# Modèles statistiques du développement de tumeurs cancéreuses

Mathieu Emily

*INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE*

*N° attribué par la bibliothèque*

|_|_|_|_|_|_|_|_|_|_|

***THÈSE***

pour obtenir le grade de **DOCTEUR DE L'INPG**

**Spécialité : « Mathématiques Appliquées »**

préparée au laboratoire Techniques de l'Imagerie, de la Modélisation et de la Cognition (TIMC) dans le cadre de **l'École Doctorale « Mathématiques, Sciences et Technologies de l'Information, Informatique »**

préparée et soutenue publiquement par

Mathieu EMILY

le 22 Septembre 2006

***Titre :***

# Modèles statistiques du développement de tumeurs cancéreuses

---

***sous la direction de Olivier François***

---

**JURY**

| | |
|---|---|
| Jacques Demongeot | Président |
| Avner Bar-hen | Rapporteur |
| Carsten Wiuf | Rapporteur |
| Florence Forbes | Examinateur |
| Olivier François | Directeur de thèse |
| Jean-Michel Billiot | Co-encadrant |

A la mémoire d'Etienne Bertin

## Remerciements

Je tiens en tout premier lieu à remercier le professeur Jacques Demongeot qui m'a fait l'honneur de présider mon jury. Je souhaite également lui faire part de toute ma reconnaissance de m'avoir accueilli dans son laboratoire et d'avoir toujours suivi du coin de l'œil mes travaux.

Un grand merci à mes deux rapporteurs, le professeur Avner Bar-Hen et le professeur Carsten Wiuf, pour leurs lectures minitieuses de mon manuscrit de thèse. J'ai tout particulièrement apprécié leurs présences à ma soutenance de thèse malgré les kilomètres qui séparent Grenoble de leurs universités. Leurs questions et remarques constituent pour moi une source d'inspiration intarissable.

Je souhaite témoigner toute ma gratitude à Florence Forbes pour sa participation, en tant qu'examinatrice, à ma soutenance. Ses questions pertinentes sur la partie spatiale ont apporté un œil nouveau sur mes travaux.

A l'origine de ce projet, Etienne Bertin aura malgré tout été à mes côtés pendant les trois années de mon doctorat. Je profite de l'occasion qui m'est donnée pour lui dédier l'ensemble de ces travaux. En quatre de mois de stage de DEA, il m'aura donné l'envie de devenir chercheur en mathématiques appliquées à la biologie et aura marqué un tournant profond dans ma vie professionnelle.

Ses deux complices dans l'étude de processus ponctuels de Gibbs sur graphes aléatoires m'ont permis de rester fidèle à l'esprit initial de cette thèse. Je vous remercie donc profondément, Jean-Michel et Rémy, les complémentaires du trio BBD, pour m'avoir aiguillé dans ce monde fait de marques et de points en reprenant une partie du flambleau. J'ai grandement apprécié l'ensemble des discussions qui ont façonnées le modèle spatial et je suis extrèmement reconnaissant de tout ce que vous avez réalisé pour l'aboutissement de cette thèse.

L'autre partie du flambeau fut portée par Olivier. Son investissement à 200%, son sérieux, son perfectionnisme, son efficacité, sa curiosité scientifique, sa passion pour la recherche, son pragmatisme, son dynamisme, son calme, sa patience, sa disponiblité, son anglais, son goût immodéré pour le café (noir ou rouge !), etc. sont autant d'atouts qui m'ont

aidé à mener à bien cette thèse. Il m'a « formé » en me transmettant toute son expérience académique et je lui suis infinimement reconnaissant pour son encadrement. Merci également de m'avoir permis de voyager pour présenter mes travaux et par la même de découvrir d'autres lieux de recherche.

Une thèse est également le fruit de discussions et d'échanges au sein d'une équipe de recherche. J'ai eu la chance de faire partie de l'équipe TIMB du laboratoire TIMC, où la bonne humeur et le dynamisme m'ont « forcé » à investir le bureau B40 ! Je tiens à remercier les membres permanents, dont le nombre a augmenté exponentiellement au cours des dernières années. Un merci tout particulier à Jean-Louis pour sa gestion de l'équipe et à Hervé tant pour son regard éclairé de probabiliste que pour être amateur de sport.

Mille remerciements à tous les intermittents qui sont passés par le TIMB pendant ces trois ans : Nicolas mon premier stagiaire, Eric qui prend la relève, Nicolas pour avoir bombardé les machines de calculs, Nicolas le chercheur bénévole pour son imagination débordante, Julien pour son indécision devenue légendaire à Grenoble, Michael pour ses histoires toutes plus invraissemblables les unes que les autres, et enfin mes trois co-bureaux Loïc, Adrien et David qui ont réussi à me supporter tant bien que mal. Loïc, mon oncle en recherche que je considère plus comme un grand frère de recherche, a su me montrer la voie pour appréhender au mieux ce métier de thésard. Sa science footbalistique, ainsi que ses achats réguliers de café, furent des atouts majeurs dans la vie de l'équipe. Adrien, le roi de la compo, aura bien essayé de m'apprendre le ski hors-piste mais c'est moi qui l'ai perverti à la gestion d'équipes. En tant que « type populaire » il a beaucoup influé sur « l'esprit d'équipe ». David, le descendant caché de Markov, après m'avoir appris la recherche d'apparts puis d'emplois a réussi à me détourner de ma rédaction de thèse en organisant des video projections de matchs de la coupe du monde. Cette dernière année de thèse, souvent décrite pour être la plus compliquée, aura été pour ma part très agréable en partie grâce à lui.

Je tiens à remercier l'ensemble des membres de l'équipe GMCAO ainsi que tous ceux qui ont participé aux différents tournois de foot inter-labo. Je souhaite également remercier le personnel admistratif et plus particulièrement Céline, Corinne et Kim pour leurs aides

précieuses. Que les membres des réunions dites « 8 à 9 » du vendredi matin reçoivent toute ma considération pour m'avoir fait découvrir de multiples rouages du monde de la recherche.

Bien que la thèse soit une activité de recherche, celle-ci n'aurait peut-être pas aboutie si je n'avais pas découvert en parallèle l'enseignement supérieur. Merci aux membres de l'équipe pédagogique de mathématiques du CPP de Grenoble pour m'avoir acceuilli. Un merci tout particulier à Claude pour m'avoir fait confiance en m'offrant la possibilité de gérer un module de manière autonome et pour m'avoir fait l'honneur de sa présence à ma soutenance de thèse. Merci également à l'équipe pédagogique de proba-stats de l'ENSIMAG et notamment Jean-Baptiste avec qui j'ai été très heureux d'enseigner.

Le monitorat, grâce aux formations, m'aura permis de rester en contact avec de nombreuses personnes. Je tiens donc à remercier ceux qui ont pris part à cette expérience unique qu'à constituer le stage de deuxième année. Un grand merci également à Eric, Laurent, Jérôme et Emmanuel ainsi que Sylvain, notre tuteur, pour tous les bons moments associés au projet de troisième année.

Je tiens à remercier les amis qui m'ont soutenu pendant cette thèse. Un merci tout particulier à Guillaume et Julia pour m'avoir souvent hébergé pour des conférences et avoir inlassablement essayé de comprendre ce que je faisais pendant ces trois ans. Leur aide précieuse n'aura jamais faibli : même quelques minutes avant le début de ma soutenance ils aidaient à préparer le pot en ayant même recruté quatre nouveaux bras, ceux de Sylvain et Méla.

Un immense merci également à Isabelle dont le déplacement pour ma soutenance m'a beaucoup touché, Philippe, Vincent et Olivier pour tous ces bons moments passés en Bretagne et à la montagne entre la plage de St Enogat et les pistes du Christomet.

Je ne remercierai jamais assez mes parents pour tout ce qu'ils ont fait pour moi au cours de cette thèse mais également avant, en me laissant toujours décider de mes orientations. Les allers-retours Bretagne-Grenoble se sont multipliés au cours de cette troisième année et ils ont, comme à leur habitude, toujours répondu présent malgré les heures de route. Les rillettes de maquereaux, fraîchement pêchés, ainsi que les mini-crêpes accompagnées

d'un cidre du producteur ont donné à mon pot la touche bretonne qui me tenait à cœur.

Merci également à mon frère, Sébastien, et sa petite famille, Floriane, Benjamin et Nathan. En plus de leur soutien pendant ces trois années, ils m'ont procuré l'immense joie d'être tonton tout d'abord et ensuite parrain.

Enfin, je souhaite par dessus tout remercier Maud pour tant de choses qu'un second manuscrit serait nécessaire. Cette thèse est également la sienne et elle n'aurait eu aucun mal à la soutenir à ma place. Son investissement dans la rédaction malgré les milliers de kilomètres et le décalage horaire a donné une autre dimension à ce manuscrit. Merci également pour son soutien et ses encouragements incessants et surtout merci pour ces trois années à mes côtés.

# Modèles statistiques du développement de tumeurs cancéreuses

**Résumé**

Les nombreux mécanismes biologiques à l'origine du cancer restent aujourd'hui encore mal compris. L'amélioration de leurs connaissances peut s'effectuer par le biais de modèles mathématiques. Dans ce travail de thèse, nous nous sommes focalisés sur la mise en place d'outils statistiques pour la détection précoce de tumeurs. Nous avons proposé deux modèles stochastiques portant sur le développement de tumeurs cancéreuses. Le premier modèle s'intéresse à la détection de l'instabilité génétique dans une population de cellules. Nous nous sommes attachés à détecter l'événement initiateur de cette instabilité génétique en modélisant la généalogie des cellules par un arbre coalescent. Dans le deuxième modèle, nous nous sommes intéressés aux liens entre l'adhésion cellulaire et la croissance d'une tumeur. Nous avons intégré l'hypothèse d'adhésion différentielle dans un modèle d'interaction de Gibbs afin de quantifier le dysfonctionnement de l'adhésion cellulaire dans un tissu cancéreux.

**Mots Clés :** Cancer, Instabilité génétique, Perte de Mismatch Repair, Coalescent, Hypothèse d'adhésion différentielle, Diagramme de Dirichlet, Modèles de Gibbs, Estimateur de pseudo-vraisemblance.

# Statistical modeling of tumorigenesis

**Abstract**

Biological mecanisms responsible for carcinogenesis are nowadays poorly understood. Mathematical modeling may improve the knowledge of these mecanisms. The present work is dedicated to the elaboration of statistical procedures for early detection of tumors. Two stochastic models, inspired respectively from the initiation phase and the progression phase, have been proposed. The first model focuses on the detection of genetic instability in cell population. the unknown genealogy of a sample of cells within a tumoral tissue has been modeled by a coalescent tree. The second model analyses the links between cellular adhesion and the spread of a tumor. In order to quantify alterations in cellular adhesion of a cancerous tissue, the Differential Adhesion Hypothesis has been integrated in a Gibbsian model.

**Keywords :** Cancer, Genetic instability, Loss of Mismatch Repair, Coalescent, Differential Adhesion Hypothesis, Dirichlet tilling, Gibbsian model, Pseudo-Likelihood.

# Table des matières

# Table des figures

# Liste des tableaux

# Chapitre 1

# Introduction générale

Le travail de recherche présenté dans ce manuscrit s'inscrit dans le contexte général des probabilités et de la statistique appliquées à la biologie. Les travaux se sont focalisés sur la modélisation probabiliste des processus de cancérisation et la mise au point d'outils statistiques permettant une détection précoce du cancer. Dans la première partie de l'introduction, l'historique des liens entre la biologie et les mathématiques est retracé. Dans une seconde partie, nous rappelons brièvement la manière dont a été envisagé le cancer dans l'histoire avant d'expliquer biologiquement son origine et sa formation. La troisième partie présente quelques modélisations mathématiques dédiées à des processus de cancérisation. Enfin nous introduirons les travaux réalisés au cours de cette thèse.

## 1.1 Mathématiques et Biologie

Depuis de nombreuses années, des liens se sont tissés entre les mathématiques et la biologie. L'interaction entre les deux disciplines s'est considérablement renforcée durant ces deux dernières décennies. Plusieurs raisons à cet important rapprochement sont avancées par les spécialistes des deux domaines de recherche. Tout d'abord, l'explosion des masses de données à caractère biologique a conduit les mathématiciens à développer des outils d'analyse, d'exploitation et d'interprétation de ces données. Ensuite, l'accroissement phénoménal de la puissance de calcul des ordinateurs a également permis la résolu-

tion théorique de certains modèles mathématiques ainsi que la simulation de nombreux phénomènes biologiques. Enfin, l'amélioration des outils techniques et des connaissances en biologie a rendu possible de nouveaux protocoles d'expérimentation plus massifs. Les mathématiques ont tenu un rôle dans l'élaboration de ces protocoles pour leur assurer notamment certaines propriétés de reproductivité. Le séquençage complet du génôme humain est un exemple représentatif de la collaboration qui s'est tissée entre les mathématiques et la biologie. L'aboutissement d'un tel projet a nécessité d'importants moyens de calcul et l'utilisation de nombreuses analyses mathématiques a contribué efficacement à ce travail.

L'impact des mathématiques sur la biologie s'est fait à plusieurs niveaux. Nous pouvons citer tout d'abord la biologie moléculaire et cellulaire pour laquelle certains concepts mathématiques ont permis des avancées spectaculaires. Par exemple, l'analyse de la structure de l'ADN a eu recours à des outils de géométrie différentielle [Dickerson, 1989]. L'étude du développement d'un organisme a donné lieu à de nombreuses modélisations de processus comme l'embryogénèse ou la morphogénèse. Ces modélisations sont pour la plupart issues d'équations mathématiques de type « Equations aux Dérivées Partielles » ou de concepts stochastiques [Turing, 1952]. Le meilleur exemple de collaboration reste celui concernant l'écologie et la biologie des populations. La modélisation mathématique de ces phénomènes correspond à ce que certains appellent « l'âge d'or de la biologie théorique ». Sans que l'on ne puisse à proprement parler d'écoles, ces modèles se sont concentrés autour de deux courants incarnés par Sir Ronald Fisher et Alfred James Lotka. Les Equations aux Dérivées Partielles sont à la base de la dynamique des populations (Equation de Lotka-Voltera). De son côté Fisher, père fondateur de la génétique quantitative, reste aujourd'hui l'un des statisticiens les plus célèbres. Dans les deux cas, les mathématiques ont permis de comprendre certains phénomènes, d'élaborer de nouvelles hypothèses et de vérifier certaines lois.

A l'inverse, les mathématiques ont souvent trouvé, depuis plusieurs années, une source de développement importante dans la biologie. Cependant, l'utilisation d'outils purement mathématiques pour résoudre un problème biologique s'avère très souvent inefficace. En effet, l'application biologique nécessite la plupart du temps un raffinement du modèle mathématique et peut donc engendrer une modification voire une complexification du

modèle. Dans d'autres cas, les problèmes biologiques ne trouvent aucun formalisme mathématique qui puisse leur être appliqué. Ces problèmes nécessitent donc le développement de nouveaux outils mathématiques par la généralisation d'autres concepts voire même par l'élaboration de nouvelles théories. Par exemple, l'apparition de la théorie des systèmes dynamiques, ainsi que le développement des équations différentielles partielles ont été inspirés par des problèmes biologiques. La statistique et les probabilités reste certainement les domaines mathématiques les plus influencés par la biologie. Le mouvement brownien, encore largement étudié aujourd'hui, fut, par exemple, découvert par un botaniste du nom de Robert Brown en observant l'évolution des grains de pollen dans l'eau. La théorie des bifurcations proposée notamment par René Thom [Thom, 1975] fut, quant à elle, largement inspirée par les biologistes et notamment par le concept de paysage épigénétique formulé par Waddington [Waddington, 1957].

Malgré l'apport des mathématiques, de nombreux phénomènes biologiques restent à ce jour inexpliqués. Dans certains cas, l'inadéquation entre des phénomènes biologiques et des solutions mathématiques ont révélé l'extraordinaire complexité de certaines situations. Certains de ces phénomènes, comme le cancer par exemple, restent à ce jour des problèmes à l'origine de nombreux travaux.

## 1.2 Le cancer

Les travaux de recherche présentés dans cette thèse se focalisent sur les processus de cancérisation. Dans cette partie, nous donnons un bref état des lieux des connaissances actuelles sur le cancer et leur évolution au cours du temps avant d'aborder dans les parties suivantes les problèmes biologiques autour desquels s'articuleront cette thèse.

La mortalité liée au cancer tient malheureusement d'aujourd'hui une place de tout premier rang. Dans les pays industrialisés, et notamment en France, le cancer est même en passe de devenir la première cause de mortalité devant les maladies cardio-vasculaires. Mais qui est-il vraiment ? Cette question, qui reste à ce jour sans réponse définitive, est pourtant l'une des plus vieilles questions abordées par la biologie et la médecine. En effet, les traces les plus anciennes du cancer se trouvent dans des fragments de squelettes d'animaux préhistoriques et ont été datées d'environ un million d'années avant notre ère.

Des tumeurs ont également été trouvées sur des momies découvertes dans des pyramides égyptiennes. Des tablettes recouvertes de caractères cunéiformes de la bibliothèque de Ninive font également mention du cancer. Cette maladie est également évoquée lors de la découverte de monuments funéraires étrusques ou sur des momies péruviennes. Le plus ancien texte connu à ce jour semble être un papyrus chirugical de Edwin Smith qui daterait de l'ancien empire égyptien et qui serait attribué à Imouthe, grand prêtre d'Héliopolis et Premier Ministre du roi Djoser vers 2800 avant J-C. Les anciens hindous, 2000 ans avant notre ère, ont tenté de détruire les cancers en y appliquant des cataplasmes de pâte corrosive contenant de l'arsenic. Pour la petite histoire, les personnes qui ne mourraient pas du cancer mourraient empoisonnées par l'arsenic.

Quelques temps plus tard, vers 525 avant J-C, en Grèce, Hérédote nous apprend qu'Atossa, fille de Cyprus et femme de Darius, fit appeler Démocedes, le médecin grec, pour une tumeur ulcérée du sein qu'il réussit à guérir sans que le traitement employé ne soit connu. Hippocrate, dans de nombreux écrits qui lui sont attribués, fait également plusieurs fois allusion au cancer. Pour Hippocrate, le « carcinome » était une tumeur envahissante conduisant à une mort inéluctable. Hippocrate emploie le mot « carcinos » qui signifie crabe en Grec et qui évoque le crabe dévorant les tissus.

Au Moyen-Age, en France, Henri de Mondeville écrivait en 1320 : « aucun cancer ne guérit, à moins d'être radicalement extirpé tout entier. En effet, si peu qu'il en reste, la malignité augmente dans la racine ». Vers 1585, Ambroise Paré, dans son traité des « tumeurs contre nature », décrit la tumeur du sein d'une dame d'honneur de la reine Catherine de Medicis. Au XVII$^{ème}$ siècle, Gendron, médecin du frère de Louis XIV, conçoit le cancer comme « une modification tissulaire localisée qui s'étend par prolifération, curable si elle est extirpée dans sa totalité », jetant ainsi un pont de deux siècles et demi entre son époque et la nôtre. Anne d'Autriche, reine de France, fut également atteinte d'un cancer du sein, dont elle ne put en être guérie.

C'est à Bichat et Laënnec, vers 1802, que l'on doit la conception anatomique de la maladie cancéreuse et la théorie cellulaire moderne du cancer. La notion de cancer tissulaire est apportée par l'allemand Müller en 1826 tandis que Rudolph Vichow prouve quant à lui que la cellule cancéreuse naît d'autres cellules.

L'histoire du cancer ne s'arrête malheureusement pas au XIX$^{ème}$ siècle. La maladie

va se répandre dans les populations. La lutte contre le cancer est d'ailleurs devenue dans de nombreux pays industrialisés une cause nationale. Les causes de cet accroissement du nombre de cas restent cependant difficiles à évaluer. Il semble néanmoins que le vieillissement de la population, notamment grâce aux progrès de la médecine et à une amélioration de l'hygiène de vie, soit en grande partie responsable de cet accroissement. Ainsi, environ la moitié des cancers est diagnostiquée après 65 ans. Or la part de la population dépassant 60 ans était de 15% avant 1750 et de 30% vers le milieu du XIX$^{ème}$ siècle alors qu'elle atteint aujourd'hui 80%.

Les avancées sur la compréhension du cancer depuis le milieu du XIX$^{ème}$ siècle sont considérables. Il est donc difficile de retracer tout cet historique et nous allons nous intéresser dans cette section à l'état d'avancement de ces recherches à ce jour.

Aujourd'hui, en France, environ 278 000 nouvelles personnes sont touchées par un cancer chaque année et 150 000 en meurent, soit 25 000 décès supplémentaires en 20 ans. Pour le cancer du sein, en France, 34 000 nouveaux cas sont déclarés tous les ans. Ce chiffre s'élève à 33 000 pour le cancer du colon tandis que pour le cancer du poumon, 19 750 nouveaux cas sont déclarés en France tous les ans. En voyant ces chiffres, il n'est pas étonnant de constater que la probabilité d'avoir un cancer au cours de sa vie soit estimée à une sur deux pour un homme et à une sur trois pour une femme.

De nombreuses études en santé publique et en épidémiologie ont été menées au cours des dernières années afin de déterminer certains facteurs de risque liés au cancer et ainsi en améliorer le pronostic. Selon ces études, les facteurs de risque peuvent se classer en deux groupes distincts : les facteurs comportementaux et les facteurs environnementaux. Parmi les facteurs comportementaux, une des grandes découvertes de ces dernières années est la très forte corrélation entre les fumeurs et les personnes atteintes d'un cancer. Même si le tabac constitue le facteur de risque majeur, son association avec d'autres facteurs comme l'alcool, l'inactivité physique ou le surpoids augmentent considérablement le risque de développer la maladie. Ces facteurs comportementaux sont parfois associés à des facteurs environnementaux comme la pollution, l'exposition aux radiations ionisantes ou ultra-violettes qui multiplient d'autant plus le risque.

Le diagnostic d'un cancer peut se faire de plusieurs manières. D'une part il existe des signes et symptômes caractérisés pour certains types de cancers. Pourtant, à ce jour, le

diagnostic définitif est prononcé suite à une biopsie. Cet examen qui consiste à prélever des échantillons de tissu dans l'organe suspecté d'être atteint permet d'indiquer ensuite la nature des cellules en prolifération. Il permet également de déduire le grade de la tumeur et ainsi de proposer un traitement approprié. Parmi les traitements pratiqués pour contrer le développement d'une tumeur nous pouvons citer les plus utilisés : la chirurgie, la radiothérapie, la chimiothérapie, l'immunothérapie, etc... Le but de chacun de ces traitements est de faire disparaître toute trace de la tumeur dans l'organisme infecté et ainsi minimiser le risque de rechute.

Au cours des dix dernières années, de nombreuses études épidémiologiques ont montré que plus la détection d'une lésion pré-cancéreuse est précoce, plus le pronostic est favorable. Le principe est de chercher, par dépistage, chez une personne qui ne présente pas de symptômes, à mettre en évidence la maladie de manière précoce. Les examens nécessaires sont soit cliniques (palpation des seins pour le cancer du sein, toucher rectal pour le cancer de la prostate, etc ...) soit paracliniques (radiographies, dosages biologiques). Traiter un cancer à un stade précoce présente l'avantage d'offrir un traitement moins lourd au patient. De plus ces traitements sont souvent beaucoup moins onéreux et beaucoup plus efficaces que des traitements de cancer à un stade avancé.

Même si les campagnes de dépistage permettent de soigner efficacement et à moindre coût de nombreux cas de cancer, il est nécessaire de connaître la biologie du cancer pour améliorer les méthodes de détection ainsi que l'efficacité des traitements. Les moyens importants alloués à la recherche contre le cancer ont permis, par exemple, à la biologie de décrypter certains phénomènes mis en jeu dans le développement d'un cancer.

Comprendre la biologie du cancer est essentiel pour modéliser mathématiquement ce phénomène. Dans cette section nous allons brièvement rappeler certaines hypothèses biologiques sur le développement d'une tumeur.

Pour commencer, il est difficile de parler du cancer de manière générale et il est à présent usuel de préciser : le cancer du sein, le cancer du poumon, le cancer du colon, etc..., du fait des spécificités de chacun. Il n'existe d'ailleurs pas de véritable définition du cancer. Le cancer correspond à une prolifération anarchique de cellules, qui échappent aux mécanismes normaux de différenciation et de régulation de leur multiplication. Ces cellules sont capables d'envahir le tissu normal avoisinant, en le détruisant, puis de migrer

à distance pour former des métastases.

Chaque type de cancer a probablement des facteurs de déclenchement, de promotion et de progression différents, qui doivent être détaillés en étudiant chacun des cancers. Cependant, un schéma général de l'histoire naturelle d'un cancer peut être décrit. Ce schéma comporte trois étapes principales :

1. l'initiation,
2. la promotion,
3. la progression.

L'initiation correspond au tout premier phénomène de la transformation cancéreuse. On ne la connaît que grâce aux études expérimentales et aux études épidémiologiques qui ont permis de faire le lien entre les cancers et certains facteurs déclenchants. L'initiation correspond à une lésion (ou mutation) irréversible du matériel génétique d'une cellule contenu dans ses molécules d'ADN.

La promotion correspond à une exposition prolongée, répétée ou continue, à une substance qui entretient et stabilise la lésion initiée. L'agent promoteur va exercer son action pendant de nombreuses années, et ainsi faciliter la multiplication des cellules initiées. L'initiation et la promotion sont des phénomènes qui se déroulent à l'échelle cellulaire, il est donc difficile de les observer en détails par les moyens actuels.

La progression correspond quant à elle à l'acquisition des propriétés de multiplication non contrôlée, à l'acquisition de l'indépendance, à la perte de la différenciation et à l'invasion locale et métastatique. La progression constitue une étape observable du développement d'un cancer. Malheureusement, le diagnostic à cette étape est bien souvent trop tardif et ne permet pas, à l'aide des techniques actuelles, de vaincre la tumeur.

Ces découvertes sont le résultat d'expérimentations biologiques qui ont été pour la plupart confrontées à des modèles mathématiques. Un exemple récent d'application des mathématiques s'est manifesté dans une étude sur le cancer du sein. En effet, certains chercheurs ont constaté que les tissus cancéreux étaient trois à quatre fois plus conducteurs que les tissus sains. Guidés par cette idée, des chercheurs du CMAP (Centre de mathématiques appliquées de l'Ecole Polytechnique) ont résolu un problème théorique vieux de deux décennies sur des problèmes inverses appliqués à la tomographie par impédance électrique. En appliquant cette technique aux patientes atteintes d'un cancer du

sein, ces chercheurs ont proposé un examen ne comportant aucun risque pour la patiente, efficace et peu coûteux [Ammari et al., 2004].

L'amélioration des procédures de détection précoce et de prévention passe par une meilleure compréhension des phénomènes impliqués dans la cancérogénèse. L'outil mathématique est donc devenu aujourd'hui incontournable pour modéliser ces phénomènes. Une étude récente retrace les différents concepts moléculaires impliqués dans le cancer et explique l'apport d'une modélisation mathématique de ces concepts pour améliorer le pronostic [Hahn et Weinberg, 2002]. Cette modélisation mathématique du cancer a déjà donné certains résultats intéressants et laisse entrevoir de nouvelles perspectives.

