une manière de nous affranchir de cet étirement.

Enfin, bien que cela ne soit pas visible dans cet exemple, la taille de chaque segment varie en fonction du bruit sur la position d'hybridation pour chaque position. Cette variation est, en proportion, beaucoup plus importante sur les segments de petite taille.

3.2.1.2 Une méthode d'alignement pour la cartographie

Pour répondre aux enjeux évoqués dans la section précédente, j'ai développé une approche qui cherche à réduire chaque segment composant une signature en le classant dans des catégories qui regroupent les segments de tailles proches. Cette approche permet ensuite d'utiliser ces catégories comme un alphabet, et ainsi d'appliquer des stratégies d'alignement pour cartographier les fragments. Ces catégories sont construites de manière à limiter l'effet de l'étirement sur la classification, et les effets de classification incorrecte sont atténués par une approche similaire aux matrices de substitution employées dans les algorithmes d'alignement de séquences biologiques comme BLAST (Altschul 1991) et Smith-Waterman (T. F. Smith et Waterman 1981b). La fusion de segments est assimilée à un *gap* dans la séquence.

[2] Plus d'information dans la partie sec. 5 ici il s'agit de PS825HP, lors d'une expérience réalisée le 2 août 2017, molécule n°0

[3] Les critères de sélection sont les suivants pour cette expérience. Taux d'hybridation > 22%, suppression des hybridations dont la position est à moins de 50Bp de la molécule en position fermée, suppression de l'hybridation la plus haute, correspondant à la molécule ouverte. Ces critères ont été sélectionnés de manière à minimiser le nombre d'hybridations surnuméraires, et à éviter le bruit supplémentaire entre les cycles qui apparaissent quand la molécule approche de la position fermée, et quand la bille paramagnétique approche de la surface.

Nous allons aborder ces éléments constitutifs, mais auparavant, appliquons notre approche à notre exemple.

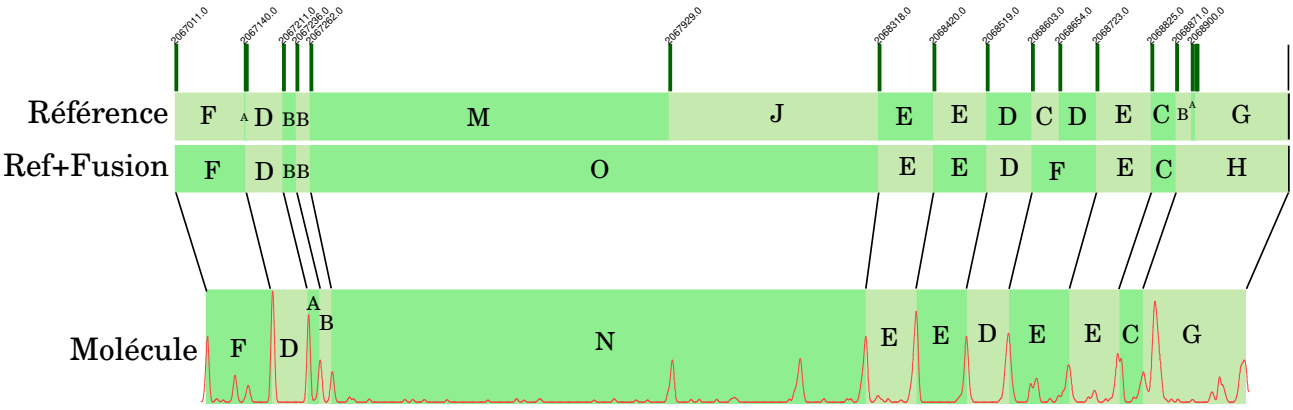


Figure 3.8: Chaque segment est classé dans la catégorie qui correspond à sa longueur, chaque catégorie étant symbolisée par une lettre. La séquence ordonnée de ces catégories sera comparée aux séquences déduites du génome de référence, simplifiant le processus de recherche. La construction des catégories limite l'effet de l'étirement sur les positions d'hybridation. Les deux teintes de verts utilisés ne sont utilisées qu'à titre de visualisation des différents segments.

Chaque segment est représenté par une lettre symbolisant la catégorie dans laquelle l'algorithme a décidé de la classer. La signature modifiée par l'algorithme, maximisant la similitude entre le fragment de référence et les données expérimentales, est représentée au centre de la figure *(Ref+Fusion)*. Entre la séquence de référence et la séquence expérimentale, le calcul du score est effectué par l'algorithme.

Pour calculer le score de l'alignement d'une séquence par rapport à une autre, l'algorithme va comparer la classe des segments successifs et leur attribuer un score basé sur la probabilité d'une substitution de la classe du segment de la première signature par la classe du segment dans la seconde. Le score est maximum lorsque les segments sont de classe identique, et le score diminue lorsque les segments de la séquence expérimentale s'alignent sur une classe du génome de référence de taille plus éloignée. Par exemple, dans la comparaison du début des deux séquences, le segment de classe F s'aligne à un segment de classe identique sur le génome de référence, produisant un score très élevé de 3.08. En revanche la comparaison au locus 2067211 d'un segment de classe B avec un segment de classe A ne contribue qu'un score de 0.40

Le score global de l'alignement entre les deux séquences est la somme des scores de chaque segment expérimental contre la référence.

Lorsque deux segments consécutifs de la séquence expérimentale s'alignent individuellement sur des segments de classe très différente, l'algorithme peut proposer leur fusion en un segment plus grand à condition que cela augmente le score de l'alignement. Dans ce cas, une pénalité (ici de -0.5) est appliquée sur le score global pour chaque fusion, ou chaque extension d'une fusion. Par exemple, en position 2067262 deux segments de classe M et J sont fusionnés en un segment de classe O qui pourra être comparé au segment N de la signature expérimentale résultant en un score de $0.23 - 0.5 = -0.27$.

3.2.2 Présentation détaillée de l'approche

À présent, nous avons une idée globale du fonctionnement de l'approche par segmentation, nous allons explorer plus en détail ses caractéristiques.

Pour commencer, nous allons explorer la stratégie qui permet le choix des catégories de segment, puis nous parlerons des matrices de substitutions et de méthodes de fusions de segments, nous terminerons sur un exemple plus concret d'utilisation de sequences de segments catégorisées pour l'alignement de deux signatures.

3.2.2.1 Catégorie des segments

L'intuition, lors de la conception de cette méthode, est que la spécificité d'une signature expérimentale est très supérieure à la complexité des génomes. En partant de cette hypothèse, nous pouvons envisager de dégrader notre signal sans perdre la spécificité minimale requise pour assigner avec confiance une signature donnée à une position dans le génome de référence. Ce compromis a pour objectif de nous affranchir des problématiques de bruit et d'étirement variable de la molécule sous-jacente.

Néanmoins, cette dégradation ou simplification du signal présente le risque de perdre des informations cruciales pour la cartographie. Il s'agit donc de trouver un équilibre dans la recherche de similarité entre les signatures attendues et les signatures expérimentales afin de rendre la comparaison plus facile à mettre en oeuvre tout en conservant le maximum d'informations.

L'idée de créer des catégories de segments est de convertir la succession de distances entre deux hybridations, qui sont des mesures continues en paires de bases, en un ensemble de symboles discrets. Par exemple, tous les segments compris entre des hybridations éloignées de 0 à 50 bp seront assignés à la catégorie "A", tous les segments compris entre des hybridations éloignées de 51 à 150 bp seront assignés à la catégorie "B", comme décrit dans la Figure 3.7. Cette opération revient à transformer une succession de positions d'hybridation en une séquence de symboles discrets appartenant à un alphabet. Il faut alors que cet alphabet soit construit de manière astucieuse pour maximiser le nombre de segments classés dans la bonne catégorie, même lorsqu'ils sont étirés aux valeurs extrêmes de la distribution de l'étirement ou affectés par un bruit important.

3.2.3 Détermination des catégories de segments

On souhaite déterminer les bornes de cette catégorie de telle manière que toutes les versions étirées (sur la gamme d'étirement maximal) d'un segment situé au centre de cette catégorie soient classées dans cette catégorie et non une catégorie adjacente, avec un niveau de confiance α (par exemple 99,73% soit 3σ). Pour la catégorie décalée de deux positions, ce niveau de confiance sera α' (par exemple $99,99\%$ soit 4σ). Pour ce faire, on commence par estimer les bornes de la catégorie qui contiendra les plus petits fragments, nommée A. On place la borne gauche de la deuxième catégorie (nommée B) adjacente à la borne droite de la catégorie A, et on écarte à nouveau la borne droite de B afin de satisfaire la même règle que pour A et ainsi de suite pour toutes les classes. L'étirement étant linéaire, on est assuré que les contraintes de niveaux de confiance α et α' sont toujours respectés. Les variables α et α' du bruit et d'étirement sont considérées simultanément pour chaque catégorie et le niveau de confiance doit être respecté pour ces deux paramètres.

Alors que l'amplitude de l'étirement est proportionnelle à la taille des segments, le bruit est constant, car il concerne chaque position d'hybridation. Le bruit est donc le paramètre dominant sur les petites distances, mais négligeable sur les grandes.

Nous définissons un ensemble d'intervalles de confiance qui nous serviront lors de la construction.

Table 3.1: En pratique, on construit les catégories en partant des segments les plus courts. On étend une catégorie tant que les contraintes α et α' ne sont pas respectées à gauche de la catégorie. L'étirement étant linéaire. Nous avons l'assurance par construction que les contraintes à droite seront respectées.

Intervalles de confiances	Définition
α_{noise}	la confiance dans le fait qu'un élément au centre du groupe ne sera pas classé dans une catégorie adjacente, du fait du bruit de mesure
α'_{noise}	la confiance dans le fait qu'un élément au centre du groupe ne sera pas classé dans une catégorie à deux écarts standards de sa catégorie d'origine, du fait du bruit de mesure
$\alpha_{stretching}$	La confiance dans le fait qu'un élément au centre du groupe ne sera pas classé dans un groupe adjacent, du fait de l'étirement de la signature.
$\alpha'_{stretching}$	la confiance dans le fait qu'un élément au centre du groupe ne sera pas classé dans une catégorie à deux écarts standards de sa catégorie d'origine, du fait de l'étirement de la molécule

Une fois ces catégories de segments constituées, nous pouvons considérer chaque catégorie comme la lettre d'un alphabet.

Lors de la cartographie d'une signature, un segment sera remplacé par sa catégorie correspondante.

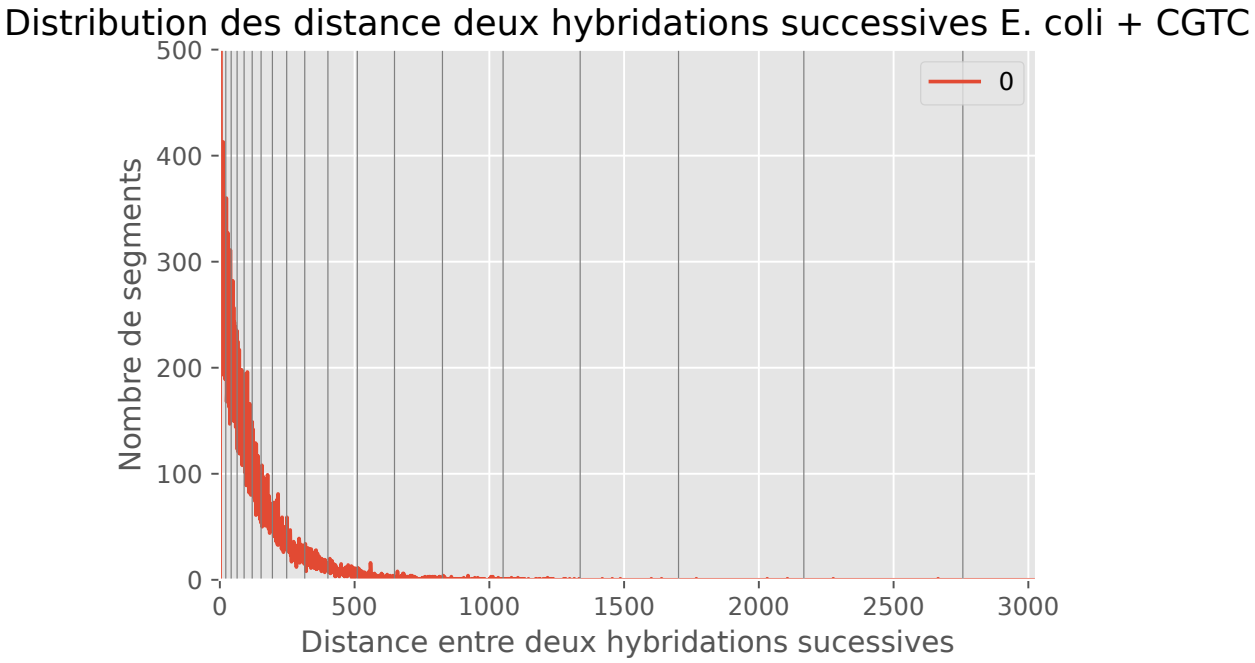


Figure 3.9: Nous observons ici le nombre de segments classés par leur longueur dans le génome d'E. coli NEB 5α (en rouge) Chaque ligne verticale grise représentent la délimitation entre deux catégories a été déterminée par la méthode présentée ci-dessus, pour $\sigma_{bruit} = 5$nm, $Etirement = N(\mu = 0.88, \sigma = 0.04)$.

Table 3.2: Liste des catégories et de leur probabilité d'apparition dans le génome d'E. coli

Symbole	Début (Bp)	Fin (Bp)	Fréquence
A	1	21	15,94%
B	22	41	11,15%
C	42	63	10,89%
D	64	89	10,56%
E	90	119	9,99%
F	120	152	8,29%
G	153	194	8,09%
H	195	247	7,56%
I	248	314	6,04%
J	315	409	4,66%
K	401	509	3,24%
L	510	647	1,99%
M	648	825	0,90%
N	826	1050	0,43%
O	1051	1336	0,19%
P	1337	1702	0,05%
Q	1702	2166	0,0083%
R	2167	2756	0,0083%
S	> 2757		0,0028%

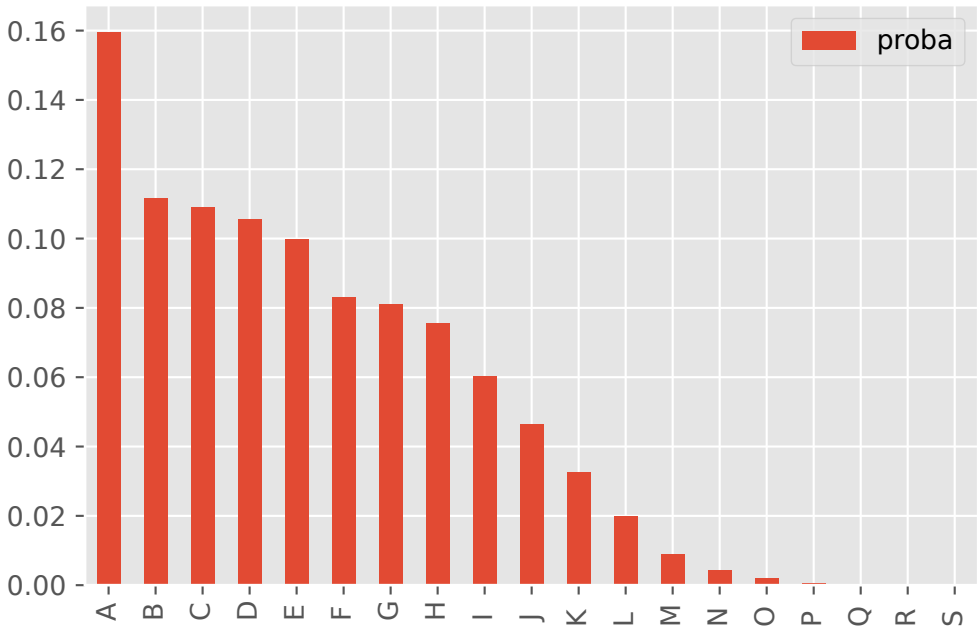


Figure 3.10: Représentation de la probabilité d'un segment aléatoire d'E. coli obtenu avec l'oligonucléotide CGTC, d'appartenir à une des catégories définies plus haut.

3.2.4 Matrice de substitution

Maintenant que nous avons à notre disposition un alphabet, nous pouvons définir un ensemble de métriques qui permet de quantifier les similitudes ou *a contrario* les différences entre deux séquences. Cela inclut des opérations de base à l'échelle du symbole telles que décrite pour la première fois avec la distance d'édition (Levenshtein 1966) qui définit, en analyse de texte, la notion de distance entre deux mots par le nombre minimal d'addition et de suppression de lettres qui permet la transition d'un mot à un autre. Alors que la distance d'édition ne tient compte que de l'addition et de la suppression d'une lettre, Damereau ajoutera la notion de substitution d'un symbole (Damerau 1964). L'ensemble des travaux sur les séquences prend racine dans cette approche. Plus tard, les auteurs de BLAST (Altschul 1991) proposent une manière de calculer la probabilité qu'a un acide aminé d'être remplacé par un autre à partir de deux modèles: celui attendu, composé de l'ensemble des probabilités de substitution d'un symbole par un autre p_{ab}, et l'aléatoire, composé des probabilités d'apparition indépendantes des symboles $q_a q_b$ Le rapport entre ces deux modèles à l'échelle logarithmique introduit dans une matrice, constitue la matrice de substitution ou matrice de score telle que $s(a,b) = log(\frac{p_{ab}}{q_a q_b})$.

Ce concept de matrice de substitution est aujourd'hui largement utilisé dans les algorithmes d'alignement de séquences de nucléotides ou d'acides aminés.

Dans le problème qui nous concerne, la matrice liée strictement à la substitution est de nature légèrement différente. En effet, la corrélation entre les symboles est principalement fonction de la distance entre les centres des catégories tbl. 3.2, alors que dans le cas de matrices de substitution liées aux acides nucléiques la corrélation est généralement liée aux mutations provoquant la transformation d'une base en une autre. Dans le cas des acides aminés, la corrélation concerne soit la similarité des codons qui les encodent, leurs caractéristiques chimiques ou leur conservation au cours de l'évolution. Cependant, ces différences n'altèrent pas les postulats théoriques proposés par (Altschul 1991) dans le cadre de l'alignement des séquences pour BLAST, dont nous reprenons les principes généraux ici. Une approche similaire est utilisée.

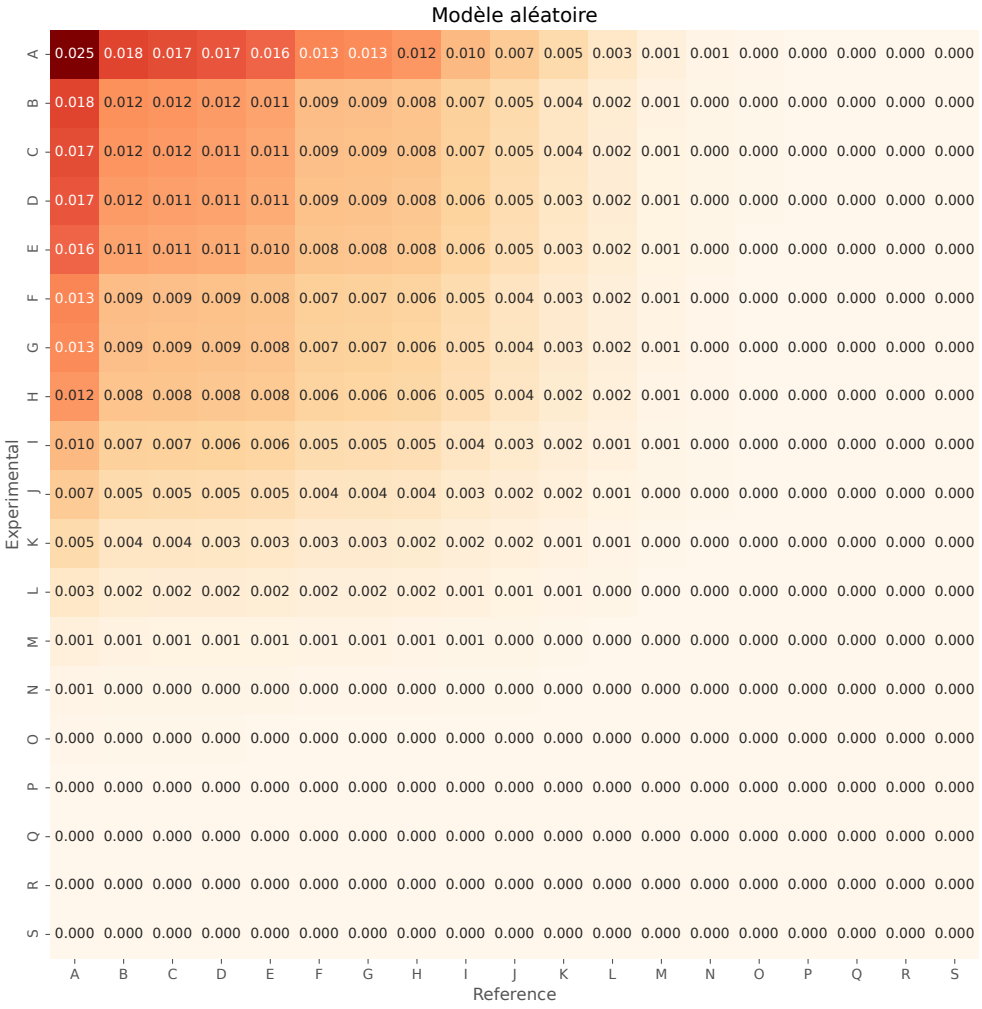


Figure 3.11: Probabilité de coïncidence pour chaque paire de catégories, en utilisant la probabilité d'un segment d'appartenir à une catégorie pour le génome d'E. coli et l'oligonucléotide CGTC

La probabilité de coïncidence des fragments x et y multiplié par la probabilité de présence pour chacune des catégories des segments x_i du fragment x et y_j de y pris indépendamment, comme présenté dans la méthode d'altshul

$$P(x, y|R) = \prod_i q_{Cat(x_i)} \prod_j q_{Cat(y_j)}$$

La probabilité de coïncidence entre x_i et y_i formant la matrice fig. 3.11, c'est-à-dire le produit de la probabilité d'appartenance à une catégorie par celle d'une autre catégorie.