## 1.3 La modélisation mathématique de la cancérogénèse

La modélisation des processus cancéreux s'effectue aujourd'hui selon deux axes principaux. Tout d'abord, les avancées en biologie moléculaire ont permis de mettre au point des modèles de phénomènes intracellulaires. Ces modèles, liés bien souvent à des hypothèses d'initiation du cancer, permettent d'améliorer la compréhension des phénomènes initiateurs. Un autre axe de modélisation s'intéresse au développement spatial d'une tumeur au sein d'un tissu. Les agents mis en jeu dans cette approche sont des agents intercellulaires et environnementaux.

L'investigation mathématique pour la cancérogénèse a commencé dans les années 50 par des théories développées notamment par [Nordling, 1953], [Armitage et Doll, 1954], [Armitage et Doll, 1957] et [Fisher, 1958]. Ces études pionnières ont permis de développer la théorie selon laquelle l'initiation d'un cancer est induite par une suite d'événements génétiques. Cette théorie, dite théorie « multiple hits », a été étudiée en détails par Ashley [Ashley, 1969]. Une étude statistique menée par Knudson [Knudson, 1971] sur 48 personnes atteintes d'un cancer de la rétine (retinoblastome) a montré que dans certains cas de cancer, deux mutations sont nécessaires pour développer la maladie : ce modèle est appelé le modèle « two-hit ». Pour une forme héréditaire, une mutation provient des parents tandis que l'autre est aléatoire. Pour une forme non-héréditaire, les deux mutations se produisent aléatoirement. Cette théorie est résumée à la Figure 1.1. Ce travail de Knudson est la base de nombreux modèles décrivant les processus d'initia-

tion d'un cancer. Une extension stochastique a été proposée au début des années 1980 [Moolgavkar et Knudson, 1981]. Dans ce modèle, les événements de mutation sont considérés aléatoires et permettent la transition des cellules d'un état à un autre : l'évolution d'un cancer sera donc une succession d'événements aléatoires. Cette vision explique en partie le fait que la majorité des cancers sont détectés après 60 ans. De plus, par cette hypothèse, certains chercheurs ont évalué à 99% la probabilité de développer un cancer pour un être humain atteignant l'âge de 150 ans.

Ce point de vue présente la génétique comme ayant un rôle fondamental dans le développement d'un cancer. Cet aspect « tout génétique » a donné lieu à de nombreux modèles mathématiques concentrés sur les processus intracellulaires permettant le développement d'une tumeur. En particulier ces travaux ont discuté en détail le rôle très controversé du phénotype de mutation exprimé par les cellules cancéreuses dans la tumorogénèse. En 1995, Tomlinson et Bodmer ont étudié un modèle mathématique markovien à états discrets [Tomlinson et Bodmer, 1995]. Le nombre d'états est fixé à quatre, et chaque état représente un état cellulaire : $F_0$ représente les cellules souches, $F_1$ les cellules semi-différenciées et $F_2$ les cellules complètement différenciées, tandis que $F_3$ est un état puits pour les cellules en apoptose. Les taux de transitions entre états dépendent des paramètres suivants : le nombre de cellules dans chacun des états, le taux de division cellulaire associé à chaque état et les taux de changement d'état. Une étude mathématique du modèle a permis aux auteurs de conclure qu'un dysfonctionnement du programme de mort cellulaire peut être suffisant mais n'est pas nécessaire à la génèse d'une tumeur. En 2002, Cairns propose un modèle mathématique expliquant le lien entre les phénotypes des cellules cancéreuses et les agents cinétiques particuliers impliqués dans la cancérogénèse [Cairns, 2002]. Le rôle de l'instabilité chromosomique a été étudié par l'équipe du professeur Nowak [Nowak et al., 2002] et [Michor et al., 2005]. Ce modèle présente l'évolution d'une crypte de cellules contenant six types de cellules différents. Le processus d'évolution est décrit à l'aide d'équations de Kolmogorov linéaires, qui peuvent être résolues analytiquement. Les auteurs, par une analyse mathématique du modèle, fournissent des conditions pour que l'instabilité chromosomique soit responsable de l'inactivation de gènes suppresseurs de tumeurs et donc de l'initiation d'une tumeur. Un autre type de modélisation de l'évolution des cryptes colorectales au stade prétumoral a été proposé par

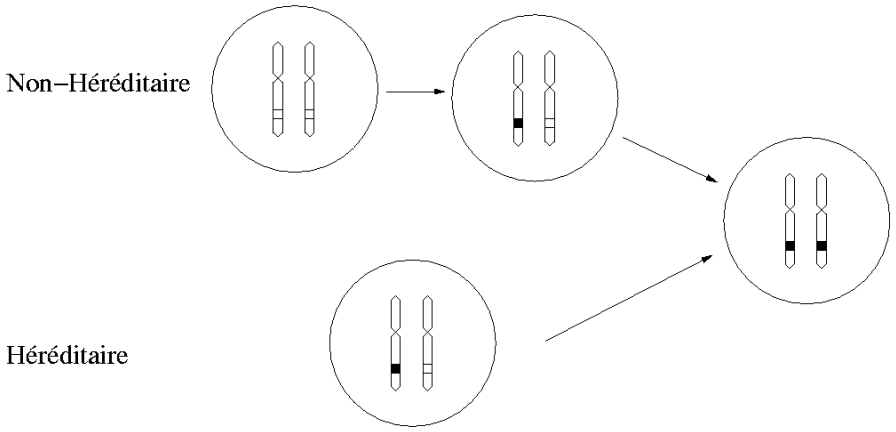



**Fig. 1.1:** *Schéma de la théorie dite « two-hit », développée par Knudson. Dans le cas non-héréditaire, deux mutations doivent survenir, une sur chacun des chromosomes, pour que le cancer soit initié. Dans le cas héréditaire, par contre, une mutation est héritée sur l'un des chromosomes, ce qui implique qu'une seule mutation est nécessaire sur l'autre chromosome pour développer un cancer.*

Calabrese [Calabrese et al., 2004]. Enfin, en 2003, l'instabilité génétique a été incorporée au modèle initial de Knudson [Little et Wright, 2003]. D'autres approches, issues de la théorie des jeux, ont été proposées pour étudier les processus intracellulaires mis en jeu dans la carcinogénèse [Tomlinson, 1997] et [Root-Bernstein et Bernstein, 1999].

Une fois l'excitation liée à l'époque du « tout génétique » retombée, certaines découvertes biologiques ont montré l'existence de signaux extracellulaires. Ces signaux jouent, en particulier, un rôle important dans la diffusion du cancer au sein d'un tissu.

En 1990, un modèle phénoménologique du cancer du colon mettant en corrélation certains événements génétiques spécifiques et la morphologie du tissu a été développé [Faeron et Vogelstein, 1990]. Cette approche a ouvert de nouvelles perspectives de modélisation mettant l'accent sur le fait qu'une mutation particulière donne à une cellule un avantage sélectif si celle-ci se trouve dans un environnement favorable. La plupart des données expérimentales obtenues ont, par la suite, été intégrées dans divers modèles. Par exemple, certains modèles déterministes étudient à l'aide d'équations de réaction-diffusion, la propagation d'une tumeur dans les premiers temps du développement [Ward et King, 1999] et [Sheratt et Nowak, 1992] puis dans la phase invasive de la tumeur [Orme et Chaplain, 1996] et [Gatenby et Gawlinski, 1996]. Une approche par automates cellulaires a également été largement utilisée pour modéliser la progression spatiale du cancer [Qi et al., 1993], [Düchting et al., 1996]. D'autres modèles mathématiques se sont focalisés sur d'autres phénomènes cancéreux spécifiques comme l'angiogenèse [Anderson et Chaplain, 1998], la réponse immunitaire de l'organisme contre le développement d'une tumeur [Owen et Sherratt, 2004] ou le processus d'invasion d'une tumeur [Turner et Sherratt, 2002]. D'autres concepts mathématiques, comme les fractales, ont été également appliqués à la recherche de motifs spatiaux caractéristiques dans un tissu cancéreux [Baish et Jain, 2000].

Nous constatons donc que, malgré la complexité des processus cancéreux, tant au niveau génétique qu'au niveau cellulaire, de nombreux modèles mathématiques ont permis de faire avancer les connaissances sur cette maladie. Des travaux de synthèse récents résument brièvement les études marquantes dans le domaine de la modélisation mathématique du cancer [Komarova, 2005], [Wodarz et Komarova, 2005] et [Quaranta et al., 2005]. Ces travaux insistent sur le fait que le dialogue entre les théoriciens et les expérimenta-

teurs doit se renforcer pour rendre la collaboration encore plus performante. Ce constat est souligné par un article de May traitant de la modélisation mathématique en biologie [May, 2004]. Dans cet article, l'auteur met en garde contre certains modèles pour lesquels un des aspects du phénomène est modélisé dans les moindres détails (par une forte paramétrisation, par exemple) au détriment d'autres aspects parfois éludés.

## 1.4 Contenu de ce travail de thèse

Ce travail de thèse s'inscrit dans la tradition des modèles présentés à la section précédente. L'objectif est de proposer des outils statistiques permettant la détection précoce du cancer. Ces outils statistiques sont issus de la modélisation probabiliste de phénomènes impliqués dans le développement d'un cancer. Deux axes de recherche principaux ont été développés. D'une part une procédure d'estimation du phénotype de mutation, résultant de l'initiation d'un cancer, est proposée. Cette procédure est fondée sur une modélisation probabiliste de la généalogie des cellules d'un tissu. La modélisation du cancer par un processus multi-cascades, est issue des théories développées dans les années 50 [Armitage et Doll, 1954] et [Armitage et Doll, 1957]. Ces théories sont fondées sur des études concernant la distribution de l'âge d'un cancer. Etre capable de dater un cancer reste une préoccupation forte et permet de traiter la maladie plus efficacement. Dans un premier temps, nous nous sommes placés au niveau des séquences d'ADN de chacune des cellules d'un tissu cancéreux. La généalogie des cellules d'un tissu peut se modéliser par un processus de branchement appelé modèle coalescent [Kingman, 1982a]. Ce type de processus est couramment utilisé dans le domaine de la génétique des populations. Une des applications les plus marquantes du coalescent consiste à dater l'âge d'un allèle dans une population. Une approche similaire, appliquée à un échantillon de cellules d'un tissu cancéreux, peut permettre de dater l'événement initiateur d'un cancer. De plus, l'utilisation de modèles de mutation comme le modèle à infinité de sites [Watterson, 1975] fournit un cadre théorique intéressant pour détecter une hausse brutale du taux de mutations dans la généalogie d'un échantillon. Cette hausse brutale est biologiquement associée à l'initiation d'un cancer. Des mesures de la diversité génétique sont couramment utilisées pour estimer des taux de mutation au sein de populations. Dans notre problème, l'uti-

lisation de ces concepts biologiques peut servir à la détection de la hausse brutale d'un taux de mutation et ainsi aider au diagnostic précoce d'un cancer.

D'autre part, en se basant sur le fait que la phase de développement spatial d'un cancer est le lieu d'interactions entre cellules voisines, une modélisation markovienne de ces interactions est proposée. Cette modélisation permet l'estimation de paramètres pertinents pour la détection d'un cancer comme l'évaluation de la force d'adhésion intercellulaire. Les études biologiques nous ont appris qu'une altération génétique ne pouvait suffire à elle seule à expliquer le développement d'une tumeur. L'environnement cellulaire joue également un rôle prépondérant dans ce phénomène : les cellules évoluent les unes en fonction des autres par des voies de signalisation extracellulaire extrêmement complexes. L'évolution d'un tissu doit prendre en compte ces réseaux d'interactions en ce sens qu'un tissu n'est pas le résultat de l'évolution de cellules indépendantes mais bien l'œuvre de cellules en société. Ce concept de sociologie cellulaire a été introduit dès 1977 [Chandebois, 1977]. Il fut ensuite largement repris dans des modèles d'évolution de tissu et notamment de tissus cancéreux [Marcelpoil, 1993] et [Bigras et al., 1996]. Ces modèles font appel à des concepts mathématiques tels que la géométrie aléatoire et les processus d'interaction « plus proches voisins ». Ces modèles mathématiques ont été introduits vers la fin des années 80 [Baddeley et Møller, 1989]. Cette classe de modèles a été par la suite bien étudiée et élargie [Bertin et al., 1999a]. Le rapprochement de ces deux domaines, sociologie cellulaire et modèles d'interaction de Gibbs, permet une modélisation rigoureuse et fine des processus d'interaction intercellulaires. Ces processus biologiques sont aujourd'hui assez bien caractérisés [Wheelock et Johnson, 1989]. De plus certaines procédures d'estimation, par maximum de pseudo-Vraisemblance, sont adaptées à ce type de modèle [Besag, 1975]. La modélisation mathématique des processus impliqués dans le développement d'un cancer autorise l'extraction de paramètres importants comme la force d'adhésion entre les cellules. Ces paramètres peuvent s'estimer par maximum de pseudo-vraisemblance et permettent de discrimer différents types tissulaires.

D'une manière générale, ce travail de thèse propose des procédures statistiques afin d'estimer certains paramètres associés au développement d'un cancer. Ces outils peuvent aider à améliorer les moyens de détection actuels. Ces aides interviennent à deux niveaux bien précis : le niveau intracellulaire par l'utilisation de données génétiques et le niveau

intercellulaire par l'utilisation de données d'expression.

Dans un premier chapitre, nous allons décrire le modèle coalescent à deux taux de mutation développé pour détecter la perte de MisMatch Repair. Dans le second chapitre, nous allons détailler le modèle d'interactions entre cellules proposé pour intégrer l'hypothèse d'adhésion différentielle dans l'étude du développement d'une tumeur.

## Chapitre 2

# Coalescent avec deux taux de mutation et application à l'instabilité génomique

## 2.1 Introduction

Le développement d'un cancer est connu pour être un processus très complexe. Depuis le 20ème siècle, il est communément admis par la communauté scientifique que la présence d'une zone tumorale est le résultat d'un processus à plusieurs étapes. Ce processus est généralement décomposé en trois étapes majeures : l'initiation, la promotion puis la progression.

Les processus d'initiation d'un cancer présente la caractéristique d'être difficile à identifier. Certains spécialistes s'accordent à dire qu'au cours de l'initiation, le tissu est considéré en phase pré-tumorale : la manifestation phénotypique du cancer n'est pas effective, ce qui rend sa détection délicate.

La compréhension des mécanismes de l'initiation du cancer représente pourtant un challenge particulièrement intéressant dans la lutte contre le cancer. Au cours de cette étape, l'état cancéreux du tissu n'est en effet pas encore irréversible et l'organisme peut dans certains cas encore empêcher le développement de la maladie. La détection d'un dé-

veloppement tumoral au cours de la phase d'initiation, souvent appelé détection précoce d'un cancer, permettrait la mise en place de protocoles médicaux encore plus efficaces.

### 2.1.1 L'initiation d'un cancer

Certains mécanismes de l'initiation cancéreuse ont pu être identifiés grâce notamment aux études expérimentales et aux études épidémiologiques. L'initiation correspond à une modification au niveau de l'ADN, le plus souvent par altérations génétiques (mutations) sur une ou plusieurs bases des séquences d'ADN [Boveri, 1929]. Dans les années 50, Armitage et Doll ont mis en avant qu'un nombre spécifique de mutations sont nécessaires pour qu'une cellule normale devienne cancéreuse ([Armitage et Doll, 1954, Armitage et Doll, 1957]). Les auteurs ont, de plus, associé le degré de malignité d'une cellule au nombre d'altérations génétiques subies par la cellule.

Ces résultats ont permis de caractériser les mutations responsables de l'initiation. En effet une cellule subit de nombreuses mutations génétiques associées à des facteurs internes ou des facteurs environnementaux au cours de ses différents cycles. Toutes ces mutations n'induisent cependant pas l'initiation d'un cancer. La plupart des mutations ont lieu dans les immenses segments d'ADN qui ne codent pour aucun gène et n'ont ainsi aucune répercussion biologique.

Parfois, les mutations affectent un gène impliqué dans le « fonctionnement » d'une cellule, comme par exemple dans la division cellulaire. Dans ce cas, la cellule possède tout un matériel enzymatique destiné à réparer les mutations produites : plusieurs points de contrôle sont présents tout au long du cycle cellulaire. A ces points de contrôle certains gènes garantissent le maintien et l'intégrité du génome de la cellule. Il arrive cependant que l'action de ces gènes, appelés *care takers* échoue. Dans ce cas, une autre catégorie de gènes, appelés *gènes suppresseurs de tumeurs*, force la cellule à se suicider par le phénomène d'apoptose, par exemple.

Tous ces mécanismes de contrôle et de réparation d'ADN expliquent la rareté des cancers si on les compare au nombre infiniment grand de mutations et de divisions cellulaires ayant lieu dans la vie d'un organisme vivant. Ces résultats mettent aussi en avant l'importance du bon fonctionnement de certains gènes spécifiques : les gènes « care takers » et les gènes « suppresseurs de tumeurs ». L'atteinte de ces gènes par une (ou plusieurs)

mutation a donc une grande importance puisqu'elle inhibe les fonctions de contrôle et de réparation d'une cellule. La grande majorité des cancers humains est d'ailleurs associée à une altération de ces gènes, ce qui laisse à penser que l'initiation d'un cancer est directement liée à ce phénomène.

Ainsi le développement d'un cancer s'accompagne d'une part d'un dysfonctionnement des gènes régulant la fidélité de la réplication de l'ADN (gènes care takers) et d'autre part d'altérations des gènes suppresseurs de tumeurs (gènes codant par exemple pour les protéines p53 ou pRb du rétinoblastome). Pour certains cancers, les gènes care takers ont été identifiés. Par exemple, l'initiation du cancer du colon HNPCC (Human Non Polyposic Colon Cancer) est causée par une altération des gènes MSH2 et MSH6 chargés de surveiller l'intégrité du génome, associée à un défaut du gène MLH1, codant pour la réparation de l'ADN [Fishel et al., 1993] et [Lindblom et al., 1993]. Le développement de certains cancers du sein et des ovaires est associé à un dysfonctionnement des gènes BRCA1 et BRCA2 responsables du contrôle et de la réparation au cours de la réplication [Wooster et Weber, 2003].

### 2.1.2 L'instabilité génétique et ses conséquences

Un dérèglement des gènes de réparation associé à un dysfonctionnement des gènes suppresseurs de tumeurs a pour conséquence l'instauration d'une instablité génétique dans une ou plusieurs lignées cellulaires. En effet, sous ces conditions, une cellule ancestrale va, par division, générer deux cellules filles ayant les mêmes propriétés de réplication. Par conséquent, le nombre de mutations sur cette lignée va devenir anormalement élevé. Ce constat est en accord avec de nombreuses expérimentations biologiques qui exhibent une très grande quantité de mutations au sein de tumeurs (voir [Loeb, 1991] par exemple).

Parmi les hypothèses biologiques formulées pour rendre compte de ce nombre très élevé de mutations dans les cellules cancéreuses, l'hypothèse de **phénotype de mutation** proposée par Loeb et collaborateurs [Loeb et al., 1974] s'avère être très robuste aux expérimentations. Un phénotype de mutation est l'expression d'un dysfonctionnement du système MMR (Mismatch Repair). Le système MMR est un système de contrôle et de réparation de l'ADN mettant en jeu des gènes care takers et des gènes suppresseurs de tumeurs. En exprimant un phénotype de mutation, les cellules cancéreuses manifestent de

manière sous-jacente un taux de mutation anormalement fort, que nous nommerons par la suite le taux de mutation élevé. Ce taux de mutation élevé est transmis par divisions successives à l'ensemble des cellules de la lignée cancéreuse.

Cette hypothèse phénotypique a été largement éprouvée au cours d'expérimentations biologiques [Loeb, 1991], [Jackson et Loeb, 1998a] et [Jackson et Loeb, 1998b]. Cependant, ces expérimentations ne permettent pas de déterminer si l'hypothèse de phénotype de mutation est une condition nécessaire et suffisante à l'initiation d'un cancer. A cet effet plusieurs modèles biologiques, résumés dans un article récent, ont été développés [Beckman et Loeb, 2005]. De plus, d'après une autre étude, le phénotype de mutation doit se manifester très tôt dans la génèse d'une tumeur pour que celle-ci puisse ensuite se développer sous l'action de promoteurs [Bielas et Loeb, 2005].

En conséquence, si nous nous plaçons sous l'hypothèse de phénotype de mutation, proposée et développée par Loeb, un cancer serait initié au sein d'une cellule par une perte de MMR. Cette perte de MMR, assimilable à un dysfonctionnement des gènes régulant la réplication de l'ADN et des gènes suppresseurs de tumeurs, engendre une hausse brutale du taux de mutation dans la cellule, dite cellule ancestrale. Puis par divisions successives, cette cellule ancestrale engendre une lignée de cellules filles, également affectée par une perte de MMR et donc sujette à un taux de mutation élevé.

### 2.1.3 Détection de l'instabilité génétique

Il est bien connu qu'une détection précoce d'un cancer favorise largement sa guérison. Bielas et Loeb [Bielas et Loeb, 2005] estiment à 20 ans le temps séparant l'initiation d'une tumeur à la manifestation du cancer. La détection d'un phénotype de mutation représente donc un challenge particulièrement intéressant dans la lutte contre le cancer.

Dans ce chapitre, nous proposons une procédure statistique de détection d'instabilité génétique reposant sur l'estimation du taux de mutation élevé. Cette procédure s'applique à un jeu de données constitué par les séquences d'ADN issues d'un échantillon de cellules du tissu. A partir de ces séquences, la généalogie du tissu peut se modéliser à l'aide d'un processus mathématique de branchement, connu sous le nom de coalescent [Kingman, 1982a]. Les mutations survenues au cours du temps suivent alors le modèle à infinité de sites le long des branches de l'arbre de coalescence. Faire l'hypothèse d'insta-

bilité génétique revient à supposer qu'un événement de mutation particulier s'est produit une et une seule fois dans la généalogie de l'échantillon. Cet événement a pour conséquence la hausse du taux de mutation pour tous les descendants de la mutation. Dans la suite de ce chapitre, nous proposons une modélisation mathématique de la généalogie d'une population de gènes. Le modèle mathématique de généalogie que nous proposons est un modèle de coalescence conditionnée à l'événement de mutation engendrant la hausse du taux de mutation.

Dans un premier temps, nous allons introduire mathématiquement le modèle du coalescent. Nous allons rappeler certains fondements du modèle coalescent, à partir du modèle de Wright-Fisher, ainsi que les principaux résultats existants sur la topologie des arbres coalescents dans un cas neutre.

Par la suite, nous rappellerons comment un processus de mutation peut se superposer à un modèle généalogique coalescent. Nous évoquerons alors deux techniques classiques d'estimation de taux de mutation, l'estimateur de Watterson et l'estimateur de Tajima.

A l'issue de ces rappels mathématiques, nous proposerons alors un modèle de généalogie de gènes à deux taux de mutation, inspiré directement du phénomène biologique d'instabilité génétique. Nous montrerons comment les contraintes biologiques de l'instabilité génétique se traduisent mathématiquement dans un modèle de coalescence conditionnel, introduit par [Wiuf et Donnelly, 1999]. Ce modèle, dit « coalescent conditionnel », suppose qu'un événement de mutation particulier est survenu une et une seule fois dans la généalogie de l'échantillon de gènes. Ce conditionnement induit une modification de la topologie de l'arbre coalescent puisque nous ne pouvons plus considérer que les temps inter-coalescents sont indépendants.

Nous étudierons alors certaines propriétés topologiques sur les arbres de coalescence conditionnés. Ces propriétés se focaliseront essentiellement sur la distribution des temps de coalescence. Grâce à ces propriétés nous fournirons un algorithme de simulation d'arbres coalescent conditionnés.

Nous présenterons ensuite des processus d'estimation du taux de mutation élevé. Nous nous attacherons à corriger le biais des estimateurs de Watterson et de Tajima. Ces corrections nécessitent l'obtention de nombreux résultats intermédiaires portant notamment sur les longueurs des branches de l'arbre.

Nous proposerons enfin des tests statistiques permettant la détection d'une instabilité génétique dans le tissu. Ces tests se baseront d'une part sur la loi des temps de coalescence et d'autre part sur les deux statistiques correspondant aux estimateurs corrigés de Watterson et de Tajima.

## 2.2 Le coalescent

Dans cette section, nous allons introduire le modèle mathématique du coalescent proposé par Kingman au début des années 80 [Kingman, 1982b] et [Kingman, 1982a]. Nous rappellerons ensuite certaines propriétés fondamentales de ce processus, portant sur les temps de coalescence et la longueur de l'arbre. Enfin, nous expliquerons, par l'intermédiaire de l'exemple de l'âge de l'allèle, comment ce modèle peut aider à répondre au problème biologique de détection d'un taux de mutation élevé.

Dans un premier temps, la construction du modèle mathématique du coalescent nécessite la description théorique du modèle de Wright-Fisher.

### 2.2.1 Le modèle de Wright-Fisher

Le modèle de Wright-Fisher originel ([Fisher, 1922], [Wright, 1931]), décrit l'évolution au cours du temps d'une population diallélique de taille constante en ignorant les effets des mutations et de la sélection. Dans cette section, nous allons rappeler les principales hypothèses simplificatrices du modèle de Wright-Fisher. Ce modèle présente l'intérêt de satisfaire une propriété markovienne facilitant son étude. Nous rappellerons l'approximation de la diffusion dans ce modèle qui suppose que la population d'étude est de très grande taille $N \to \infty$. Enfin, nous nous intéresserons à l'étude de la généalogie dans le modèle de Wright-Fisher. Nous montrerons, en particulier, comment ce modèle permet de modéliser, à partir de l'état présent d'une population, l'évolution des générations futures ainsi que la généalogie de cette population. Ce dernier point nous permettra d'introduire le modèle coalescent à partir de modèle de Wright-Fisher.

#### Modèle de Markov

Une particularité du modèle de Wright-Fisher est que la population est de taille

constante et égale à $N$ au cours des générations. De plus, les générations sont supposées non recouvrantes, c'est à dire qu'aucun individu de la génération $r$ n'est plus vieux qu'un individu de la génération suivante $r+1$. Chaque individu est caractérisé par un locus pour lequel deux allèles $A$ et $B$ sont présents. La variable $M_r$ est définie comme le nombre d'individus exprimant l'allèle $A$ à la génération $r$. La population à la génération $r+1$ est directement issue de la population à la génération $r$ de la manière suivante : chacun des $N$ allèles de la génération $r+1$ est tiré aléatoirement et de façon indépendante parmi les $N$ allèles de la génération $r$. Ainsi, la loi de $M_{r+1}$ sachant que $M_r = i$ est une loi binomiale de paramètre $i/N$, quelque soit la génération $r$ :

$$p_{ij} = \mathbb{P}(M_{r+1} = j | M_r = i) = \begin{pmatrix} N \\ j \end{pmatrix} \left(\frac{i}{N}\right)^j \left(1 - \frac{i}{N}\right)^{N-j} \quad 0 \leq i, j, \leq N.$$

Nous en déduisons donc que le processus $\{M_r, r = 0, 1, \ldots\}$ est une chaîne de Markov homogène, de matrice de transition $P = (p_{ij})_{i=0,\ldots,N;j=0\ldots,N}$ et d'espace d'états $\mathcal{S} = \{0, 1, \ldots, N\}$.

Pour des populations réelles, comme la population de gènes chez un individu, le coefficient $N$ est souvent « très grand ». Ainsi, pour prendre en compte mathématiquement ce phénomène, il est usuel de se placer sous l'approximation de la diffusion.