Le modèle aléatoire utilise simplement le produit des fréquences définies dans le tableau tbl. 3.2. La matrice est constitué du produit des fréquences des catégories prises deux à deux. Elle représente la probabilité de rencontrer deux catégories prises aléatoirement dans le génome.

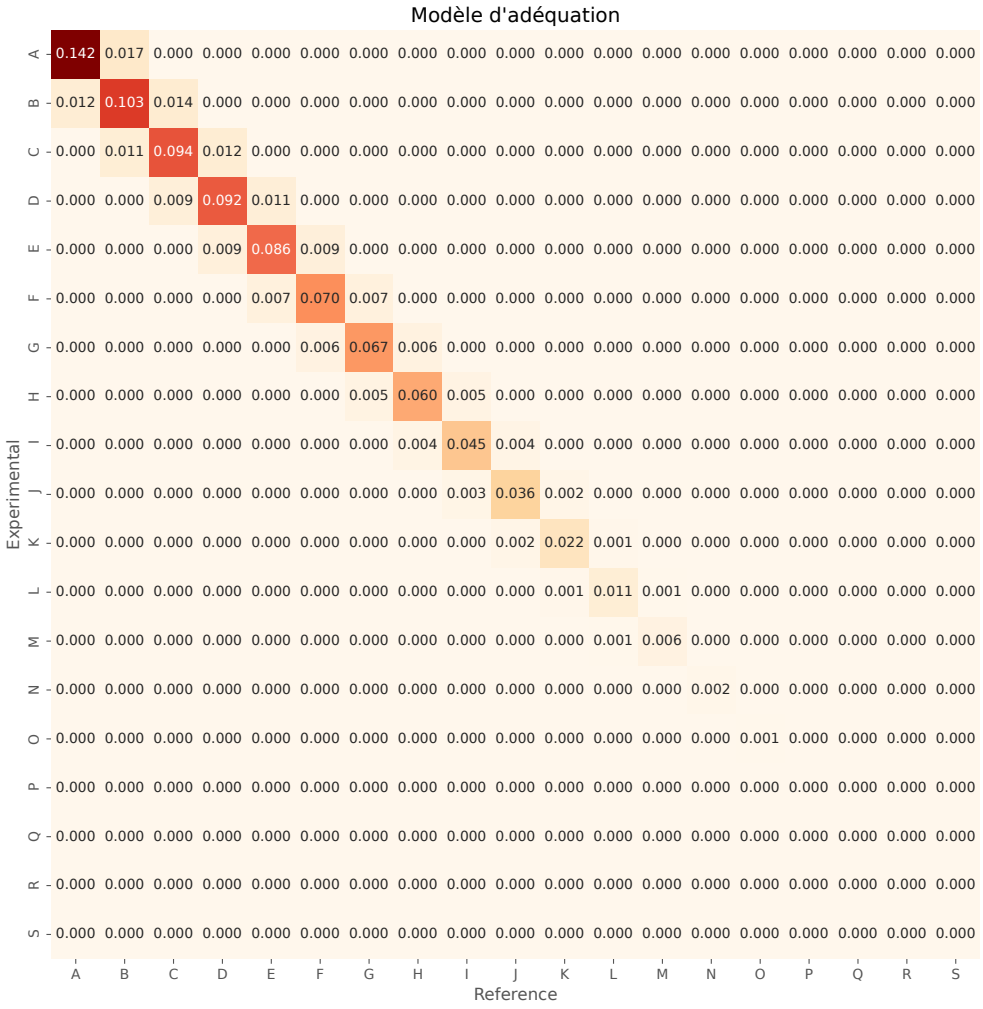


Figure 3.12: Probabilité d'adéquation : sachant qu'un segment appartient à une catégorie donnée, la probabilité qu'un autre segment lui correspondant appartienne à une catégorie. Construit en utilisant la probabilité d'existence d'un segment d'une longueur donnée sur le génome d'E. coli avec l'oligonucléotide CGTC

Le modèle d'adéquation est lui basé sur la probabilité de la catégorie des x_i et y_i de x et y sachant que x et y proviennent à l'origine du même contexte. Pour cela, nous utilisons la méthode du Kernel Density Estimator dont le noyau est basé sur $X \sim \mathcal{N}(\mu \simeq 0.88; \sigma \simeq 0.04)$ de l'étirement de chaque segment, et $Y \sim (\mu \simeq 0, \sigma \simeq 3nm)$, sur chaque position d'hybridation, de la même manière que pour la construction des catégories. Ces noyaux sont basés sur l'observation empirique de données expérimentales déjà acquises au laboratoire ; ces paramètres sont susceptibles de varier en fonction du bruit du système pour une expérience donnée, et de l'uniformité de la distribution de l'aimantation des billes.

Pour construire la matrice, non comptons le nombre de segments d'une taille donnée dans le génome d'E. coli (ou de tout autre génome de référence) et nous assignons la probabilité d'un segment de cette

taille d'appartenir à une catégorie donnée dans la matrice multipliée par leur nombre.

Nous répétons cette opération pour les segments de chaque taille. À la fin de ce processus, nous divisons la matrice par le nombre total de segments du génome de références.

$$P(x, y|M) = \prod_i p_{Cat(x_i)Cat(y_i)}$$

Enfin, nous effectuons le logs odds ratio des deux matrices, formant une matrice de score à partir des deux matrices de probabilité aléatoire, et d'adéquation.

$$s(a, b) = log(\frac{p_{ab}}{q_a q_b})$$

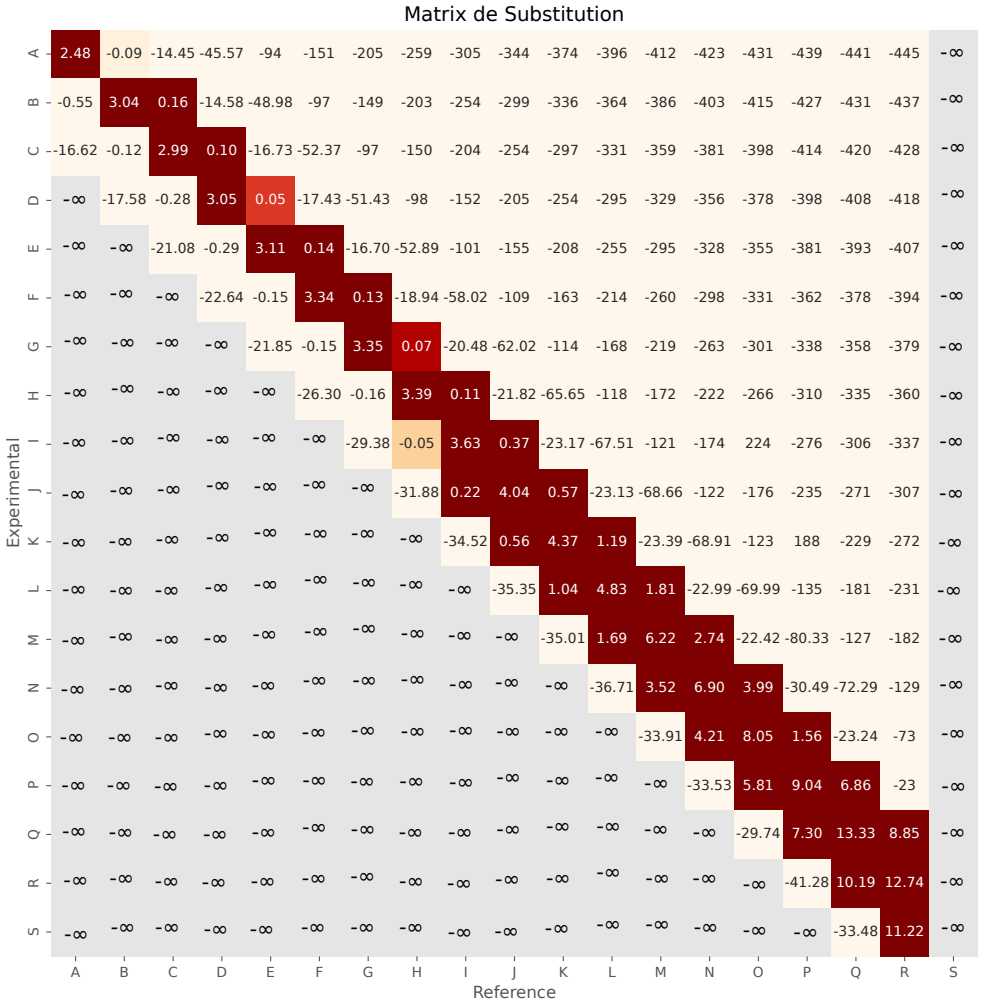


Figure 3.13: Matrix de score tel que $s(a,b) = log(\frac{p_{ab}}{q_a q_b})$ soit le log de la matrice du modèle d'adéquations sur la matrice du modèle aléatoire. Cette matrice avantage très fortement les adéquations parfaites entre deux catégories; plus deux catégories sont éloignées les unes des autres, plus le score devient mauvais. Atteignant un seuil d'impossibilité, c'est à dire $-\infty$ qui correspondent aux cases grises. Sans atteindre, $-\infty$ un score très négatif aura pour conséquence le rejet immédiat d'une adéquation.

3.2.5 Fusion de segments

Comme nous avons pu le constater dans l'exemple fig. 3.7, en comparant la signature théorique aux données expérimentales, certaines hybridations ne sont pas retrouvées..[4] De même, nous pouvons constater des hybridations supplémentaires à des positions non attendues.

Pour nous accommoder de ces problèmes d'hybridations manquantes / surnuméraires, nous introduisons la notion de « fusion de segments ». Une fusion considère deux segments voisins comme

[4]Soit il n'y a pas d'hybridation à la position indiquée, soit le taux d'hybridation trop faible nous oblige ne pas en tenir compte

un seul, et classe celui-ci selon la catégorie à laquelle il appartiendrait moyennant un coût de fusion fixé à -0.5, par test successif de valeur. Plusieurs fusions peuvent être effectuées successivement, s'accompagnant d'un coût supplémentaire à chaque ajout d'un segment au segment fusionné lui aussi fixé à -0.5. Par comparaison, en moyenne, un match parfait ajoute 5.64 au score.[5] Deux limites supérieures sont fixées : une limite de l'extension de fusion, permettant de restreindre le nombre de segments successifs pouvant être fusionnés, et une limite du nombre total d'évènements de fusion dans une comparaison donnée. Ces deux limites permettent de minimiser la perte de spécificité des signatures du fait de l'augmentation du nombre de fusions, et de diminuer le taux de faux positif.

3.2.5.1 Paramètres de l'algorithme et valeurs usuelles

Paramètre	Description	Valeur
Catégories de segments	Définition des catégories dans lesquelles est classé chaque segment	tbl. 3.2
Matrice de score	$s(a,b) = log(\frac{p_{ab}}{q_a q_b})$	fig. 3.13
Cout d'une fusion	Coût déduit du score de la solution suite à une fusion de deux segments	0.5
Nombre de fusions successives acceptées	Nombre maximum de fusions successives avant qu'une solution candidate soit éliminée	6
Nombre de fusions globales acceptées	Nombre maximum de fusions avant qu'une solution candidate soit éliminée	10
Ratio de rejet	Score minimal (normalisé par le nombre d'éléments) pour qu'une solution candidate ne soit pas rejetée au cours de son évaluation	0
Ratio d'acceptation	Score minimal (normalisé par le nombre d'éléments) pour qu'une solution candidate soit acceptée comme solution valide	0.4
Distribution de l'étirement acceptable	Basée sur les valeurs d'étirements définies dans la partie sec. 1.3.4	$\sigma = 0.04, \mu = 0.88, Ic =$ 99.9%

3.2.6 Fonctionnement général de l'algorithme

Maintenant que nous avons présenté l'ensemble des éléments permettant l'utilisation de l'algorithme de cartographie, parlons de son fonctionnement lui-même.

Lors d'une recherche ou préalablement, les catégories de segments sont construites à partir de l'oligonucléotide utilisé et du génome de référence, Celui-ci est encodé sous forme de distances et la matrice de score est ainsi construite.

Par la suite, la cartographie consiste à parcourir du génome de référence.

Tout au long de ce parcours nous étudions des solutions candidates que nous éliminons si leur score est trop mauvais, et acceptons si il est suffisant et que la solution est complète (à savoir tous les éléments de la signature expérimentale sont explorés).

[5] Pour rendre plus cohérent les scores, et les coûts de fusion, nous aurions pu normaliser la matrice des scores par le score moyen d'un match parfait.

Pour cela, à chaque segment du génome de référence, on crée une solution candidate, ajoutée à la liste de traitement.

Ensuite nous prenons l'ensemble des solutions acceptables à l'itération précédente (c'est-à-dire lors de l'observation du segment précédent) que nous ajoutions également à la liste de traitement. Nous évaluons alors le score de ces différentes solutions candidates. Si celui-ci est supérieur au score de rejet, nous vérifions si la solution est complète, si c'est le cas, nous l'ajoutons aux solutions acceptées. Sinon, nous l'ajoutons à la liste des solutions acceptables pour la prochaine itération. Ensuite, nous ajoutons à la liste des solutions acceptables l'ouverture d'une fusion si le score offert par la fusion est supérieur au score actuel.

Nous répétons cette opération tout au long du génome.

```
pending = []
nextpending = []
accepted = []
# Pour tous les segments du génome de référence
for rsegment in refSegments:
    # On ajoute la locus actuel comme une solution candidate
    pending += newCandidate(locus(rsegment))

    # On traite chaque solution candidate
    for candidate in pending:
        # On rejette les candidats ne respectant pas les critères de
        # * score minimal
        # * nombre de fusions maximum
        # * dont la taille n'est pas compatible avec le fragment expérimental
     if !candidateRejected?(candidate):
            # Si il est possible de créer une nouvelle fusion
            # et qu'elle offre un score supérieur au candidat actuel
            if fusionAccepted?(candidate):
                # On crée un nouveau candidat avec fusion en cours
                fusionCandidate = newFusionCandidate(candidate, rsegment)
             # On ajoute le nouveau candidat à la nouvelle liste des candidats
             nextPending += fusionCandidate
        # Dans tout les cas, on crée un nouveau candidat avec le segment courant,
        # en fermant la fusion courante
        candidate = newCandidate(candidate, rsegment)

        # Si le candidat a une longueur, un score suffisant,
        # il est accepté comme solution, sinon
        # il retourne dans la liste de candidats
        if candidateAccepted?(candidate):
            accepted += candidate
        else:
            nextPending += candidate

        # On met à jour la liste des solutions candidates, et on recommence
        pending = nextPending
        nextPending = []
```

```
# On trie les solutions par leur score
sort(accepted)
```

À la fin de ce processus, nous trions l'ensemble des solutions acceptées par leur score.

Le calcul du score d'alignement est effectué au fur et à mesure de la construction d'une solution candidate. Cependant, pour expliquer le calcul d'un score, nous allons utiliser un exemple en montrant le calcul complet du score.

Pour cela nous reprenons la figure fig. 3.10, qui correspond à un exemple de la cartographie d'une molécule PS825HP à son locus de référence attendu.

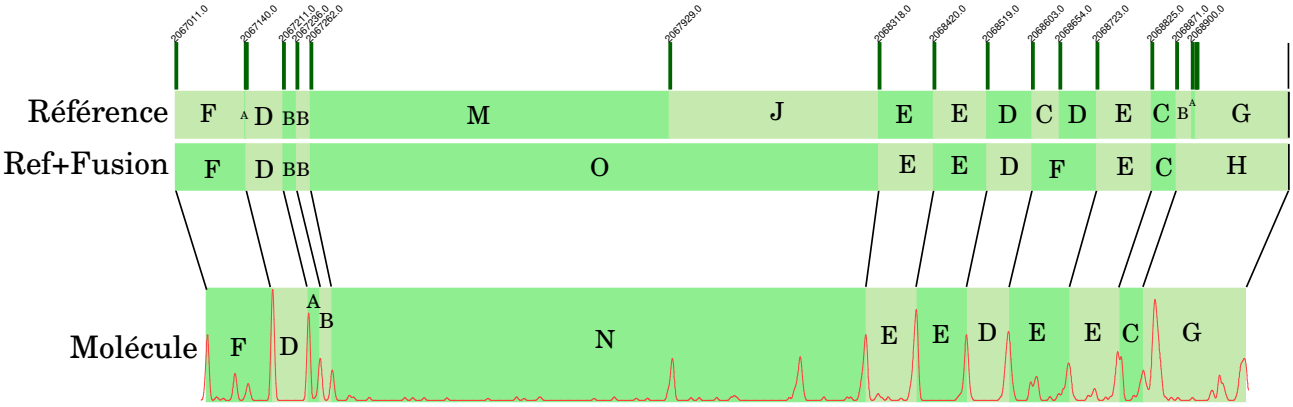


Figure 3.14: Reprise de la figure fig. 3.10

Calcul du score de la signature de référence face à la signature expérimentale

```
   F     A
   |     |
(3.34) - inf
   |     |
   F     D

= inf
```

Immédiatement, du fait de l'hybridation manquante le calcul s'arrête puisque la solution candidate actuelle est inférieure à 0. Par contre, la fusion du F avec le A possède un score supérieur. Nous continuons donc en fusionnant F et A.

```
   F+A=F      D     A     B    M     J
   |          |     |     |    |     |
(3.34-0.5)+3.05-0.09+3.04+3.52-155
   |          |     |     |    |     |
   F          D     A     B    N     E

= -142,64
```

De nouveau, alors que plusieurs segments se sont bien alignés, la solution est éliminée à cause d'un match trop mauvais. On continue donc avec la solution qui fusionne les segments M et J. De proche

en proche, la plupart des solutions sont éliminées au fur et à mesure laissant uniquement les meilleurs alignements.

Voici l'exemple du meilleur alignement trouvé pour cette séquence :

```
   F+A=F    D    A     B       M+J=O    E E     D    C+D=F     E    C      B+A+G=H
     |      |    |     |         |      | |     |      |       |    |         |
(3.34-0.5)+3.05-0.09+3.04+(3.99-0.5)+(3.11*2)+3.05+(0.14-0.5)+3.11+2.99+(0.07-0.5-0.5)
     |      |    |     |         |      | |     |      |       |    |         |
     F      D    B     B         N      E E     D      E       E    C         G

= 23,48
```

À présent nous avons une vision d'ensemble du fonctionnement de la méthode de classification de segments.

3.3 Conclusion

Durant ma thèse, j'ai exploré de nombreuses possibilités algorithmiques pour réaliser la cartographie de signature. J'ai présenté ici les deux principales méthodes ayant abouti et démontré leur efficacité. Elles sont issues d'un certain nombre d'itérations et d'une démarche d'amélioration successive tout au long de ma thèse.

Les méthodes de classifications et par régression sont en réalité encore des preuves de concepts qui ne sont pas parfaitement adaptés à l'utilisation à grande échelle. Cependant, elle offre de grandes possibilités d'améliorations que nous traiterons notamment lors de la conclusion de la thèse.

Ayant pour l'instant expliqué uniquement le fonctionnement des algorithmes en tant que tel, nous allons pouvoir, dans les deux prochains chapitres, tester et comparer ces méthodes et étudier la manière dont elles permettent de résoudre le problème de la cartographie de molécules.

Chapitre 4

Estimation des performances de l'identification de signature

Sommaire du chapitre

Lors de la partie précédente, nous avons présenté deux méthodes permettant l'identification de signatures. Afin d'évaluer leurs performances, j'ai développé une plateforme de tests. Celle-ci permet tout d'abord de simuler des signatures pseudo-expérimentales basées sur un génome de référence et un jeu d'oligonucléotides répondant aux contraintes présentées dans la sec. 1.3 notamment la possibilité de faire varier l'étirement, le bruit de mesure, la non-détection d'hybridations ou la présence d'hybridations non spécifiques, ainsi que le choix des oligonucléotides. L'ensemble de ces paramètres peut être rejoué sur une même signature. Cette plateforme de tests permet également de tester les méthodes d'identification en faisant varier l'ensemble des paramètres des algorithmes, seuls ou conjointement avec les paramètres de simulation des signatures pseudo-expérimentales.

Toutes ces actions sont configurables par des fichiers au format YAML, exécutables en ligne de commande dans un environnement Python adéquat, après compilation du code critique en C++. À partir de cette plateforme, il est possible d'évaluer les différentes propriétés des algorithmes, de lancer des tests et de les visualiser.

L'ensemble des tests sont effectués en parallèle sur les quatres cœurs d'un processeur Intel Core i7-3770 CPU @ 3.40GHz avec 32Go de RAM

4.1 Simulation des données pseudo-expérimentale

Nous cherchons à étudier la capacité des méthodes explorées dans le Chapitre 3 ec à établir la correspondance entre une signature de molécule et une séquence spécifique du génome de référence.

Pour obtenir des valeurs informatives pour un génome donné, il convient d'étudier toute la diversité des signatures possibles pour ce génome et de s'assurer de la robustesse des méthodes aux différents bruits. Pour cela, j'ai adopté deux stratégies selon le type de simulations effectuées. La première consiste à découper l'ensemble du génome en fragments se chevauchant, par exemple avec des fragments de 2000 Bp dont les points de départ sont séparés de 200 Bp tout le long de la séquence de référence. L'échantillon des signatures obtenues de cette manière couvre une très large diversité des signatures possibles pour un génome donné. La seconde méthode consiste à y piocher aléatoirement un grand nombre de fragments dans le génome. Si le nombre de signatures est suffisant, la diversité du génome est correctement représentée. La première méthode est systématique, peut impliquer un grand nombre de tests et permet en théorie de se confronter au génome entier. La seconde permet de choisir plus directement le nombre de tests et de se confronter à un échantillon aléatoire du génome.