**Approximation de la diffusion**

L'idée basique de la diffusion dans le modèle de Wright-Fisher consiste à s'intéresser à la proportion d'allèles $A$ dans la population, c'est à dire $M_r/N$, et non pas au processus $M_r$ du nombre total d'allèles $A$ [Neuhauser et Tavaré, 2001]. Afin d'éviter une limite dégénérée pour ce processus, le temps doit également être renormalisé avec comme unité de temps $N$ générations. Cette approximation joue un rôle très important dans la suite de ce chapitre.

Nous pouvons rappeler ici que notre problème initial nécessite de modéliser, à partir des séquences d'ADN, la généalogie de cellules au sein d'un tissu humain. Nous allons à présent nous intéresser à la modélisation de la généalogie d'une population grâce au modèle de Wright-Fisher. Cette modélisation nous sera utile pour étudier le problème de l'instabilité génétique.

**Généalogie dans le modèle de Wright-Fisher : processus « forward » et « backward »**

Dans ce paragraphe, nous considérons le modèle de Wright-Fisher sous une perspective généalogique : nous allons chercher, selon les cas, à modéliser les générations futures ou la généalogie passée si celle-ci est inconnue, en fonction de l'état présent [Tavaré, 2004]. Cette reconstruction est considérée dans cette partie pour un échantillon de gènes au sein d'un tissu.

Dans cette section, nous allons tout d'abord nous intéresser à la modélisation des générations futures (vision « forward »), puis nous verrons comment passer à la modélisation d'une généalogie passée (vision « backward »).

En l'absence de recombinaison, une séquence, présente à la génération $r$, peut être vue comme une copie d'une séquence choisie aléatoirement parmi les séquences présentes à la génération $r-1$, elle-même étant une copie d'une séquence de la génération $r-2$, etc...Ainsi, chaque séquence $s$ d'ADN à la génération $r$ peut être considérée comme un individu qui a un parent (la séquence de la génération $r-1$ dont $s$ est une copie) et des enfants (les séquences de la génération $r+1$ qui sont des copies de $s$).

Afin d'étudier en détail ce processus, il est assez usuel de numéroter les individus d'une génération donnée : $1, 2, \ldots, N$. De plus, $\nu_i$ $(i = 1, 2, \ldots, N)$ représente le nombre d'enfants à la génération $r+1$ issus de l'individu $i$ de la génération $r$. Par le processus décrit ci-dessus, pour lequel chaque enfant choisit son parent aléatoirement dans la population de la génération précédente, nous obtenons que :

$$\mathbb{P}(\nu_1 = m_1, \nu_2 = m_2, \ldots, \nu_N = m_N) = \frac{N!}{m_1! m_2! \ldots m_N!} \left(\frac{1}{N}\right)^N, \tag{2.1}$$

avec la contrainte que $\nu_1 + \nu_2 + \cdots + \nu_N = N$.

L'équation 2.1 nous permet donc de simuler la généalogie d'une population. Cette vision « forward », c'est à dire qui suit le sens du temps, permet de construire les générations futures d'une population à partir du présent.

Un des nombreux avantages du modèle de Wright-Fisher est qu'il autorise également une vision « backward » de la généalogie d'une population, pour laquelle le processus remonte le temps. Pour cette vision, nous essayons de reconstruire la généalogie passée d'un échantillon de séquences d'ADN séquencée au temps présent.

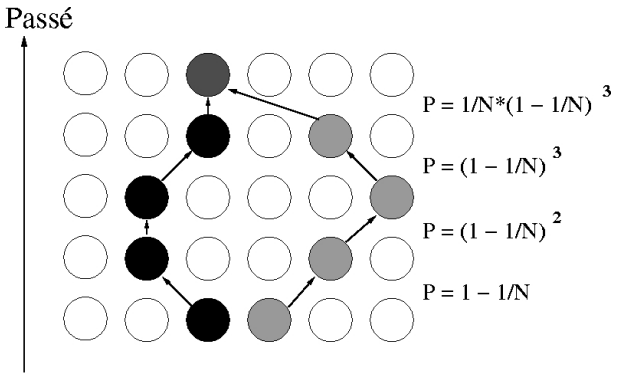



**FIG. 2.1:** *Exemple de généalogie dans un modèle de Wright-Fisher à $N = 6$ individus.*

Pour décrire le processus « backward » du modèle de Wright-Fisher, nous pouvons remarquer que chaque individu de la génération $r$ choisit aléatoirement un parent parmi les $N$ individus de la génération $r-1$. De plus, les choix successifs d'une génération à une autre sont indépendants. Nous constatons également que tous les individus de la génération $r$ ne possède nt par forcément un « enfant », ce qui modélise le fait que certaines lignées s'éteignent à partir d'une génération.

Dans un premier temps nous allons déterminer le nombre de générations nécessaires pour que deux individus au hasard dans la population aient un ancêtre commun. Comme les parents sont choisis au hasard et de façon indépendante, nous avons :

$$\mathbb{P}(\text{2 individus ont 2 parents distincts}) = \left(1 - \frac{1}{N}\right).$$

Comme les parents choisissent eux aussi leurs parents aléatoirement et de façon indépendante, nous obtenons que :

$$\begin{aligned}&\mathbb{P}(\text{2 individus ont un ancêtre différent à la génération } r)\\ &\quad = \mathbb{P}(\text{Le premier ancêtre commun de 2 individus est d'une génération antérieure à } r)\\ &\quad = \left(1 - \frac{1}{N}\right)^r.\end{aligned}$$

La figure 2.1 donne un exemple d'application de la formule précédente. En renormalisant le temps par unité de $N$ générations, par l'approximation de la diffusion, nous

obtenons que :

$$\mathbb{P}(\text{Le premier ancêtre commun de 2 individus apparaît après un temps } t) = \left(1 - \frac{1}{N}\right)^{Nt}.$$

Ce calcul consititue une des bases du modèle coalescent décrivant la généalogie d'un échantillon de taille $n$ pris dans l'ensemble de la population de taille $N$. Comme nous l'avons remarqué pour la diffusion dans le modèle de Wright-Fisher, la taille de la population $N$ peut être considérée infinie. En pratique, l'étude d'un échantillon d'individus ne peut donc pas satisfaire les hypothèses du modèle de Wright-Fisher : il n'est pas raisonnable de considérer que l'échantillon dont nous disposons est la population totale. Ainsi, lorsque nous étudions un échantillon de séquences d'ADN, nous devons supposer que cet échantillon de taille $n$ est un sous-ensemble d'une population totale de taille $N$ qui elle suit la dynamique de Wright-Fisher. Dans la suite de cette section, nous allons nous intéresser à la généalogie d'un sous-échantillon de taille $n$, choisie dans une population de taille $N$ suivant la dynamique de Wright-Fisher ou Moran.

### 2.2.2 Le coalescent : processus ancestral dans le modèle de Wright-Fisher pour une population de grande taille

Dans ce paragraphe, nous nous intéressons au processus ancestral de la généalogie d'un échantillon d'individus, pris dans une population de taille $N$. Une variable d'intérêt, caractérisant la généalogie de l'échantillon, est le nombre d'ancêtres de l'échantillon de taille $n$ au temps $t$ (« backward » dans le temps). Pour cela nous définissons le processus $\{A_n(t) : t = 0, 1, \ldots, n, \ldots\}$, appelé processus ancestral, tel que :

$A_n(t) \equiv$ Nombre d'ancêtres à la génération $t$ d'un échantillon de taille $n$ à la génération 0.

Afin d'étudier le processus $A_n$, il est utile de calculer la probabilité que $k$ individus aient $j$ parents distincts. Nous obtenons, par analyse combinatoire [Tavaré, 2004] :

$$\mathbb{P}(k \text{ individus ont } j \text{ parents distincts}) = N(N-1)\ldots(N-j+1)\mathcal{S}_k^{(j)}N^{-k},$$

où $\mathcal{S}_k^{(j)}$ est le nombre de Stirling de deuxième espèce, c'est à dire le nombre de partitions en $j$ classes d'un ensemble de $k$ individus.

Nous pouvons donc constater que le processus ancestral $A_n(t)$ est une chaîne de Markov d'espace d'états $\{1, 2, \ldots, n\}$ dont les probabilités de transition sont données par :

$$\mathbb{P}(A_n(t+1) = j | A_n(t) = k) =$$
$$\mathbb{P}(k \text{ individus ont } j \text{ parents disctincts})$$
$$N(N-1)\ldots(N-j+1)\mathcal{S}_k^{(j)} N^{-k} \quad \text{pour } j = 1, \ldots, k \quad \text{et } k \leq n.$$

Nous allons à présent nous intéresser au cas où la population est de grande taille. Ceci signifie donc que nous nous intéressons au processus ancestral d'un échantillon de taille $n$ dans une population de taille $N$ avec $N \to +\infty$. Pour cette approximation nous obtenons alors, pour $k = 2, \ldots, n$ :

$$\begin{aligned} \mathbb{P}(A_n(t+1) = k | A_n(t) = k-1) &= \mathcal{S}_k^{(k-1)} \frac{N(N-1)\ldots(N-k+2)}{N^k} \\ &= \begin{pmatrix} k \\ 2 \end{pmatrix} \frac{1}{N} + O(N^{-2}), \end{aligned}$$

car $S_k^{(k-1)} = \begin{pmatrix} k \\ 2 \end{pmatrix}$.

De plus, pour $1 \leq j < k-1$, nous obtenons que :

$$\begin{aligned} \mathbb{P}(A_n(t+1) = k | A_n(t) = j) &= \mathcal{S}_k^{(j)} \frac{N(N-1)\ldots(N-k+2)}{N^k} \\ &= O(N^{-2}). \end{aligned}$$

Et enfin, pour $k = 1, \ldots, n$ :

$$\begin{aligned} \mathbb{P}(A_n(t+1) = k | A_n(t) = k) &= \frac{N(N-1)\ldots(N-k+2)}{N^k} \\ &= 1 - \begin{pmatrix} k \\ 2 \end{pmatrix} \frac{1}{N} + O(N^{-2}). \end{aligned}$$

Nous posons $I$ la matrice identité, et $Q$ la matrice dont les coefficients non nuls sont donnés par :

$$q_{k,k} = -\begin{pmatrix} k \\ 2 \end{pmatrix}, \quad q_{k,k-1} = \begin{pmatrix} k \\ 2 \end{pmatrix}, \quad k = n, n-1, \ldots, 2.$$

Ainsi, la matrice de transition, $G_N$, du processus ancestral peut s'écrire :

$$G_N = I + N^{-1}Q + O(N^{-2}).$$

En renormalisant, par approximation de la diffusion, le temps par unité de $N$ générations, nous obtenons :

$$G_N^{Nt} = (I + N^{-1}Q + O(N^{-2})) \to e^{Qt}.$$

Nous en déduisons donc que le nombre distinct d'ancêtres à la génération $Nt$ est déterminée par la chaîne de Markov à temps continu $A_n$ dont le comportement est décrit par la matrice $Q$. Plus précisément, $A_n$ est un processus de mort pure, avec pour état initial $A_n(0) = n$, et décroissant par sauts successifs d'une, et une seule, unité. De plus, lorsque $A_n$ est dans l'état $k$ $(1 < k < n)$, le temps d'attente, $X_k$, avant le prochain saut suit une loi exponentielle de paramètre $\lambda_k = k(k-1)/2$, les $X_k$ étant indépendants, et satisfaisant l'approximation de la diffusion :

$$X_k \to \mathcal{E}(\lambda_k) \quad \text{pour } k = 2, \ldots, n, \tag{2.2}$$

où $\mathcal{E}$ représente la loi exponentielle.

Le processus ancestral a été largement étudié dans [Karlin et McGregor, 1972] [Cannings, 1974] [Watterson, 1975] [Griffiths, 1980] et [Tavaré, 1984].

Nous allons à présent nous intéresser aux propriétés topologiques d'un arbre de coalescence. Nous nous focaliserons essentiellement sur la loi du temps jusqu'à l'ancêtre commun le plus récent et la longueur totale de l'arbre. Ces deux quantités résument la topologie de l'arbre.

### 2.2.3 Propriétés d'un arbre de coalescence

Dans cette section, nous allons rappeler certaines propriétés fondamentales du processus de coalescence dans le cas d'une population de grande taille. Ces propriétés permettent de fournir un algorithme très simple de simulation d'un arbre coalescent.

Dans un premier temps, nous allons nous intéresser au temps nécessaire qu'il faut attendre pour que l'échantillon n'ait plus qu'un seul ancêtre dans la population totale. Cet ancêtre est appelé l'ancêtre commun le plus récent comme traduction de « most recent

common ancestor (MRCA) ». Cette quantité détermine l'âge de la racine de l'arbre de coalescence. Puis nous nous intéresserons à la longueur totale de l'arbre, c'est à dire à la somme des longueurs de toutes les branches de l'arbre. Cette quantité est essentielle pour étudier le nombre de mutations qui surviennent au cours de l'histoire de l'échantillon.

Notons $W_n$ la variable représentant le temps nécessaire pour qu'un échantillon de taille $n$ n'ait plus qu'un seul ancêtre. $W_n$ représente le temps minimal pour lequel $A(W_n) = 1$. En rappelant que le temps est mesuré par unité de $N$ générations, nous obtenons que :

$$W_n = X_n + X_{n-1} + \cdots + X_3 + X_2,$$

où les $X_k$ sont des variables indépendantes, de lois exponentielles de paramètres $k(k-1)/2$.

Nous déduisons donc que :

$$\mathbb{E}[W_n] = \sum_{k=2}^{n} \mathbb{E}[X_k] = \sum_{k=2}^{n} \frac{2}{k(k-1)} = 2\left(1 - \frac{1}{n}\right).$$

Puisque les $X_k$ sont indépendants, nous obtenons également que :

$$\begin{aligned} Var(W_n) &= \sum_{k=2}^{n} Var(X_k) = \sum_{k=2}^{n} \left( \begin{array}{c} k \\ 2 \end{array} \right)^{-2} \\ &= 8 \sum_{k=1}^{n-1} \frac{1}{k^2} - 4\left(1 - \frac{1}{n}\right)\left(3 + \frac{1}{n}\right). \end{aligned}$$

Notons $L_n$ la variable représentant la longueur totale d'un arbre coalescent. Cette longueur peut s'écrire de la manière suivante :

$$L_n = 2X_2 + \cdots + nX_n = \sum_{k=2}^{N} kX_k.$$

Ainsi nous obtenons que [Tavaré, 2004] :

$$\mathbb{E}[L_n] = 2 \sum_{j=1}^{n-1} \frac{1}{j} \approx 2 \log n$$

et :

$$Var(L_n) = 4 \sum_{j=1}^{n-1} \frac{1}{j^2} \approx 2\pi^2/3.$$

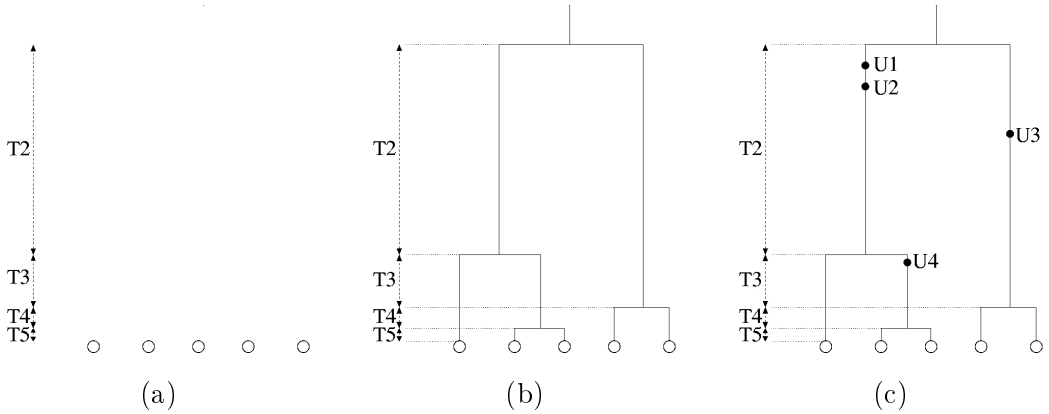



**Fig. 2.2:** *Simulation d'un arbre de coalescence avec mutations pour un échantillon de 5 individus. La figure (a) représente le tirage des temps $X_k$ selon leurs lois respectives et de façon indépendante. La figure (b) nous montre le choix aléatoire des individus qui coalescent à chaque événement de coalescence. La figure (c) nous fait apparaître le processus de simulation des mutations le long de chaque branche de l'arbre.*

Enfin, la loi de $L_n$ est également explicite [Tavaré, 2004] :

$$\mathbb{P}(L_n \leq t) = (1 - \exp(-t/2))^{n-1}, \quad \mathrm{t} \geq 0$$

La prochaine section est dédiée à la simulation d'un arbre de coalescence neutre. Ce processus de simulation est relativement immédiat au regard de la formule 2.2.

### 2.2.4 Simulation d'un arbre de coalescence

Le processus de simulation se concentre sur la topologie de l'arbre. Pour cette étape, les temps de coalescence $X_n, X_{n-1}, \ldots, X_2$ sont tout d'abord simulés selon leurs lois respectives ($X_i \rightarrow \mathcal{E}(\lambda_k)$), voir Equation 2.2). Puis pour chaque événement de coalescence, deux individus (ou nœuds de l'arbre) sont choisis au hasard. Ainsi, la topologie complète de l'arbre est déterminée. Un exemple de réalisation, pour un échantillon de taille $n = 5$, est présenté à la figure 2.2.

Nous allons à présent introduire une application classique du modèle coalescent. Cette application, appelée l'âge de l'allèle, constitue un problème très important en génétique.

### 2.2.5 Un problème classique : l'âge de l'allèle

Grâce à l'explosion de la génétique, il est aujourd'hui possible d'étudier des événements du passé à l'aide d'observations présentes. L'étude de l'âge d'un allèle est en ce sens d'un intérêt tout particulier dans de nombreuses études génétiques. Par exemple, l'utilisation de techniques de séquençage pour cartographier certaines maladies nécessite de connaître notamment l'âge des allèles.

De nombreuses techniques, résumées dans un article de synthèse, ont été proposées pour résoudre ce problème statistique [Slatkin et Rannala, 2000]. L'une des techniques les plus utilisées et les plus étudiées consiste à inférer l'âge d'un allèle à partir de sa fréquence d'observation dans une population donnée. Ainsi, Kimura et Ohta ont déterminé l'espérance de l'âge d'une mutation, sous l'hypothèse de neutralité et en supposant que la fréquence d'observation vaut $x$ [Kimura et Ohta, 1973]. Cette espérance s'écrit :

$$\mathbb{E}[\text{Age}|x] = -2\frac{x}{1-x}\log(x).$$

Ce résultat très connu a par la suite été étendu dans de nombreux travaux. Nous pouvons remarquer que, dans le cas de l'instabilité génétique, l'estimation de l'âge de l'allèle revient, dans une certaine mesure, à estimer l'âge de la perte de MMR. Nous allons nous intéresser ici à une extension consistant à utiliser une modélisation coalescente de la généalogie de l'échantillon pour inférer l'âge de l'allèle. En 1998, Griffiths et Tavaré ont proposé une étude sur des arbres de coalescence généraux, pour lesquels les temps intercoalescences $X_i$ sont inconnus [Griffiths et Tavaré, 1998].

En rappelant que les variables $X_k$, pour $k = 2, \ldots, n$, représentent les temps de coalescence de l'échantillon étudié, nous allons citer les six résultats principaux déterminés par [Griffiths et Tavaré, 1998].

Tout d'abord, sous le modèle à infinité de sites, pour un arbre de coalescence général,

la probabilité $q_{n,b}$ qu'un site de ségrégation ait $b$ bases mutantes est donnée par :

$$q_{n,b} = \frac{(n-b-1)!(b-1)!\sum_{k=2}^{n} k(k-1)\begin{pmatrix} n-k \\ b-1 \end{pmatrix}\mathbb{E}(X_k)}{(n-1)!\sum_{k=2}^{n} k\mathbb{E}(X_k)}. \tag{2.3}$$

De plus, dans le cadre d'une population à taille constante, l'espérance de l'âge d'un allèle ayant $b$ copies dans un échantillon de taille $n$ vaut :

$$2\begin{pmatrix} n-1 \\ b \end{pmatrix}^{-1}\sum_{k=2}^{n}\begin{pmatrix} n-j \\ b-1 \end{pmatrix}\frac{n-j+1}{n(j-1)}. \tag{2.4}$$

L'espérance du temps jusqu'au plus récent ancêtre commun à l'échantillon, conditionnellement au fait que l'allèle ait $b$ copies au sein de l'échantillon, vaut :

$$2\left(1-\frac{1}{n}\right)+2\begin{pmatrix} n-1 \\ b \end{pmatrix}^{-1}\sum_{k=2}^{n}\begin{pmatrix} n-j \\ b-1 \end{pmatrix}\frac{1}{j(j-1)}. \tag{2.5}$$

Le temps moyen jusqu'à l'ancêtre commun le plus récent, conditionnellement à une mutation de fréquence $x$ est donné par :

$$2-\frac{2x}{1-x}\left(1+\frac{2-x}{1-x}\log x\right). \tag{2.6}$$

L'âge moyen d'une mutation dans un arbre général de coalescence vaut :

$$\frac{\frac{1}{2}\sum_{k=2}^{\infty} k(k-1)(1-x)^{k-2}\mathbb{E}[S_k^2 - S_{k+1}^2]}{\sum_{k=2}^{\infty} k(k-1)(1-x)^{k-2}\mathbb{E}[X_k]}, \tag{2.7}$$

où $S_k = \sum_k^{\infty} X_j$.

Enfin, la probabilité qu'un allèle, observé avec un fréquence $x$, soit le plus vieux dans la population est donnée par :

$$\frac{\sum_{k=2}^{\infty} k(k-1)x^{k-2}\mathbb{E}[X_k]}{\sum_{k=2}^{\infty} k(k-1)(1-x)^{k-2}\mathbb{E}[X_k] + \sum_{k=2}^{\infty} k(k-1)x^{k-2}\mathrm{E}[X_k]}. \tag{2.8}$$

Ces six résultats, démontrés dans un coalescent général, nous seront utiles dans la suite de ce chapitre. Nous y ferons référence dès que nous les utiliserons.

Dans cette section, nous nous sommes intéressés à la modélisation de la généalogie d'un échantillon de gènes pris dans un tissu. Cette modélisation s'articule autour de la

topologie de l'arbre généalogique représenté par un processus coalescent : les variables d'intérêt sont essentiellement les temps de coalescence, le temps jusqu'à l'ancêtre commun le plus récent et la longueur totale de l'arbre.

L'ensemble des résultats que nous venons de rappeler ne tient pas compte des mutations qui surviennent au cours de l'histoire de la population de gènes. Dans la section suivante, nous allons rappeler comment les mutations sont classiquement intégrées dans un modèle coalescent. Cette technique suppose que le processus de mutation soit indépendant du processus généalogique. Sous cette hypothèse, nous allons rappeler deux estimateurs du taux de mutation : l'estimateur de Watterson et l'estimateur de Tajima.

## 2.3 Estimation d'un taux de mutation sous le modèle à infinité de sites

Dans un premier temps, nous allons rappeler comment les mutations sont superposées à un arbre de coalescence. Nous nous intéresserons, par la suite, à l'estimation du taux de mutation dans un modèle coalescent neutre.

### 2.3.1 Mutations dans le coalescent

Nous pouvons noter $\mu$ la probabilité d'une mutation par base et par génération. Ainsi, le nombre moyen de mutation pour une lignée, au cours de $g$ générations, vaut $g\mu$. En rappelant que le temps est mesuré en unité de $N$ générations, nous obtenons que le nombre moyen de mutations pour une lignée pendant un temps $t$ vaut : $tN\mu$. Cette dernière quantité est finie lorsque $\mu$ est de l'ordre de $1/N$. Sous ces conditions, nous pouvons donc poser :

$$\theta = 4N\mu.$$

En nous plaçant dans le cas d'une population de grande taille, c'est à dire $N \to \infty$, l'hypothèse $\mu \to 0$ doit être considérée. Ainsi, sous ces conditions limites, les mutations surviennent le long des branches de l'arbre coalescent selon un processus de Poisson de paramètre $\theta/2$.

Nous allons à présent rappeler certaines propriétés mathématiques du modèle coalescent. Ces propriétés focalisent exclusivement sur la topologie des arbres de coalescence.

L'estimation du taux de mutation d'une population est associée à une mesure de diversité génétique de cette population. Actuellement, deux estimateurs sont couramment utilisés : l'estimateur de Watterson et l'estimateur de Tajima. L'analyse mathématique de ces estimateurs est grandement simplifiée par l'hypothèse du modèle à infinité de sites.

### 2.3.2 Le modèle à infinité de sites

Le modèle à infinité de sites a été introduit à la fin des années 60 par Kimura [Kimura, 1969]. Ce modèle postule que deux mutations ne peuvent affecter le même site. Cette hypothèse vient du fait que le taux de mutation est faible, ainsi la probabilité que deux mutations touchent le même site peut être considérée comme nulle. Ce modèle a été formalisé par Watterson [Watterson, 1975].

### 2.3.3 L'estimateur de Watterson

L'estimateur de Watterson s'interprète comme le nombre de sites de ségrégation présents sur l'ensemble de l'échantillon étudié. Sous l'hypothèse du modèle à infinité de sites, le nombre de sites de ségrégation correspond exactement au nombre total de mutations, $S_n$, survenues au cours de la généalogie de l'échantillon. En supposant que cette généalogie se modélise par un processus ancestral comme décrit à la section précédente, nous obtenons que, conditionnellement à la longueur de l'arbre $L_n$, $S_n$ suit une loi de Poisson de taux $\theta L_n/2$.

Nous en déduisons donc que :

$$\begin{aligned} \mathbb{E}[S_n] &= \mathbb{E}[\mathbb{E}[S_n|L_n]] \\ &= \mathbb{E}[\theta L_n/2] \\ &= \frac{\theta}{2}\sum_{j=1}^{n-1}\frac{2}{j}. \end{aligned}$$

Ainsi :

$$\mathbb{E}[S_n] = \theta\sum_{j=1}^{n-1}\frac{1}{j} \sim \theta\log(n). \tag{2.9}$$

Pour déterminer la variance du nombre de sites de ségrégation, nous pouvons nous référer à Watterson [Watterson, 1975] qui a montré que :

$$Var(S_n) = \theta \sum_{j=1}^{n-1} \frac{1}{j} + \theta^2 \sum_{j=1}^{n-1} \frac{1}{j^2}. \tag{2.10}$$

D'après l'équation 2.9, nous avons :

$$\theta_W = \frac{S_n}{\sum_{j=1}^{n-1} \frac{1}{j}}$$

est un estimateur sans biais du taux de mutation $\theta$. De plus, d'après l'équation 2.10, nous avons $Var(\theta_W) \to 0$ quand $n \to +\infty$.