L'algorithme classe la liste des locus qui présentent le meilleur score selon leur qualité. Le locus réel de la signature simulée est alors recherché dans ce classement, le cas idéal étant qu'il soit classé premier.

Tout au long du chapitre, nous allons utiliser la notion d'exactitude de la cartographie. Il s'agit du rapport entre le nombre de signatures simulées pour lesquelles l'algorithme de cartographie classe en premier le locus réel (le locus attendu) et le nombre total de signatures testées.

4.2 Hypothèse de simulation

Trois paramètres principaux peuvent être la cause de faux positifs dans nos analyses : le bruit sur la mesure (intrinsèque aux pinces magnétiques et aux conditions d'acquisition des données, en nm), l'étirement variable des molécules (proportionnelle à la magnétisation des billes superparamagnétique) et l'existence potentielle d'hybridations manquantes ou surnuméraires.

De par leur conception, les deux méthodes de cartographie ne sont pas équivalentes du point de vue de leur sensibilité aux différentes sources d'erreur. La méthode de régression est très peu sensible à l'étirement et à l'existence d'hybridations manquantes. Le choix des classes de la méthode de classification est, quant à elle, conçu pour très bien résister au bruit et aux variations d'étirement, dans la limite des conditions expérimentales.

Par ailleurs, lors de l'utilisation des méthodes de cartographie, un fonctionnement optimal nécessite que l'utilisateur précise comme paramètres de l'algorithme le niveau attendu de bruit, d'étirement de molécules et d'hybridations manquantes. Dans nos tests, lorsque nous ferons varier un de ces paramètres sur les échantillons nous ajusterons donc les paramètres de cartographie en conséquence.

Enfin, notons que, dans les premières simulations, nous ne prendrons pas en compte la possibilité d'hybridations manquantes pour la méthode de classification, afin de mieux isoler l'effet du bruit.

4.2.1 Bruits de mesure

Nous comparons ici la sensibilité au bruit de mesure des deux méthodes de cartographie présentées dans le chapitre précédent. Pour cela, nous utilisons la première stratégie de simulation, avec des fragments de 2000 Bp extraits tout le long du génome, positionnés toutes les 200 Bp pour la méthode de classification et toutes les 2000 Bp pour celle par régression. Dans le même temps, l'étirement est fixé à la valeur la plus probable dans les expériences (0.88 Bp par nanomètre) et nous n'avons pas supprimé d'hybridations. Le bruit indiqué est exprimé en nanomètres. Si nous exprimions ce bruit en termes de base, il serait en moyenne de 14% ($\frac{1}{0.88} = 1.14$) plus important en raison de l'étirement des molécules.

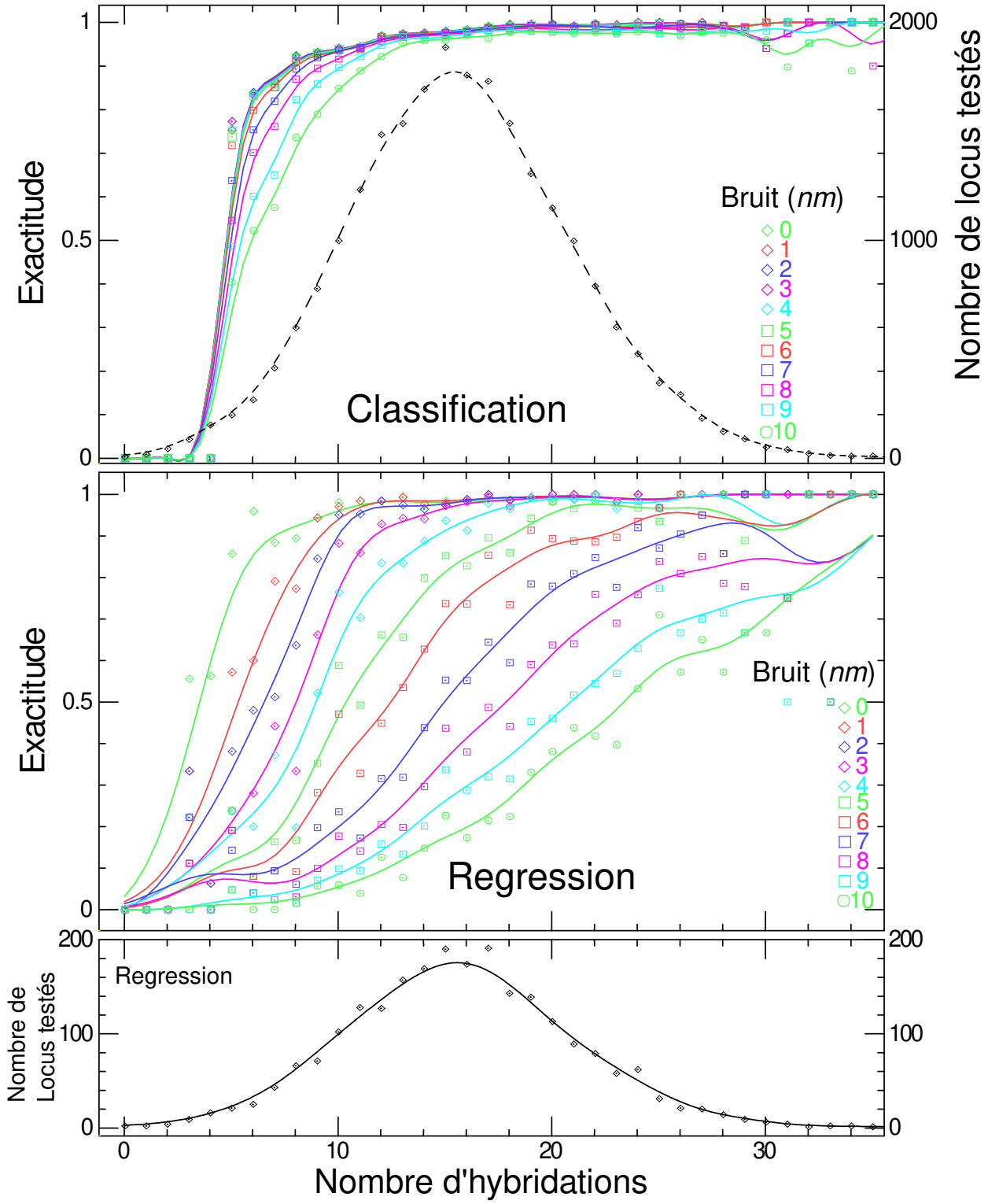


Figure 4.1: Taux de succès des deux méthodes en fonction du nombre d'hybridations de la signature pour des valeurs de bruit croissante. Il faut un peu moins de dix hybridations pour trouver une séquence, mais les deux méthodes se comportent de façon différente par rapport au bruit. La méthode par régression y est très sensible. Le test a été réalisé en tirant au hasard des séquences qui ont en moyenne 16 hybridations sur 2000 bps, le nombre d'essais pour chaque valeur du nombre d'hybridations est donné par la courbe noire en pointillée. Le petit nombre de séquences avec beaucoup d'hybridations fait que la statistique au-delà de 25 n'est pas très bonne.

Méthode par classification des segments: L'algorithme parvient à retrouver la position d'origine des molécules simulées lorsqu'elles possèdent typiquement plus de 7 hybridations, ce qui est attendu pour le génome d'E. coli. À partir de 10 hybridations, le taux d'exactitude dépasse les 80% et ce dans l'ensemble de l'intervalle de bruit considéré. Le bruit ajouté à la position des hybridations apparaît compensé par l'existence d'un grand nombre d'hybridations.

Méthode par régression: Contrairement à la méthode par classification des segments, la méthode par régression se montre beaucoup plus sensible au bruit qui dégrade significativement la capacité de cartographie de l'algorithme. Alors que, sans bruit, elle présente des résultats similaires à l'algorithme de classification, à 5nm de bruit nous observons que pour atteindre une confiance proche de 99%, il est nécessaire de disposer de 20 hybridations sur une signature. À partir de 6 nm, il n'est plus envisageable d'atteindre une exactitude supérieure à 90%

Les molécules utilisées faisant 2000Bp, nous attendons en moyenne 16 hybridations pour l'oligonucléotide CGTC (4 Bp), ce que nous retrouvons sur le panel du bas. Il faut noter que le nombre de molécules ayant plus de 20 hybridations décroît rapidement, ce qui conduit à un manque de puissance statistique pour ces molécules.

Comme le montre la Figure 4.1, la reconnaissance du bon locus dépend en premier lieu du nombre d'hybridations observées pour un oligonucléotide : dans le cas d'E. coli et de l'oligonucléotide CGTC, avec la méthode de classification ,il faut environ 10 hybridations pour que la chance de succès atteigne 90%. Le bruit dégrade cette reconnaissance: dans le cas de la méthode par classification la dégradation de la qualité des résultats reste faible. En revanche, avec la méthode par régression l'effet est beaucoup plus marqué et la reconnaissance devient vite problématique. De nouveau, notons que dans les paramètres de la méthode par classification, aucune hybridation manquante n'est tolérée, ce qui aurait un impact sur ses performances, alors, que par construction la méthode par régression accepte les hybridations manquantes.

Lors d'une expérience, nous ne maîtrisons pas le nombre d'hybridations qui dépend du locus de la séquence et de sa longueur. Alors que précédemment nous observions l'exactitude en fonction du nombre d'hybridations, dans la Figure 4.2 nous observons l'exactitude uniquement en fonction du bruit pour des séquences de 2000Bp (les mêmes séquences que dans fig. 4.1) , quelque soit le nombre d'hybridations sur ces séquences. Cela donne ainsi un aperçu du taux de succès global pour une méthode donnée, dans une situation réelle, ou la quantité d'hybridation n'est pas maîtrisée.

Avec la méthode de classification, on peut espérer un taux de succès de 95% pour un niveau de bruit correspondant à celui des expériences. La méthode de régression n'est utilisable que pour des niveaux de bruit assez faible, pour les niveaux de bruit très bas elle atteint un taux de succès plus élevé que la méthode de classification.

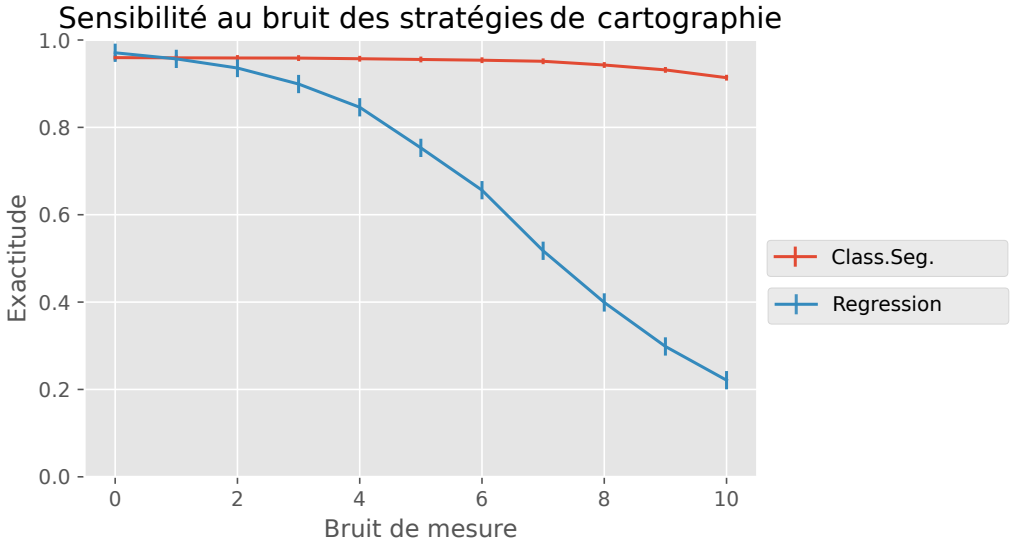


Figure 4.2: L'algorithme par classification se montre exact à 95% pour un bruit jusqu'à 7 nm. La méthode de régression est beaucoup plus sensible à ce paramètre.

4.2.2 Étirement

Nous testons à présent la sensibilité de nos approches à l'étirement. Pour cela, nous fixons un bruit nul, ne considérons aucune hybridation manquante, et ne faisons varier que l'étirement.

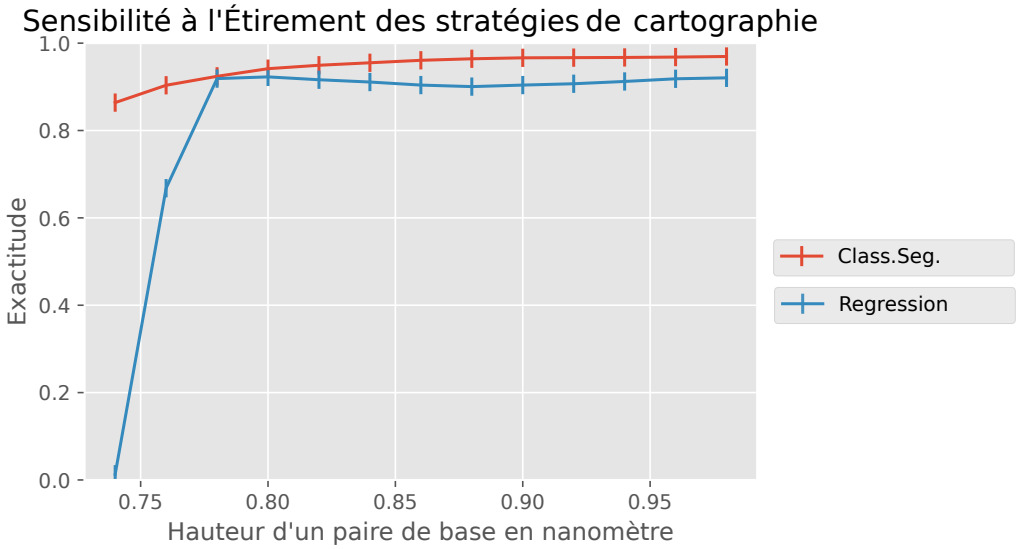


Figure 4.3: Taux de succès des deux méthodes en fonction de l'étirement. Le paramètre d'étirement est assez facile à prendre en compte dans les deux méthodes.

Grâce à leurs conceptions, les deux méthodes sont assez peu sensibles à la variabilité de l'étirement. Dans la méthode de régression, nous fixons les bornes de l'étirement possible afin de restreindre l'espace de recherche. La décroissance constatée en dessous d'un étirement de 0.78 nm/Bp est directement liée à cette limite.

4.2.2.1 Premières conclusions :

La méthode par régression, conçue pour accommoder un étirement variable, montre de bonnes performances. La méthode par classification montre également d'excellentes performances et une faible sensibilité aux variations d'étirement.

Néanmoins, il faut noter ici que les espaces de recherche des deux algorithmes ne sont ici pas comparables. En effet, la méthode de classification est paramétrée pour ne pas tenir compte des hybridations manquantes. La méthode par régression ne possède pas cette possibilité, elle explore donc un espace de recherche plus important. Ceci peut influencer négativement ses performances comparativement à la méthode par classification. À présent, nous allons prendre en compte des hybridations manquantes pour les deux algorithmes, rendant la comparaison plus juste.

4.2.3 Hybridations manquantes

Lors de nos expériences, nous observons des hybridations manquantes. La façon la plus simple de comprendre ce phénomène et que chaque hybridation est décrite par un phénomène aléatoire que nous avons simulé par une loi binomiale. On considère que nous effectuons 100 cycles d'ouverture et de fermeture de nos molécules, avec une probabilité d'hybridation de 8% à chaque cycle. Donc en moyenne nous devrions avoir 8 hybridations, mais ce nombre est sujet à des fluctuations importantes. Dans les expériences, une hybridation est considérée comme valide dès lors qu'elle est observée plusieurs fois. Dans nos tests, nous avons fixé ce seuil de détection à 6 hybridations.

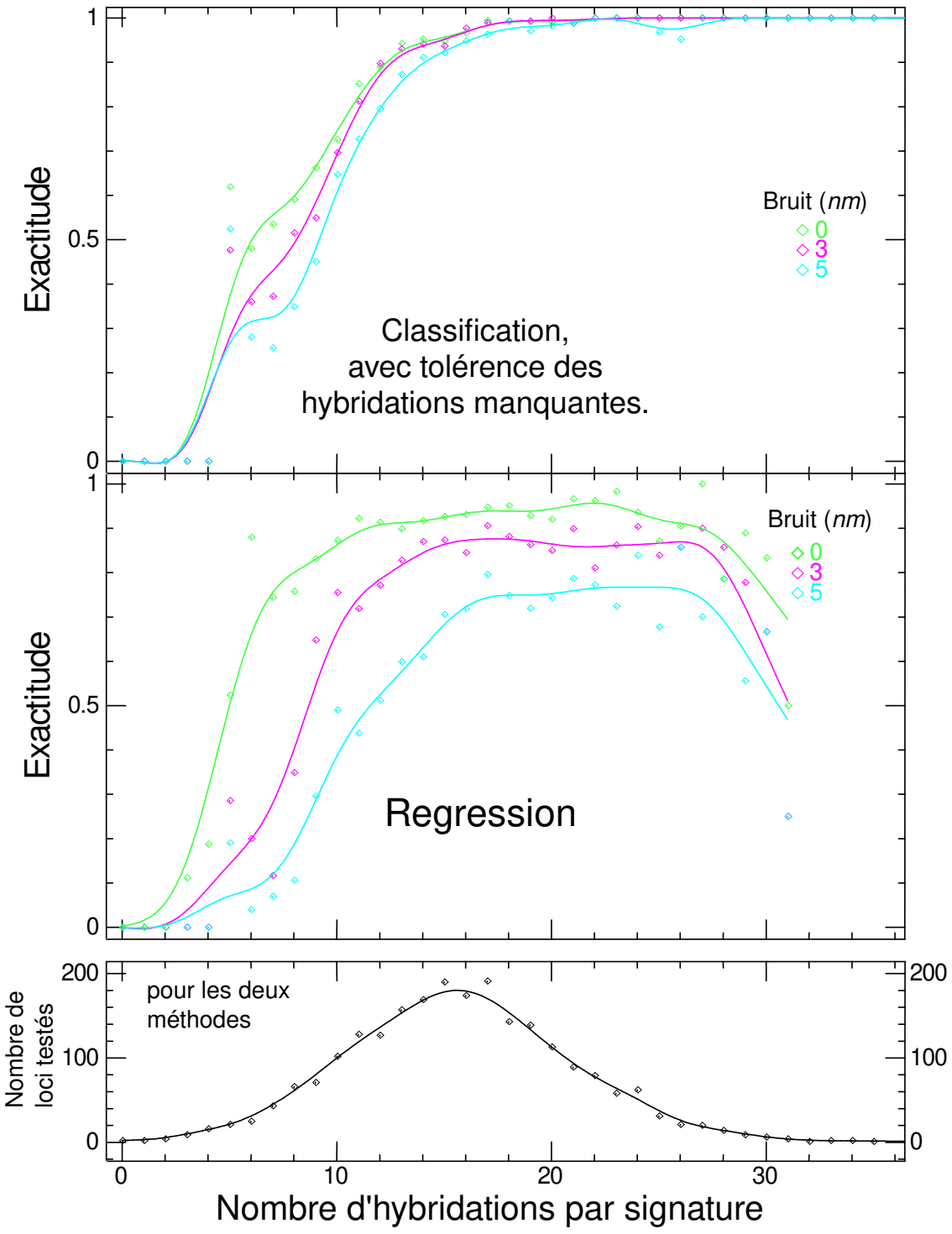


Figure 4.4: Taux de succès des deux méthodes, sans hybridations manquantes dans les signatures simulées, mais en paramétrant les méthodes de telle façon qu'elles admettent la possibilité d'hybridations manquantes. Le calcul est fait pour un étirement aléatoire dans la gamme expérimentale avec trois niveaux de bruit différents. La possibilité de traiter des signatures possédant des hybridations manquantes ne modifie pas les performances de la méthode par régression, l'algorithme gérant par défaut les hybridations manquantes (les résultats sont identiques à fig. 4.1), par contre elle dégrade le taux de succès de la méthode de classification qui doit considérer un jeu de solutions possibles plus grand.

Pour cette simulation, nous avons utilisé un étirement variable suivant une loi normale centrée autour de $0.88bp/nm$ avec un écart-type de 0.04. Nous avons simulé trois niveaux de bruits : $0nm$, $3nm$ et $5nm$, sachant que ces deux dernières valeurs sont proches de celle observée sur nos microscopes.

Dans un premier temps, avec ces réglages nous avons mesuré les performances dans des conditions où les signatures ne possédaient pas d'hybridations manquantes, mais en indiquant au programme qu'il pouvait y en avoir. Pour l'algorithme par régression, les résultats sont équivalents aux précédents. En revanche, pour la méthode par classification, l'existence potentielle d'hybridations manquantes paramétrées dans l'algorithme dégrade légèrement les performances. Elle restent toutefois au-dessus de 90% d'exactitude si les fragments d'ADN possèdent au moins 12 hybridations. On constate que la méthode de classification reste moins sensible au bruit que celle par régression.

À présent nous testons sur les mêmes échantillons, mais avec des hybridations manquantes, ce qui dégrade évidemment les résultats pour les deux méthodes. Sur les 16 positions d'hybridations, nous observons en moyenne, avec les paramètres de notre loi binomiale, en moyenne 3 hybridations manquantes.