### 2.3.4 L'estimateur de Tajima

Par formalisme, supposons que l'échantillon, $\mathbf{y}$ soit composé de $n$ séquences d'ADN, chacune de longueur $s$. L'échantillon peut alors s'écrire :

$$\mathbf{y} = (y_i)_{i=1\ldots n} = (y_{i1}, y_{i2}, \ldots, y_{is})_{i=1\ldots n}.$$

Le nombre de sites qui diffèrent entre la séquence $i$ et la séquence $j$, noté $\Pi(i,j)$, peut s'écrire :

$$\Pi(i,j) = \sum_{l=1}^{s} \mathbb{1}(y_{il} \neq y_{jl}) \quad i \neq j.$$

L'estimateur de Tajima pour un échantillon de taille $n$, est alors défini comme le nombre moyen de différences entre deux séquences de l'échantillon, soit :

$$\Pi_n = \frac{1}{n(n-1)} \sum_{i \neq j} \Pi(i,j).$$

Afin de mesurer le biais de l'estimateur de Tajima, nous pouvons étudier la variable aléatoire $\Pi_n$ sous l'hypothèse d'infinité de sites et en supposant que la généalogie de l'échantillon suit un modèle coalescent. Rappelons ici que les mutations suivent un processus de Poisson de paramètres $\theta/2$ le long des branches de l'arbre de coalescence.

Nous rappellerons uniquement l'espérance et la variance de $\Pi_n$ :

$$\begin{aligned} \mathbb{E}[\Pi_n] &= \frac{1}{n(n-1)} \sum_{i \neq j} \mathbb{E}[\Pi(i,j)] \\ &= \mathbb{E}[\Pi(1,2)] \\ &= \theta \mathbb{E}[X_2], \end{aligned}$$

où $X_2$ est le temps de coalescence pour un échantillon de taille 2. Nous avons vu précédemment que $X_2$ suit une loi exponentielle de paramètre 1. Nous en déduisons donc que, dans le cas d'une population de taille constante (c'est à dire $N$ constant), nous avons :

$$\mathbb{E}[\Pi_n] = \theta. \tag{2.11}$$

La variance de $\Pi_n$ a été calculée par Tajima [Tajima, 1983] :

$$Var(\Pi_n) = \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2. \tag{2.12}$$

D'après l'équation 2.11, $\Pi_n$ est un estimateur sans biais du taux de mutation $\theta$. Cependant, l'équation 2.12 nous montre que $\Pi_n$ ne converge pas lorsque $n \to +\infty$.

Dans cette section, nous avons rappelé comment les événements de mutation sont intégrés à un modèle coalescent. L'estimation du taux de mutation, dans un modèle neutre à un seul taux, a également été évoquée.

## 2.4 Un modèle à deux taux de mutation

Dans cette section, nous allons proposer un modèle à deux taux de mutation, satisfaisant les conditions de modélisation liées à l'instabilité génétique. Dans le cadre de l'instabilité génétique, qui est le propos de notre étude, les cellules d'un tissu peuvent être de deux types disctincts : cellules cancéreuses ou cellules normales. Les cellules cancéreuses manifestent un phénotype de mutation caractérisé par un taux de mutation anormalement élevé pour ces cellules. Nous rappelons que l'hypothèse d'instabilité génétique postule que la hausse du taux de mutation est brutale : cette hausse est la conséquence d'un événement unique, caractérisant la perte de Mismatch Repair et noté $\Delta$. Par conséquent la généalogie des cellules d'un tissu peut se représenter comme dans la Figure 2.3.

Dans cette section, nous allons montrer que les hypothèses topologiques de l'instabilité génétique sur la généalogie des cellules sont similaires aux hypothèses émises pour construire le modèle du coalescent conditionnel. Nous rappellerons ensuite certaines propriétés du modèle coalescent conditionnel nécessaires à notre étude. Ces résultats nous

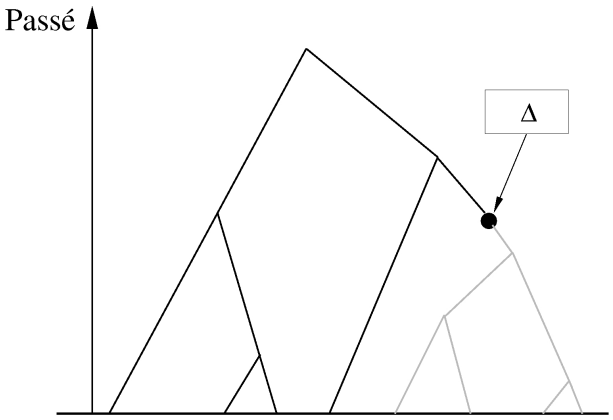



**Fig. 2.3:** *Représentation de la généalogie d'un échantillon sous l'hypothèse d'instabilité génétique. La généalogie est conditionnée à l'événement $\Delta$. Toutes les branches issues de $\Delta$, en gris sur le schéma ont un taux de mutation élevé. A l'inverse les autres branches, en noir, ont un taux de mutation normal.*

permettront d'une part de fournir un algorithme de simulation d'un processus de coalescence conditionnelle et d'autre part d'étudier mathématiquement la correction des estimateurs de Watterson et Tajima dans un modèle à deux taux.

### 2.4.1 Coalescent conditionnel et notations

Par la suite, nous considérons que la taille de l'échantillon de gènes vaut $n$. Cet échantillon est divisé en deux sous-échantillons que nous noterons respectivement $\mathcal{B}$ et $\mathcal{C}$. $\mathcal{B}$ caractérise le sous-échantillon de cellules cancéreuses tandis que $\mathcal{C}$ symbolise les cellules normales. D'après nos hypothèses, les cellules de $\mathcal{B}$ sont les descendantes de la cellule affectée par l'événement $\Delta$ (perte de MMR). Nous noterons $B$ la variable aléatoire caractérisant le cardinal de $\mathcal{B}$, c'est à dire le nombre de cellules ayant un taux de mutation élevé.

L'étude de modèles de coalescent conditionnel nécessite deux niveaux de conditionnement. Tout d'abord, l'arbre de coalescent est conditionné au fait que l'ensemble des séquences de $\mathcal{B}$ sont des descendants de l'événement $\Delta$. Ce niveau de conditionnement

est un conditionnement topologique que nous noterons par la suite $E$. Le second niveau de conditionnement présuppose que l'événement $\Delta$ se produit une et une seule fois dans la généalogie de l'échantillon. Cet événement est noté $M$. Le conditionnement $E$ affecte la topologie générale de l'arbre tandis que l'événement $M$ influe sur les longueurs des branches de l'arbre. Il s'avère que le double conditionnement, $E \cap M$, revient à considérer l'occurrence d'un événement unique de polymorphisme (UEP) dans l'arbre de coalescence [Wiuf et Donnelly, 1999] et [Tavaré, 2004]. Cette remarque fait le lien entre notre modèle et les études sur l'âge de l'allèle présentées à la section précédente.

L'avantage sélectif des cellules prétumorales étant encore mal décrit dans la littérature, nous avons négligé les effets de ce phénomène et fait l'hypothèse de neutralité pour les mutations. Les taux de mutation sont notés $\theta_0$ pour le taux de mutation normal et $\theta_1$ pour le taux de mutation élevé. Le temps est compté en unité de $N$ générations de sorte que $\theta_0/2 = 2N\mu_0$ et $\theta_1/2 = 2N\mu_1$ où $\mu_0$ est le taux de mutation normal par base et par génération et $\mu_1$ le taux de mutation élevé par base et par génération.

Conditionner à l'événement $E \cap M$ revient à supposer que seule la généalogie des séquences du sous-échantillon $\mathcal{B}$, *i.e.* les descendants de l'événement de mutations causant la perte de MMR, est affectée par le taux de mutation $\theta_1$. Ainsi, les mutations suivent un processus de Poisson de paramètre $\theta_1/2$ le long des branches issues de l'événement $\Delta$ et un processus de Poisson de paramètre $\theta_0$ sur les autres branches de l'arbre. Ces notations sont résumées à la Figure 2.4.

Pour l'étude des processus coalescents conditionnels, certaines variables sont particulièrement utiles. Afin de décrire la topologie de l'arbre, nous allons étudier la distribution du nombre d'ancêtres de l'échantillon total dans la généalogie de l'échantillon.

Dans un premier temps, nous allons rappeler la loi du nombre d'individus de l'échantillon sujets à un taux de mutation élevé (*i.e.* la loi de $B$). Cette loi va jouer un rôle fondamental dans la suite de l'étude. Pour des raisons mathématiques, la plupart des résultats seront obtenus conditionnellement à $B = b$. Déterminer la loi de $B$ nous permettra d'intégrer les résultats conditionnels sur l'ensemble des valeurs de $B$.

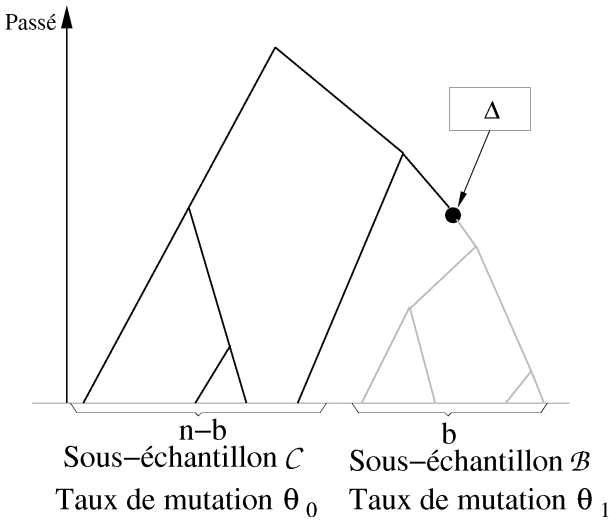



**FIG. 2.4:** *Exemple d'arbre coalescent conditionnel avec $n = 8$ et $b = 4$. Cet arbre satisfait la condition $E \cap M$ : l'ensemble des individus du sous-échantillon $\mathcal{B}$ coalesce avant qu'un individu de $\mathcal{B}$ ne coalesce avec un individu de $\mathcal{C}$. Ainsi la seule portion de l'arbre touchée par le taux de mutation élevé $\theta_1$ est la généalogie de $\mathcal{B}$ jusqu'à $\Delta$.*

### 2.4.2 Le spectre de fréquences

La loi de probabilité pour la variable $B$, est communément appelée *spectre de fréquences*, comme traduction de *frequency spectrum*. L'analyse de cette loi a été faite dans différentes études [Griffiths et Tavaré, 1998], [Stephens, 2000], [Griffiths et Tavaré, 2003]. Nous rappelons cette distribution par la proposition suivante.

**Proposition 1**

$$\mathbb{P}(B = b|E \cap M) = \frac{1}{bH_{n-1}}, \quad b = 1, \ldots, n-1, \tag{2.13}$$

*où* $H_{n-1}$ *est le* $(n-1)^{ème}$ *nombre harmonique :*

$$H_n = \sum_{i=1}^{n-1} \frac{1}{i}.$$

La preuve de cette proposition se trouve dans [Griffiths et Tavaré, 1998], Equation (8.3), page 285.

Nous allons à présent nous intéresser aux effets du conditionnement $E$ sur la topologie de l'arbre de coalescence. Ces effets seront quantifiés sur les variables modélisant le nombre total d'ancêtres à chaque événement de coalescence.

### 2.4.3 Le conditionnement $E$

En tout premier lieu, nous allons rappeler la probabilité de l'événement $E$, qui traduit le fait que l'ensemble des descendants de l'événement perte de MMR coïncide exactement avec les séquences du sous-échantillon $\mathcal{B}$ :

$$\mathbb{P}(E) = \frac{2}{b+1} \begin{pmatrix} n-1 \\ b-1 \end{pmatrix}^{-1}. \tag{2.14}$$

Ce résultat a été démontré par Wiuf et Donnelly [Wiuf et Donnelly, 1999] en utilisant les propriétés markoviennes du coalescent, et par Tavaré [Tavaré, 2004] qui a utilisé l'aspect combinatoire.

Comme nous l'avons fait remarquer précédemment, le conditionnement $E$ affecte uniquement la topologie de la généalogie et n'a aucun effet sur les temps de coalescence. La topologie d'un arbre de coalescence peut se caractériser par le nombre d'ancêtres des

sous-échantillons à chaque événement de coalescence. Nous allons donc introduire certaines notations qui nous permettront d'étudier le nombre d'ancêtres du sous-échantillon $\mathcal{B}$ et du sous-échantillon $\mathcal{C}$.

**Définition 1** ($J_i$) *Conditionnellement à $B = b$, nous définissons $J_r$ ($r = 1, \ldots, b-1$) le nombre total d'ancêtres lorsque le sous-échantillon $\mathcal{B}$ a pour la première fois exactement $r$ ancêtres.*

Cette définition implique que $J_r$ appartient à la plage de valeurs $\{r+1, \ldots, n-b+r\}$. De plus, en notant $J_0$ le nombre total d'ancêtres à l'instant où le sous-échantillon $\mathcal{B}$ coalesce pour la première avec un (ou plusieurs) individus du sous-échantillon $\mathcal{C}$, nous obtenons que :

$$1 \leq J_0 < J_1 < \cdots < J_{b-1} < J_b \equiv n.$$

De façon analogue, nous allons nous intéresser au nombre total d'ancêtres aux temps de coalescence du sous-échantillon $\mathcal{C}$.

**Définition 2** ($K_i$) *Conditionnellement à $B = b$, nous définissons $K_r$ ($r = 1, \ldots, c-1$) le nombre total d'ancêtres lorsque le sous-échantillon $\mathcal{C}$ a pour la première fois exactement $r$ ancêtres.*

Nous obtenons que :

$$K_1 < K_2 < \cdots < K_{c-1} < K_c \equiv n.$$

Enfin une dernière information topologique importante dans notre étude est le nombre total d'ancêtres, noté $J_\Delta$, lorsque la mutation $\Delta$ survient.

**Définition 3** ($J_\Delta$) *Conditionnellement à $B = b$, nous définissons $J_\Delta$ le nombre total d'ancêtres lorsque la mutation $\Delta$ survient.*

Ces notations sont résumées sur un exemple présenté à la Figure 2.5. Nous allons tout débord étudier la loi des variables $J_r$ conditionnellement à $E$. Puis nous nous intéresserons aux lois des variables $K_r$. Enfin, nous déterminerons la loi de $J_\Delta$. Cette dernière loi permettra d'étudier la loi des $J_r$ conditionnellement à $J_\Delta$. L'ensemble des résultats que

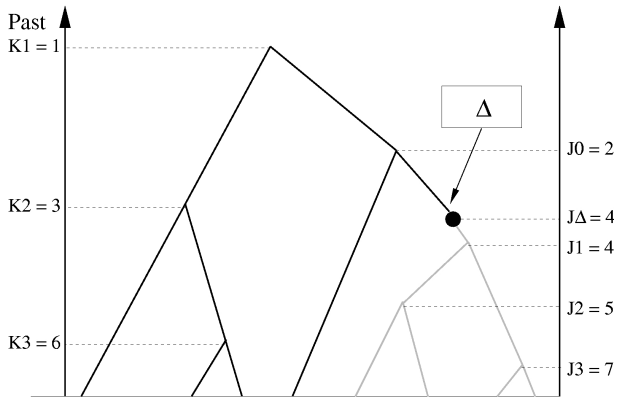



**Fig. 2.5:** *Résumé des notations sur un échantillon de $n = 8$ séquences dont $B = 4$ ont un taux de mutation élevé. Cet exemple nous donne les valeurs du nombre total d'ancêtres à chaque événement de coalescence.*

nous allons montrer nous sera utile d'une part pour étudier les corrections des estimateurs de Watterson et de Tajima et d'autre part pour proposer un algorithme de simulation d'arbre coalescent conditionnel.

Grâce au résultat fourni par l'équation 2.14, concernant la probabilité de $E$, nous pouvons expliciter de nombreuses probabilités conditonnellement à $E$.

En rappelant le fait que dans un arbre de coalescence, les branchements sont choisis au hasard nous avons après simplification ([Tavaré, 2004], équation 8.1.1, page 370) :

$$\mathbb{P}(J_r = j_r, r = b-1, \ldots, 0) = \frac{2b!(b-1)!(n-b)!(n-b-1)!j_0}{n!(n-1)!}. \tag{2.15}$$

En combinant les équations 2.14 et 2.15, nous obtenons la loi jointes des $J_r$, conditionnellement à $E$ ([Tavaré, 2004], équation 8.1.3, page 371) :

$$\mathbb{P}(J_r = j_r, r = b-1, \ldots, 0|E) = j_0 \begin{pmatrix} n \\ b+1 \end{pmatrix}^{-1}. \tag{2.16}$$

En remarquant que l'équation précédente (Equation 2.16) est indépendante des $j_i$ pour $i = b-1, \ldots, 1$, la loi de $J_0$ correspondant au nombre total d'ancêtres lorsque le

sous-échantillon $\mathcal{B}$ coalesce avec une lignée de $\mathcal{C}$, s'écrit conditionnellement à $E$, par le lemme suivant.

**Lemme 1** *Pour tout* $j \in [1, \ldots, n-b+1]$ *:*

$$\mathbb{P}(J_0 = j|E) = j \begin{pmatrix} n-j-1 \\ b-1 \end{pmatrix} \begin{pmatrix} n \\ b+1 \end{pmatrix}^{-1}. \tag{2.17}$$

Ce résultat s'explique par le fait qu'il y a $\begin{pmatrix} n-j-1 \\ b-1 \end{pmatrix}$ façons de choisir les $j_r$ $(r = b-1, \ldots, 2)$.

A partir des équations 2.16 et 2.17, nous obtenons la loi jointes des $J_r$, $r > 1$, conditionnellement à $E$ et $J_0 = j_0$ par le lemme suivant.

**Lemme 2** *Pour tout* $j \in [1, \ldots, n-b+1]$*, nous avons :*

$$\mathbb{P}(J_r = j_r, r = b-1, \ldots, 1|J_0 = j, E) = \begin{pmatrix} n-j-1 \\ b-1 \end{pmatrix}^{-1}. \tag{2.18}$$

Ce résultat nous sera très utile pour étudier l'estimateur de Watterson dans le modèle à deux taux de mutation (voir section 2.5).

Certains résultats similaires peuvent être montrés pour les $K_i$. Nous allons proposer ici un résultat qui nous sera utile dans l'étude de l'estimateur de Tajima dans le modèle à deux taux de mutation (voir section 2.6). Ce résultat nous donne la loi des $K_r$ (*i.e.* le nombre total d'ancêtres aux instants d'événements de coalescence pour les individus de $\mathcal{C}$) conditionnellement à $E$ et $J_0 = j_0$.

**Proposition 2** *Supposons que* $B = b$*, et notons* $c = n - b$*. Pour* $r = j, \ldots, c-1$ *et* $k = r+1, \ldots, r+b$*, nous avons :*

$$\mathbb{P}(K_r = k|J_0 = j, E) = \frac{\begin{pmatrix} k-j-1 \\ r-j \end{pmatrix} \begin{pmatrix} n-k-1 \\ c-r-1 \end{pmatrix}}{\begin{pmatrix} n-j-1 \\ b \end{pmatrix}}.$$

**Preuve.**

Nous pouvons remarquer que le vecteur $(J_0, \ldots, J_{b-1}, K_0, \ldots, K_{n-b-1})$ est obtenu par une permutation du vecteur $1, \ldots, n-1$. Ainsi conditionnellement à $J_0 = j$, le vecteur $J_1, \ldots, J_{b-1}, K_j, \ldots, K_{n-b-1}$ constitue aussi une permutation du vecteur $(j+1, \ldots, n-1)$. Nous pouvons donc en déduire, d'après l'équation 2.15, que :

$$\mathbb{P}(K_r = j_r, r = b-1, \ldots, 1 | J_0 = j, E) = \begin{pmatrix} n-j-1 \\ b \end{pmatrix}^{-1}.$$

De plus, en remarquant que l'expression ci-dessus est indépendante des $k_i$, nous en déduisons que :

$$\begin{aligned}
&\mathbb{P}(K_r = k \mid J_0 = j\,;\, E \cap M) \\
&= \sum_{j<k_j<\cdots<k_{r-1}<r\ r<k_{r+1}<\cdots<k_{c-1}<n} \mathbb{P}(K_r = k_r,\, r = 1, \ldots, c-1 \mid J_0 = j\,;\, E \cap M) \\
&= \frac{\begin{pmatrix} k-j-1 \\ r-j \end{pmatrix} \begin{pmatrix} n-k-1 \\ c-r-1 \end{pmatrix}}{\begin{pmatrix} n-j-1 \\ b \end{pmatrix}}.
\end{aligned}$$

■

Afin d'expliciter complètement la topologie de l'arbre de coalescence conditionnel, il nous reste à déterminer le nombre d'ancêtres de l'échantillon, $J_\Delta$ lorsque l'événement $\Delta$ survient.

Les travaux de Stephens ont permis l'étude de $J_\Delta$ [Stephens, 2000]. En particulier lorsque $\delta$ tend vers 0, où $\delta$ est le taux de mutations pour la mutation particulière $\Delta$, nous obtenons la propriété suivante

**Proposition 3** *Supposons que $B = b$ et conditionnellement à $E$ nous avons, pour $k = n-b+1, \ldots, 2$ :*

$$p_k^\Delta \equiv \mathbb{P}(J_\Delta = k | E) = \begin{pmatrix} n-k \\ b-1 \end{pmatrix} \begin{pmatrix} n-1 \\ b \end{pmatrix}^{-1}. \tag{2.19}$$

Cette propriété est démontrée dans [Stephens, 2000] (Equation (33) page 117).

Ce résultat aide par exemple à retrouver l'âge de la mutation, $\tau_\Delta$ grâce à la formule suivante :

$$\tau_\Delta = 2 \sum_{k=2}^{n-b+1} \frac{n-k+1}{n(k-1)} p_k^\Delta.$$

Cette formule se retrouve dans différentes études, parmi celles-ci, nous pouvons citer [Griffiths et Tavaré, 1998], [Wiuf et Donnelly, 1999] et [Stephens, 2000]. En 2003, une version simplifiée de cette formule a été proposée par [Griffiths et Tavaré, 2003] :

$$\tau_\Delta = \frac{2b}{n-b} \sum_{j=b+1}^{n} \frac{1}{j}.$$

Nous pouvons également obtenir la probabilité conditionnelle pour $J_0$ sachant $J_\Delta = k$, par la proposition suivante. Cette proposition sera utile pour l'algorithme de simulation.

**Proposition 4** *Supposons que $B = b$ et conditionnellement à $E$ et $J_\Delta = k$ ($k = n-b+1, \ldots, 2$), nous avons, pour $j = n-b, \ldots, r-1$ :*

$$\mathbb{P}(J_0 = j | J_\Delta = k, E) = \frac{2j}{k(k-1)}. \tag{2.20}$$

**Preuve :**

En rappelant que les choix de coalescences sont uniformes et par des arguments de combinatoire nous avons, pour $j = n-b, \ldots, r-1$ :

$$\begin{aligned}
\mathbb{P}(J_0 = j | J_\Delta = k, E) &= \frac{\binom{k-1}{2}}{\binom{k}{2}} \frac{\binom{k-2}{2}}{\binom{k-1}{2}} \cdots \frac{\binom{j}{2}}{\binom{j+1}{2}} \frac{j}{\binom{j}{2}} \\
&= \frac{2j}{k(k-1)}.
\end{aligned}$$

■

De façon similaire à la proposition 4, nous pouvons déduire la loi conditionnelle de $J_r$, pour $r > 0$, sachant $J_\Delta = k$. Cette quantité nous sera utile dans l'étude de l'estimateur de Watterson sous le modèle à deux taux de mutation.

**Proposition 5** *Supposons que $B = b$. Soit $r = 1, \ldots, b-1$ et $k \in [2, n-b+1]$. Conditionnellement à $E$ et $J_\Delta = k$, nous avons, pour $j \in [k+r-1, n-b+r]$ :*

$$\mathbb{P}(J_r = j | J_\Delta = k, E) = \frac{\begin{pmatrix} j-k \\ r-1 \end{pmatrix} \begin{pmatrix} n-j-1 \\ b-r-1 \end{pmatrix}}{\begin{pmatrix} n-k \\ b-1 \end{pmatrix}}. \tag{2.21}$$

**Preuve.**

D'après l'équation 2.16, la loi jointe des $J_r$ sachant $E$ et $B = b$ est donnée par :

$$\mathbb{P}(J_r = j_r, r = b-1, \ldots, 0 | E) = j_0 \begin{pmatrix} n \\ b+1 \end{pmatrix}^{-1}.$$

En remarquant que cette expression est indépendante de $j_1, \ldots, j_{r-1}$, nous obtenons que :

$$\begin{aligned} &\mathbb{P}(J_r = j \mid J_\Delta = k\,;\, E \cap M) \\ &= \sum_{k \le j_1 < \cdots < j_{r-1} < j} \mathrm{P}(J_1 = j_1, \ldots, J_{r-1} = j_{r-1}, J_r = j \mid J_\Delta = k;\, E \cap M) \\ &= \begin{pmatrix} j-k \\ r-1 \end{pmatrix} \begin{pmatrix} n-j-1 \\ b-r-1 \end{pmatrix} \begin{pmatrix} n-k \\ b-1 \end{pmatrix}^{-1}. \end{aligned}$$

■

Dans cette partie, nous avons étudié la loi du nombre total d'ancêtres de l'échantillon à chaque événement de coalescence, ainsi qu'au moment de l'événement $\Delta$. Ces lois ont été étudiées conditionnellement à l'événement $E$. Nous allons à présent nous intéresser aux effets du conditionnement $M$ sur l'arbre de coalescence. Rappelons que l'événement $M$ signifie que l'événement $\Delta$ survient une et une seule fois dans la généalogie.

### 2.4.4 Le conditionnement $E \cap M$

Dans cette section, nous allons nous intéresser à l'effet du conditionnement $M$ sur la longueur des branches de l'arbre. Nous rappelons ici que les temps inter-coalescences, pour un échantillon de taille $n$, sont notés $X_i$, $i = 2, \ldots, n$. De plus, dans le modèle

coalescent neutre (voir section 2.2.2), les $X_i$ suivent une loi exponentielle de paramètre $\lambda_i$ où :

$$\lambda_i = \frac{i(i-1)}{2}.$$

Nous proposons alors le théorème suivant qui nous donne la loi jointe des temps de coalescence sous l'hypothèse $E \cap M$, et conditionnellement à $B = b$.

**Théorème 1** *Supposons que la mutation $\Delta$ a $B = b$ descendants. La loi jointe des temps inter-coalescences, $(X_2, \ldots, X_n)$, conditionnellement à $E \cap M$ est donnée par :*

$$f(x_2, \ldots, x_n) = \sum_{k=2}^{n-b+1} p_k^{\Delta} \lambda_k x_k \prod_{\ell=2}^{n} f_\ell(x_\ell), \tag{2.22}$$

*où, $f_\ell(x_\ell)$ est la densité d'une variable de loi exponentielle de taux $\lambda_\ell$.*

**Preuve.**

Cette preuve est inspirée des travaux de [Tavaré, 2004] (Chap. 8, p. 110).