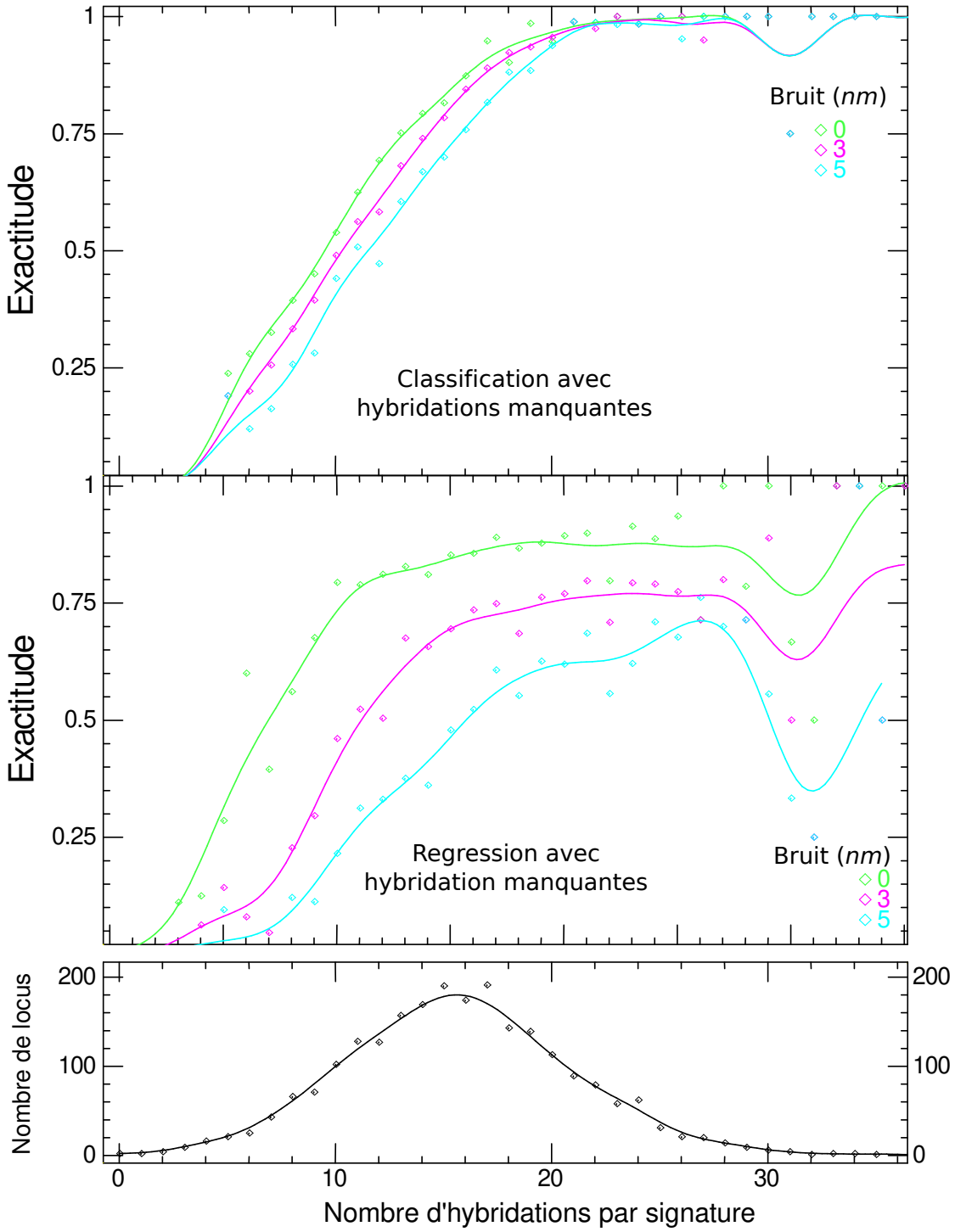


Figure 4.5: Taux de succès des deux méthodes pour des signatures ayant un étirement variable, un bruit donné et ayant en moyenne trois hybridations manquantes en fonction du nombre total d'hybridations. Les hybridations manquantes dégradent la qualité des deux méthodes d'abord mécaniquement en réduisant le nombre total de celles-ci, mais aussi en diminuant la spécificité des signatures. Toutefois si le nombre total d'hybridations est suffisant le taux de succès devient très acceptable pour la méthode par classification.

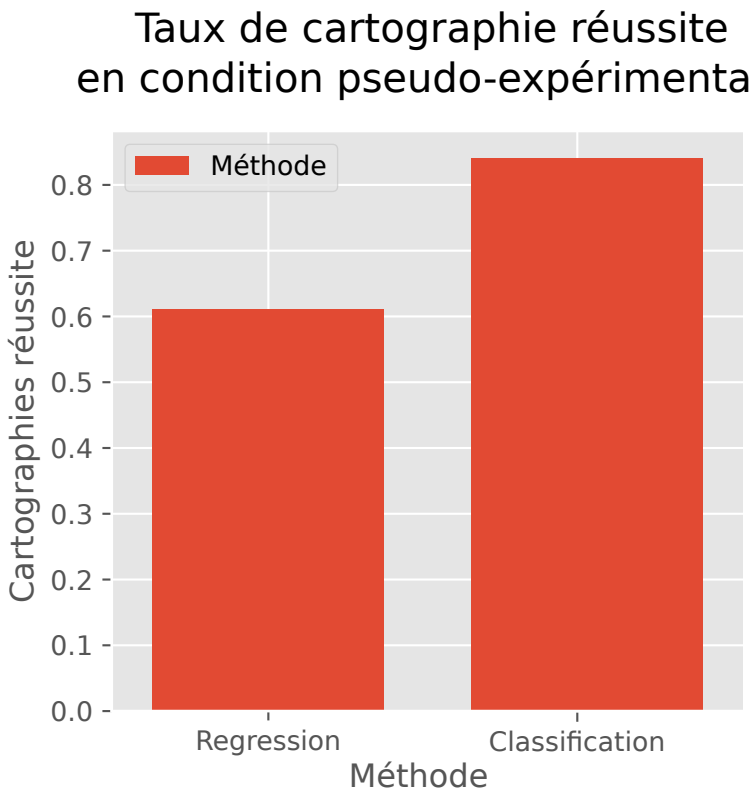


Figure 4.6: Taux de succès des deux méthodes pour des signatures ayant un étirement variable, un bruit donné et avec en moyenne trois hybridations manquantes, quel que soit le nombre d'hybridations sur la signature considérée. On constate que le taux de cartographie réussite de la méthode par régression atteint 61%, et pour la classification 84%

Ces deux derniers graphes résument le message que l'on peut tirer sur les deux méthodes présentées:

Les deux méthodes fournissent un taux de succès intéressant. Cependant, la classification a un comportement plus intéressant : elle est peu sensible au bruit. Et dès que le nombre d'hybridations est suffisant, elle offre le meilleur taux de succès, souvent proche de 100%. Si l'on regarde son taux de succès quel que soit le nombre d'hybridations, 84% des échantillons sont correctement cartographiés.

Cette sensibilité importante de la méthode par régression au bruit dégrade énormément les résultats. Ainsi seul 61% des échantillons sont correctement cartographiés avec la méthode par régression. Une amélioration dans la robustesse au bruit pourrait lui permettre d'avoir des performances supérieures à la méthode par classification.

La méthode de simulation provoque un effet de bord : en effet, en moyenne, pour le génome d'*E. coli* et l'oligonucléotide CGTC, un fragment de 2000 Bp possède 16 hybridations. L'exactitude est donc calculée sur un nombre beaucoup plus important d'échantillons pour une signature dans cette moyenne, que pour une signature possédant très peu ou beaucoup d'hybridations. Cela nous amène à penser qu'étant donné le nombre d'échantillons, les points à partir de 27 hybridations ne peuvent être considérés comme significatifs.

Cependant, de premiers résultats nous amènent à penser que, pour les deux méthodes, lorsque la densité d'hybridations augmente de manière trop importante, le taux de succès décroît. On peut comprendre cet effet, car, à haute densité, le bruit diminue la qualité de l'information. On considère qu'à partir d'une densité moyenne supérieure à une hybridation toutes les 50 bases, les performances

des algorithmes commencent diminuer.

4.3 Conclusion

Dans ce chapitre, nous avons tout d'abord pu découvrir la méthodologie ayant permis de tester et de comparer les résultats donnés par les deux algorithmes de cartographie, confrontés à des données pseudo-expérimentales. Ensuite, nous avons pu analyser les données utiles issues de ces tests.

Durant cette thèse, j'ai cherché à produire un logiciel permettant de tester le plus de paramètres possible sur chaque algorithme en fournissant des échantillons stables et reproductibles. Ce logiciel devrait pouvoir être réutilisé aisément dans le futur pour tester d'autres algorithmes étant donnée sa généricité. Il a permis de réaliser un très grand nombre de simulations d'optimiser les paramètres de recherche des algorithmes, de comparer les versions successives et trouver les points faibles et forts de chaque méthode. C'est un outil robuste.

Les résultats issus de ce logiciel nous ont permis d'avoir une estimation au plus près possible des conditions expérimentales des capacités des algorithmes. À la fin de cette phase d'expérimentation, nous pouvons conclure qu'en l'état actuel, l'algorithme par catégorie possède les caractéristiques les plus intéressantes pour la cartographie de fragments sur E. coli. En possession de cette information, c'est cet algorithme que nous allons utiliser pour les tests expérimentaux de la partie suivante.

Alors que ces tests semblent montrer que la méthode par régression est inutilisable en condition réelle. Il faut nuancer ce propos. En effet, l'algorithme possède encore des améliorations simples à mettre en place qui n'ont pas été mises en place durant ce travail de thèse, et mérite d'être évalué, tel que nous pourrons le voir dans la discussion.

Chapitre 5

Tests expérimentaux sur molécules uniques

Sommaire du chapitre

5.1 Démarche expérimentale

Afin de vérifier expérimentalement la spécificité des signatures et la capacité de nos algorithmes à déterminer de manière spécifique l'origine d'un fragment d'ADN parmi un ensemble de référence, un protocole en aveugle a été établi. Nous décidons de travailler sur l'organisme Escherichia coli, souche K-12 NEB5-alpha (Escherichia coli strain K-12 NEB5-alpha Anton et Raleigh 2016), dont nous souhaitons extraire 11 Fragments d'ADN compris entre 1000 et 3000 Bp. Un collaborateur du laboratoire, Jimmy Ouellet, a construit des jeux d'amorces PCR à partir du génome de référence qui permettent d'amplifier *a priori* spécifiquement 11 fragments avec comme unique contrainte l'absence de site de restriction de l'enzyme Bsa1. Je reçois les paires d'amorces PCR sans mention de leur séquence ni de leurs coordonnées dans le génome. Nous avons ensuite construit les molécules en épingle à cheveux à partir de l'ADN obtenu par PCR sur le génome d'E. coli en utilisant chaque jeu d'amorces. Ma connaissance des molécules se limite donc à leur longueur mesurée sur gel d'électrophèse après PCR, et en l'absence du site Bsa1. Une petite faille dans notre protocole est apparue au fur à mesure des expériences: la numérotation des amorces correspondait à l'ordre des locus dans le génome. Les détails du protocole se trouvent en annexe sec. 7.2

La construction des molécules en épingles à cheveux s'est faite en plusieurs étapes. Tout d'abord, j'ai effectué la construction d'une première molécule pour vérifier les étapes du protocole. Après avoir confirmé expérimentalement avec les pinces magnétiques la structure en épingle à cheveux et l'efficacité de la construction, j'ai construit 5 autres molécules en suivant le même protocole, puis les 5 molécules

restantes ont été construites par Jimmy Ouellet et Sylwia Gorlach, à Depixus. La conception du protocole permettant à la construction des molécules en épingle à cheveux est attribuable à Jimmy Ouellet ainsi qu'à son équipe au sein de Depixus.

À partir de ces 11 fragments d'ADN de séquence inconnue, l'obtention d'une signature nécessite un jeu d'oligonucléotides afin de réaliser des hybridations. Pour les premières expériences, c'est la méthode présentée dans la sec. 2 qui est utilisée et résumée dans le tableau tbl. 2.4. Par la suite, à l'aide des nouvelles méthodes développées à Depixus, des oligonucléotides de 4 bases, inaccessibles au préalable, ont été utilisés. Leur taux d'hybridation moyen sur le génome étant équivalent aux jeux de 4 oligonucléotides s'hybridant sur 5 bases, ils peuvent être utilisés seuls.

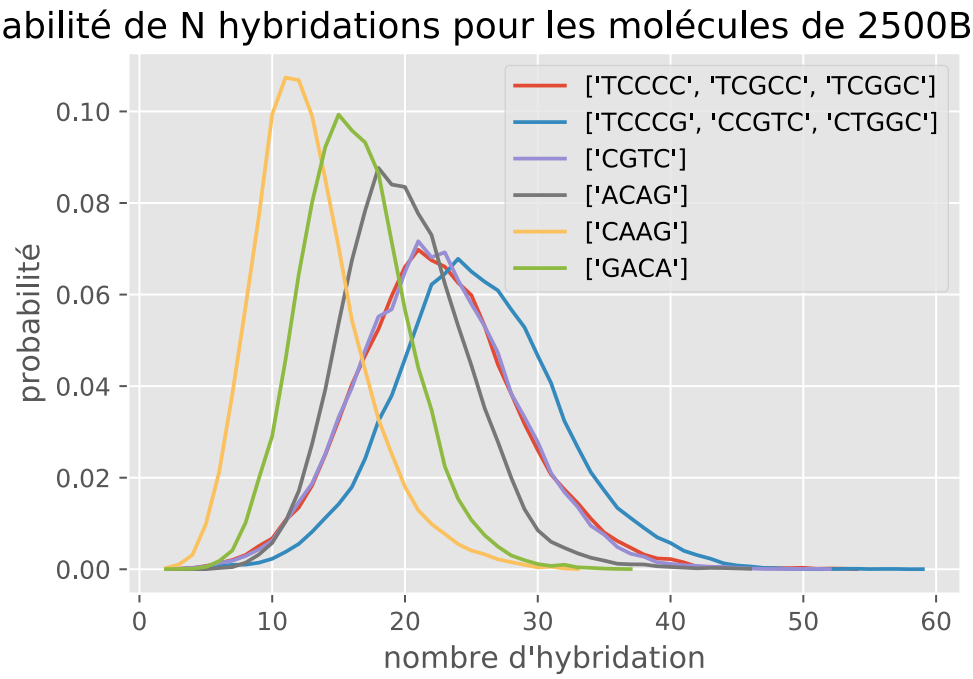


Figure 5.1: Nombre d'hybridations sur des molécules de 2500 Bps pour différents oligonucléotides ou groupes d'oligonucléotides. Ce paramètre est directement relié à la chance de succès de nos méthodes.

Une fois les molécules construites et les oligonucléotides sélectionnés, l'acquisition des signatures est effectuée sur un microscope à pinces magnétiques. Pour les molécules PS813HP à PS823HP, les expériences sont réalisées à l'ENS, à 27°C. Une petite erreur de manipulation nous a fait réaliser des expériences sur différentes molécules en épingle à cheveux successivement dans la même cellule microfluidique. Il s'est avéré que le rinçage des cellules microfluidiques n'éliminait pas l'intégralité des molécules des expériences précédentes, ce que nous avons découvert en identifiant des molécules en épingle à cheveux des tests précédents qui se sont retrouvées mélangées avec les molécules de la dernière expérience, formant ainsi un mélange de molécules. Les molécules PS825HP à PS833HP ont été analysées à Depixus dans de nouvelles cellules microfluidiques en évitant ce mélange. Ainsi, nous n'observons plus de molécules mélangées au sein d'une même expérience. Les expériences sont cette fois-ci effectuées à 23°C.

Avant chaque expérience en présence d'oligonucléotides, les molécules en épingle à cheveux sont ouvertes et fermées à plusieurs reprises, en augmentant la force appliquée relativement lentement dans le temps, c'est-à-dire en faisant une rampe. Ces expériences permettent d'estimer à quelle position des

aimants la molécule en épingle à cheveux s'ouvre, et ce pour chaque bille. Cette position correspond à une force appliquée de 15 pN, et nous disposons ainsi d'une calibration approchée de chaque bille. De plus, ces cycles lents dits «rampes» permettent d'avoir des informations structurelles sur la molécule. En effet, certaines molécules peuvent avoir des défauts structuraux (des bases manquantes sur un brin, par exemple) provoquant un retard de l'ouverture ou de la fermeture, de manière systématique pour tous les cycles. Les rampes nous informent également de manière marginale sur la composition des nucléotides de la molécule : une région contenant des bases *fortes* (G et C, 3 liaisons hydrogène) s'ouvrira et se fermera à une force légèrement plus importante. Au contraire pour les régions riches en bases *faibles* (A et T, 2 liaisons hydrogène) pour lesquelles l'ouverture et la fermeture s'effectueront à une force moins importante. L'énergie libre d'une séquence GC est typiquement de 4.38 k_BT (soit environ $2.65kcal/mol$) par base, tandis que celle d'une séquence AT de 1.46 k_BT (soit environ $0.86kcal/mol$).

Les paramètres des expériences ont été choisis pour limiter au maximum les effets de dérive au cours de l'expérience et ainsi limiter la nécessité de correction des données.

5.2 Traitement des données

L'acquisition des données s'effectue en utilisant un logiciel nommé pico, développé au laboratoire. L'ensemble des paramètres et des données brutes de l'expérience (notamment la position en hauteur des molécules) est stocké dans un fichier "track". Les données brutes des acquisitions ne peuvent pas être directement exploitées par les algorithmes de recherche de signatures dont la donnée d'entrée est une succession de positions d'hybridation. Il est donc nécessaire d'effectuer une série de traitements normalisés pour parvenir à des données de cette forme.

Durant la période de ma thèse, avec le développement de Depixus et le partenariat avec l'ABCDLab, les outils et méthodes de traitement des données ont beaucoup évolué. Au début de ma thèse, J'ai participé à ces travaux notamment par le développement d'une méthode de correction de la dérive thermique, et le développement d'un pipeline de traitement qui a ensuite servi de toute première base pour le développement des outils chez Depixus. Lors de mes premières analyses, j'ai continué à utiliser mon pipeline de traitement d'origine, ainsi qu'un grand nombre d'améliorations fourni par Vincent Croquette.

J'ai ensuite commencé à utiliser le pipeline totalement réécrit au sein de Depixus, par Pol d'Avezac ainsi que David Salthouse. Les évolutions à la fois proposées par Vincent Croquette et l'équipe de Depixus ont permis d'améliorer significativement la qualité des données extraites, en récupérant de manière plus précise la position des hybridations, ainsi que leur dynamique. De plus, un gros travail d'automatisation a été effectué.

J'ai lié le pipeline de Depixus à des traitements manuels pour l'intégrer à mon programme qui orchestre le traitement des données ainsi que la recherche des locus des signatures et la présentation des résultats. Tout ceci est entièrement paramétrable au travers de fichier YAML.

5.2.1 Préparation des données

Le prétraitement des données s'effectue tout d'abord à l'aide de traitement manuel sur PlayItAgain-Sam, développé au laboratoire, suivi du traitement par le pipeline de Depixus.

Tout d'abord, pour chaque molécule, on découpe et on superpose l'ensemble des cycles d'ouverture/fermeture des molécules fig. 1.33. Sont supprimés les cycles pour lesquels la molécule ne s'ouvre / ne se ferme pas correctement, ainsi que les cycles pour lesquels la qualité du signal n'est pas celle attendue, par exemple lorsque le suivi de la bille n'est plus effectué correctement, provoquant des

sauts dans sa position. À ce stade est éliminé l'ensemble des molécules ne possédant pas suffisamment de cycles corrects ou dont les données présentent des caractéristiques qui ne permettront pas la recherche des signatures. Par exemple un blocage systématique masquant la plupart des hybridations ou un bruit des données manifestement trop important dénotent un problème dans le système de suivi ou une construction de la molécule en épingle à cheveux incorrecte. À ce stade, les critères de sélections sont subjectifs et permettent d'éliminer les données qui seront manifestement inutilisables. Notons que l'on considère qu'en dessous d'environ 30 cycles, les données ne seront pas exploitables.