Nous posons $s = (s_2, \ldots, s_n)$ et $X = (X_2, \ldots, X_n)$. Rappelons que :

$$s.X = \sum_{i=2}^{n} s_i X_i.$$

Conditionnellement à $E$, la transformée de Laplace multidimensionnelle est égale à :

$$\begin{aligned}
\mathbb{E}[e^{-s.X} \mid E] &= \sum_{k=2}^{n-b+1} \mathbb{E}[e^{-s.X} \mathbb{1}_{J_\Delta = k} \mid E] \\
&= \sum_{k=2}^{n-b+1} \mathbb{E}[\mathbb{E}[e^{-s.X} \mathbb{1}_{J_\Delta = k} \mid X, E]] \\
&= \sum_{k=2}^{n-b+1} \mathbb{E}[e^{-s.X} \mathbb{P}(J_\Delta = k \mid X, E)] \\
&= \sum_{k=2}^{n-b+1} \mathbb{E}[e^{-s.X} X_k \frac{\delta}{2} e^{-\frac{L_n \delta}{2}}] \lambda_k \begin{pmatrix} n-k \\ b-1 \end{pmatrix} \begin{pmatrix} n \\ b+1 \end{pmatrix}^{-1} \\
&= \frac{\delta}{2} \sum_{k=2}^{n-b+1} \mathbb{E}[e^{s.X} X_k] \lambda_k \begin{pmatrix} n-k \\ b-1 \end{pmatrix} \begin{pmatrix} n \\ b+1 \end{pmatrix}^{-1} + o(\delta) \\
&= \frac{\delta}{2} \sum_{k=2}^{n-b+1} \mathbb{E}[e^{-s.X}] \frac{\lambda_k}{s_k + \lambda_k} \begin{pmatrix} n-k \\ b-1 \end{pmatrix} \begin{pmatrix} n \\ b+1 \end{pmatrix}^{-1} + o(\delta).
\end{aligned}$$

En conditionnant à $M$, et en considérant le cas où $\delta$ tend vers 0 nous obtenons que :

$$\mathbb{E}[e^{-s.X} \mid E \cap M] = \sum_{k=2}^{n-b+1} \mathbb{E}[e^{-s.X}] \frac{\lambda_k}{s_k + \lambda_k} \frac{\begin{pmatrix} n-k \\ b-1 \end{pmatrix}}{\begin{pmatrix} n-1 \\ b \end{pmatrix}}.$$

Pour conclure, il suffit de remarquer que la transformée de Laplace multidimensionnelle de la densité donnée à l'équation 2.22 coïncide parfaitement avec l'expression ci-dessus.

■

D'après le théorème 1, nous pouvons déduire l'espérance des temps de coalescence sous un modèle de coalescence conditionnel, donné par le corollaire suivant

**Corollaire 1** *Supposons que la mutation $\Delta$ a $B = b$ descendants. Nous avons, pour $\ell = 2, \ldots, n$ :*

$$\mathbb{E}[X_\ell | E \cap M] = \begin{cases} \left(1 + p_\ell^\Delta\right)/\lambda_\ell & \text{si } \ell \leq n - b + 1 \\ 1/\lambda_\ell & \text{sinon.} \end{cases}$$

**Preuve :** En supposant que l'événement $\Delta$ a $B = b$ descendants, nous pouvons déduire du théorème 1 les distributions marginales des temps inter-coalescence.

Deux cas peuvent être distingués. Tout d'abord, pour $\ell = 2, \ldots, n-b+1$, nous avons :

$$f(x_\ell) = \left( \sum_{k=2, k \neq \ell}^{n-b+1} p_k^\Delta + p_\ell^\Delta \lambda_\ell x_\ell \right) f_\ell(x_\ell).$$

Puis pour $\ell = n - b + 2, \ldots, n$, nous avons :

$$f(x_\ell) = f_\ell(x_\ell).$$

Dans les deux expressions précédentes, $f_\ell$ représente la densité d'une variable de loi exponentielle de paramètre $\lambda_\ell$. L'espérance des distributions ci-dessus nous donne :

$$\mathbb{E}[X_\ell | E \cap M] = \begin{cases} \left(1 + p_\ell^\Delta\right)/\lambda_\ell & \text{si } \ell \leq n - b + 1 \\ 1/\lambda_\ell & \text{sinon.} \end{cases}$$

■

En conséquence du théorème 1 et du corollaire 1, nous pouvons remarquer que, conditionnellement à $E \cap M$, les temps inter-coalescence $X_i$ ne sont plus indépendants. Cependant, d'après le théorème 1, le conditionnement à $J_\Delta = k$ (*i.e.* le nombre total d'ancêtres lorsque la mutation $\Delta$ survient vaut $k$), nous donne que la variable $X_k$ suit une loi Gamma de paramètres 2 et $\lambda_k$, tandis que pour $\ell = 2, \ldots, n$, tel que $\ell \neq k$, les variables $X_\ell$ suivent une loi Gamma de paramètres 1 et $\lambda_\ell$ (*i.e.* une loi exponentielle de paramètre $\lambda_\ell$).

Dans cette section, nous avons donc déterminé la loi du nombre d'ancêtres à chaque événement de coalescence. Nous nous sommes également intéressés à la loi jointe des temps de coalescence.

### 2.4.5 Simulation du modèle à deux taux de mutation

Cette étude théorique sur la topologie des arbres coalescents conditionnels ainsi que sur la distribution des temps inter-coalescence nous a permis de fournir un algorithme de simulation de ce type d'arbre. L'algorithme proposé se décompose en six étapes spécifiées ci-dessous.

1. Le nombre d'individus, $B$, touchés par la mutation $\Delta$ est choisi selon la loi du *spectre de fréquences*, donnée par l'équation 2.13 :

$$\mathbb{P}(B = b | E \cap M) = \frac{1}{bH_{n-1}}.$$

2. Le nombre total d'ancêtres au moment de l'événement $\Delta$ est choisi selon la loi décrite à l'équation 2.19 :

$$\mathbb{P}(J_\Delta = k | E \cap M) = \begin{pmatrix} n-k \\ b-1 \end{pmatrix} \begin{pmatrix} n-1 \\ b \end{pmatrix}^{-1}.$$

3. Le nombre total d'ancêtres lorsque le sous-échantillon $\mathcal{B}$ coalesce avec une lignée du sous-échantillon $\mathcal{C}$ est alors déterminé conditionnellement à $J_\Delta = k$, selon l'équation 2.20 :

$$\mathbb{P}(J_0 = j | J_\Delta = k, E \cap M) = \frac{2j}{k(k-1)}.$$

4. Le nombre total d'ancêtres au moment des coalescences du sous-échantillon $\mathcal{B}$, représenté par les variables $J_i$, est ensuite choisi. Pour cela, $b-1$ valeurs sont choisies uniformément dans l'intervalle $[J_0, \ldots, n-1]$. Ces valeurs sont ensuite ordonnées.

5. Le nombre total d'ancêtres au moment des coalescences du sous-échantillon $\mathcal{C}$ est choisi en prenant le complémentaire des $J_i$ dans $[1, \ldots, n]$.

6. Les temps de coalescence sont alors simulés selon les lois Gamma respectives.

Des exemples de simulations sont proposés à la figure 2.6.

Nous allons à présent nous intéresser à l'estimation du taux de mutation élevé, $\theta_1$. Rappelons tout d'abord que dans notre modélisation les mutations suivent un processus de Poisson le long des branches de l'arbre. Plus spécifiquement, les mutations suivent un processus de Poisson de paramètre $\theta_1/2$ pour toutes les branches qui descendent de l'événement de mutation $\Delta$ et un processus de Poisson de paramètre $\theta_0/2$ sur toutes les autres branches. Ces notations sont résumées à la figure 2.4.

Comme nous l'avons vu à la section 2.3, dans un contexte génétique, l'estimation d'un taux de mutation est souvent associée à une mesure de la diversité génétique dans l'échantillon. Deux mesures de diversité sont couramment utilisées. La première mesure consiste à compter le nombre de sites de ségrégation. Dans le cadre du modèle à infinité de sites, cette mesure est directement proportionnelle au taux de mutation et conduit à l'estimateur de Watterson [Watterson, 1975]. La deuxième mesure de diversité génétique qui nous a intéressé consiste à compter le nombre de bases différentes, considérées sur l'ensemble des paires d'individus. Sous l'hypothèse du modèle à infinité de sites, la moyenne de cette mesure produit un estimateur du taux de mutations. Cet estimateur a été étudié en détail par Tajima [Tajima, 1983].

Dans cette section nous allons étudier ces deux mesures de diversité génétique dans le modèle à deux taux de mutations présenté précédemment. Plus précisemment, nous allons nous focaliser sur la correction du biais de l'estimateur de Watterson puis de l'estimateur de Tajima.

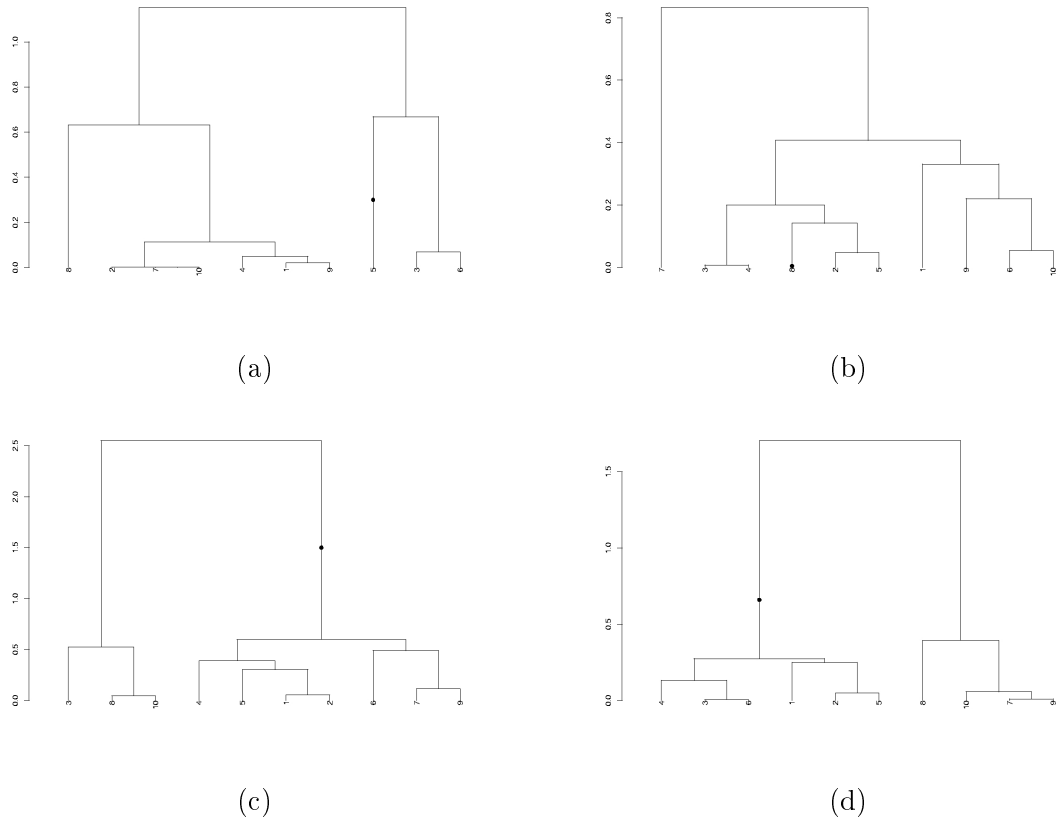


**Fig. 2.6:** *Résultats de simulations utilisant l'algorithme proposé. L'événement $\Delta$ est matérialisé par un point noir sur une branche. Chaque simulation a été faite pour $n = 10$. Dans le cas (a) et (b), nous avons obtenu $B = 1$. Pour (a), lorsque la mutation est apparue, l'échantillon comptait $J_\Delta = 4$ ancêtres. Dans le cas (b), l'échantillon comptait $J_\Delta = 9$ ancêtres. Pour la simulation (c), nous avons obtenu $B = 2$ et $J_\Delta = 7$. Enfin pour la simulation (d), nous avons obtenu $B = 6$ et $J_\Delta = 2$.*

## 2.5 Correction de l'estimateur de Watterson

Certaines propriétés de l'estimateur de Watterson, dans le cas d'un modèle à un seul taux, ont été rappelées à la section 2.3.3.

### 2.5.1 L'estimateur de Watterson sous le modèle à deux taux de mutation

Afin de corriger l'estimateur de Watterson, nous allons déterminer la loi du nombre de sites de ségrégation dans le modèle à deux taux de mutation. De nouvelles notations vont être introduites par l'intermédiaire des quatre définitions suivantes.

**Définition 4** *Nous notons $L_n^{\Delta}$ la longueur totale d'un arbre de coalescence conditionnel.*

**Définition 5** *La longueur de la généalogie du sous-échantillon $\mathcal{B}$, est notée $L_n^1$. Cette longueur est considérée jusqu'à la racine de cette sous-généalogie, c'est-à-dire jusqu'à l'ancêtre commun le plus récent du sous échantillon $\mathcal{B}$.*

**Définition 6** *En empruntant les notations introduites par [Wiuf et Donnelly, 1999], nous notons $\eta_n$ le temps séparant la racine de la généalogie du sous-échantillon $\mathcal{B}$ et l'événement de mutation $\Delta$.*

**Définition 7** *Notons $L_n^0$ la longueur totale des branches sujettes à un taux de mutation normal, $\theta_0$. Nous pouvons remarquer que :*

$$L_n^0 = L_n^{\Delta} - L_n^1 - \eta_n.$$

Les notations introduites dans les définitions précédentes sont résumées sur un exemple proposé à la figure 2.7.

Rappelons que sous le modèle à infinité de sites, le nombre de sites de ségrégation est égal au nombre total de mutations. Ainsi, dans le modèle à deux taux de mutation, le nombre de sites de ségrégation, S, peut être vu comme la somme de deux termes :

$$S = S^0 + S^1,$$

où $S^0$ représente le nombre de mutations survenues avec un taux $\theta_0$ et $S^1$ avec un taux $\theta_1$. Nous en déduisons donc que $S^0$ suit une loi de Poisson de paramètre $L_0\theta_0/2$ tandis que

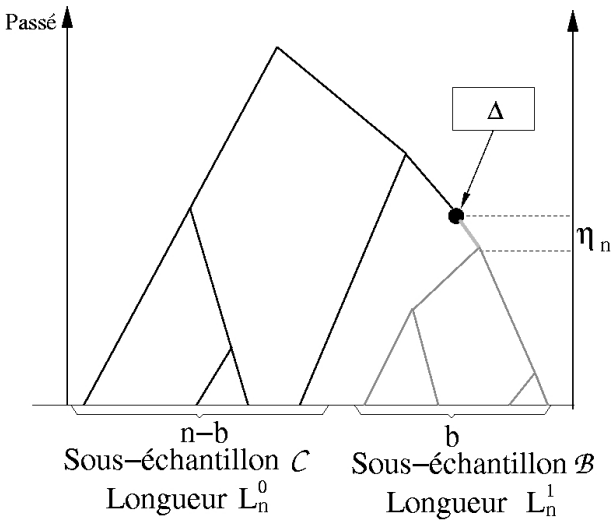



**FIG. 2.7:** *La longueur, jusqu'à sa racine, de la généalogie du sous-échantillon $\mathcal{B}$ est notée $L_n^1$. Le temps entre la racine de cette sous-généalogie et l'événement de mutation $\Delta$ est noté $\eta_n$. La longueur du reste de l'arbre est noté $L_n^0$, avec la propriété que $L_n^0 = L_n^\Delta - L_n^1 - \eta_n$. La longueur $L_n^1$ est représentée en gris foncé sur la figure, la longueur $\eta_n$ en gris clair et la longueur $L_n^0$ en noir.*

$S^1$ suit une loi de Poisson de paramètre $(L_n^1+\eta_n)\theta_1/2$. Ainsi la variable $S_O$ suit une loi de Poisson de paramètre $L_n^0\theta_0/2$ et $S_1$ suit une loi de Poisson de paramètre $(L_n^1+\eta_n)\theta_0/2$.

Ces notations nous permettent de fournir un estimateur sans biais du taux de mutation élevé, $\theta_1$, par le théorème suivant.

**Théorème 2** *Notons :*

$$B_n = \frac{1}{2}(\mathbb{E}[L_n^1] + \mathbb{E}[\eta_n])$$

*et*

$$A_n = \frac{1}{2}[L_n^\Delta] - B_n.$$

*Soit $\hat{\theta_1}$ la quantité définie par :*

$$\hat{\theta_1} = \frac{S - A_n\theta_0}{B_n}.$$

*Nous avons alors :*

$$\mathbb{E}[\hat{\theta_1}] = \theta_1.$$

**Preuve.**

Nous avons, par définition de $\hat{\theta_1}$ :

$$\mathbb{E}[\hat{\theta_1}] = \frac{\mathbb{E}[S] - A_n\theta_0}{B_n}. \tag{2.23}$$

De plus :

$$\mathbb{E}[S] = \mathbb{E}[S^0] + \mathbb{E}[S^1]. \tag{2.24}$$

Comme $S^0$ (resp. $S^1$) suit une loi de Poisson de paramètre $L_n^0\theta_0/2$ (resp. $(L_n^1 + \eta_n)\theta_0/2$), nous avons :

$$\mathbb{E}[S^0] = \mathbb{E}[L_n^0]\theta_0/2 = (\mathbb{E}[L_n^\Delta] - \mathbb{E}[L_n^1] - \mathbb{E}[\eta_n])\theta_0/2 \tag{2.25}$$

et

$$\mathbb{E}[S^1] = (\mathbb{E}[L_n^1] + \mathbb{E}[\eta_n])\theta_1/2. \tag{2.26}$$

En substituant les équations 2.25 et 2.26 dans l'équation 2.24, elle-même substituée dans l'équation 2.23, nous obtenons que :

$$\mathbb{E}[\hat{\theta_1}] = \theta_1$$

■

L'utilisation pratique de cet estimateur nécessite cependant la connaissance des quantités : $\mathbb{E}[L_n^{\Delta}]$, $\mathbb{E}[L_n^1]$ et $\mathbb{E}[\eta_n]$. Ces quantités permettent le calcul des coefficients $A_n$ et $B_n$.

### 2.5.2 Calcul des coefficients $A_n$ et $B_n$

Afin de proposer un calcul des coefficients correctifs $A_n$ et $B_n$, nous allons formuler explicitement des expressions de $\mathbb{E}[L_n^{\Delta}]$, $\mathbb{E}[L_n^1]$ et $\mathbb{E}[\eta_n]$. L'expression de $\mathbb{E}[L_n^{\Delta}]$ est donnée par la proposition 6. La quantité $\mathbb{E}[L_n^1]$, conditionnellement à $B = b$, est explicitée par l'intermédiaire de la proposition 7. Enfin, l'expression de $\mathbb{E}[\eta_n]$, conditionnellement à $B = b$, est donnée par la proposition 8.

Commençons par l'expression de l'espérance de $L_n^{\Delta}$, la longueur totale de l'arbre de coalescence conditionnel.

**Proposition 6** *Nous avons :*

$$\frac{1}{2}\mathbb{E}[L_n^{\Delta}] = H_{n-1} + \frac{1}{H_{n-1}} \sum_{b=1}^{n-1} \sum_{k=2}^{n-b+1} \frac{p_k^{\Delta}}{b(k-1)}.$$

**Preuve.**

$L_n^{\Delta}$ représente la longueur totale de l'arbre qui peut également s'écrire de la façon suivante :

$$L_n^{\Delta} = \sum_{k=2}^{n} kX_k$$

D'après le corollaire 1, nous avons, conditionnellement à $B = b$ :

$$E(X_\ell) = \begin{cases} \left(1 + p_\ell^{\Delta}\right)/\lambda_\ell & \text{si } \ell \leq n - b + 1 \\ 1/\lambda_\ell & \text{sinon.} \end{cases}$$

Ainsi, nous obtenons que :

$$\begin{aligned}\mathbb{E}[L_n^{\Delta}|B=b] &= \sum_{k=2}^{n} k\mathbb{E}[X_k|B=b] \\ &= \sum_{k=2}^{n-b+1} 2\frac{(1+p_k^{\Delta})}{(k-1)} + \sum_{k=n-b+2}^{n} 2\frac{1}{k-1} \\ &= 2H_{n-1} + 2\sum_{k=2}^{n-b+1} \frac{p_k^{\Delta}}{k-1}\end{aligned}$$

Nous en déduisons, en sommant sur l'ensemble des $b$ grâce à la formule des probabilités totales, que :

$$\begin{aligned}\frac{1}{2}\mathbb{E}[L_n^{\Delta}] &= \sum_{b=1}^{n-1} \mathbb{E}[L_n^{\Delta}|B=b]P(B=b) \\ &= H_{n-1} + \sum_{b=1}^{n-1}\sum_{k=2}^{n-b+1} \frac{p_k^{\Delta}}{k-1}\frac{1}{bH_{n-1}} \\ &= H_{n-1} + \sum_{b=1}^{n-1}\sum_{k=2}^{n-b+1} \frac{1}{H_{n-1}}\frac{p_k^{\Delta}}{b(k-1)}.\end{aligned}$$

■

Nous pouvons à présent nous attacher à expliciter la formule, conditionnellement à $B=b$, de l'espérance de $L_n^1$, la longueur du sous-arbre de coalescence du sous-échantillon $\mathcal{B}$.

**Proposition 7** *Nous avons pour* $b = 1\ldots, n-1$ *:*

$$\mathbb{E}[L_n^1|B=b] = \sum_{j=2}^{n-b+1} p_j^{\Delta} \sum_{k=j+1}^{n} \frac{2}{k(k-1)} c_{jk},$$

*où*

$$c_{jk} = b - (b-1)\frac{n-k}{n-j} - \frac{(n-k)!(n-j-b+1)!}{(n-j)!(n-k-b+1)!},$$

*pour* $j = 2, \ldots, n-b+1$ *et* $k = j+1, \ldots, n$.

Le résultat de cette propriété a été démontré par Griffiths et Tavaré (Paragraphe 6, pages 405-406, Equation 6.7) [Griffiths et Tavaré, 2003].

Enfin, la propriété suivante nous donne l'expression, conditionnellement à $B=b$, de $\eta_n$, le temps séparant l'âge de l'ancêtre le plus récent du sous-échantillon $\mathcal{B}$ de l'événement de mutation $\Delta$.

| $n$ | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
|---|---|---|---|---|---|---|---|---|---|---|
| $A_n$ | 2.171 | 2.693 | 3.024 | 3.265 | 3.455 | 3.612 | 3.747 | 3.864 | 3.967 | 4.061 |
| $B_n$ | 0.595 | 0.68 | 0.713 | 0.732 | 0.746 | 0.756 | 0.764 | 0.771 | 0.776 | 0.781 |

**Tab. 2.1: Table des coefficients correctifs $A_n$ et $B_n$.** *Ce tableau récapitule les valeurs numériques des coefficients correctifs $A_n$ et $B_n$ intervenant dans la statistique $\hat{\theta_1} = (S - A_n\theta_0)/B_n$ pour $n$ dans l'intervalle $5 - 50$.*

**Proposition 8** *Nous avons, pour $b = 1, \ldots, n-1$ :*

$$\mathbb{E}[\eta_n | B = b] = 2 \sum_{k=2}^{n-b+1} \frac{p_k^{\Delta}}{k}.$$

Le résultat de cette propriété a été démontré par Wiuf et Donnelly (paragraphe 4, page 192, Equation 25) [Wiuf et Donnelly, 1999].

Ainsi en sommant sur $b$, les équations des propriétés 7 et 8, nous obtenons les expressions littérales de $\mathbb{E}[L_n^1]$ et $\mathbb{E}[\eta_n]$. Ces expressions, associées à la propriété 6, nous permettent d'obtenir une expression des coefficients $A_n$ et $B_n$ intervenant dans la définition de $\hat{\theta_1}$ (voir théorème 2).

L'expression des coefficients $A_n$ et $B_n$ n'ayant, en elle-même, que peu d'intérêt, nous avons choisi de reporter une table de valeurs de ces coefficients (voir Tableau 2.5.2).

### 2.5.3 Performances de $\hat{\theta_1}$

Les performances de l'estimateur $\hat{\theta_1}$ ont été testées sur des données simulées. Ces données ont été obtenues grâce à l'algorithme proposé à la section 2.4.5. Le taux de mutation normal a été fixé à 1 : $\theta_0 = 1$. Le choix de cette valeur s'explique par le fait que le taux de mutation normal a été estimé à $\mu_0 \approx 10^{-10}$. La valeur de $\theta_0 = 1$ correspond donc à un nombre total de cellules égal à $N = 2.5 \times 10^9$ dans le tissu considéré, ce qui constitue un ordre de grandeur tout à fait acceptable. De plus, notre approche se focalise sur l'estimation du rapport entre $\theta_0$ et $\theta_1$, ce qui justifie également la normalisation du paramètre $\theta_0$.

| | $\theta_1 = 10$ | | $\theta_1 = 100$ | | $\theta_1 = 1,000$ | |
|---|---|---|---|---|---|---|
| $n$ | $\mathbb{E}$ | Ecart-type | $\mathbb{E}$ | Ecart-type | $\mathbb{E}$ | Ecart-type |
| 10 | 9.9 | 12.0 | 97.4 | 112.4 | 947.5 | 1109.7 |
| 20 | 10.3 | 12.5 | 99.7 | 122.4 | 991.9 | 1211.1 |
| 30 | 10.2 | 12.8 | 102.9 | 126.1 | 1060.3 | 1286.1 |
| 40 | 10.2 | 13.2 | 100.9 | 128.9 | 1018.2 | 1286.2 |
| 50 | 10.4 | 13.5 | 102.0 | 131.7 | 1045.7 | 1235.9 |

**Tab. 2.2: Résultats pour $\hat{\theta_1}$.** *Le biais et l'écart-type de $\hat{\theta_1}$ ont été calculés à partir de 10.000 simulations pour des échantillons de tailles : $n = 10 - 50$. Le taux de mutation normal a été fixé à $\theta_0 = 1$.*

Trois valeurs pour le taux de mutation élevé ont été considérées : $\theta_1 = 10, 10^2, 10^3$. De plus nous nous sommes intéressés aux tailles d'échantillon suivantes : $n = 10, 20, 30, 40, 50$. Le biais et l'écart type ont été estimés empiriquement sur 10000 simulations dans chacun des cas. Les résultats ont été reportés dans le tableau 2.2.

Les résultats obtenus nous confirment que l'estimateur $\hat{\theta_1}$ est sans biais. Cependant, les écart-types sont assez élevées. Ceci s'explique par le fait que les distributions empiriques montrent une queue de distribution lourde. De plus, il semble que l'erreur soit due à un terme d'ordre $\theta_1^2$. Nous avons ajusté un modèle de régression de la forme $\alpha_n\theta_1 + \beta_n\theta_1^2$ à la variance. Pour $n = 20$, l'ajustement est presque parfait avec $\text{Var} = 1.47\theta_1^2$. Nous avons en effet obtenu $R^2 = 0.999$ et une $p - valeur < 10^{-12}$. Pour $n = 40$, l'ajustement nous donne $\text{Var} = 1.68\theta_1^2$ avec $R^2 = 0.997$ et $p - valeur < 10^{-12}$. En conséquence, l'écart-type n'exhibe pas une forte décroissance lorsque la taille de l'échantillon augmente.

## 2.6 Correction de l'estimateur de Tajima

Dans cette section, nous allons mener une étude similaire pour corriger l'estimateur de Tajima sous le modèle à deux taux de mutation. L'estimateur de Tajima est défini comme le nombre moyen de différences entre deux séquences de l'échantillon. Certaines

propriétés de l'estimateur de Tajima, dans le cas d'un modèle à un seul taux de mutation, ont été rappelées à la section 2.3.4.