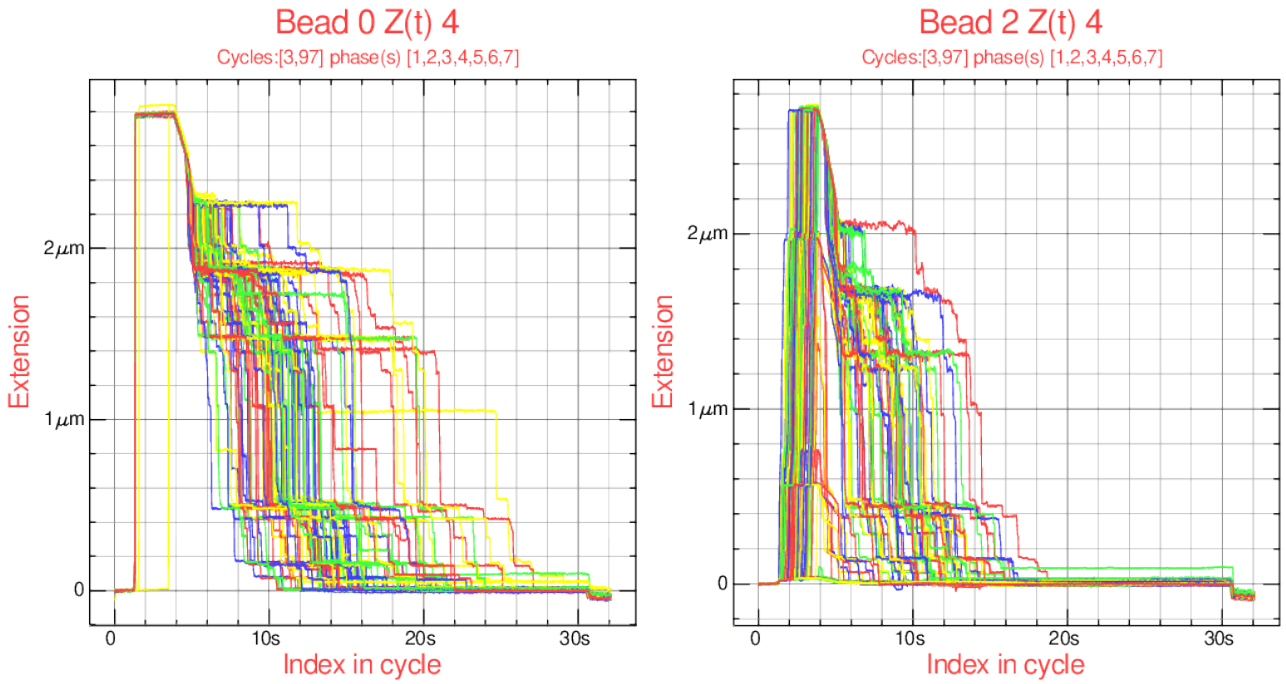


Figure 5.2: Observation de deux molécules en épingle à cheveux, l'ensemble des cycles sont superposés et se distinguent par des couleurs différentes. La molécule de gauche possède suffisamment de cycles, la molécule s'ouvre et se ferme correctement, le bruit semble d'un niveau acceptable. La molécule de droite semble s'ouvrir correctement pour la plupart des cycles, mais l'ouverture se fait avec un retard dû au fait que la bille colle un peu à la surface, le bruit sur la position est très important, et ne permet pas de distinguer les hybridations. Cette molécule sera éliminée lors de la sélection préliminaire.

5.2.2 Correction des données

L'étape suivante consiste à tenter de réduire le bruit systématique lié notamment à une dérive de la température induite par le mouvement des aimants qui ne sont pas thermalisés. Pour cela, la méthode la plus simple consiste à suivre une ou plusieurs billes fixées à la surface de l'échantillon. De cette manière, il est possible de simplement soustraire la moyenne des données de hauteurs des billes fixes aux hauteurs des molécules utilisables. En l'absence de bille fixe, il est possible d'inférer une pseudo bille fixe à partir des données de l'ensemble des billes paramagnétiques, et de la soustraire de manière similaire. fig. 1.31

5.3 Résultats

Dans cette partie, nous allons nous pencher sur les résultats d'une expérience et de sa cartographie. Étant donné les résultats de l'étude comparative présentés dans la partie précédente, nous avons choisi d'utiliser l'algorithme de classification de segments pour la cartographie. À l'heure actuelle l'étude des résultats n'a pas été menée avec la méthode par régression.

5.4 Résultats détaillés d'une expérience

Examinons d'abord les résultats d'une expérience, de manière à découvrir les enjeux expérimentaux liés à l'acquisition et comprendre les causes de l'échec de la cartographie sur certaines molécules. Nous prendrons l'exemple d'une expérience effectuée avec la molécule en épingle à cheveux PS817HP et l'oligonucléotide CGTC, effectuée le 9 juin 2017. Cette partie repose sur l'étude a posteriori des résultats. Ayant acquis par notre algorithme la connaissance du locus d'origine des différentes molécules, nous utilisons cette information pour effectuer une analyse détaillée des problèmes.

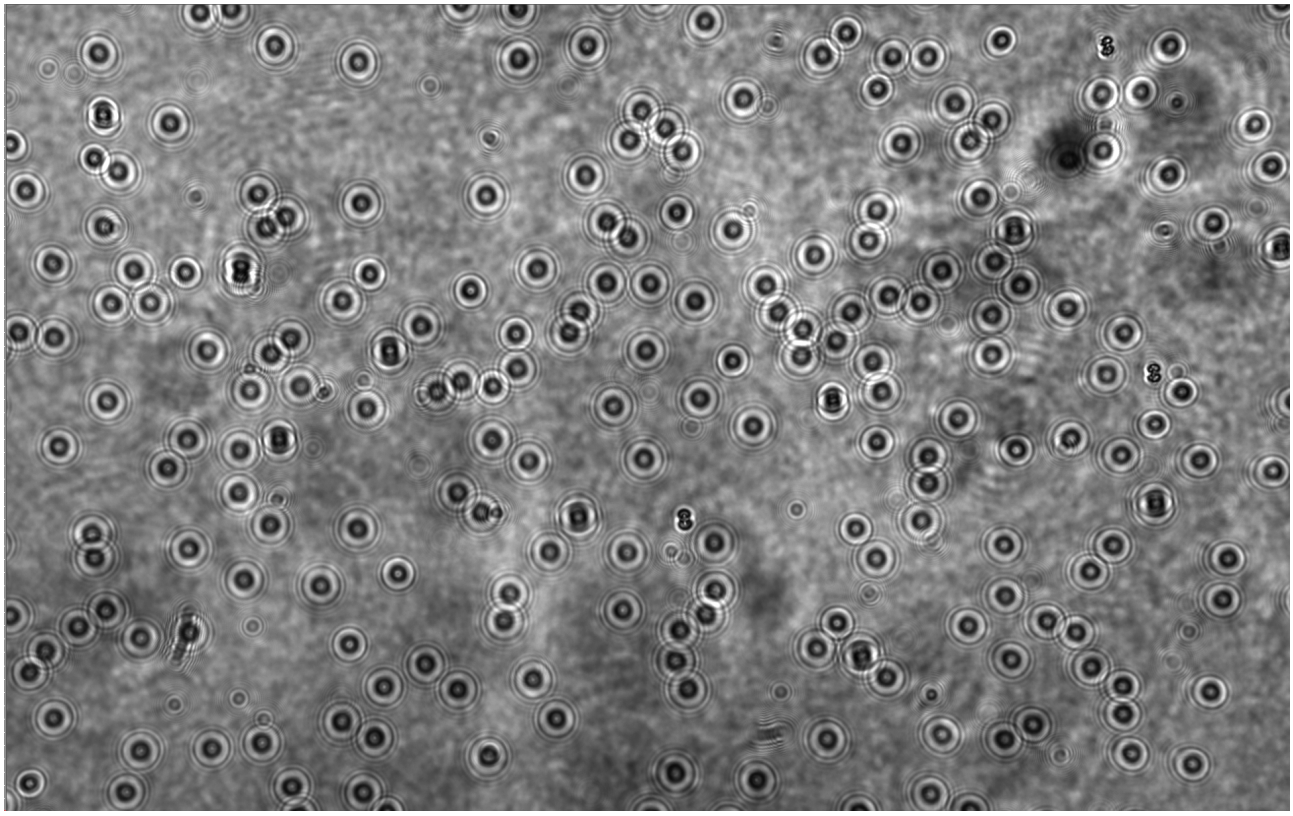


Figure 5.3: Champ de vue de l'expérience PS817HP-CGTC. nous observons ici que la qualité du fond de l'image est relativement mauvaise. Le problème peut provenir de poussières dans la ligne optique de l'instrument et quelques artéfacts qui peuvent à la fois provenir d'une mauvaise uniformité du recouvrement de la surface permettant d'attacher les molécules ainsi que d'impuretés présentes dans l'huile de microscope utilisé à l'interface entre l'objectif et la lamelle de microscope. Le contraste étant suffisant, l'impact sur la résolution est minime.

Table 5.1: Statut des billes paramagnétiques visibles sur le champ de vue

Statuts des billes / molécules	Quantité
Billes paramagnétiques visibles	~160
Molécules en épingle à cheveux fonctionnelles et qui peuvent être suivies	40
Molécules en épingle à cheveux sélectionnées lors du prétraitement	27

Lors de cette expérience, environ 160 billes magnétiques sont présentes à l'écran. Sur ces 160 billes, 40 molécules sont distinguables et se comportent comme une molécule en épingle à cheveux lors du début de l'expérience. L'importante perte de molécules (120 billes) est explicable de plusieurs façons.

Liées aux limites biologiques

- La bille paramagnétique s'est fixée directement à la surface de l'échantillon par des interactions non spécifiques.
- Une molécule est bien attachée à la bille, mais:
  - La quantité de matériau paramagnétique n'est pas suffisante pour permettre aux aimants d'exercer une force assez grande sur la bille, pour ouvrir la molécule en épingle à chevaux. Son extension reste quasi nulle;
  - La molécule est partiellement fixée à la surface;
  - La construction de la molécule est défectueuse, elle n'est pas une simple épingle à cheveux;
  - Plusieurs molécules sont attachées à la bille et à la surface. La force appliquée par les aimants n'est pas assez grande pour ouvrir plusieurs molécules simultanément.

Liées au suivi des billes paramagnétiques

- Les billes magnétiques sont trop proches, ce qui rend le suivi des billes difficiles voir impossible;
- Des taches sur le fond de l'échantillon réduisent la qualité de l'image (en haut à droite de la Figure 5.3), ce qui empêche un suivi stable des billes paramagnétiques.

Lors de l'acquisition des données, les cycles hautes forces / basses forces, permettent de rendre fonctionnelles de nouvelles molécules, qui au départ ne le semblaient pas. Elles ne prennent pas part à l'acquisition des données et ne sont pas considérées dans les données subséquentes.

Suite à l'acquisition des données, les prétraitements présentés dans la Section 5.2.1 entraînent l'élimination des données de 13 molécules, de qualités insuffisantes. 27 molécules sont donc exploitées pour la cartographie en utilisant la méthode de catégorisation de segments.

Table 5.2: Statut de cartographie des molécules analysées

Statut de la cartographie des molécules	Nombre de molécules
Cartographies ayant échoué	**9** (33%)
Cartographies ayant réussi	**18** (67%)
Total	**27**

L'analyse manuelle des molécules pour lesquelles il n'a pas été possible de retrouver la position dans cette expérience nous donne des informations sur les raisons de ces échecs. Les principales causes proviennent de défauts lors de l'acquisition et du prétraitement des données ainsi que dans les choix des paramètres d'optimisation de l'algorithme et la possibilité de celui-ci d'accepter des positions d'hybridation surnuméraires. En effet, d'une part, lorsque les données acquises comportent des dérives,

le prétraitement peut provoquer la création deux positions d'hybridations proches là où une position unique devrait être trouvée. D'autre part, des positions d'hybridations qui n'étaient pas attendues sont présentes sur les molécules en épingles à cheveux observées. L'hypothèse de base, lors de la conception des algorithmes, était que certaines hybridations pouvaient manquer, mais nous n'avions pas vraiment prévu la présence d'hybridations supplémentaires. Par conséquent pour des raisons de temps de calcul, la possibilité d'avoir des positions surnuméraires n'a pas été ajoutée, bien qu'elle soit facile à implémenter. Pour pallier ce défaut, le taux d'hybridation nécessaire pour accepter une position d'hybridation dans une signature a été artificiellement augmenté, de manière à éliminer des hybridations surnuméraires quitte à créer des hybridations manquantes, qui sont bien tolérées dans la version actuelle de l'algorithme. Cette manière de faire permet un taux de récupération satisfaisant, mais ne résout pas toutes les situations rencontrées, notamment lorsque le taux d'hybridation à une position non attendue est trop élevé.

Ces paramètres ont été optimisés pour chaque expérience. Généralement pour les expériences effectuées à 27°C, le taux d'hybridation minimal a été paramétré à 3% des cycles pour qu'une position soit acceptée, À 23°C, elle a été paramétré aux alentours de 15% des cycles.

Table 5.3: Provenance des molécules analysées, déterminées par la comparaison manuelle entre la signature des molécules et les locus candidats.

Molécule analysée	*E.coli* Locus	Nombre de molécules	Correctement Cartographiées	Nombre d'hybridations théoriques attendues dans la molécule.
PS815HP	55158	11	6 (55%)	21
PS817HP	266822	16	12 (75%)	15
Total		**27**	18 (67%)	

Comme indiqué précédemment l'utilisation d'une même cellule microfluidique sur plusieurs expériences à amené à un mélange de molécules en épingle à cheveux dans un même champ de vue, malgré le rinçage important, et l'arrachement des liaisons par passage d'un coton-tige sur la surface. Des molécules provenant des deux locus (PS815HP et PS817HP, Voir Tableau 5.3) sont donc présentes lors de cette expérience, et ont pu être identifiées. Ce problème avec le protocole expérimental à eu le mérite de montrer la capacité de l'algorithme par classification, à cartographier des molécules correctement, sans connaissance a priori de leur locus.

On peut voir que le nombre d'hybridations théoriques plus faibles (15 Hybridations) attendues au locus de PS817HP, n'empêche pas l'algorithme de cartographier correctement 75% des billes, ce qui est encourageant.

Le phénomène de mélange de molécules au sein d'une même expérience reste d'occurrence rare pour l'ensemble des expériences menées. Elle concerne uniquement une expérience PS817HP+CGTC, PS819HP+CGTC et PS821HP+CGTC. Dans le cas des deux dernières molécules, les expériences ont été réeffectuées sur des cellules neuves pour que le biais disparaisse.

Les expériences pour les molécules à partir de PS825HP ont été effectuées en utilisant une nouvelle méthode développée à Depixus qui élimine tout risque de mix de molécules en épingle à cheveux.

5.5 Résultats sur l'ensemble des molécules testés

Pour l'ensemble des molécules construites, 20 expériences ont été effectuées sur les molécules en épingle à cheveux, en utilisant dans un premier temps les jeux d'oligonucléotides présentés dans la partie sec. 2,

notamment le mix1 composé de `TCCCCA, TCGCCA, TCGGCA`.

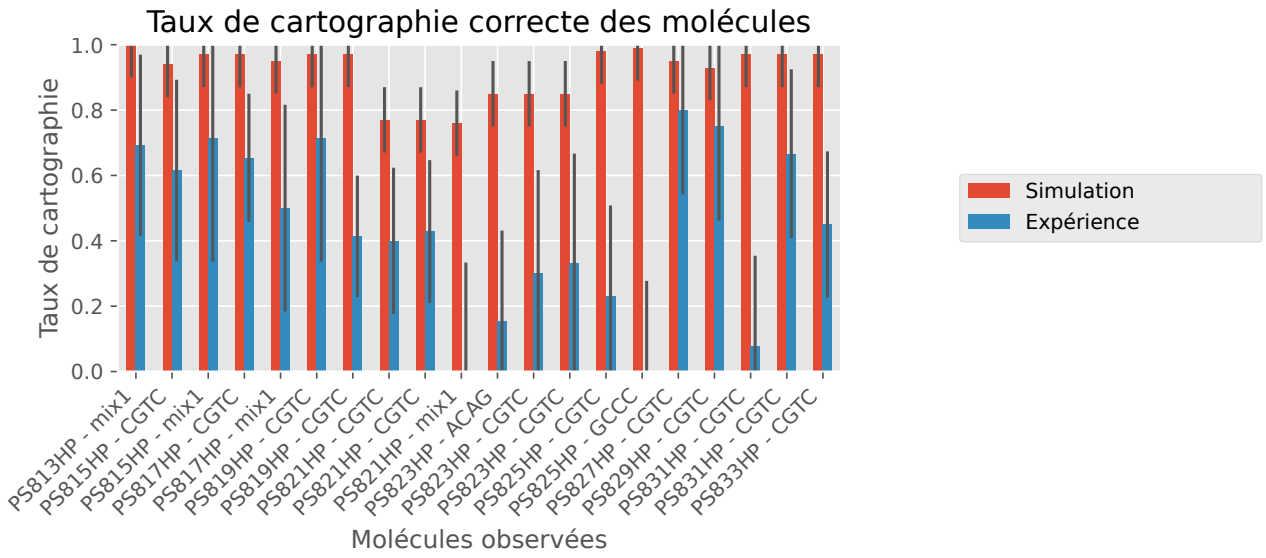


Figure 5.4: Taux de cartographie correcte pour l'ensemble des expériences effectuées sur les 11 molécules construites. En rouge les simulations au plus proche des conditions expérimentales: $5nm$ de bruits, étirement de $0.88Bp/nm, \sigma = 0.04nm, I_c = 99.9\%$ et en moyenne 3 Hybridations manquantes. En bleu les expériences effectuées sur microscope à pinces magnétiques. Ce résultat montre que des améliorations expérimentales peuvent apporter un taux de succès meilleur.

Dans la figure fig. 5.4, sur les onze locus sélectionnés, tous ont été identifiés, mais avec un taux de cartographie correcte variable, avec des taux variant de 5% à 80%, et pour certaines expériences, de moins bonne qualité n'ont pas permis la cartographie mais la répétition des expériences a permis le recouvrement de l'ensemble.

Pour les simulations, connaissant à présent le locus de chaque molécule, nous pouvons simuler la molécule au locus donné en utilisant le génome de référence. Nous faisons jouer les paramètres présentés présenté en 1.3.4 et en 4, c'est à dire avec un bruit de $5nm$, un étirement moyen de $0.88Bp/nm$, et un écart type de $0.04Bp/nm$, On prend également un taux de sélection des hybridations utilisant une loi binomiale à 100 tirages, $p = 0.08$ et $P(X > 6)$. Ces chiffres sont équivalents à effectuer 100 cycles de fermeture de molécule, avec une probabilité d'hybridation à une position donnée de 8% et l'acceptation d'une position d'hybridation si plus de 6 hybridations sont constatées à cette position

Nous répétons cette opération 100 fois au même locus, pour simuler 100 signatures différentes d'une même molécule.

Les simulations sont beaucoup plus optimistes que les résultats expérimentaux, en effet elles prévoient un taux de succès pour les expériences se situant entre 80% et 99%. La simulation ne reflète donc pas totalement la réalité expérimentale. On note néanmoins que lorsque les résultats expérimentaux sont peu performants, les simulations produisent également des performances plus faibles. Cette relative corrélation suggère que les simulations reflètent une réalité expérimentale. Les paramètres choisis pour les simulations sont supposés proches de la réalité, à l'exception de l'absence d'hybridations surnuméraires qui sont compensées par un seuil de sélection des hybridations très haut lors du prétraitement des données Voir 5.4 . Cela n'explique cependant pas l'ampleur de la différence observée entre la simulation et l'expérience. Celle-ci pourrait s'expliquer par la qualité de l'acquisition des don-

nées. Un grand nombre d'améliorations sont en cours sur le plan de la construction des molécules, des cellules microfluidiques, ainsi que du prétraitement des données dans le laboratoire ainsi qu'à Depixus.

Ces résultats valident donc notre méthode d'identification, mais il reste une marge pour augmenter le taux de succès, à la fois par l'amélioration des algorithmes, ainsi que par des optimisations de l'ensemble de la chaîne d'acquisition.

Chapitre 6

Discussion

Sommaire du chapitre

Ce travail de thèse a abordé la dynamique d'hybridations d'oligonucléotides d'ADN, de LNA, et d'autres molécules modifiées. J'y ai développé des méthodes de cartographie et je les ai testées à la fois au travers de simulations numériques et d'expériences sur des molécules issues de la bactérie E. coli.

Alors que mon travail démontre la faisabilité de l'utilisation d'un système de signatures utilisant des pinces magnétiques ainsi que des sondes telles que des oligonucléotides. Il ouvre des perspectives nouvelles notamment pour la reconnaissance des marqueurs épigénétiques sur des molécules uniques. En effet nous avons démontré l'identification rapide de molécules sur lesquelles nous pouvons déterminer la présence de ces marqueurs épigénétiques. Cependant, il soulève également des problèmes quant à la réalisation de cartographie à grande échelle utilisant ces méthodes. Ces problèmes, qui ne mettent pas en péril l'idée de la cartographie par pince magnétique, devront être résolus pour utiliser le système en production.

Dans cette discussion, nous allons aborder les limites biologiques, instrumentales, et algorithmiques inhérentes au système actuel, et nous y apporterons de nouvelles hypothèses de recherche pour passer outre ces limitations. Nous proposerons ensuite des perspectives d'améliorations, et nous décrirons les prochaines étapes à franchir. Nous remettrons ensuite ces sujets dans le contexte global du champ de recherche.

6.1 Qualité de données expérimentales et du prétraitement des données

Au fur et à mesure du déroulement de ce projet de recherche, des problèmes liés aux données expérimentales sont apparus. Ceux-ci ont généralement pu être partiellement corrigés par des méthodes bio-informatiques, mais ils rendent plus complexe la cartographie de molécules. Ces problèmes méritent d'être mieux compris pour adapter les méthodes d'analyses ainsi que pour les corriger en amont durant la phase de préparation des molécules et des oligonucléotide, ou mieux durant la phase d'acquisition des données.

6.1.1 Problématiques liées aux hybridations surnuméraires

Durant les premières années de ma thèse, les données expérimentales montraient qu'il pouvait arriver que nous n'observions pas expérimentalement une hybridation qui aurait dû être constatée théoriquement, au regard du génome de référence. Nous les appelons ci-après des hybridations manquantes.

En revanche, il a été estimé peu probable que nous observions le problème inverse, c'est-à-dire une position supplémentaire sur la signature expérimentale, mais absente dans le génome de référence. Nous les appelons ci-après des hybridations surnuméraires. Cependant, l'expérience a montré par la suite que ces hybridations surnuméraires étaient en réalité fréquentes dans certaines conditions.

Tout d'abord nous avons pu observer dans la partie sec. 2.2.2, que des oligonucléotides LNA 6-mer pouvait s'hybrider uniquement sur 5 bases. La dernière base formant un mismatch provoquant ainsi des hybridations supplémentaires. Cependant, nous avons pu caractériser de manière relativement exhaustive le comportement de ses oligonucléotides, et simplement les considérer comme des 5-mer, au lieu de 6-mer.

Cependant, le problème des hybridations surnuméraires est réapparu quand nous avons commencé à utiliser des oligonucléotides 4-mer, possédant un renforcement supplémentaire des liaisons. Ceux-ci ont engendré des problèmes de cartographie des molécules pour des raisons que nous avons comprises plus tard comme étant des hybridations surnuméraires .

Au-delà des hybridations surnuméraires, un deuxième problème a été constaté qui semble lui être relié. En effet, en étudiant rétrospectivement les molécules dont nous avons établi la cartographie et notamment les cas les plus difficiles, nous avons remarqué que la probabilité d'hybridations mesurée (le rapport du nombre de cycles présentant un blocage à la position considérée au nombre total de cycles qui définit la hauteur du pic) est très variable, que ce soit pour les hybridations attendues ou pour les pics surnuméraires. Par ailleurs, les pics surnuméraires montraient parfois un nombre d'hybridations supérieur à celui des pics aux positions attendues fig. 6.1. Cette variabilité est plus grande que ce que l'on pourrait attendre avec la distribution de Poisson telle que présentée dans sec. 1.3.3.2.1.

L'observation d'une molécule issue de l'expérience basée sur la construction PS825HP et l'oligonucléotide CGTC illustre ces deux problèmes.

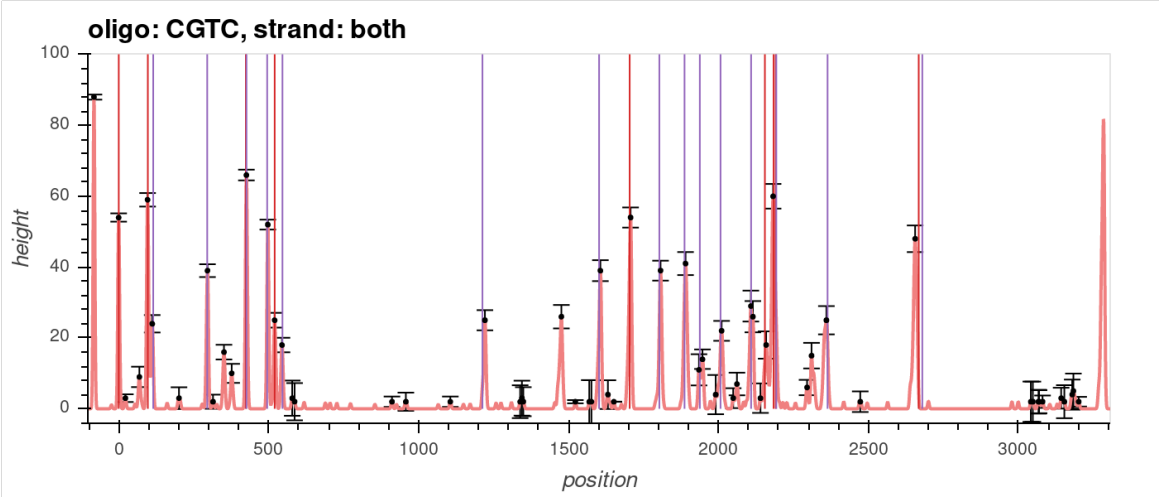


Figure 6.1: En rouge, l'histogramme du nombre d'hybridations observées de l'oligonucléotide CGTC sur la molécule en épingle à cheveux PS825HP (signifié par le terme *height* dans la figure) et les lignes verticales rouge et violettes correspondantes aux positions d'hybridation attendues sur les deux brins en paires de bases, en alignant *a posteriori* avec le génome de référence (Les lignes rouges indique une hybridation sur le brin sens, et les lignes violettes sur le brin anti-sens). Le nombre d'hybridations varie fortement d'un site à l'autre, certains pics sont très forts ($\simeq$ 60) d'autres, très significatifs ($\simeq$ 20) et ne peuvent correspondre à une erreur et un nombre important de pics sont petits avec juste quelques hybridations. Les prédictions correspondent bien à la majorité des pics observés, mais il est difficile de comprendre pourquoi certains sont très forts et d'autres sont faibles. Ce qui est plus étonnant c'est que des pics significatifs ne correspondent pas à un pic prédit (par exemple les pics à $\approx 350Bp, \approx 370Bp, \approx 1470Bp$). Les barres d'erreurs, orientées verticalement, correspondent en réalité à la largeur horizontale d'un pic d'hybridation. Une petite barre d'erreur correspond à une hybridation très résolue, alors qu'une barre d'erreur large correspond à un *pic* peu résolu, qui pourrait être le résultat de plusieurs hybridations très proches.

En alignant dans le génome de référence le profil des positions d'hybridations expérimentales sur les positions attendues pour l'oligonucléotide CGTC, nous obtenons une très bonne correspondance entre les nanomètres que nous avons mesurés et les paires de bases du génome. Ceci nous permet d'aller identifier la séquence sous les pics surnuméraires. Par exemple le pic à la position 1470 bp correspond ainsi à une séquence TACG , suggérant que nous observons un mésappariement de la base C terminale de l'oligonucléotide CGTC avec un T sur la séquence de la molécule en épingle à cheveux, et ce avec une affinité suffisante pour que le mésappariement soit observable. L'oligonucléotide qui devrait s'hybrider en face de TACG étant CGTA, nous décidons d'appeler ce mésappariement *CGTA*. Cette convention d'écriture est celle qui sera utilisée dans la suite de la section.

En recherchant les positions de mésappariement CGTA sur l'ensemble de la séquence fig. 6.2, nous constatons que toutes les positions possibles de mésappariement ne se concrétisent pas par un pic d'hybridation. Cependant, nous n'observons pas de biais relatif au brin sur lequel s'observent les mésappariements fig. 6.3. Ces observations nous amènent à conclure que le mésappariement d'un T avec le C-3' de l'oligonucléotide CGTC ne suffit pas seul à expliquer la présence des hybridations surnuméraires, indépendamment du brin sur lequel s'hybride l'oligonucléotide.

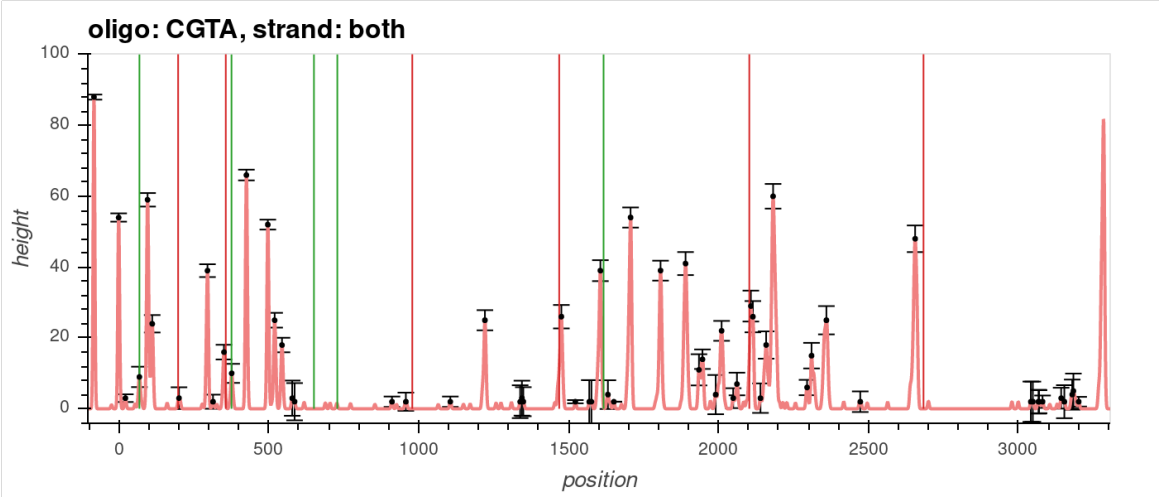


Figure 6.2: En établissant la carte des hybridations de CGTA sur les deux brins (en rouge et vert), nous observons que des positions d'hybridations semblent expliquer la présence d'un *pic* expérimental (comme à la position 1470 bp). Cependant, à plusieurs endroits, et notamment aux abords de 600 bp, et 1 000 bp, nous observons des cibles de CGTA sur le génome de références qui ne se traduisent pas en *pic* expérimental. Le simple mésappariement d'un C par un A ne semble pas être une explication suffisante pour expliquer l'ensemble des positions d'hybridations sur PS825HP.

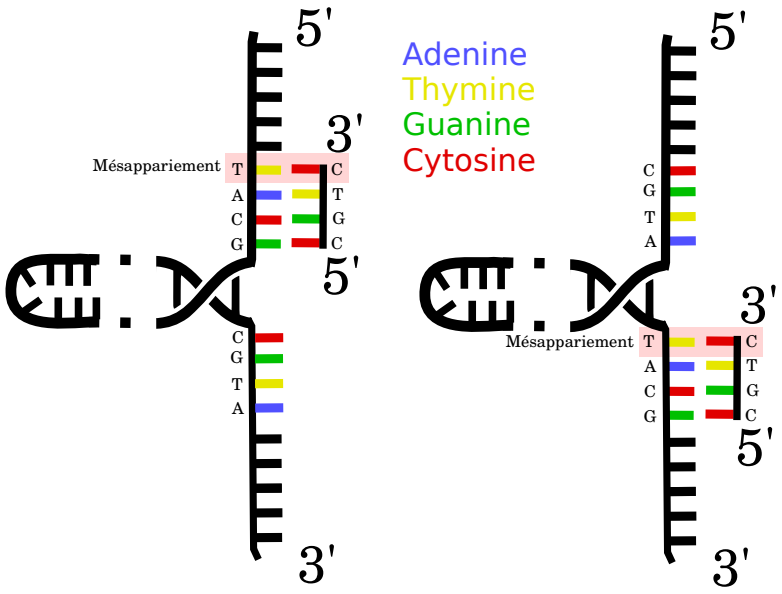


Figure 6.3: Hybridation d'un oligonucléotide de 4 bases CGTC avec un mésappariement d'un côté ou de l'autre de la fourche. En fait la majorité de ces évènements ne provoque pas de blocage détectable. oligo-binding-near-missmatch-CGTA

En observant l'ensemble des sites de mésappariement et en faisant des hypothèses quant aux séquences qui les causent[1] plusieurs motifs candidats ont pu être déterminés. De la même manière que `CGTA` ces

[1]L'hypothèse principale, basée sur des données empiriques antérieures, est que les mésappariements se produisent principalement aux extrémités des oligonucléotides, et que seules les bases de faible taille (A et G) devraient permettre un mésappariement sur une base interne.

motifs ne sont pas suffisants pour expliquer tous les mésappariements.

En observant les différentes possibilités de mésappariement, on constate que les cas où il y a un oligonucléotide mésapparié simultanément sur chaque brin à la fourche expliquent de manière satisfaisante la plupart des hybridations surnuméraires.

Pour bien comprendre ce principe des mésappariements sur les deux brins, nous prenons les deux exemples de mésappariements simultanés qui sont les plus courants pour la molécule PS825HP, à savoir la combinaison de CGTA et CGTT ainsi que CGTA et CGAC.

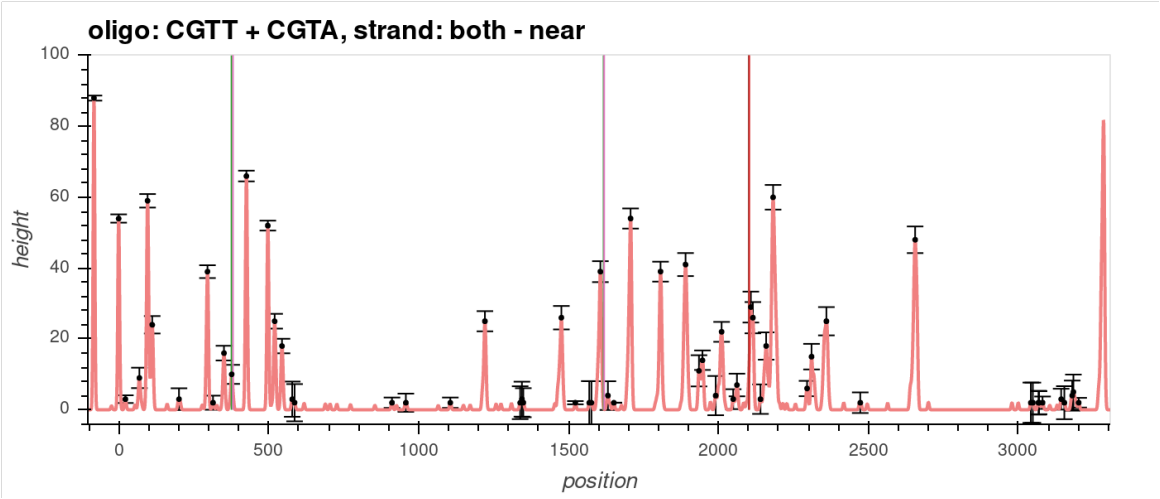


Figure 6.4: Dans cette figure, nous observons les 3 sites théoriques possibles de mésappariements CGTT et CGTA quand ils surviennent à des positions quasi symétriques par rapport à la fourche (lignes verticales). On observe que tous correspondent effectivement à des hybridations surnuméraires.

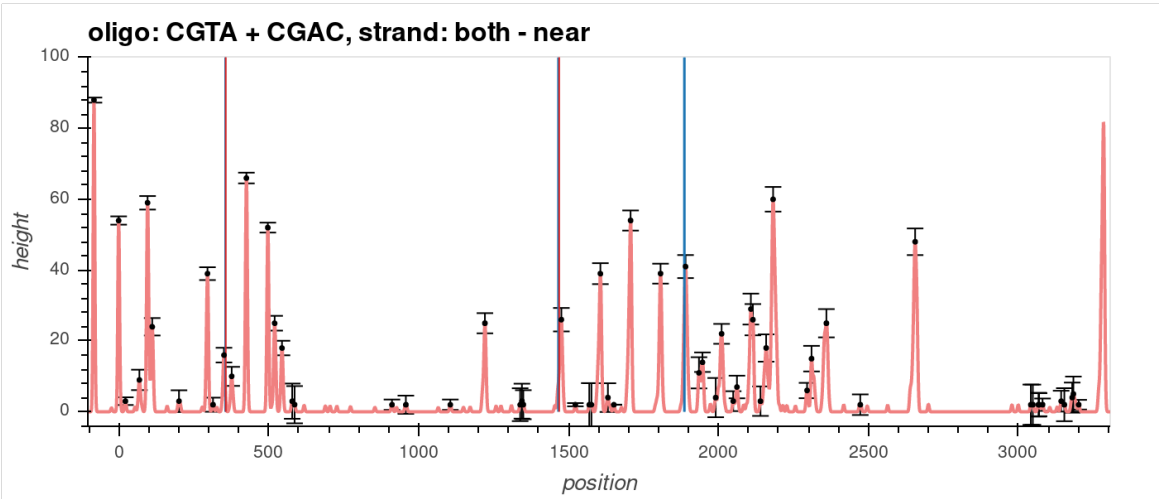


Figure 6.5: De la même façon, nous observons que le couple CGTA et CGAC génèrerait 3 sites possibles de mésappariements quasi symétriques par rapport à la fourche (lignes verticales). Ce couple permet d'expliquer quelques-uns des autres pics surnuméraires, en particulier celui à 350 bp et celui à 1470 bp.

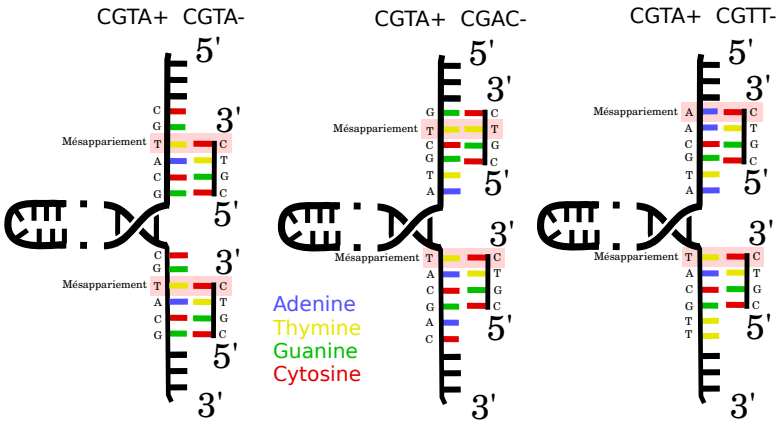


Figure 6.6: Schéma décrivant trois jeux de mésappariements simultanés sur chacun des brins de part et d'autre de la fourche qui provoque systématiquement un blocage visible sous forme d'une hybridation surnuméraire. Ici dans les trois cas la position relative des oligonucléotides par rapport à la fourche présentent un décalage de deux bases.

En utilisant un petit algorithme (cf. ci-dessous), nous avons pu déterminer un ensemble de combinaisons de mésappariements proches sur les deux brins qui expliquent la plupart des hybridations surnuméraires. Basées sur un petit jeu de données expérimentales, nous n'avons pas pu tester toutes ces combinaisons de mésappariements. On ne peut donc pas considérer que ces couples de mésappariements soient exhaustifs.

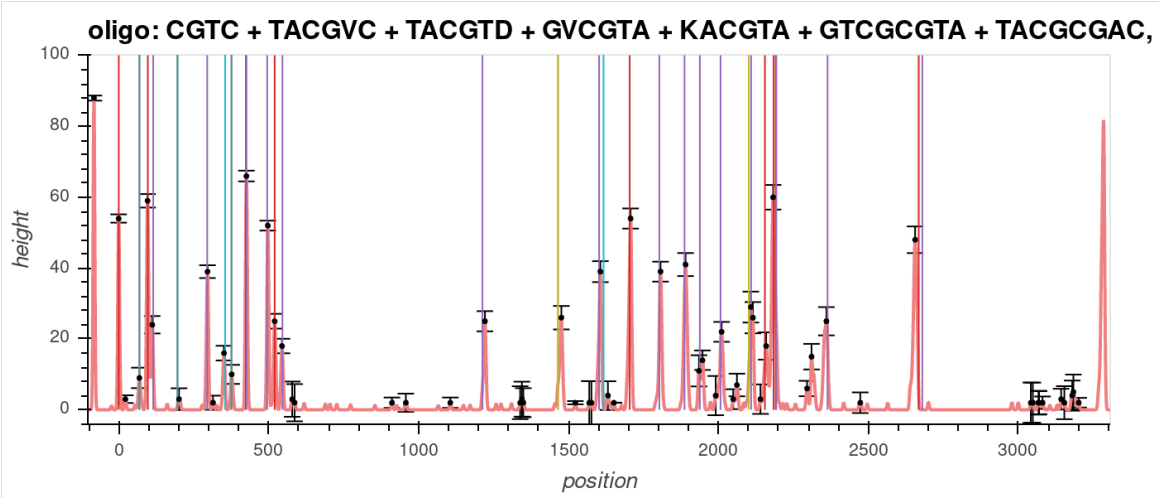


Figure 6.7: Comparaison de l'histogramme expérimental avec une prédiction des positions de blocage correspondant à celles de la séquence exacte CGTC sur les deux brins et aux différentes combinaisons impliquant deux positions d'hybridations voisines sur les deux brins avec les mésappariements que nous avons sélectionnés. Nous voyons que l'accord est bon. Pour exploiter ce résultat, il reste à trouver une méthode automatique de prédiction des mésappariements pertinents. Les mésappariements sont ici représentés condensés, c'est-à-dire que seul le complémentaire de la séquence sur laquelle les deux oligonucléotides sont hybridés est représenté. De plus, des bases étendues sont utilisées (V,D,K,etc.) quand c'est possible.

L'algorithme suivant permet de construire les paires de mésappariements possibles :

1. On groupe toutes les motifs qui peuvent expliquer une position d'hybridation en examinant la séquence de référence aux abords de la position d'hybridation.
2. On calcule le poids de chaque motif selon les critères suivant:
   - On augmente son poids en fonction de la quantité d'hybridation qu'il pourrait potentiellement expliquer (c'est-à-dire que ce motif est à proximité d'une position d'hybridation);
   - On diminue son poids en fonction du nombre d'occurences du motif qui ne sont à proximité d'aucune position d'hybridation constatée.
3. Réduction des groupes : On essaye de trouver le couple de motifs sur les brins + et - ayant le plus haut poids qui explique la position d'hybridation. Ce faisant, on élimine les motifs les moins probables de la liste des motifs qui pourraient expliquer une position d'hybridation.
4. On modifie le poids des différents motifs en fonction du résultat à l'étape 3. Puis on recommence à l'étape 2 jusqu'à stabilité des résultats.

Pour certaines hybridations surnuméraires, malgré la recherche automatique puis manuelle de motifs pouvant être des mésappariements, il n'a pas été possible d'expliquer cette observation par la méthode de mésappariements sur les deux brins. Des recherches supplémentaires seront donc nécessaires pour expliquer ces hybridations.

L'hypothèse de base posée au début des tests expérimentaux et du développement des algorithmes de cartographie était que, bien que des sites d'hybridation théoriques étaient présents dans les molécules étudiées, ils pouvaient ne pas être détectés à cause de la dynamique aléatoire d'hybridation. En pratique, la plupart du temps, nous détectons toutes les hybridations. Le problème principal est que nous détectons également des hybridations supplémentaires telle qu'expliqué dans la partie précédente. La méthode par régression n'est pas particulièrement sensible à ces problèmes d'hybridations surnuméraires, tant que la densité d'hybridation reste relativement faible. Concernant l'algorithme des distances, il manque quelques modifications permettant d'accommoder ces hybridations surnuméraires. Pour pallier ce problème, la stratégie que nous avons adoptée a été de choisir un seuil de détection d'hybridation assez élevé, ce qui revient, en réalité ,à effacer ces pics surnuméraires et à créer des hybridations manquantes pour les pics faibles. Ces paramètres particuliers nous ont permis de retrouver efficacement la séquence, mais en perdant des informations qui peuvent être utiles, par exemple lors qu'une hybridation surnuméraire possède un taux d'hybridation supérieur à une hybridation attendue. L'hybridation attendue, est alors éliminée alors qu'elle est un signal qui devrait effectivement faire parti de la signature.