### 2.6.1 L'estimateur de Tajima sous le modèle à deux taux de mutation

Afin de corriger l'estimateur de Tajima, nous allons déterminer la loi du nombre de différences entre les bases de deux séquences. Trois cas particuliers doivent être distingués. Nous allons, tout d'abord, nous intéresser au nombre de différences entre deux séquences appartenant toutes les deux au sous-échantillon $\mathcal{B}$, puis entre deux séquences appartenant, toutes les deux, au sous-échantillon $\mathcal{C}$. Le dernier cas est celui pour lequel une des deux séquences appartient à $\mathcal{B}$ et l'autre à $\mathcal{C}$.

En tout premier lieu, certaines notations doivent être introduites par l'intermédiaire des trois définitions suivantes.

**Définition 8** *Nous notons $\tau_{\mathcal{B}}$ le temps moyen de coalescence entre deux séquences appartenant au sous-échantillon $\mathcal{B}$.*

**Définition 9** *Nous notons $\tau_{\mathcal{C}}$ le temps moyen de coalescence entre deux séquences appartenant au sous-échantillon $\mathcal{C}$.*

**Définition 10** *Nous notons $\tau_{\mathcal{B},\mathcal{C}}$ le temps moyen de coalescence entre une séquence appartenant au sous-échantillon $\mathcal{B}$ et une séquence appartenant au sous-échantillon $\mathcal{C}$.*

Nous pouvons également rappeler que $\tau_\Delta$ caractérise l'âge moyen de la mutation, c'est-à-dire, le temps moyen nécessaire à remonter pour atteindre l'événement de mutation $\Delta$.

Ces définitions nous permettent d'introduire un estimateur, $\Pi_1$, du taux de mutation élevé, $\theta_1$. Le théorème suivant nous montre que $\Pi_1$ est sans biais.

**Théorème 3** *Posons :*

$$\Pi_1 = \frac{\Pi - C_n\theta_0}{D_n}$$

*tel que les coefficients $C_n^b$ (correspondant au coefficient $C_n$ sachant $B = b$) et $D_n^b$ (correspondant au coefficient $D_n$ sachant $B = b$) sont définis par :*

$$C_n^b = \tau_{\mathcal{C}}\left(1 - \frac{b}{n}\right)^2 + (2\tau_{\mathcal{B},\mathcal{C}} - \tau_\Delta)\frac{2b(n-b)}{n^2}$$

*et*

$$D_n^b = \tau_{\mathcal{B}} \left(\frac{b}{n}\right)^2 + \tau_{\Delta} \frac{2b(n-b)}{n^2}.$$

$\Pi$ *représente le nombre moyen de différences paire à paire entre deux séquences de l'échantillon.*

*Nous avons alors :*

$$\mathbb{E}[\Pi_1] = \theta_1.$$

**Preuve.**

Nous allons tout d'abord montrer que $\mathbb{E}[\Pi_1|B = b] = \theta_1$.

Conditionnellement à $B = b$, nous avons, par définition de $\Pi_1$ :

$$\mathbb{E}[\Pi_1|B = b] = \frac{\mathbb{E}[\Pi|B = b] - C_n^b \theta_0}{D_n^b}. \tag{2.27}$$

Le terme $\mathbb{E}[\Pi]$ peut se décomposer en trois termes :

$$\mathbb{E}[\Pi|B = b] = \mathbb{E}[\Pi_{\mathcal{B}}] \left(\frac{b}{n}\right)^2 + \mathbb{E}[\Pi_{\mathcal{C}}] \left(1 - \frac{b}{n}\right)^2 + \mathbb{E}[\Pi_{\mathcal{B},\mathcal{C}}] \frac{2b(n-b)}{n^2}. \tag{2.28}$$

$\mathbb{E}[\Pi_{\mathcal{B}}]$ représente le nombre moyen de différences entre deux séquences issues du sous-échantillon $\mathcal{B}$ sachant $B = b$. Ce cas apparaît avec une probabilité $(b/n)^2$. $\mathbb{E}[\Pi_{\mathcal{C}}]$ représente le nombre moyen de différences entre deux séquences issues du sous-échantillon $\mathcal{C}$ sachant $B = b$. Ce cas apparaît avec une probabilité $(1-b/n)^2$. Le dernier terme $\mathbb{E}[\Pi_{\mathcal{B},\mathcal{C}}]$ correspond au cas d'un séquence issue de $\mathcal{B}$ et l'autre de $\mathcal{C}$ sachant $B = b$. Ce cas survient avec une probabilité $2b(n-b)/n^2$.

Ainsi, d'après les notations introduites concernant les temps moyens de coalescence entre deux séquences (Définitions 8, 9 et 10) et le fait que les mutations suivent un processus de Poisson le long des branches, nous avons :

$$\mathbb{E}[\Pi_{\mathcal{B}}] = \tau_{\mathcal{B}} \theta_1, \tag{2.29}$$

$$\mathbb{E}[\Pi_{\mathcal{B},\mathcal{C}}] = (2\tau_{\mathcal{B},\mathcal{C}} - \tau_{\Delta})\theta_0 + \tau_{\Delta}\theta_1, \tag{2.30}$$

$$\mathbb{E}[\Pi_{\mathcal{C}}] = \tau_{\mathcal{C}} \theta_0. \tag{2.31}$$

Ainsi, en combinant les équations 2.29, 2.30, 2.31 et 2.28, nous obtenons que :

$$\mathbb{E}[\Pi|B=b] = C_n^b\theta_0 + D_n^b\theta_1.$$

Ceci nous permet, d'après l'équation 2.27, d'obtenir le résultat attendu :

$$\mathbb{E}[\Pi_1|B=b] = \theta_1.$$

Nous en déduisons donc, en intégrant $b$ sur la loi du spectre de fréquence que :

$$\mathbb{E}[\Pi_1] = \theta_1.$$

■

### 2.6.2 Calcul des coefficients $C_n$ et $D_n$

L'utilisation pratique de l'estimateur de Tajima corrigé nécessite de connaître l'expression des quantités $\tau_{\mathcal{B}}$, $\tau_{\mathcal{B},\mathcal{C}}$ et $\tau_{\mathcal{C}}$. Ces expressions sont données respectivement par les propriétés suivantes : propriété 9, pour $\tau_{\mathcal{B}}$, propriété 10, pour $\tau_{\mathcal{C}}$, et les propriétés 11-12 pour l'expression de $\tau_{\mathcal{B},\mathcal{C}}$.

Avant de proposer les expressions de $\tau_{\mathcal{B}}$, $\tau_{\mathcal{C}}$ et $\tau_{\mathcal{B},\mathcal{C}}$, nous allons nous intéresser au temps nécessaire pour que l'échantillon ait $k$ ancêtres.

**Définition 11** *Nous notons $T_j$, pour $j = 2,\dots,n$, le temps par lequel l'échantillon a pour la première fois $j-1$ ancêtres. Ainsi :*

$$T_j = X_n + \cdots + X_j, \quad j = 2,\dots,n.$$

Dans le lemme suivant, nous allons établir la loi sachant $B=b$ des variables $T_j$. Ce résultat sera déterminant pour les démonstrations à venir.

**Lemme 3** *Conditionnellement à $B=b$, et pour $k = 2,\dots,n-b+1$, nous avons :*

$$\mathbb{E}[T_j|J_\Delta = k] = \begin{cases} \frac{2(n-j+1)}{n(j-1)} & \text{si } j \geq k \\ \frac{2(n-j+1)}{n(j-1)} + \frac{2}{k(k-1)} & \text{sinon.} \end{cases}$$

**Preuve.** Ce résultat se déduit du corollaire 1, qui nous donne, conditionnellement à $B = b$ :

$$\mathbb{E}[X_\ell | J_\Delta = k] = \begin{cases} \frac{2}{\ell(\ell-1)} & \text{si } \ell \neq k \\ 2\frac{2}{\ell(\ell-1)} & \text{sinon.} \end{cases}$$

En considérant dans un premier temps $j \geq k$, nous avons :

$$\mathbb{E}[T_j | J_\Delta = k] = \sum_{\ell=j}^{n} \mathbb{E}[X_\ell | J_\Delta = k].$$

Par une récurrence simple nous pouvons montrer que, si $j \geq k$ alors nous avons :

$$\mathbb{E}[T_j | J_\Delta = k] = \frac{2(n-j+1)}{n(j-1)}.$$

Dans le cas où $j < k$, nous pouvons remarquer que, conditionnellement à $J_\Delta = k$, le temps de coalecence $X_k$ est, en espérance, doublé. Ainsi, comme les autres temps restent, en espérance, inchangés, nous obtenons que pour $j < k$ :

$$\mathbb{E}[T_j | J_\Delta = k] = \frac{2(n-j+1)}{n(j-1)} + \frac{2}{k(k-1)}.$$

■

**Cas 1 : Coalescence au sein de $\mathcal{B}$**

Nous allons commencer par expliciter la formule de $\tau_\mathcal{B}$, représentant l'espérance du temps de coalescence pour deux séquences du sous-échantillon $\mathcal{B}$.

**Lemme 4** *Sachant que $B = b$, nous avons :*

$$\tau_\mathcal{B} = \frac{b+1}{b-1} \sum_{r=1}^{b-1} \frac{2}{(r+1)(r+2)} \mathbb{E}[T_{J_r+1}],$$

*où $T_{j+1} = X_n + \cdots + X_{j+1}$.*

**Preuve.**

Par définition, $T_{j+1}$ correspond au temps minimum nécessaire pour que l'échantillon ait $j$ ancêtres. Ainsi, nous pouvons remarquer que si un nœud de l'arbre apparaît lorque l'échantillon a $J$ ancêtres, alors, le temps moyen de ce nœud vaut $\mathbb{E}[T_{J+1}]$.

Lorsque $B = b$, nous pouvons constater que le sous-arbre de la généalogie de $\mathcal{B}$ contient $b-1$ nœuds, ou événements de coalescence. De plus, lorsque deux individus de $\mathcal{B}$ coalescent au nœud $r$, le nombre total d'ancêtres vaut $J_r$. Notons $\mathrm{P}_r = \mathbb{P}(2 \text{ individus coalescent au nœud } r)$. En conditionnant par rapport au nœud de coalescence lorsque deux individus de $\mathcal{B}$ coalescent, le nœud $r$, nous obtenons que :

$$\tau_{\mathcal{B}} = \sum_{r=1}^{b-1} \mathrm{P}_r \mathbb{E}[T_{J_r+1}]. \tag{2.32}$$

De plus, comme les coalescences se font au hasard parmi les individus, nous avons :

$$\begin{aligned}
\mathrm{P}_r &= \frac{\binom{b}{2}-1}{\binom{b}{2}} \times \frac{\binom{b-1}{2}-1}{\binom{b-1}{2}} \times \cdots \times \frac{\binom{r+2}{2}-1}{\binom{r+2}{2}} \times \frac{1}{\binom{r+1}{2}} \\
&= \frac{(b-1)(b-2)}{b(b-1)} \times \frac{b(b-3)}{(b-1)(b-2)} \times \cdots \times \frac{(r+3)r}{(r+2)(r+1)} \times \frac{2}{(r+1)r} \\
&= \frac{b+1}{b-1} \frac{2}{(r+2)(r+1)}.
\end{aligned} \tag{2.33}$$

Ainsi, en remplaçant ce résultat dans l'équation 2.32, nous obtenons que :

$$\tau_{\mathcal{B}} = \frac{b+1}{b-1} \sum_{r=1}^{b-1} \frac{2}{(r+1)(r+2)} \mathbb{E}[T_{J_r+1}].$$

■

Afin d'expliciter la formule proposée au lemme précédent, nous allons donner une formule de l'espérance des $T_{J_r+1}$ par le lemme suivant.

**Lemme 5** *Pour $r = 1, \ldots, b-1$, nous avons :*

$$\mathbb{E}[T_{J_r+1}] = \sum_{k=2}^{n-b+1} \sum_{j=k+r-1}^{c+r} \mathrm{P}(J_r = j \mid J_\Delta = k) \frac{2(n-j)}{jn} p_k^\Delta.$$

**Preuve.** En conditionnant à $J_\Delta = k$, puis à $J_r = j$, nous avons :

$$\begin{aligned}
\mathbb{E}[T_{J_r+1}] &= \sum_{k=2}^{n-b+1} \mathbb{E}[T_{J_r} | J_\Delta = k] p_k^\Delta \\
&= \sum_{k=2}^{n-b+1} \sum_{j=k+r-1}^{n-b+r} \mathbb{P}(J_r | J_\Delta = k) \mathbb{E}[T_{j+1} | J_\Delta = k] p_k^\Delta.
\end{aligned}$$

Enfin, d'après le lemme 3, nous avons, pour $j > k$ :

$$\mathbb{E}[T_{j+1}|J_\Delta = k] = \frac{2(n-j)}{jn}.$$

Nous en déduisons donc que :

$$\mathbb{E}[T_{J_r+1}] = \sum_{k=2}^{n-b+1} \sum_{j=k+r-1}^{c+r} \mathrm{P}(J_r = j \mid J_\Delta = k) \frac{2(n-j)}{jn} p_k^\Delta.$$

■

La proposition suivante nous permet de proposer une formule de $\tau_\mathcal{B}$ facilement calculable pour $n$ de taille raisonnable.

**Proposition 9** *Nous avons :*

$$\tau_\mathcal{B} = \frac{b+1}{b-1} \sum_{r=1}^{b-1} \frac{2}{(r+1)(r+2)} \sum_{k=2}^{n-b+1} \sum_{j=k+r-1}^{c+r} \mathbb{P}(J_r = j \mid J_\Delta = k) \frac{2(n-j)}{jn} p_k^\Delta.$$

**Preuve.**

En combinant les deux lemmes précédents, lemmes 4 et 5, nous obtenons directement le résultat.

■

Nous avons donc formulé une expression explicite de $\tau_\mathcal{B}$ en rappelant que $\mathbb{P}(J_r = j \mid J_\Delta = k)$ est donnée par la proposition 5.

**Cas 2 : Coalescence entre $\mathcal{B}$ et $\mathcal{C}$**

Nous allons à présent expliciter la formule de $\tau_{\mathcal{B},\mathcal{C}}$, représentant l'espérance du temps de coalescence pour une séquence du sous-échantillon $\mathcal{B}$ et une séquence du sous-échantillon $\mathcal{C}$. Cette formule est donnée conditionnellement à $B = b$.

**Proposition 10** *Conditionnellement à $B = b$, nous avons :*

$$\tau_{\mathcal{B},\mathcal{C}} = 2 \sum_{k=2}^{n-b+1} \frac{k+1}{k-1} \sum_{j=2}^{k} \left( \frac{2(n-j+1)}{(j-1)n} + \frac{2}{k(k-1)} \right) \frac{1}{j(j+1)}.$$

**Preuve.**

En conditionnant à $J_\Delta$, nous avons :

$$\tau_{\mathcal{B},\mathcal{C}} = \sum_{k=2}^{n-b+1} \mathbb{E}[\tau_{\mathcal{B},\mathcal{C}}|J_\Delta = k]p_k^\Delta.$$

Notons $\mathrm{P}_j$ la probabilité que les deux individus coalescent au $j^{\mathrm{e}}$ événement de coalescence. Nous pouvons tout d'abord remarquer que le temps de coalescence entre un individu de $\mathcal{B}$ et un individu de $\mathcal{C}$ est supérieur à $\tau_\Delta$. Donc, d'après l'équation précédente, nous obtenons alors :

$$\tau_{\mathcal{B},\mathcal{C}} = \sum_{k=2}^{n-b+1} \sum_{j=2}^{k} \mathbb{E}[T_j|J_\Delta = k]\mathrm{P}_j p_k^\Delta. \tag{2.34}$$

De manière similaire à la démonstration précédente, pour l'équation 2.33, nous avons :

$$\mathrm{P}_j = \frac{k+1}{k-1} \times \frac{2}{j(j+1)}. \tag{2.35}$$

De plus, comme $j \leq k$, nous avons, d'après le lemme 3 :

$$\mathbb{E}[T_j|J_\Delta = k] = \frac{2(n-j+1)}{(j-1)n} + \frac{2}{k(k-1)}. \tag{2.36}$$

Ainsi en remplaçant les équations 2.36 et 2.35 dans l'équation 2.34, nous obtenons que :

$$\tau_{\mathcal{B},\mathcal{C}} = 2\sum_{k=2}^{n-b+1} \frac{k+1}{k-1} \sum_{j=2}^{k} \left( \frac{2(n-j+1)}{(j-1)n} + \frac{2}{k(k-1)} \right) \frac{1}{j(j+1)}.$$

■

**Cas 3 : Coalescence au sein de $\mathcal{C}$**

Enfin, nous allons expliciter la formule de $\tau_{\mathcal{C}}$, représentant l'espérance du temps de coalescence pour une séquence du sous-échantillon $\mathcal{B}$ et une séquence du sous-échantillon $\mathcal{C}$. Cette formule est donnée également conditionnellement à $B = b$. Rappelons en tout premier lieu que, conditionnellement à $B = b$, nous notons $c$ le nombre d'individus du sous-échantillon $\mathcal{C}$, $c = n - b$.

**Proposition 11** *Conditionnellement à $B = b$, nous avons, pour $r = 1, \ldots, c-1$ :*

$$\tau_{\mathcal{C}} = 2\frac{c+1}{c-1} \sum_{r=1}^{c-1} \frac{\mathbb{E}[T_{K_r+1}]}{(r+1)(r+2)}.$$

**Preuve.**

Le schéma de cette preuve est très ressemblant à la preuve de la propriété 9. En rappelant que $\mathrm{P}_r$ caractérise la probabilité que les deux individus coalescent au nœud $r$ de la sous-généalogie de $\mathcal{C}$, nous avons :

$$\tau_{\mathcal{C}} = \sum_{k=1}^{c-1} \mathrm{P}_r \mathbb{E}[T_{K_r+1}].$$

De plus, nous avons :

$$\mathrm{P}_r = \frac{c+1}{c-1} \times \frac{2}{(r+1)(r+2)}.$$

Nous en déduisons donc que :

$$\tau_{\mathcal{C}} = 2\frac{c+1}{c-1} \sum_{r=1}^{c-1} \frac{\mathbb{E}[T_{K_r+1}]}{(r+1)(r+2)}.$$

■

L'expression de $\mathbb{E}[T_{K_r+1}]$ étant compliquée, nous avons choisi de la décliner dans la propriété suivante.

**Proposition 12** *Conditionnellement à $B = b$, nous avons :*

$$\begin{aligned}\mathbb{E}[T_{K_r+1}] &= \sum_{j=1}^{r} \sum_{k=r+1}^{b+r} \mathbb{P}(K_r|J_0 = j) \left( \frac{2(n-k)}{nk} + \epsilon_{jk} \right) \mathbb{P}(J_0 = j) \\ &+ \sum_{j=r+1}^{c-1} \left( \frac{2(n-r)}{nr} + \epsilon_j \right) \mathbb{P}(J_0 = j)\end{aligned}$$

*où, pour $k = r+1, \ldots, b+r$ et $j = 1, \ldots, r$ :*

$$\epsilon_{jk} = \sum_{\ell=j+1}^{k} \frac{2}{\ell(\ell-1)} \mathbb{P}(J_\Delta = \ell | J_0 = j),$$

$$\epsilon_j = \sum_{\ell=j+1}^{n-b+1} \frac{2}{\ell(\ell-1)} \mathbb{P}(J_\Delta = \ell | J_0 = j),$$

$$P(J_0 = j) = \sum_{k=2}^{n-b+1} \frac{2j}{k(k-1)} p_k^\Delta.$$

*Pour $\ell = j+1, \ldots, n-b+1$ :*

$$\mathbb{P}(J_\Delta = \ell | J_0 = j) = \frac{\frac{2j}{\ell(\ell-1)} p_\ell^\Delta}{\sum_{m=2}^{n-b+1} \frac{2j}{m(m-1)} p_m^\Delta}.$$

**Preuve.**

Nous remarquons facilement que :

$$\mathbb{E}[T_{K_r+1}] = \sum_{j=1}^{c} \mathbb{E}[T_{K_r+1}|J_0 = j]P(J_0 = j).$$

Le calcul de $\mathbb{E}[T_{K_r+1}|J_0 = j]$ nécessite de distinguer deux cas : $j > r$ et $j \leq r$.

**Cas 1 : $j > r$.**

Nous pouvons remarquer que dans ce cas $K_r + 1 = r$. Ainsi, en conditionnant par rapport à la valeur de $J_\Delta$ et en utilisant le corollaire 1, nous obtenons que :

$$\begin{aligned}
\mathbb{E}[T_{K_r+1}|J_0 = j] &= \sum_{\ell=r}^{n} \mathbb{E}[X_\ell|J_0 = j] \\
&= \sum_{\ell=r}^{j} \frac{2}{\ell(\ell-1)} + \sum_{\ell=j+1}^{n-b+1} 2\frac{2}{\ell(\ell-1)}\mathbb{P}(J_\Delta = \ell|J_0 = j) + \sum_{\ell=n-b+2}^{n} \frac{2}{\ell(\ell-1)} \\
&= \frac{2(n-r)}{nr} + \epsilon_j.
\end{aligned}$$

**Cas 2 : $j \leq r$.**

Dans ce cas, nous sommes obligés de conditionner à la valeur de $K_r$. Une fois ce conditionnement fait, la suite est identique au cas 1.

$$\begin{aligned}
\mathbb{E}[T_{K_r+1}|J_0 = j] &= \sum_{k=r+1}^{b+r} \mathbb{P}(K_r = k|J_0 = j)\mathbb{E}[T_k|J_0 = j] \\
&= \sum_{k=r+1}^{b+r} \mathbb{P}(K_r = k|J_0 = j)\left(\frac{2(n-k)}{nk} + \epsilon_{jk}\right).
\end{aligned}$$

■

Ainsi en rappelant que l'expression de $\mathbb{P}(K_r|J_0 = j)$ est donnée à la proposition 2 et que l'expression de $\mathbb{P}(J_0 = j)$ est obtenue à l'équation 2.17, nous obtenons une formule explicite de $\mathbb{E}[T_{K_r+1}]$. En reportant cette formule dans l'expression de la proposition 11, nous obtenons ainsi une formule explicite pour $\tau_{\mathcal{C}}$ sachant $B = b$.

En conséquence, en sommant les expressions des propriétés 10 et 11 par rapport à $b$, nous obtenons une expression littérale pour $\tau_{\mathcal{B},\mathcal{C}}$ et $\tau_{\mathcal{C}}$. En utilisant de plus la propriété 9, nous en déduisons l'expression des coefficients $C_n$ et $D_n$. Etant donné que ces expressions sont particulièrement difficiles à déchiffrer, nous avons choisi de reporter une table de valeurs pour ces deux coefficients, voir tableau 2.3.

| $n$ | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
|---|---|---|---|---|---|---|---|---|---|---|
| $C_n$ | 0.996 | 1.019 | 1.021 | 1.02 | 1.02 | 1.019 | 1.019 | 1.018 | 1.018 | 1.018 |
| $D_n$ | 0.253 | 0.218 | 0.199 | 0.187 | 0.178 | 0.171 | 0.166 | 0.161 | 0.156 | 0.154 |

**Tab. 2.3: Table des coefficients correctifs $C_n$ et $D_n$.** *Ce tableau récapitule les valeurs numériques des coefficients correctifs $C_n$ et $D_n$ intervenant dans la statistique $\Pi_1 = (\Pi - C_n\theta_0)/D_n$ pour $n$ dans l'intervalle $5 - 50$.*

### 2.6.3 Performances de $\Pi_1$

Les performances de l'estimateur $\Pi_1$ ont été testées sur des données simulées. Ces données ont été obtenues grâce à l'algorithme proposé à la section 2.4.5. Le taux de mutation normal a été fixé à 1 : $\theta_0 = 1$. Comme nous l'avons vu précédemment, la valeur de $\theta_0 = 1$ correspond donc à un nombre total de cellules égal à $N = 2.5 \times 10^9$ dans le tissu considéré. De plus, notre approche s'intéresse à l'estimation du rapport entre $\theta_0$ et $\theta_1$, ce qui justifie également la normalisation du paramètre $\theta_0$.

Nous avons considéré trois valeurs pour le taux de mutation élevé : $\theta_1 = 10, 10^2, 10^3$. De plus nous nous sommes intéressés aux tailles d'échantillon suivantes : $n = 10, 20, 30, 40, 50$. Le biais et l'écart-type ont été estimés empiriquement à partir de 10000 simulations dans chacun des cas. Les résultats ont été reportés dans le tableau 2.4.

Les résultats obtenus nous confirment que l'estimateur $\Pi_1$ est sans biais. Cependant, les écart-types sont assez élevés, ce qui s'explique par le fait que les distributions empiriques montrent une queue de distribution lourde. De plus, il semblerait que l'erreur provienne d'un terme d'ordre $\theta_1^2$. Nous avons alors ajusté un modèle de régression de la forme $\alpha_n\theta_1 + \beta_n\theta_1^2$ à la variance. Pour $n = 20$, l'ajustement est presque parfait avec $\text{Var} = 1.47\theta_1^2$. Nous avons en effet obtenu $R^2 = 0.999$ et $P < 10^{-12}$. Pour $n = 40$, l'ajustement nous donne $\text{Var} = 1.68\theta_1^2$ avec $R^2 = 0.997$ et $P < 10^{-12}$. En conséquence, l'écart-type n'exhibe pas une forte décroissance lorsque la taille de l'échantillon augmente.

| | $\theta_1 = 10$ | | $\theta_1 = 100$ | | $\theta_1 = 1,000$ | |
|---|---|---|---|---|---|---|
| $n$ | $\mathbb{E}$ | Ecart-type | $\mathbb{E}$ | Ecart-type | $\mathbb{E}$ | Ecart-type |
| 10 | 9.9 | 13.7 | 107.342 | 133.9 | 1006.2 | 1243.5 |
| 20 | 10.2 | 14.7 | 100.91 | 136.2 | 1030.5 | 1458.9 |
| 30 | 9.5 | 15.5 | 100.875 | 147.9 | 1040.0 | 1589.5 |
| 40 | 10.7 | 17.8 | 95.763 | 159.0 | 998.4 | 1538.1 |
| 50 | 10.3 | 17.6 | 106.478 | 164.6 | 1039.7 | 1598.1 |

**Tab. 2.4: Résultats pour $\Pi_1$.** *Le biais et l'écart-type de $\Pi_1$ ont été calculés à partir de 10000 simulations pour des échantillons de tailles : $n = 10 - 50$. Le taux de mutation normal a été fixé à $\theta_0 = 1$.*

## 2.7 Tests de l'occurrence de l'événement $\Delta$

Dans cette section, nous allons présenter des tests portant sur l'occurrence de l'événement $\Delta$. Nous proposons tout d'abord un test basé sur les temps de coalescence. Puis nous proposons plusieurs tests issus des statistiques de Watterson et Tajima corrigées ou non. Ces résultats ont été obtenus par simulations, en utilisant notamment l'algorithme proposé à la section 2.4.5.

### 2.7.1 Test sur les temps de coalescence

Dans cette section, nous allons nous intéresser à un test de rapport de vraisemblance construit sur les temps de coalescence permettant de rejeter l'absence de la mutation $\Delta$. Nous posons comme hypothèse nulle pour ce test $H_0$ : « absence de la mutation $\Delta$ ». L'hypothèse alternative, notée $H_1$, est la suivante : « occurrence de l'événement de mutation $\Delta$ ».

Pour ce test, nous supposons que la généalogie de l'échantillon est connue. Rappelons que les temps de coalescence sont notés $(X_k)_{k=2,\dots,n}$. Sous l'hypothèse $H_0$, pour laquelle, l'événement $\Delta$ n'est pas survenu au cours de l'histoire de l'échantillon, nous avons déjà constaté (voir Equation 2.2), que les temps inter-coalescence sont indépendants et suivent

une loi exponentielle, de paramètres respectifs $\lambda_k = k(k-1)/2$ :

$$X_k \rightsquigarrow \mathcal{E}(\lambda_k).$$

Ainsi sous l'hypothèse $H_0$, la loi jointe des temps inter-coalescence, s'écrit de la manière suivante :

$$f(x_2, \ldots, x_n) = \prod_{\ell=2}^{n} f_\ell(x_\ell),$$

où $f_\ell(x_\ell)$ est la densité d'une variable de loi exponentielle de paramètre $\lambda_\ell$. Sous l'hypothèse $H_1$, le théorème 1, nous donne la loi jointe des temps inter-coalescence. Nous pouvons rappeler que, conditionnellement à $B = b$ et à $E \cap M$, nous avons :

$$f(x_2, \ldots, x_n) = \sum_{k=2}^{n-b+1} p_k^{\Delta} \lambda_k x_k \prod_{\ell=2}^{n} f_\ell(x_\ell).$$

Le test statistique proposé consiste à prendre le rapport des vraisemblances. Ce test, théoriquement optimal pour des échantillons de grande taille, peut s'écrire de la façon suivante :

$$r = \frac{L(x, H_1)}{L(x, H_0)} = \sum_{k=2}^{n-b+1} p_k^{\Delta} \lambda_k x_k.$$

Nous proposons comme critère de rejet que $r$ soit plus grand que le quantile 95% des données neutres. La puissance de ce test a été étudiée numériquement à partir de 10000 simulations de temps de coalescence sous l'hypothèse $H_0$ et sous l'hypothèse $H_1$. Les résultats que nous avons obtenus montrent que la puissance de ce test ne dépasse pas 0.2 pour $n = 10, 20, 50, 100$ et dans le cas où $b \approx n$. Dans le cas où $b$ a des valeurs plus faible, la faible puissance du test est encore plus marquée : la puissance est environ égal à 0.1 lorsque $b/n \approx 0.5$.

Puisqu'il suppose que la topologie de l'arbre de coalescence, ainsi que les temps de coalescence (*i.e.* la longueur des branches de l'arbre) sont connus, ce test est difficile à utiliser en pratique. Cependant, la faible puissance de ce test met en avant le fait que la détection de l'occurrence de l'événement de mutation $\Delta$ est extrêment délicate.

### 2.7.2 Test à l'aide des estimateurs de Watterson et de Tajima corrigés

Dans cette section, nous allons proposer plusieurs tests statistiques permettant d'étudier l'occurrence de l'événement de mutation $\Delta$. Ces tests sont construits à partir des

| $n$ | $\theta_1 = 10$ | $\theta_1 = 100$ | $\theta_1 = 1,000$ |
|---|---|---|---|
| 10 | 0.10 | 0.29 | 0.90 |
| 20 | 0.06 | 0.18 | 0.70 |
| 30 | 0.13 | 0.29 | 0.65 |
| 40 | 0.11 | 0.24 | 0.59 |
| 50 | 0.09 | 0.21 | 0.55 |

**Tab. 2.5: Puissance pour $\hat{\theta_1}$.** *Puissance du test construit à partir de la statistique $\hat{\theta_1}$, pour lequel* $H_0$ *est l'occurrence de* $\Delta$ *et* $\theta_1 > \theta_0$ *contre l'hypothèse alternative* $H_1$ *: absence de* $\Delta$*. Le taux de mutation normal a été fixé à* $\theta_0 = 1$ *et nous avons fait varier le taux de mutation élevé de* $\theta_1 = 10$ *à* $\theta_1 = 1000$*.*

statistiques de Watterson et de Tajima corrigées, proposées respectivement dans les théorèmes 2 et 3.

Le premier test que nous proposons utilise la statistique de Watterson corrigée, $\hat{\theta_1}$. L'hypothèse nulle, $H_0$, se définit comme l'occurrence de $\Delta$ associée à une hausse du taux de mutation $\theta_1 > \theta_0$, tandis que l'hypothèse alternative, $H_1$, est l'absence de l'événement $\Delta$. La puissance de ce test a été étudiée par simulations en fixant le taux de mutation normal $\theta_0 = 1$ et en choisissant le taux de mutation élevé $\theta_1 = 10, 100, 1000$. La puissance du test a alors été calculée pour des tailles d'échantillon variant entre $n = 10$ et $n = 50$. Les résultats obtenus ont été reportés dans le tableau 2.5.

Nous pouvons constater que des puissances raisonnables sont obtenues pour des valeurs du taux de mutation élevé $\theta_1$ supérieure au taux de mutation normal $\theta_0$ d'un facteur 1000. De plus, la puissance ne semble pas augmenter lorsque la taille de l'échantillon augmente de $n = 10$ à $n = 50$.

Le second test que nous proposons utilise la statistique de Watterson, $\hat{\theta}$. L'hypothèse nulle, $H_0$, est l'absence de l'événement $\Delta$, tandis que l'hypothèse alternative, $H_1$, se définit comme l'occurrence de $\Delta$ associée à une hausse du taux de mutation $\theta_1 > \theta_0$. Ce test a été étudié par simulations en fixant le taux de mutation normal $\theta_0 = 1$ et en choisissant

| $n$ | $\theta_1 = 10$ | $\theta_1 = 100$ | $\theta_1 = 1,000$ |
|---|---|---|---|
| 10 | 0.44 | 0.75 | 0.93 |
| 20 | 0.44 | 0.74 | 0.90 |
| 30 | 0.48 | 0.75 | 0.89 |
| 40 | 0.42 | 0.73 | 0.88 |
| 50 | 0.43 | 0.72 | 0.87 |

**Tab. 2.6: Puissance pour $\hat{\theta}$.** *Puissance du test construit à partir de la statistique $\hat{\theta}$, pour lequel* $H_0$ *est l'absence de* $\Delta$ *contre l'hypothèse alternative,* $H_1$*, se définissant comme l'occurrence de* $\Delta$ *et* $\theta_1 > \theta_0$*. Le taux de mutation normal a été fixé à* $\theta_0 = 1$ *et nous avons fait varier le taux de mutation élevé de* $\theta_1 = 10$ *à* $\theta_1 = 1000$*.*

le taux de mutation élevé $\theta_1 = 10, 100, 1000$. La puissance du test a alors été calculée pour des tailles d'échantillon variant entre $n = 10$ et $n = 50$. Les résultats obtenus ont été reportés dans le tableau 2.6.

Les résultats obtenus nous montrent que la puissance de ce test se situe entre 0.43 et 0.93. Nous pouvons constater que pour des valeurs de $\theta_0$ inférieure à 10, le test est très peu puissant. Lorsque le taux de mutation élevé atteint des valeurs supérieures à 1000 ($\theta_1 > 1000$), le gain en puissance est significatif. De plus, le tableau 2.6 nous indique que la détection de l'événement $\Delta$ est meilleure avec un taux de mutation fort et une taille d'échantillon faible.

Le troisième test que nous proposons est similaire au premier. La seule différence est que ce test est construit à partir de la statistique de Tajima corrigée. Les résultats obtenus ont été reportés dans le tableau 2.7.

Enfin, le quatrième test que nous proposons est similaire au deuxième test proposé en utilisant la statistique de Tajima $\Pi$ pour rejeter l'hypothèse $H_0$ d'absence de $\Delta$. Les résultats obtenus ont été reportés dans le tableau 2.8.

Les deux tableaux 2.7 et 2.8 nous montrent que les résultats obtenus avec la statistique de Tajima et la statistique de Watterson sont similaires. Dans chaque cas, des

| $n$ | $\theta_1 = 10$ | $\theta_1 = 100$ | $\theta_1 = 1000$ |
|---|---|---|---|
| 10 | 0.09 | 0.32 | 0.72 |
| 20 | 0.12 | 0.29 | 0.54 |
| 30 | 0.14 | 0.24 | 0.44 |
| 40 | 0.12 | 0.19 | 0.35 |
| 50 | 0.13 | 0.20 | 0.40 |

**Tab. 2.7: Puissance pour $\Pi_1$.** *Puissance du test construit à partir de la statistique $\Pi_1$, pour lequel $H_0$ est l'occurrence de $\Delta$ et $\theta_1 > \theta_0$ contre l'hypothèse alternative $H_1$ : absence de $\Delta$. Le taux de mutation normal a été fixé à $\theta_0 = 1$ et nous avons fait varier le taux de mutation élevé de $\theta_1 = 10$ à $\theta_1 = 1000$.*

| $n$ | $\theta_1 = 10$ | $\theta_1 = 100$ | $\theta_1 = 1000$ |
|---|---|---|---|
| 10 | 0.44 | 0.73 | 0.91 |
| 20 | 0.44 | 0.69 | 0.84 |
| 30 | 0.39 | 0.64 | 0.80 |
| 40 | 0.34 | 0.64 | 0.79 |
| 50 | 0.34 | 0.62 | 0.76 |

**Tab. 2.8: Puissance pour $\Pi$.** *Puissance du test construit à partir de la statistique $\Pi$, pour lequel $H_0$ est l'absence de $\Delta$ contre l'hypothèse alternative, $H_1$, se définissant comme l'occurrence de $\Delta$ et $\theta_1 > \theta_0$. Le taux de mutation normal a été fixé à $\theta_0 = 1$ et nous avons fait varier le taux de mutation élevé de $\theta_1 = 10$ à $\theta_1 = 1000$.*

puissances raisonnables sont obtenues lorsque le taux de mutation élevé est supérieur au taux de mutation normal d'un facteur d'au moins 1000. Ce facteur 1000 est en accord avec certaines expérimentations biologiques selon lesquelles la manifestation d'un phénotype de mutation ne peut se détecter que pour une hausse du taux de mutation d'au moins 1000 [Loeb, 2001].

## 2.8 Discussion et conclusion

Le travail développé dans ce chapitre est la modélisation d'une hypothèse biologique décrivant l'initiation d'un cancer. Cette hypothèse biologique suppose que le développement d'un cancer débute par un événement de mutation particulier. Partant du constat que les cellules cancéreuses sont caractérisées par un nombre anormalement élevé de mutations, Loeb et ses collaborateurs ont proposé que le nombre de mutations anormalement élevé soit la conséquence d'un phénotype de mutation exprimé par les cellules cancéreuses [Loeb et al., 1974]. Ce phénotype de mutation serait dû à une déficience du système de contrôle de la réplication de l'ADN au cours de la mitose. Le système de contrôle de la réplication est souvent appelé le système MMR (MisMatch Repair), et met en jeu des gènes codant à la fois pour contrôler la fidélité de la réplication de l'ADN et pour réparer l'ADN répliqué. Certains de ces gènes, les gènes MLH1 et MSH2, ont d'ailleurs été identifiés pour leur implication dans le développement d'un cancer du colon.

La conséquence d'une perte de MMR est la hausse du taux de mutation. En effet, en cas de perte de MMR, l'organisme « laisse passer » plus de mutations. La détection d'une hausse dans le taux de mutation semble donc être un bon outil pour améliorer la détection précoce du cancer.

Dans ce chapitre, nous avons proposé un modèle intégrant la perte de MMR. A cet effet, nous avons considéré un échantillon de $n$ cellules, à partir desquels $n$ séquences d'ADN on été extraites. Nous avons supposé que ces cellules provenaient d'un tissu cancéreux. L'échantillon de $n$ séquences d'ADN est donc composé d'un nombre aléatoire de séquences ayant un taux de mutation élevé, complété par des séquences ayant un taux de mutation normal. Les séquences ayant un taux de mutation élevé forment un sous-échantillon, noté $\mathcal{B}$, complété par le sous-échantillon $\mathcal{C}$.

Nous avons choisi de modéliser la généalogie de cet échantillon de séquences en intégrant l'événement de mutation $\Delta$. Le modèle de généalogie que nous avons utilisé est un modèle coalescent conditionnel introduit initialement pour estimer l'âge de l'allèle [Wiuf et Donnelly, 1999], [Griffiths et Tavaré, 2003]. Les hypothèses de ce modèle de généalogie correspondent exactement aux hypothèses de la perte de MMR. Ainsi, la partie de la généalogie affectée par le taux de mutation élevé correspond au sous-arbre qui descend directement de l'événement $\Delta$.

Nous nous sommes ensuite placés sous l'hypothèse du modèle à infinité de sites [Watterson, 1975]. Ce modèle suppose que deux mutations ne peuvent se produire sur le même site. Nous avons alors utilisé pour les mutations un modèle de Poisson le long des branches de la généalogie. Plus précisemment, les mutations suivent un processus de Poisson de paramètre $\theta_1/2$ le long des branches qui descendent de $\Delta$ et de paramètre $\theta_0/2$ ailleurs. Les paramètres $\theta_0$ et $\theta_1$ correspondent respectivement aux taux de mutation normal et élevé.

Sous ce modèle à deux taux, nous avons corrigé deux statistiques classiques : l'estimateur de Watterson [Watterson, 1975] et l'estimateur de Tajima [Tajima, 1983] pour le taux de mutation $\theta_1$. Ces corrections ont nécessité l'étudé détaillée de la topologie de l'arbre de coalescence conditionnel. Certains résultats, comme l'âge de la mutation $\Delta$, avaient déjà été étudiés et nous ont été très utiles.

Nous nous sommes efforcés de déterminer des résultats complémentaires sur l'arbre de coalescence conditionnel. Ces résultats ont porté notamment sur la loi jointe des temps inter-coalescence, l'espérance de la longueur totale de l'arbre, les temps moyens de coalescence entre deux individus de $\mathcal{B}$, entre deux individus de $\mathcal{C}$ et entre un individu de $\mathcal{B}$ et un individu de $\mathcal{C}$.

Ces résultats nous ont permis de corriger le biais des statistiques de Watterson et de Tajima. Nous avons alors proposé plusieurs tests, à partir des statistiques de Watterson et de Tajima corrigées ou non, permettant de tester d'une part l'occurrence de l'événement $\Delta$ et d'autre part l'absence de $\Delta$. Par ailleurs, les résultats théoriques que nous avons obtenus, associés à certaines propriétés issues de l'étude de [Tavaré, 2004], nous ont permis de proposer un algorithme de simulation d'arbre de coalescence conditionnel.

Les estimateurs de Watterson et de Tajima corrigés ont été testés à l'aide de si-

mulations. Cette étude a confirmé que ces estimateurs sont effectivement sans biais. Cependant, l'écart-type de ces estimateurs reste relativement élevé. C'est la raison pour laquelle les tests proposés montrent une puissance significative uniquement lorsque le rapport entre le taux de mutation normal et le taux de mutation élevé est au moins égal à 1000. Nous avons également remarqué que la puissance optimale des tests pour rejeter l'hypothèse d'instabilité génétique est obtenue pour de faibles tailles d'échantillon ($n = 10$). Ce constat vient du fait que notre modèle est conditionné à l'occurrence de $\Delta$. Ainsi, lorsque la taille de l'échantillon est faible, l'information relative concernant le taux de mutation élevé est la plus importante. Ce résultat suggère qu'il est préférable de choisir judicieusement les loci à séquencer, ou même de développer de nouveaux marqueurs spécifiques, plutôt que d'augmenter le nombre de cellules de l'échantillon. Cette expertise biologique pourrait améliorer la puissance des tests [Boland et al., 1998].

Les simulations que nous avons fournies ont été obtenues sous l'hypothèse de neutralité. Au cours des premières étapes du développement prétumoral, cette hypothèse de neutralité paraît compatible avec la théorie proposée par Loeb. De nombreuses hypothèses alternatives à celle de Loeb ont cependant été avancées. L'une d'entre elles explique notamment qu'une cellule doit exhiber un avantage selectif pour se transformer en une cellule prétumorale. Cette cellule pourra devenir maligne par selection clonale [Cairns, 1975], [Nowell, 1976], [Tomlinson et al., 1996]. Les travaux présentés dans ce chapitre permettent de fournir un test pour cette hypothèse. En effet, une façon classique de tester l'influence de la sélection est d'étudier la statistique D de Tajima [Tajima, 1989]. Dans le contexte de ce chapitre la statistique D de Tajima s'obtient comme la différence $\hat{\theta_1} - \Pi_1$.

La prise en compte de la sélection dans le modèle à deux taux proposé dans ce chapitre complique sensiblement le modèle. Cependant certaines contributions récentes peuvent aider dans cette perspective [Krone et Neuhauser, 1997], [Stephens et Donnelly, 2003], [Coop et Griffiths, 2004].

Enfin, certaines difficultés concernant la mesure de l'effet précis de l'instabilité dans le génome des cellules cancéreuses ont été avancées [Anderson, 2001]. Ces problèmes expérimentaux justifient notre approche. En effet, l'estimation rigoureuse de l'instabilité génétique nécessite d'utiliser des outils mathématiques construits en ce sens. Le contrôle

du biais des estimateurs de Watterson et de Tajima permet donc de minimiser l'erreur de prédiction.

## Chapitre 3

# Modèle de Gibbs pour l'organisation tissulaire

Le processus de développement d'un cancer est connu pour être un processus multi-cascades. Bien que la succession des étapes de ce processus soit à ce jour mal connue, chaque type de cancer semble présenter des caractéristiques différentes quant à son développement. Un des facteurs de développement tumoral est la perte d'adhésion cellulaire. L'adhésion cellulaire correspond à l'ensemble des forces mises en jeu entre les cellules voisines d'un tissu pour assurer le maintien et l'intégrité d'un tissu. Depuis quelques années, la perte d'adhésion cellulaire est considérée comme une étape importante du processus de cancérisation, notamment au cours de la phase d'invasion générant des métastases.

Depuis les années 50, les mécanismes mis en jeu dans l'adhésion cellulaire ont fait l'objet de nombreuses recherches. Avant de trouver une application directe au cancer, ces recherches se sont concentrées sur l'explication de phénomènes biologiques traditionnels dans le développement d'un organisme multicellulaire : l'embryogénèse ou la morphogénèse, par exemple. Il a fallu attendre les années 90 pour que le lien entre adhésion cellulaire et cancer soit mis en avant. Cette découverte a suscité ainsi une extension des recherches sur l'adhésion cellulaire.

Dans une première partie de ce chapitre, le contexte précis d'adhésion cellulaire sera défini. Puis nous présenterons le complexe Cadhérine-Caténine qui est une composante

importante de l'adhésion cellulaire et nous rappelerons l'implication de ce complexe dans la cancérisation. Nous enchaînerons avec la présentation de modèles mathématiques construits pour étudier par simulation l'effet de l'adhésion cellulaire sur le développement d'organismes pluri-cellulaires. Nous proposerons alors une classe de modèles, la classe $CC$ (pour Cadhérines-Caténines), construite sur une géométrie continue du tissu. Après avoir rappelé le contexte mathématique du diagramme de Dirichlet et des processus ponctuels de Gibbs marqués, nous montrerons certaines propriétés théoriques de la classe $CC$. Nous formaliserons ensuite un algorithme de Metropolis-Hastings pour la classe $CC$, puis nous étudierons les propriétés de convergence de cet algorithme. Enfin, nous proposerons deux estimateurs du paramètre d'adhésion basés sur la pseudo-vraisemblance.

## 3.1 L'adhésion cellulaire : un régulateur du développement d'un organisme multicellulaire

### 3.1.1 Contexte général

Le développement d'un organisme multicellulaire se fait par l'intermédiaire d'un programme de régulation complexe. La principale conséquence phénotypique de ce programme est l'émergence de configurations particulières des cellules d'un tissu. Par exemple, selon les cas, les cellules d'un tissu vont avoir tendance à s'agréger de façon homotypique ou hétérotypique, à migrer, à grossir, etc... Ces différentes configurations sont considérées comme la signature d'un stade (ou phase) de développement d'un tissu présent dans l'embryogénèse ou la morphogénèse par exemple [Cowin, 2000].

La connaissance précise des mécanismes physico-biologiques gérant l'organisation cellulaire d'un tissu reste, à ce jour, un sujet de recherche en plein essort. De nombreuses expériences biologiques, menées essentiellement dans les années 60, par Steinberg et son équipe ont permis de mettre en avant l'importance de la cellule comme entité primordiale dans la formation d'un tissu [Steinberg, 1962a], [Steinberg, 1962b], [Steinberg, 1962c], [Steinberg, 1963]. Plus précisemment, ces expériences ont prouvé le rôle fondamental joué par les interactions intercellulaires dans l'émergence, le développement et le maintien d'un tissu. Par conséquent, un tissu peut être vu comme un ensemble de cellules (entité unitaire) contraint par des interactions.

Les interactions au sein d'un tissu sont généralement classées en deux catégories : les interactions entre les cellules d'une part (interactions de type Cellule-Cellule) et les interactions entre les cellules et le substrat extra-cellulaire d'autre part (interactions de type Cellule-Substrat). Par le biais de ces interactions, les cellules vont pouvoir adopter des configurations très diverses et induire de multiples configurations de cellules. Le bon fonctionnement de ces interactions va ainsi permettre au tissu de se développer normalement.

La compréhension de ces interactions, ainsi que leurs implications dans le développement d'un organisme, présente donc un intérêt majeur pour de multiples phénomènes : les organisations cellulaires caractéristiques des tissus sains mais aussi certains aspects mécaniques impliqués dans de nombreux cancers et autres maladies [Steinberg et Foty, 1997].

Les phénomènes cellulaires les plus étudiés sont l'auto-organisation spontanée des cellules embryonnaires soit sous forme de damier, caractérisant des interactions hétérotypiques, soit en agrégats homogènes, caractérisant des interactions homotypiques. L'observation de ces phénomènes a donné lieu à de nombreuses hypothèses sur la nature des interactions entre les cellules.

### 3.1.2 Hypothèses sur les interactions cellulaires

Les expériences biologiques réalisées dans les années 60 sur les mouvements auto-induits de certaines cellules ont donné lieu à de nombreuses théories explicatives. Bien que ces théories soient fondées sur différents concepts d'adhésion cellulaire, elles se concentrent sur les forces mises en jeu au niveau des surfaces de contact entre les cellules. Il est assez usuel de classer ces hypothèses en 4 grandes classes [Brodland, 2002] : l'hypothèse DAH, l'hypothèse DSCH, l'hypothèse specific-CAMs et l'hypothèse DITH.

#### DAH : Differential Adhesion Hypothesis

Cette hypothèse, issue d'un parallèle entre un tissu et un fluide, a été formulée par Steinberg à la suite de ses fameuses expériences [Steinberg, 1970]. L'hypothèse DAH suppose que

1. les cellules adhèrent les unes aux autres,
2. la force d'adhésion varie d'un type de cellule à l'autre,

3. Un agrégat de cellules va avoir tendance à se positionner dans une configuration qui minimise l'énergie du système.

L'hypothèse DAH est aujourd'hui la plus reconnue biologiquement et c'est en partie sur cette hypothèse que nous allons fonder notre étude.

**DSCH : Differential Surface Contraction Hypothesis**

Harris a effectué un travail parallèle sur le comportement des fluides et celui des tissus [Harris, 1976]. Il a ainsi pu mettre en évidence un certain nombre de différences qui l'ont amené à énoncer des hypothèses sur les interactions cellule-cellule. En particulier Harris, a formulé l'hypothèse de la contraction de surface différentielle. L'hypothèse DSCH suppose que

1. les cellules présentent des contractions de surface,
2. la force de contraction est
   - élevée lorsque la cellule est en contact avec le milieu extracellulaire,
   - faible lorsque la cellule est en contact avec une cellule d'un autre type histologique,
   - la plus faible lorsque la cellule est en contact avec une cellule du même type histologique.
3. les cellules de différents types exercent différents degrés de contraction de surface lorsqu'elles sont en contact avec le domaine extracellulaire.

Cette hypothèse a permis d'élaborer des modèles d'interactions hétérotypiques exhibant essentiellement des configurations en damier.

**Specific CAM-based hypothesis**

Dans les années 80, les recherches se sont concentrées sur les phénomènes chimiques liés à l'adhésion cellulaire. Ces recherches ont notamment permis de découvrir de nouvelles molécules dédiées spécifiquement à l'adhésion cellulaire appelées CAMs (Cell Adhesion Molecules). Friedlander et ses collaborateurs ont mis en avant que les CAMs, et plus particulièrement les cadhérines (section 3.2), influencent l'organisation spatiale des cellules [Friedlander et al., 1989]. L'hypothèse « Specific CAM-based » s'est développée en

parallèle de ces recherches et prétend que l'organisation cellulaire d'un tissu dépend de la nature et de la concentration des structures moléculaires à la surface des cellules.

Il s'avère que cette hypothèse est complémentaire de l'hypothèse DAH proposée par Steinberg. En effet, elle permet de définir la force d'adhésion entre deux cellules voisines qualitativement en décrivant les molécules exprimées au sein des deux cellules considérées, et quantitativement en mesurant la quantité de ces protéines exprimées par les deux cellules.

**DITH : Differential Interfacial Tension Hypothesis**

Au début des années 2000, certaines expérimentations biologiques ont montré l'incohérence de l'hypothèse DAH [Brodland et Chen, 2000]. Par ailleurs, le concept de tension de surface différentielle, qui a débouché sur l'hypothèse DITH, s'est montré cohérent avec ces expérimentations biologiques. L'hypothèse DITH, proposée par Brodland [Brodland, 2002], suppose que

1. la tension de surface agit le long des frontières entre deux cellules et entre une cellule et le domaine extracellulaire,
2. la valeur de cette tension dépend des types respectifs des cellules ou du domaine extracellulaire formant l'interface,
3. cette tension induit des mouvements locaux au niveau des jonctions entre cellules qui mènent à des configurations spécifiques de réarrangement cellulaire comme le damier.

L'hypothèse DITH offre une alternative intéressante aux autres hypothèses. Son utilisation pratique reste pour l'instant difficile [Brodland, 2002].

Les différentes hypothèses proposées sur le fonctionnement de l'adhésion cellulaire ont été, et sont encore de nos jours, la source de controverses. Malgré les progrès technologiques dans de nombreux domaines tels que la biologie moléculaire, la biochimie ou l'imagerie, il semblerait qu'aucune hypothèse générale sur l'adhésion cellulaire ne puisse être confirmée. Cependant, l'hypothèse biologique la plus couramment admise reste l'hypothèse d'adhésion différentielle (DAH) proposée par Steinberg. Dans la suite de ce chapitre nous nous placerons sous l'hypothèse DAH qui, dans sa forme originelle, est trop

conceptuelle pour pouvoir être utilisée directement dans un modèle. Ainsi, nous essaierons de coupler l'hypothèse DAH et les connaissances récentes obtenues sur les molécules d'adhésion cellulaires (CAMs).

Le rôle précis des CAMs (Cellular Adhesion Molecules) a été particulièrement étudié au cours de années 90. Geiger et Ayalon ont montré que, parmi les CAMs, les **cadhérines** et les **caténines** tiennent un rôle prépondérant chez les vertébrés, du fait de leurs implications dans les jonctions adhérentes [Geiger et Ayalon, 1992]. Ces molécules sont responsables de l'adhésion et de la force de cette adhésion entre deux cellules voisines, par la formation d'un complexe appelé **complexe Cadhérine-Caténine**.

Afin de mieux comprendre les mécanismes d'adhésion induits par les complexes Cadhérine-Caténine, il est nécessaire de faire un bilan des connaissances sur la composition et le rôle précis joué par ces complexes.