Ajouter la modification au programme pour y introduire la possibilité de traiter des hybridations surnuméraires devrait avoir un coût en termes de performance, en augmentant l'espace de recherche. Mais cet effet devrait être compensé par une réduction du seuil de détection et, du même coup, cela devrait diminuer le nombre d'hybridations manquantes à tolérer.

On peut toutefois considérer que tenir compte des hybridations surnuméraires n'est qu'un palliatif. La solution idéale serait une caractérisation complète des mésappariements. Pour cela, la bonne stratégie consisterait à identifier les doublets de mésappariements donnant un blocage détectable et incorporer cette donnée lors de la construction des signatures du génome de référence.

Malheureusement, l'étape d'identification de ces doublets n'est pas encore finalisée. Deux stratégies sont envisagées. La première, expérimentale, consisterait à tester les différents oligonucléotides sur des séquences de molécules en épingle à cheveux variées afin de généraliser la démarche adoptée dans la partie précédente et de cataloguer les mésappariements provoquant des hybridations surnuméraires . Elle demande beaucoup d'acquisitions, et doit être effectuée pour tous les oligonucléotides concernés, ce qui demande beaucoup de travail. La seconde, plus théorique, se base sur la prédiction du temps de déplacement de deux oligonucléotides de part et d'autre d'une fourche, en caractérisant la stabilité des

hybridations à partir des valeurs connues des enthalpies et des entropies des dinucléotides impliqués dans ces hybridations. La notion de temps de déplacement est expliqué dans la partie sec. 1.3.3.2.1. Cette stratégie, qui n'est que partiellement abordée, demande quelques données expérimentales afin de valider le modèle. Elle permettrait la généralisation à tous les oligonucléotides, quelles que soient leur composition et leur taille.

À ce stade, le laboratoire a mis au point un programme informatique permettant de prédire ce temps de déplacement pour un seul oligonucléotide d'ADN sur un substrat d'ADN. le code peut prédire ce qui se passe si l'on remplace l'ADN par du LNA, mais en utilisant les valeurs connues des énergies libres des dinucléotides de LNA dont les valeurs, sont malheureusement trop imprécises pour conduire à un accord expérimental. Le laboratoire a entrepris de corriger les données thermodynamiques, mais ce travail n'est réalisé qu'au trois quarts. Il reste à intégrer la correction des énergies libres pour les mésappariements LNA dont les experiences n'ont pas été commencées. Finalement, les programmes ont été modifiés pour inclure la présence d'un oligonucléotide de part et d'autre de la fourche, mais uniquement dans une configuration de symétrie parfaite, il reste à le modifier pour qu'il prenne en compte la contribution de position dissymétrique de deux oligonucléotides sur chacun des brins comme sur la figure fig. 6.6.

Pouvoir déterminer de manière plus précise les doublets de mésappariements devrait augmenter le taux de recouvrement des algorithmes (sensibilité), tout en permettant de réduire l'espace de recherche, et donc le temps de calcul.

6.1.2 Les problèmes que nous n'avons pas réellement abordés

Nous nous sommes concentrés sur les problèmes de cartographie qui correspondent à notre sujet. Cependant les expériences sont également sensibles à des problèmes de traitement du signal en amont de cette cartographie. Nous les citons ici sans chercher à être exhaustifs ni à les corriger parfaitement.

Alors qu'on arrive à détecter des blocages intrinsèques à la structure des molécules, l'élimination de ces parasites n'est pas encore entièrement maîtrisée. Plusieurs types de problèmes ont été identifiés.

Tout d'abord, certaines molécules, lors de leur fermeture, arrêtent de se refermer pendant un certain temps, sans que des oligonucléotides soient responsables de ces blocages. En effet, on observe cet effet même en l'absence d'oligonucléotides en solution. Ces blocages dits « structurels » peuvent être causés par une anomalie dans la structure de la molécule. Par exemple, un dimère de pyrimidine, ou simplement des structures secondaires qui se forment pour une séquence donnée. Alors que la plupart de ces blocages systématiques peuvent être éliminés en observant le comportement d'une molécule en l'absence d'oligonucléotides, cette suppression peut également entraîner le masquage d'une hybridation à cette position.

Des difficultés sont également rencontrées pour aligner les cycles d'ouverture et fermeture entre eux. En effet les dérives à long terme, et l'évolution du comportement des molécules au cours du temps provoquent des décalages entre les cycles et parfois une différence d'étirement. Cet effet est d'autant plus important que l'expérience est longue. D'autres effets sont parfois observés. Aux conditions proche de la force maximum permettant la fermeture des molécules, la molécule peut parfois «sauter» entre deux positions d'hybridations, s'ouvrant et se fermant légèrement entre deux ou trois positions d'hybridation. Ces effets semblent accentués par la présence d'intercalants sur les oligonucléotides.

6.2 Perspectives pour les algorithmes de cartographies

6.2.1 Évaluation des résultats

Les deux méthodes présentées dans la partie 3 permettent de récupérer une liste de solutions candidates triées par leur ajustement. Pour la méthode par régression on les trie au moyen du χ^2, et pour la méthode de classification des segments, le classement est donné par leurs scores.

Ces informations triées donnent la position cartographiée la plus probable étant donné la signature d'entrée et le génome de référence. Cependant, cela ne donne pas d'information sur la probabilité qu'un résultat soit simplement le fruit du hasard. Pour obtenir une information sur le caractère significatif d'une solution, plusieurs possibilités s'offrent à nous.

6.2.1.1 Test de robustesse vis-à-vis du bruit expérimental (rejouer les résultats)

De manière à avoir une information sur la robustesse d'un résultat, une méthode empirique consiste à utiliser les paramètres expérimentaux du système décrit dans la partie sec. 1.3.4, et à modifier l'étirement et à ajouter un peu de bruit en restant dans les limites connues de ces paramètres sur la signature expérimentale d'origine, et à relancer la recherche de cartographie sur cette signature, avec ces paramètres rejoués.

En rejouant les résultats de cette manière un nombre suffisant de fois, et en observant la consistance des résultats des algorithmes, nous obtenons une statistique sur leur robustesse. Ce score donne une indication sur la part de hasard dans la cartographie des résultats à une position donnée.

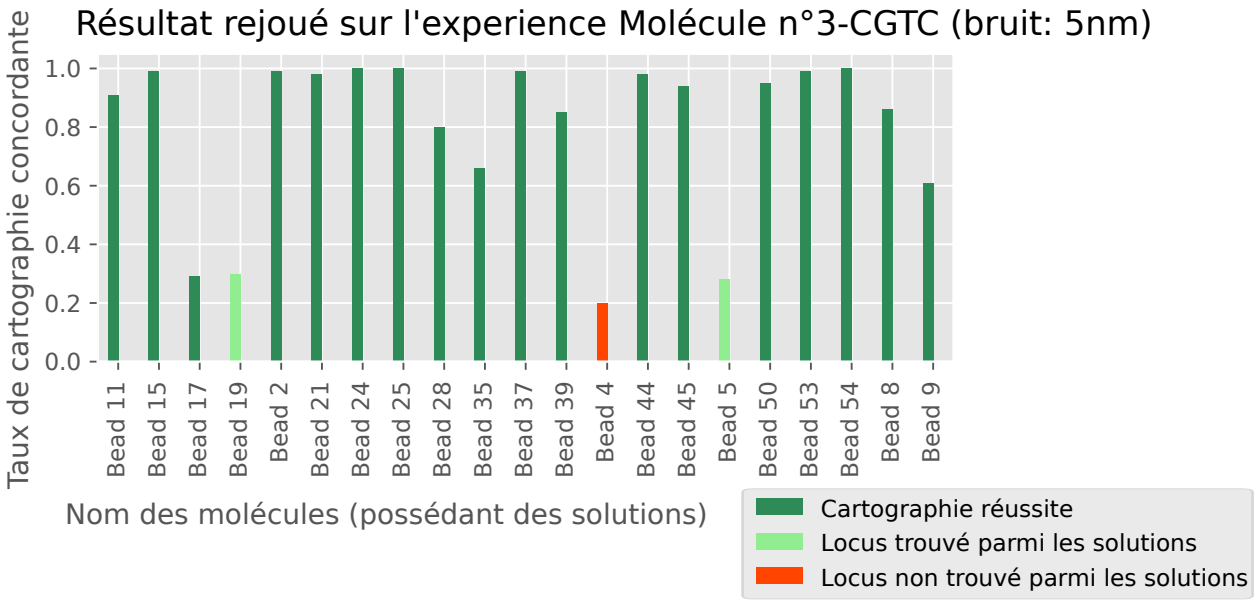


Figure 6.8: Résultats rejoués 100 fois sur l'expérience PS817HP-CGTC: L'unique paramètre rejoué dans les molécules simulées est le bruit. La modification déplace la position des hybridations selon une loi normale d'écart-type de $5nm$

Nous avons par exemple rejoué les résultats pour l'expérience PS817HP-CGTC observée précédemment dans la thèse fig. 6.8. Sur 26 molécules nous ne trouvons aucun résultat possible sur le génome d'*E. coli* pour 5 molécules, elles ne sont pas représentées sur ce graphique.

Sur les 21 Molécules restantes, les *Bead* 4, 19 et 5 n'ont pas été correctement cartographiées.

La corrélation entre le taux de cartographie correcte pour les molécules rejouées et le fait que la signature d'origine ait été cartographiée correctement est très importante. À l'exception de la *Bead 17*, l'ensemble des molécules correctement cartographiées à l'origine ont des résultats rejoués supérieures à 60% de recartographie.

Lors de nos tests expérimentaux, il était possible de s'appuyer sur le score, mais aussi sur la cohérence des résultats entre différentes molécules en épingle à cheveux pour établir la position d'origine d'un ensemble de signatures provenant de molécules identiques. Cependant cela ne correspond pas aux conditions réelles d'utilisation de la méthode cartographie. Il faut pouvoir, à partir d'un mélange de molécules aléatoires différentes provenant d'un même organisme, être capable de cartographier l'ensemble des molécules de manière indépendante. Ce test de robustesse pourrait être une solution pour s'assurer de l'exactitude des résultats de l'algorithme.

Cependant, pour cela il faudrait pousser plus loin ce mécanisme. Il faut pouvoir établir un seuil d'acceptation ou de rejet d'un résultat basé sur la qualité de résultats rejoués.

Pour faire cela de manière correcte, il serait nécessaire de calculer le bruit de fond du génome de référence pour un oligonucléotide, et ainsi, proposer un seuil non arbitraire. Nous pourrions par exemple rejouer l'ensemble du génome sur lui-même et calculer la probabilité avec un intervalle de confiance donnée d'une signature d'être cartographiée à un endroit donné au hasard.

Nous pouvons également raisonnablement proposer un seuil à 50% de situation rejoué cartographié correctement.

6.2.1.2 Validation des résultats avec un deuxième oligonucléotide

En modifiant les paramètres sur une signature expérimentale comme celle décrite ci-dessus, nous avons uniquement une information sur sa robustesse vis-à-vis de l'algorithme de recherche. En croisant les résultats de la recherche de deux signatures d'un locus donnée généré à partir d'oligonucléotides indépendants, nous pouvons obtenir un haut niveau de certitude quant à la vraisemblance de ce résultat si nous trouvons une concordance pour une position donnée et plusieurs oligonucléotides indépendants.

J'ai effectué des expériences sur les mêmes molécules avec plusieurs oligonucléotides, mais n'ai pas eu le temps d'en analyser les résultats.

Depixus a cependant continué ce travail. En faisant des expériences avec 5 oligonucléotides différents sur un ensemble de molécules aléatoires du génome d'*E. coli* et en vérifiant la concordance des résultats entre les différentes signatures pour une même molécule, ils ont pu atteindre un très haut taux de cartographie réussite. Je n'ai malheureusement pas accès aux données quantifiées pour vous les présenter ici.

Cette méthode de validation semble cependant fonctionnelle.

6.3 Des méthodes d'alignement encore insuffisantes pour une utilisation à grande échelle

6.3.1 Performance et complexité

Les deux méthodes de cartographie présentées dans cette thèse sont encore insuffisantes dans le cadre d'une utilisation à grande échelle, et ce pour plusieurs raisons.

Tout d'abord, le temps de calcul des solutions est prohibitif. Une requête par la méthode par régression prend en moyen $300ms$ et par la méthode par classification environ $30ms$ sans la gestion des fusions. Ce temps augmente jusqu'à $2secondes$ lorsque les paramètres de recherche sont les plus larges, et ce pour le génome de $4Mbp$ d'*E. coli*. La complexité des deux algorithmes étant en O($n \times m$), le temps de calcul pour le génome humain deviendrait impraticable.

Depixus a effectué des tests en utilisant la méthode par cartographie sur le génome humain. En utilisant la même stratégie par PCR en aveugle que dans la partie sec. 5. L'acquisition de données expérimentales a été effectuée pour 10 locus du génome humain. La cartographie des signatures pour un locus en aveugle a été effectuée pour une molécule. L'algorithme par classification a été capable de retrouver le locus d'origine pour cette molécule. Démontrant sa capacité à fonctionner pour un génome plus grand.

Cependant le traitement des données a pris 12h pour cette expérience sur un ordinateur de bureau. Ce temps de calcul est impraticable.

À la fin de ma thèse, j'ai développé une nouvelle stratégie basée sur la méthode de classification. Il s'agit de précalculer l'ensemble des solutions possible en se basant sur la matrice des scores, et de stocker ces résultats dans une structure optimisée telle qu'un arbre préfixe (trie) (cela pourrait également être une table de hashage). l'espace occupé par ces structures devrait être de l'ordre $O(a^x)$ a étant le nombre classe du système. et x la profondeur maximum autorisé de l'arbre. Par exemple pour 19 classes et une profondeur d'arbre de 6, l'arbre prendrait au maximum $47Mo$. La recherche serait, elle en $O(mlog(x))$, m étant la taille de la signature. Ce qui correspond à un temps d'exécution extrêmement faible.

Une version intermédiaire de cette stratégie a été implémentée, en prenant pour x (x, la profondeur de l'arbre) le nombre d'hybridations maximum pour une signature d'*E. coli* de 2000Bp. Permettant une recherche en $O(log(m))$, mais avec un espace mémoire beaucoup plus grand ($O(a^m)$). Dans les faits avec ces paramètres non optimisés, le génome d'*Ecoli* occupe un espace mémoire d'environ 500Go, dans une RAM simulée sur un SSD, avec la capacité de reconnaitre des hybridations manquantes.

Les résultats en rouge de la figure ci-dessous sont équivalents à ceux présentés à la figure 4.2. Le temps de recherche était lui d'en moyenne $60ns$.

Étant des résultats préliminaires, ils sont à prendre avec beaucoup de précautions, cependant, ils permettent d'indiquer que la méthode, après optimisation, pourrait être très prometteuse.

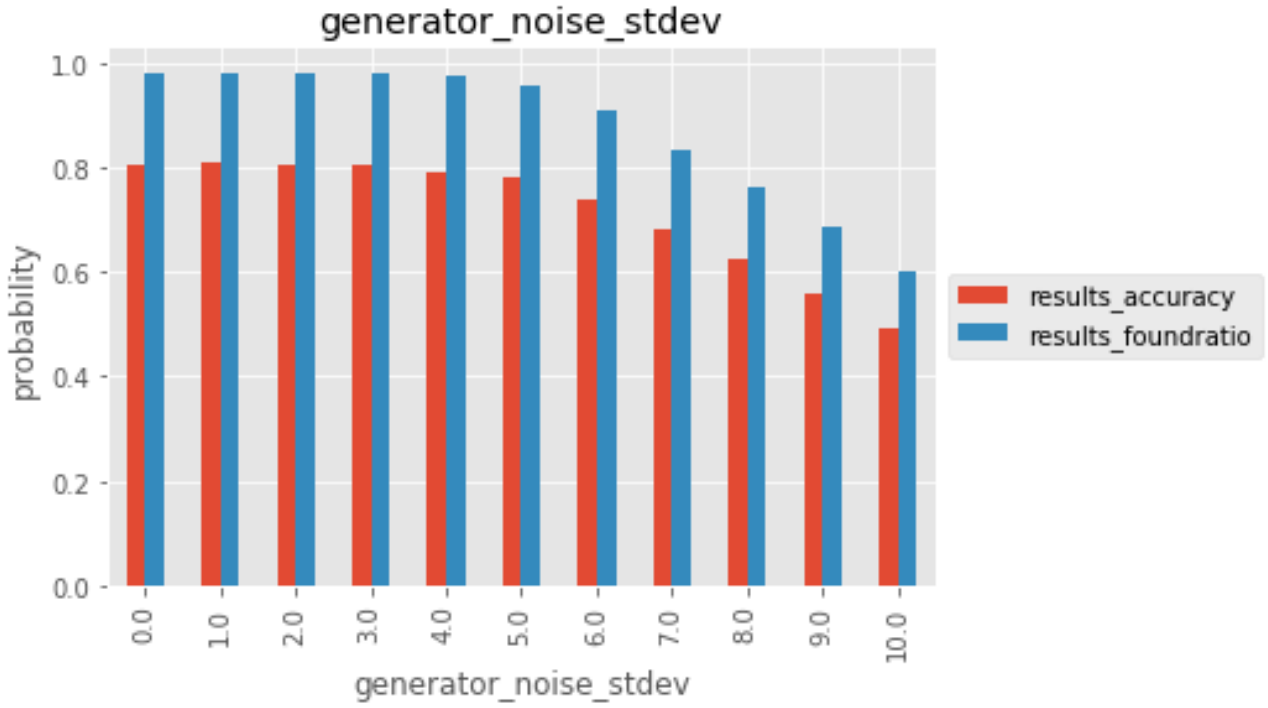


Figure 6.9: Test de la méthode Trie avec $5nm$ de bruit. Ces données ont été générées avant l'optimisation des matrices de score de l'algorithme de classification. Il devrait normalement être amélioré par les améliorations à postériori.

En plus ce cette nouvelle stratégie, de nombreuses optimisations pourraient être pensées pour les deux méthodes.

6.4 Conclusion

Cette thèse a permis de démontrer la faisabilité de la cartographie de molécules sur un génome de référence en utilisant un système de pinces magnétiques et la séquence partielle de molécule (signature) déterminée à partir d'un ou plusieurs oligonucléotides.

Les méthodes de cartographie actuelle semblent applicables sur le Génome d'*E. coli* et nécessiteront des améliorations substantielles pour permettre la cartographie sur des génomes plus grands telle que le génome humain, au moins en termes de temps de calcul.

Des pistes permettant d'atteindre ces objectifs ont été présentées dans cette discussion.

Figures

Tableaux

Chapitre 7

Annexes

- Composition Buffer
- Protocole fabrication cellule.
- Protocole de construction Hairpin PCR
- Sequences des differents hairpins GF1 - GF4 PS015HP, HP-oligoset
- Correspondance en PS8xxHP et locus.

7.1 Choix des paramètres experimentaux la signature d'une molécule en épingle à cheveux

Force à l'ouverture :

- Permet l'ouverture de la plupart des molécules
- Minimise le risque d'arrachement de molécules
- Minimise le bruit de molécule ouverte.

Temps molécule ouverte :

- À la fin de la phase, l'ensemble des molécules doivent être ouvertes.

Rampe de fermeture :

- Doit permettre de maximiser l'hybridation des oligonucléotides, qui semble s'hybrider plus efficacement sur une molécule pour laquelle la tension est moins forte.

Force test :

- Minimise la visiblilité des structures secondaire, et des zones riche en base *Forte* (Cytosine / Guanine)
- Maximise le temps d'hybridation des oligonucléotide, dont le $T_d isp$ est présumé faible du fait de la taille des oligonucléotides
- Limite le temps necessaire à la formation de la fourche, c'est à dire le blocage de la molécule en simple brin, avant le début de sa fermeture.

Temps test :

- Temps suffisant pour que la molécule puisse se refermer en entier malgré les hybridations d'oligonucléotides.

Force de fermeture:

- Permet l'éjection de l'ensemble de oligonucléotides hybridé.
- Évite à la molécule ainsi qu'à la bille paramagnétique de s'approcher trop de la surface, ou elle pourrait rester fixé.

7.1.0.0.1 Masquage des hybridations

Plusieurs paramètres permettent de diminuer l'effet de masquage des hybridations :

- Diminuer la fréquence d'hybridation des oligos : principalement en réduisant leur concentration
- Diminuer le temps d'hybridation
  - Diminuer la force température du système
  - Diminuer la force appliquée à la molécule en épingle à cheveux lors de la fermeture, pour de donner plus de *force* à la fourche
  - Changer la composition des oligonucléotides ou leur taille
  - Modifier la quantité de sels présents dans le buffer, de manière à neutraliser l'ADN et réduire l'énergie de la liaison hydrogène liant les paires de bases.
- Augmenter le nombre de cycles d'ouverture / fermeture de manière

7.2 Protocole de construction des molécule en épingle à cheveux

7.2.1 Parcours de l'ensemble des signatures

Le principe de l'agorithme de parcours de l'espace des positions possibles est relativement simple et peut-être représenté de la manière suivante.

```
pending = []
nextpending = []
accepted = []
# Pour tout les segments du genome de référence
for rsegment in refSegments:
    # On ajoute la locus actuel comme une solution candidate
    pending += newCandidate(locus(rsegment))

    # On traite chaque solution candidat
    for candidate in pending:
        # On rejete les candidats ne respectant pas les critères de
        # * score minimale
        # * nombre de fusion maximum
        # * dont la taille n'est pas compatible avec le fragment experimentale
     if !candidateRejected?(candidate):
            # Si il est possible de crée une nouvelle fusion
            # et qu'elle offre un score supérieure au candidat actuelle
            if fusionAccepted?(candidate):
                # On crée un nouveau candidat avec fusion en cours
                fusionCandidate = newFusionCandidate(candidate, rsegment)
             # On ajoute le nouveau candidat à la nouvelle liste des candidats
             nextPending += fusionCandidate
        # Dans tout les cas, on crée un nouveau candidat avec le segment courant,
        # en fermant la fusion courante
```

```
        candidate = newCandidate(candidate, rsegment)

        # Si le candidat a une longueurs, un score suffisant,
        # il est accepté comme solution, sinon,
        # il retourne dans la liste de candidats
        if candidateAccepted?(candidate):
            accepted += candidate
        else:
            nextPending += candidate

        # On met à jour la liste des solutions candidate, et on recommence
        pending = nextPending
        nextPending = []

# On tri les solution par leur score
sort(accepted)
```

7.3 Propriété intelectuelle des ressources

- Les images de hairpin utilisent https://www.onlinewebfonts.com/icon/493802 sous CC-BY, Federico Sibella

7.4 Acquisition / Prétraitement des données

De manière à rendre des données expérimentales exploitables en temps que signature, des méthodes ont été développées au sein du laboratoire et de l'entreprise.

- Élimination des dérives thermique cyclique et long terme.
- Découpe par cycles
- Alignement des données
- Limite des méthodes
  - Bruit additionnel induit par le mésalignement des cycles
  - Bruit lorsque la bille s'approche de la surface
  - Position dédoublée / fusionnée

Chapitre 8

Références

Alberts, Bruce, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, et Peter Walter. 2002. *Molecular Biology of the Cell*. 4th éd. Garland Science.

Altschul, S. F. 1991. « Amino Acid Substitution Matrices from an Information Theoretic Perspective ». *Journal of Molecular Biology* 219 (3): 555-65.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, et David J. Lipman. 1990. « Basic Local Alignment Search Tool ». *Journal of Molecular Biology* 215 (3): 403-10. https://doi.org/10.1016/S0022-2836(05)80360-2.

Anton, Brian P., et Elisabeth A. Raleigh. 2016. « Complete Genome Sequence of NEB 5-Alpha, a Derivative of Escherichia Coli K-12 DH5 ». *Genome Announcements* 4 (6). https://doi.org/10.1128/genomeA.01245-16.

Burrows, M., et D. J. Wheeler. 1994. « A Block-Sorting Lossless Data Compression Algorithm ».

Bustamante, Carlos, Zev Bryant, et Steven B. Smith. 2003. « Ten Years of Tension: Single-Molecule DNA Mechanics ». *Nature* 421 (6921): 423. https://doi.org/10.1038/nature01405.

Cherf, Gerald M., Kate R. Lieberman, Hytham Rashid, Christopher E. Lam, Kevin Karplus, et Mark Akeson. 2012. « Automated Forward and Reverse Ratcheting of DNA in a Nanopore at 5-Å Precision ». *Nature Biotechnology* 30 (4): 344-48. https://doi.org/10.1038/nbt.2147.

Clark, Matthew D., Steffen Hennig, Ralf Herwig, Sandy W. Clifton, Marco A. Marra, Hans Lehrach, Stephen L. Johnson, et the WU-GSC EST Group. 2001. « An Oligonucleotide Fingerprint Normalized and Expressed Sequence Tag Characterized Zebrafish cDNA Library ». *Genome Research* 11 (9): 1594-1602. https://doi.org/10.1101/gr.186901.

Clarke, James, Hai-Chen Wu, Lakmal Jayasinghe, Alpesh Patel, Stuart Reid, et Hagan Bayley. 2009. « Continuous Base Identification for Single-Molecule Nanopore DNA Sequencing ». *Nature Nanotechnology* 4 (4): 265. https://doi.org/10.1038/nnano.2009.12.

Consortium, International Human Genome Sequencing. 2001. « Initial Sequencing and Analysis of the Human Genome ». *Nature* 409 (6822): 860-921. https://doi.org/10.1038/35057062.

Damerau, Fred J. 1964. « A Technique for Computer Detection and Correction of Spelling Errors ». *Commun. ACM* 7 (3): 171-76. https://doi.org/10.1145/363958.363994.

Dembo, Amir, et Samuel Karlin. 1993. « Central Limit Theorems of Partial Sums for Large Segmental Values ». *Stochastic Processes and their Applications* 45 (2): 259-71. https://doi.org/10.1016/0304-4149(93)90073-D.

Ding, Fangyuan, Maria Manosas, Michelle M. Spiering, Stephen J. Benkovic, David Bensimon, Jean-François Allemand, et Vincent Croquette. 2012. « Single-Molecule Mechanical Identification and Sequencing ». *Nature Methods* 9 (4): 367-72. https://doi.org/10.1038/nmeth.1925.

Dramanac, Radoje, Ivan Labat, Ivan Brukner, et Radomir Crkvenjakov. 1989. « Sequencing of Megabase plus DNA by Hybridization: Theory of the Method ». *Genomics* 4 (2): 114-28. https://doi.org/10.1016/0888-7543(89)90290-5.

Eid, John, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, et al. 2009. « Real-Time DNA Sequencing from Single Polymerase Molecules ». *Science* 323 (5910): 133-38. https://doi.org/10.1126/science.1162986.

Ferragina, Paolo, et Giovanni Manzini. 2001. « An Experimental Study of an Opportunistic Index ». In *Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms*, 269-78. SODA '01. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics. http://dl.acm.org/citation.cfm?id=365411.365458.

Ferragina, P., et G. Manzini. 2000. « Opportunistic Data Structures with Applications ». In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, 390. FOCS '00. Washington, DC, USA: IEEE Computer Society. http://dl.acm.org/citation.cfm?id=795666.796543.

Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, et J. M. Merrick. 1995. « Whole-Genome Random Sequencing and Assembly of Haemophilus Influenzae Rd ». *Science (New York, N.Y.)* 269 (5223): 496-512.

Gallier, Jean. 2008. *Discrete Mathematics for Computer Science, Some Notes*. https://doi.org/10.1007/978-1-4419-8047-2.

Germann, Ulrich, Eric Joanis, et Samuel Larkin. 2009. « Tightly Packed Tries: How to Fit Large Models into Memory, and Make Them Load Fast, Too ». In *Proceedings of the Workshop on Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, 31-39. SETQA-NLP '09. Stroudsburg, PA, USA: Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1621947.1621952.

Gosse, Charlie, et Vincent Croquette. 2002. « Magnetic Tweezers: Micromanipulation and Force Measurement at the Molecular Level ». *Biophysical Journal* 82 (6): 3314-29. https://doi.org/10.1016/S0006-3495(02)75672-5.

Greenleaf, William J., et Steven M. Block. 2006. « Single-Molecule, Motion-Based DNA Sequencing Using RNA Polymerase ». *Science (New York, N.Y.)* 313 (5788): 801. https://doi.org/10.1126/science.1130105.

Harris, Timothy D., Phillip R. Buzby, Hazen Babcock, Eric Beer, Jayson Bowers, Ido Braslavsky, Marie Causey, et al. 2008. « Single-Molecule DNA Sequencing of a Viral Genome ». *Science (New York, N.Y.)* 320 (5872): 106-9. https://doi.org/10.1126/science.1150427.

Hodeib, Samar, Saurabh Raj, Maria Manosas, Weiting Zhang, Debjani Bagchi, Bertrand Ducos, Francesca Fiorini, et al. 2017. « A Mechanistic Study of Helicases with Magnetic Traps ». *Protein Science* 26 (7): 1314-36. https://doi.org/10.1002/pro.3187.

Holste, D., I. Grosse, et H. Herzel. 2001. « Statistical Analysis of the DNA Sequence of Human Chromosome 22 ». *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics* 64 (4 Pt 1): 041917. https://doi.org/10.1103/PhysRevE.64.041917.

Karlin, S, et S F Altschul. 1990. « Methods for Assessing the Statistical Significance of Molecular Sequence Features by Using General Scoring Schemes. » *Proceedings of the National Academy of Sciences of the United States of America* 87 (6): 2264-8. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC53667/.

Karlin, Samuel, Amir Dembo, et Tsutomu Kawabata. 1990. « Statistical Composition of High-Scoring Segments from Molecular Sequences ». *The Annals of Statistics* 18 (2): 571-81. https://doi.org/10.1214/aos/1176347616.

Kasianowicz, John J., Eric Brandin, Daniel Branton, et David W. Deamer. 1996. « Characterization of Individual Polynucleotide Molecules Using a Membrane Channel ». *Proceedings of the National Academy of Sciences* 93 (24): 13770-3. https://doi.org/10.1073/pnas.93.24.13770.

Kent, W. James. 2002. « BLAT—the BLAST-Like Alignment Tool ». *Genome Research* 12 (4): 656-64. https://doi.org/10.1101/gr.229202.

Khan, Akbar S. 2014. « Rapid Advances in Nucleic Acid Technologies for Detection and Diagnostics of Pathogens ». *Journal of Microbiology & Experimentation* 1 (2): 1-7. https://doi.org/10.15406/jmen.2014.01.00009.

Koshkin, Alexei A., Poul Nielsen, Michael Meldgaard, Vivek K. Rajwanshi, Sanjay K. Singh, et Jesper Wengel. 1998. « LNA (Locked Nucleic Acid): An RNA Mimic Forming Exceedingly Stable LNA:LNA Duplexes ». *Journal of the American Chemical Society* 120 (50): 13252-3. https://doi.org/10.1021/ja9822862.

Koster, Daniel A., Vincent Croquette, Cees Dekker, Stewart Shuman, et Nynke H. Dekker. 2005. « Friction and Torque Govern the Relaxation of DNA Supercoils by Eukaryotic Topoisomerase IB ». *Nature* 434 (7033): 671-74. https://doi.org/10.1038/nature03395.

Langmead, Ben, Cole Trapnell, Mihai Pop, et Steven L. Salzberg. 2009. « Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome ». *Genome Biology* 10 (mars): R25. https://doi.org/10.1186/gb-2009-10-3-r25.

Levene, M. J., J. Korlach, S. W. Turner, M. Foquet, H. G. Craighead, et W. W. Webb. 2003. « Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations ». *Science (New York, N.Y.)* 299 (5607): 682-86. https://doi.org/10.1126/science.1079700.

Levenshtein, V. I. 1966. « Binary Codes Capable of Correcting Deletions, Insertions and Reversals ». *Soviet Physics Doklady* 10 (février): 707. http://adsabs.harvard.edu/abs/1966SPhD...10..707L.

Lieberman, Kate R., Gerald M. Cherf, Michael J. Doody, Felix Olasagasti, Yvette Kolodji, et Mark Akeson. 2010. « Processive Replication of Single DNA Molecules in a Nanopore Catalyzed by Phi29 DNA Polymerase ». *Journal of the American Chemical Society* 132 (50): 17961-72. https://doi.org/10.1021/ja1087612.

Liu, Zhandong, Santosh S Venkatesh, et Carlo C Maley. 2008. « Sequence Space Coverage, Entropy of Genomes and the Potential to Detect Non-Human DNA in Human Samples ». *BMC Genomics* 9 (octobre): 509. https://doi.org/10.1186/1471-2164-9-509.

Manosas, Maria, Senthil K. Perumal, Vincent Croquette, et Stephen J. Benkovic. 2012. « Direct Observation of Stalled Fork Restart via Fork Regression in the T4 Replication System ». *Science* 338 (6111): 1217-20. https://doi.org/10.1126/science.1225437.

Markham, Nicholas R., et Michael Zuker. 2008. « UNAFold: Software for Nucleic Acid Folding and Hybridization ». *Methods in Molecular Biology (Clifton, N.J.)* 453: 3-31. https://doi.org/10.1007/978-1-60327-429-6_1.

Moerner, W. E., et L. Kador. 1989. « Optical Detection and Spectroscopy of Single Molecules in a Solid ». *Physical Review Letters* 62 (21): 2535-8. https://doi.org/10.1103/PhysRevLett.62.2535.

NCBI. 2017. « RefSeq NCBI Reference Sequence Database ». 2017. https://www.ncbi.nlm.nih.gov/refseq/.

Neely, Robert K., Jochem Deen, et Johan Hofkens. 2011. « Optical Mapping of DNA: Single-Molecule-Based Methods for Mapping Genomes ». *Biopolymers* 95 (5): 298-311. https://doi.org/10.1002/bip.21579.

Nelson, William, et Carol Soderlund. 2009. « Integrating Sequence with FPC Fingerprint Maps ». *Nucleic Acids Research* 37 (5): e36-e36. https://doi.org/10.1093/nar/gkp034.

Neuman, K. C., G. Charvin, D. Bensimon, et V. Croquette. 2009. « Mechanisms of Chiral Discrimination by Topoisomerase IV ». *Proceedings of the National Academy of Sciences* 106 (17): 6986-91. https://doi.org/10.1073/pnas.0900574106.

Nilsson, Adam N., Gustav Emilsson, Lena K. Nyberg, Charleston Noble, Liselott Svensson Stadler, Joachim Fritzsche, Edward R. B. Moore, Jonas O. Tegenfeldt, Tobias Ambjörnsson, et Fredrik Westerlund. 2014. « Competitive Binding-Based Optical DNA Mapping for Fast Identification of Bacteria - Multi-Ligand Transfer Matrix Theory and Experimental Applications on Escherichia Coli ». *Nucleic Acids Research* 42 (15): e118-e118. https://doi.org/10.1093/nar/gku556.

Pihlak, Arno, Göran Baurén, Ellef Hersoug, Peter Lönnerberg, Ats Metsis, et Sten Linnarsson. 2008. « Rapid Genome Sequencing with Short Universal Tiling Probes ». *Nature Biotechnology* 26 (6): 676-84. https://doi.org/10.1038/nbt1405.

Pollard, Tom, Marvin Reimer, David San, Arco Mul, Matthew Gwynfryn Thomas, Jakub Nowosad, Dennis Weissmann, et W. Caleb McDaniel. 2016. « Template for Writing a PhD Thesis in Markdown ». Zenodo. https://doi.org/10.5281/zenodo.58490.

Sanger, F., S. Nicklen, et A. R. Coulson. 1977. « DNA Sequencing with Chain-Terminating Inhibitors ». *Proceedings of the National Academy of Sciences of the United States of America* 74 (12): 5463-7.

SantaLucia, J. 1998. « A Unified View of Polymer, Dumbbell, and Oligonucleotide DNA Nearest-Neighbor Thermodynamics ». *Proceedings of the National Academy of Sciences of the United States of America* 95 (4): 1460-5.

Slater, G. S., et E. Birney. 2005. « Automated Generation of Heuristics for Biological Sequence Comparison. » *BMC Bioinformatics* 6: 31-31. https://doi.org/10.1186/1471-2105-6-31.

Smith, S. B., L. Finzi, et C. Bustamante. 1992. « Direct Mechanical Measurements of the Elasticity of Single DNA Molecules by Using Magnetic Beads ». *Science* 258 (5085): 1122-6. https://doi.org/10.1126/science.1439819.

Smith, T. F., et M. S. Waterman. 1981a. « Identification of Common Molecular Subsequences ». *Journal of Molecular Biology* 147 (1): 195-97.

———. 1981b. « Identification of Common Molecular Subsequences ». *Journal of Molecular Biology* 147 (1): 195-97. https://doi.org/10.1016/0022-2836(81)90087-5.

Soderlund, C., I. Longden, et R. Mott. 1997. « FPC: A System for Building Contigs from Restriction Fingerprinted Clones ». *Bioinformatics* 13 (5): 523-35. https://doi.org/10.1093/bioinformatics/13.5.523.

Stoddart, David, Andrew J. Heron, Ellina Mikhailova, Giovanni Maglia, et Hagan Bayley. 2009. « Single-Nucleotide Discrimination in Immobilized DNA Oligonucleotides with a Biological Nanopore ». *Proceedings of the National Academy of Sciences* 106 (19): 7702-7. https://doi.org/10.1073/pnas.0901054106.

Strick, T. R., J.-F. Allemand, D. Bensimon, A. Bensimon, et V. Croquette. 1996. « The Elasticity of a Single Supercoiled DNA Molecule ». *Science* 271 (5257): 1835-7. https://doi.org/10.1126/science.271.5257.1835.

« The Long View on Sequencing ». 2018. *Nature Biotechnology* 36 (4): 287. https://doi.org/10.1038/nbt.4125.

Turner, Douglas H. 1996. « Thermodynamics of Base Pairing ». *Current Opinion in Structural Biology* 6 (3): 299-304. https://doi.org/10.1016/S0959-440X(96)80047-9.

Vlaminck, Iwijn De, et Cees Dekker. 2012. « Recent Advances in Magnetic Tweezers ». *Annual Review of Biophysics* 41 (1): 453-72. https://doi.org/10.1146/annurev-biophys-122311-100544.

Watson, J. D., et F. H. C. Crick. 1953. « Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid ». *Nature* 171 (4356): 737-38. https://doi.org/10.1038/171737a0.

You, Yong, Bernardo G. Moreira, Mark A. Behlke, et Richard Owczarzy. 2006. « Design of LNA Probes That Improve Mismatch Discrimination ». *Nucleic Acids Research* 34 (8): e60-e60. https://doi.org/10.1093/nar/gkl175.

Résumé

Les techniques de micromanipulation de molécules d'ADN uniques offrent des perspectives nouvelles pour lire et exploiter l'information contenue dans les génomes. Cela inclut le séquençage, la cartographie, le dénombrement de molécules et l'identification de modifications chimiques de l'ADN. Dans ce contexte, l'Équipe ABCDLab de l'ENS a développé une méthode utilisant l'ouverture et la fermeture mécanique répétée d'une molécule d'ADN en épingle à cheveux par pince magnétique. Cet outil permet de déterminer la position d'hybridation de petits oligonucléotides ainsi que celle d'anticorps révélant la position de marques épigénétiques. Un avantage de cette approche est de pouvoir travailler sur la même molécule pour d'une part identifier les marques épigénétiques et d'autre part réaliser une cartographie de sa position dans le génome.

Mon travail de thèse consiste à développer un ensemble de méthodes bio-informatique visant à réaliser cette étape de cartographie. Le signal expérimental consiste en lecture des positions d'hybridation d'un, ou de plusieurs petits oligonucléotides sur la molécule étudiée. Cette mesure permet de construire une signature spécifique de la molécule que l'on peut rechercher dans le génome d'origine.

Dans ce travail de thèse, j'ai réalisé des expériences avec sur pinces magnétiques pour acquérir des signatures moléculaires sur des molécules sélectionnées en aveugle dans E. coli. J'ai développé un logiciel capable de faire la recherche de ces signatures dans un génome et ensuite effectué l'ensemble du traitement des données pour tester le logiciel. Après plusieurs étapes d'optimisation, j'ai pu retrouver la position génomique des molécules étudiées, établissant ainsi une preuve de concept de cette stratégie de cartographie.

Le travail a concerné l'ensemble de la chaîne de mesure : (1) le choix des sondes utilisées pour constituer la signature d'une molécule observée en optimisant un ensemble de critères liés aux conditions expérimentales et à la combinatoire des motifs de séquence. (2) la mise au point d'algorithmes de cartographie adaptés aux caractéristiques expérimentales des mesures. Enfin, j'ai testé ces algorithmes, à la fois sur des données simulées in silico et in vitro sur de l'ADN d'origine bactérienne.

Je discuterais en quoi les performances des solutions de cartographie développées ici sont influencées par, d'une part les limites du montage expérimental actuel, et d'autre part les limites des approches bio-informatiques. Je présenterais les voies d'amélioration possibles de ces dernières. Mes travaux établissent qu'identifier des molécules d'ADN uniques par pinces magnétiques est possible dans le contexte d'application épigénétique et en génomique.

Mots Clés

Sequençage, Empreinte Pinces Magnétiques Methylation SIMDEQ

Abstract

Single molecule micromanipulations technic offer new perspectives to read and unravel genome information. This includes sequencing, mapping, molecule counting and identification of DNA modifications. In this respect, ABCDLab team has developed a cutting edge method using repeated mechanical opening and closing of a DNA molecule with a hairpin shape using magnetic tweezers. This tool allows measuring along the DNA molecule the hybridization positions of oligonucleotides a few bases long and also to locate specific antibodies transiently bound to epigenetic markers. With this approach we can identify with the single molecule level epigenetics markers and localized them on the genome.

My PhD work consisted of developing a set of bioinformatics methods to perform DNA mapping using magnetics tweezer signal consisting of hybridization positions along the studied molecule. This measurement may be viewed as a fingerprint of the molecule which can be searched on the reference genome.

During my thesis, I have realized an experimental test using magnetic tweezers to acquire a set fingerprint data on a set of blinded selected molecules in the E. coli genome. I have developed a software performing a rapid search of these fingerprints inside the genome. Then I have performed the whole data treatment to check the software on the selected molecules. After several rounds of optimization, I have recovered the genetic position of the studied molecules, establishing a proof of concept of this cartography strategy.

The work has addressed the whole measuring chain; (1) by choosing the oligonucleotides best adapted to obtain the molecular signature by optimizing the set of experimental constrains and combinatorial motifs of the sequence. (2) by tuning the cartography algorithm to adapt to the experimental measurement constrains. Finally, I have tested these algorithms, both on simulated data in silico and on experimental fingerprint in vitro.

I shall discuss how the performances of these cartography solutions that have been developed here are impacted by the experimental limitations of the present technique, and by the bioinformatics limits. I shall present possible improvements to these methods. My studies constitute a proof of concept for genomic and epigenetic applications.

Keywords

Sequencing, Fingerprint, Magnetic Tweezers Methylation SIMDEQ