## 3.2 Le complexe Cadhérine-Caténine : un des agents principaux de l'adhésion cellulaire

Les cadhérines sont des glycoprotéines transmembranaires participant à l'adhésion de type cellule-cellule et à la reconnaissance cellule-cellule [Yap et Goodwin, 2004]. Les principaux types de cadhérines sont nommés en rapport avec le type de tissu où celles-ci furent découvertes. Par exemple les E-, N- et VE-cadhérines correspondent respectivement à des cellules épithéliales, nerveuses et endothéliales vasculaires. Une protéine cadhérine est caractérisée par un domaine extracellulaire et un domaine intracellulaire. Le domaine extracellulaire se compose exactement de cinq parties extracellulaires répétées (voir figure 3.1). Les ligants entre deux domaines extracellulaires d'une même cadhérine sont composés d'ions de calcium (Ca2+). Le domaine intracellulaire d'une cadhérine sert quant à lui d'intermédiaire entre la cellule et le domaine extracellulaire.

### 3.2.1 Domaine extracellulaire d'une cadhérine

La liaison entre deux cellules voisines est assurée par des connexions entre de nombreuses molécules cadhérines appartenant aux deux cellules. Cette liaison est plus particulièrement assurée par les domaines extracellulaires des molécules cadhérines. Deux

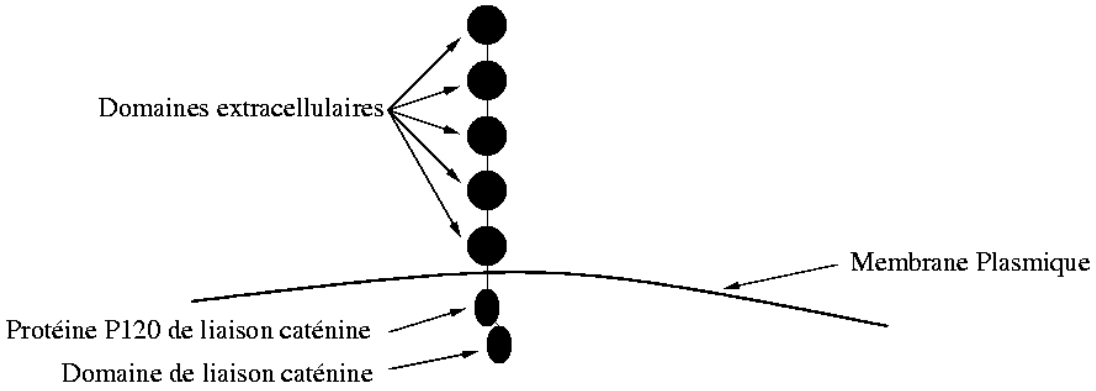



**Fig. 3.1:** *Schéma d'une molécule cadhérine. Les cadhérines sont caractérisées par la présence de cinq domaines extracellulaires répétés. Ces domaines vont assurer l'adhésion entre deux cellules voisines. Les domaines cytoplasmiques (Protéine P120 et domaine de liaison caténine) vont assurer la liaison entre la cadhérine et le cytosquelette de la cellule.*

cadhérines appartenant à deux cellules différentes vont se lier par l'intermédiaire de leurs domaines extracellulaires terminaux (voir figure 3.2). Les liaisons entre deux cadhérines favorisent principalement des interactions homotypiques. Ceci signifie que : plus deux cellules voisines se ressemblent (expriment des phénotypes proches), plus la liaison des cadhérines est forte [Wheelock et Johnson, 1989].

### 3.2.2 Domaine intracellulaire d'une cadhérine et complexe Cadhérine-Caténine

La partie intracellulaire permet le dialogue entre la cellule et les domaines extracellulaires des cadhérines. Cette partie est composée essentiellement d'une protéine de liaison caténine (protéine P120) et d'un domaine de liaison caténine (voir figure 3.1).

Le dialogue entre la cellule et la cadhérine s'effectue par l'intermédiaire de molécules appélées caténines [Ozawa et al., 1990]. Plus précisemment, une molécule $\beta$-caténine vient se lier au domaine de liaison caténine de la cadhérine. Cette molécule est également reliée à une autre molécule, nommée $\alpha$-caténine, qui vient s'attacher à un filament d'actine. Ce

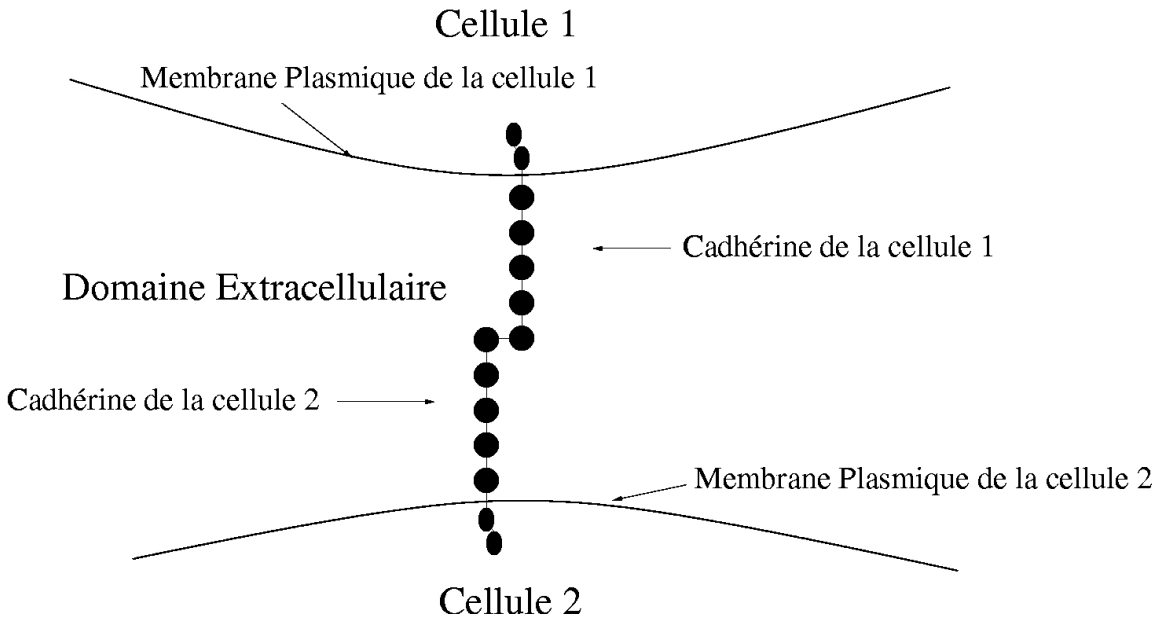



**FIG. 3.2:** *Schéma d'une jonction adhérente entre deux cadhérines issues de deux cellules voisines. Nous pouvons remarquer que la liaison se fait au niveau du dernier domaine extracellulaire de chacune des deux cadhérines.*

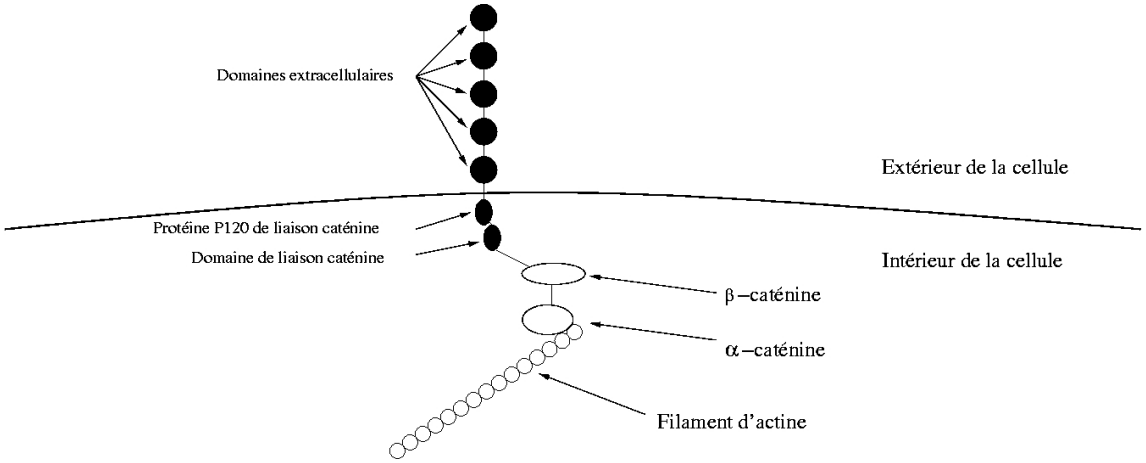



**FIG. 3.3:** *Schéma du complexe Cadhérine-Caténine au sein d'une cellule.*

complexe est schématisé à la figure 3.3. Le filament d'actine fait partie du cytosquelette, dont l'une des fonctions principales est d'assurer une certaine rigidité à la cellule.

Par l'intermédiaire du filament d'actine et du complexe Cadhérine-Caténine, il existe une interaction entre le cytosquelette d'une cellule et ses cadhérines. Cette interaction permet l'évolution et le maintien de l'architecture d'un tissu au cours de son développement (embryogénèse, morphogénèse, etc...).

Il apparaît donc que le complexe cadhérine-caténine joue un rôle majeur dans la vie de la cellule en assurant le lien entre le milieu intracellulaire et le milieu extracellulaire. Nous allons à présent résumer l'état des recherches actuelles sur la fonction précise du complexe Cadhérine-Caténine. La compréhension de ces fonctions, nous permettra par la suite de modéliser plus rigoureusement l'action de ces complexes sur l'adhésion générale des cellules d'un tissu.

### 3.2.3 Fonctions du complexe Cadhérine-Caténine

Les premières recherches sur le complexe Cadhérine-Caténine ont montré que l'agrégation cellulaire s'opérait entre cellules exprimant le même type de cadhérines. Ce phénomène est connu sous le nom de caractère homophilique des cadhérines [Nose et al., 1988]. Ces premières observations ont montré le rôle régulateur des cadhérines pour l'adhésion

cellulaire. Elles servent de joint intercellulaire « collant » entre deux cellules voisines selon leurs types respectifs. Même si ces résultats n'expliquent pas, à eux seuls, l'organisation spatiale d'un tissu, ils ont permis d'orienter les recherches sur la fonction précise du complexe Cadhérine-Caténine.

En 1994, Steinberg et Takeichi ont montré que l'organisation cellulaire, observée au cours de la morphogénèse notamment, est causée par des différences dans l'intensité des adhésions intercellulaires [Steinberg et Takeichi, 1994]. L'intensité d'une adhésion cellulaire est caractérisée par le nombre de complexes Cadhérine-Caténine assurant la liaison entre les deux cellules. Cependant, cette intensité n'est pas directement proportionnelle au nombre de molécules cadhérines, elle tient également compte du nombre de cadhérines présentes à la surface d'une cellule et non utilisées dans l'adhésion. Plus précisemment, Steinberg et Takeichi ont considéré trois situations distinctes, décrites qualitativement de la manière suivante.

1. Si deux cellules voisines ont un faible niveau d'expression de cadhérines, alors leur liaison sera relativement efficace,
2. Si deux cellules voisines ont un fort niveau d'expression de cadhérines, alors leur liaison sera très efficace,
3. Si une cellule a un faible niveau d'expression de cadhérines et l'une de ses voisines un fort niveau d'expression, alors leur liaison sera inefficace.

Ainsi, dans un système cellulaire contenant deux types de cellules, caractérisés par leurs niveaux d'expression des complexes Cadhérine-Caténine, le nombre de complexes inutilisés va avoir tendance à être minimisé [Steinberg et Takeichi, 1994]. Les complexes Cadhérine-Caténine agissent donc comme de la « colle » liant deux cellules exprimant le même type de cadhérines. Cette « colle » sera d'autant plus efficace que deux cellules adjacentes expriment le même nombre de complexes sur leur surface commune.

Il s'avère également que les cadhérines participent à l'activation de protéines en traduisant certains signaux extracellulaires [Yap et Goodwin, 2004]. Le complexe Cadhérine-Caténine joue alors un rôle de boîte de dialogue entre l'intérieur et l'extérieur de la cellule.

Kemler fut le premier à exprimer le fait que les caténines participent au lien fonctionnel entre les cadhérines et le cytosquelette [Kemler, 1993]. Les principaux signaux régulant l'activité adhésive des cadhérines sont le Wnt, la régulation de l'$\alpha$-caténine,

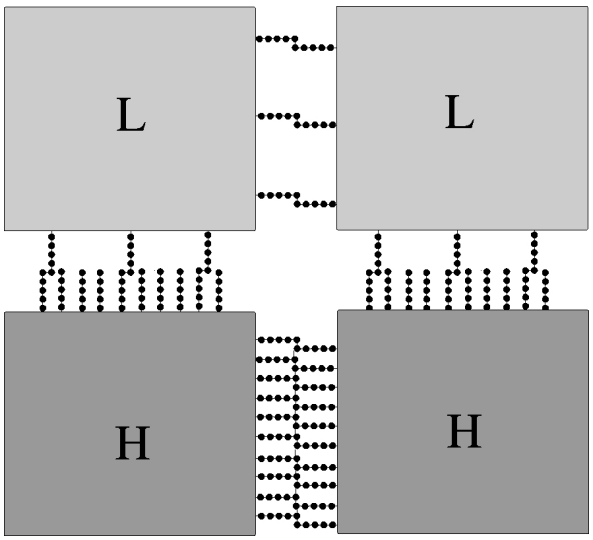



**Fig. 3.4:** *Schéma rendant compte de l'efficacité de liaisons adhérentes entre cellules voisines [Steinberg et Takeichi, 1994]. Dans le schéma, les cellules sont représentées par des carrés. Une cellule notée L exprime une faible concentration de molécules d'adhésion tandis qu'une cellule notée H exprime une forte concentration de CAMs. La liaison entre cellules H et H est très forte, la liaison entre cellules L et L est forte et la liaison entre cellules H et L est faible.*

l'expression des petites GTPases (Rac, Rho et Cdc42), l'activation de la protéine p120 et la phosphorylation de kinases [Gumbiner, 2000].

Schématiquement, les complexes Cadhérine-Caténine répercutent l'information extracellulaire en activant ou inhibant des facteurs de croissance de la cellule. Selon de nombreux biologistes moléculaires, il semblerait que ce rôle soit encore plus important, démontrant une réelle communication intercellulaire dont le vecteur serait le complexe Cadhérine-Caténine.

Pour résumer cette sous-section, nous pouvons rappeler que les complexes cadhérines-caténines jouent un rôle de boîte de dialogue entre le noyau de la cellule et le milieu extracellulaire. Ce dialogue s'effectue notamment par l'activation de voies de signalisation complexes. Il paraît beaucoup trop ambitieux de vouloir intégrer l'ensemble des voies de signalisation cellulaire dans un modèle d'adhésion. Nous devons donc à présent synthétiser ces informations pour pouvoir les intégrer le plus simplement possible dans un modèle mathématique. Nous pouvons constater que le caractère adhésif d'un tissu peut se résumer aux forces exercées entre deux cellules voisines. Ces forces peuvent se caractériser, comme un phénotype cellulaire codé par l'expression des molécules cadhérines et caténines.

### 3.2.4 Evolution des forces de l'adhésion

Shapiro et son équipe ont montré que la succession de liaisons Cadhérine-Cadhérine entre cellules voisines s'organise selon une fermeture éclair [Shapiro et al., 1995] (voir figure 3.5). Ce résultat a été obtenu en étudiant la structure crystalline des domaines extracellulaires terminaux. Il apparaît également que le caractère linéaire de la fermeture éclair se vérifie par des expériences biologiques. Cette constatation a permis à Steinberg et ses collaborateurs de raffiner l'hypothèse DAH . En effet, il apparaît que plus la surface de contact entre cellules voisines est grande, plus l'adhésion est forte car plus la fermeture éclair est fermée. Ces contraintes biologiques devront être intégrées dans un modèle d'adhésion. Elles constituent un résumé quantitatif des fonctions du complexe cadhérine-caténine dans l'adhésion cellulaire.

Nous venons de nous intéresser au rôle joué par le complexe Cadhérine-Caténine dans

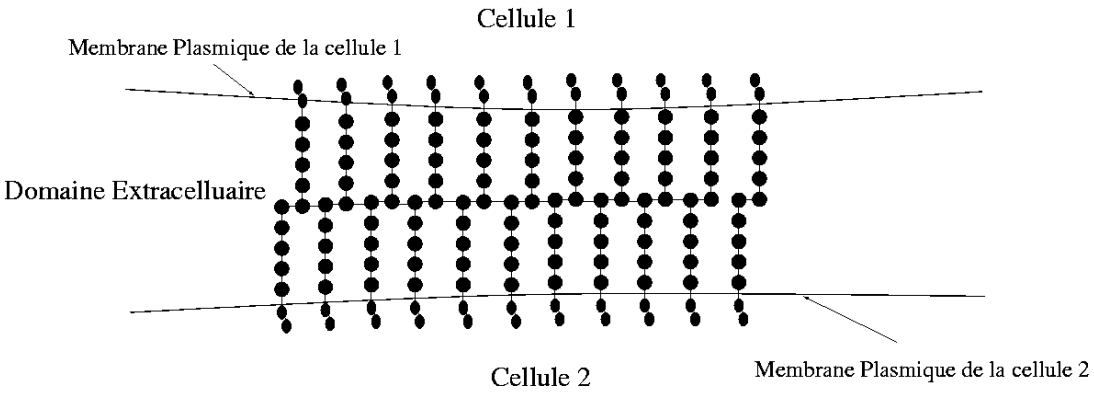



**FIG. 3.5:** *Schéma de type fermeture éclair d'une liaison cellule-cellule assurée par des cadhérines. Ce schéma est inspiré des travaux de Shapiro [Shapiro et al., 1995].*

l'adhésion cellulaire. Comme nous l'avons vu, ce rôle est prépondérant pour le maintien et l'intégrité d'un organisme pluricellulaire. Etant donné que le cancer est la manifestation du développement anormal d'un organisme, nous pouvons raisonnablement penser que le complexe Cadhérine-Caténine est directement impliqué dans le développement d'un cancer [Conacci-Sorrell et al., 2002]. Dans la prochaine section nous allons nous intéresser au lien existant entre l'adhésion cellulaire, et tout particulièrement le complexe Cadhérine-Caténine, et le développement d'un cancer.

### 3.2.5 Adhésion cellulaire et cancer

Depuis quelques années, un lien a été fait entre l'adhésion cellulaire et le développement d'une tumeur. En effet, il est apparu expérimentalement qu'un dysfonctionnement du processus d'adhésion cellulaire peut-être fortement corrélé à plusieurs types de cancer. Ces expériences biologiques se basent sur le développement de nouveaux marqueurs moléculaires qui permettent d'évaluer l'expression de protéines spécifiques au sein des cellules d'un tissu. En particulier, de nombreuses recherches se sont focalisées sur l'expression de plusieurs molécules d'adhésion (les CAMs) dont les cadhérines et caténines. Par imagerie, il est alors possible de déterminer la quantité de protéines cadhérines ou caténines exprimées par chacune des cellules composant un même tissu, et ainsi d'établir

une corrélation entre le niveau d'expression des complexes Cadhérine-Caténine dans une cellule et le type de cette cellule (saine, pré-tumorale ou tumorale, par exemple).

Pendant le développement de la plupart des cancers épithéliaux humains, la fonction adhérente des cadhérines est perdue [Birchmeier et Behrens, 1994]. Parmi les mécanismes entrant en jeu dans la perte de fonction des E-cadhérines, des mutations délétères du gène codant pour la protéine E-cadhérine ont été caractérisées [Bracke et al., 1996]. L'utilisation de $\beta$-caténines tronquées ainsi que la dérégulation des $\beta$-caténines sont associées à l'émergence de tumeurs malignes [Oyama et al., 1994]. Takayama et ses collaborateurs furent les premiers à mettre en avant la sous-expression de $\beta$-caténines dans différents types de cancers : sous-expression dans 10 cas sur 15 (67%) pour l'œsophage, 9 cas sur 19 (47%) pour l'estomac et 11 cas sur 22 (50%) pour le colon [Takayama et al., 1996]. D'autres mécanismes affectent directement l'expression et/ou la fonction des cadhérines : le réarrangement de la chromatine, l'hyperméthylation et la perte de facteurs transcripteurs [Birchmeier et Behrens, 1994], [Bracke et al., 1996].

La non-activation de certains signaux intracellulaires se répercute sur l'adhésion cellulaire gérée par les cadhérines. En particulier, les signaux mettant en jeu les petites GTPases permettent l'expression de la molécule IQGAP1. Cette molécule est en compétition avec les $\beta$-caténines pour s'associer aux $\alpha$-caténines, régulant ainsi le lien entre les E-cadhérines et le cytosquelette. Une dérégulation de ces signaux engendre une sur- ou une sous-expression des complexes $\alpha$-caténines - $\beta$-caténines, et ainsi une mauvaise régulation des jonctions adhérentes [Kuroda et al., 1998]. De même, la phosphorylation des $\beta$-caténines par une tyrosine kinase peut conduire au démontage de complexes Cadhérine-Caténine et donc à son dysfonctionnement.

Il apparait donc que les complexes Cadhérine-Caténine jouent un rôle de suppresseur de tumeurs. Cette constatation est renforcée par des études *in vivo* qui montrent que l'expression forcée de cadhérines dans des lignées cellulaires malignes induit une inversion du phénotype cellulaire d'invasif à bénin.

Une question intéressante est de savoir si la perte d'adhésion cellulaire est un prérequis à une progression tumorale ou simplement une conséquence. Selon Christofori et Stemb, la perte de fonction adhérente des cadhérines ne constituerait qu'une étape du processus d'invasion des cancers [Christofori et Semb, 1999]. La seule perte d'adhésion cellulaire

n'est pas suffisante pour induire une invasion tumorale puis des métastases.

De nombreux articles de synthèse exposent les différentes causes et conséquences d'une dérégulation de complexes Cadhérine-Caténine chez l'être humain [De-Wever et al., 2001], [Foty et Steinberg, 2004] et [Wijnhoven et al., 2000].

Il existe une très forte corrélation entre la perte d'expression de cadhérines et l'aggressivité de plusieurs types de tumeurs. Parmi les tumeurs les plus étudiées, nous pouvons citer

- le cancer du sein [Berx et VanRoy, 2001],
- le cancer du poumon [Bremnes et al., 2002],
- le cancer de la peau [McGary et al., 2002],
- le cancer du pancréas [Joo et al., 2002a],
- le cancer du col de l'utérus [Chen et al., 2003],
- le cancer de l'endomètre [Saito et al., 2003],
- les cancers gastriques [Joo et al., 2002b],
- le cancer de la tyroïde [Scheumman et al., 1995],
- le cancer de l'œsophage [Doki et al., 1993],
- le cancer du colon [Dorudi et al., 1993],
- le cancer du rein [Shimazui et al., 1995],
- le cancer de la vessie [Giroldi et al., 1999].

Pour d'autres types de tumeur, la corrélation est nettement moins évidente, et reste une source de controverse. Par exemple, pour le cancer de la prostate, une étude sur Tissue MicroArray a montré que les molécules de cadhérines ne sont pas un fort « marqueur de pronostic » même si une différence dans l'expression était généralement observée [Rubin et al., 2001]. Pour le cancer du cerveau, la corrélation est également mal caractérisée [Schweichheimer et al., 1998].

En résumé, nous avons vu dans cette section toute l'importance du complexe Cadhérine-Caténine dans l'adhésion cellulaire. D'une part, les molécules cadhérines jouent le rôle de « colle » entre deux cellules. La liaison de deux molécules cadhérines, issues de deux cellules différentes, permet l'adhésion entre ces deux cellules. D'autre part, par voies de signalisation complexes, les molécules cadhérines et caténines agissent indirectement sur

le phénotype de la cellule et en particulier sur son caractère adhérent.

Nous avons également vu que la nature de l'adhésion entre deux cellules est directement liée à l'expression des molécules du complexe Cadhérine-Caténine. En effet, une liaison adhérente entre deux cellules se schématise par une fermeture éclair, par conséquent, la force de cette liaison dépend directement de la nature des crans (les cadhérines) de la fermeture éclair.

Du point de vue de la modélisation, nous pouvons retenir qu'un tissu est une configuration de cellules. Cette configuration est le résultat d'une minimisation d'énergie, elle-même fonction de l'adhésion entre cellules voisines. Pour être relativement complet, l'adhésion entre cellules, qui se matérialise par une fermeture éclair, doit tenir compte du type (ou phénotype) des deux cellules voisines ainsi que de la longueur de la membrane entre ces deux cellules.

Nous allons à présent dresser un état de l'art des modèles de tissus biologiques mettant en jeu spécifiquement l'adhésion intercellulaire.

## 3.3 Modèles mathématiques d'adhésion cellulaire

En 1952, Turing fut l'un des premiers mathématiciens à étudier la formation de configurations spatiales dans des organismes biologiques et en particulier des organismes multicellulaires [Turing, 1952]. Il proposa alors un modèle à base d'équations de réaction-diffusion. Ce modèle a été largement commenté par Wolpert [Wolpert, 1969], puis étudié dans des cas particuliers [Gierer et Meinhardt, 1972], [Thom, 1975], [Gray et Scott, 1984] et [Lengyel et Epstein, 1991]. Ces modèles, couplés à des modèles de mouvements cellulaires induits par diffusion ont permis de rendre compte de configurations biologiques, comme le damier [Oster et al., 1983], [Ngwa et Maini, 1995]. En 2000, Painter et ses collaborateurs ont notamment considéré le cas où les cellules répondaient à l'action de deux composants chimiques dans un système de Turing [Painter et al., 2000].

Pour ces modèles, l'interaction cellule-cellule n'est pas directement prise en compte, mais s'opère par principe de diffusion. Ces modèles ne s'inscrivent donc pas dans les hypothèses d'adhésion cellulaire présentées à la section 3.1.2.

### 3.3.1 Les modèles mathématiques basés sur l'hypothèse DAH

Les expériences de Steinberg l'ont conduit à développer l'hypothèse d'adhésion DAH. Bien que n'étant vérifiée que dans des cas particuliers, cette hypothèse reste aujourd'hui la plus reconnue pour expliquer l'arrangement des cellules dans un tissu. L'hypothèse DAH postule que l'organisation cellulaire est la conséquence d'une minimisation d'énergie issue de l'adhésion entre les cellules. Cette constatation permet d'établir un lien entre la physique statistique et l'organisation cellulaire en biologie. De nombreux modèles ont donc été développés en ce sens et proposent des fonctions d'énergie, que nous appellerons *Hamiltoniens*, permettant d'inclure d'une part les connaissances biologiques et d'autre part les hypothèses biologiques à tester.

Plus précisemment, l'hypothèse DAH suppose que l'énergie d'un système cellulaire est essentiellement caractérisée par des interactions entre cellules adhérentes. Dans la plupart des travaux, « la cellule est considérée comme l'entité primordiale pour laquelle il convient de considérer son voisinage spatial comme une population organisée dans laquelle la communication joue un rôle structurant » [Wolpert, 1969]. Dans une modélisation géométrique d'un tissu, associée à une topologie de voisinage, cette remarque suppose que deux cellules vont interagir uniquement si elles sont voisines. Ainsi, les Hamiltoniens étudiés se décomposent en une somme de deux potentiels : un potentiel de singleton associé à chaque cellule du tissu, et un potentiel de paires associé à chaque relation de voisinage entre deux cellules du tissu.

Les modèles proposés sont usuellement classés en deux catégories principales : d'une part, les modèles déterministes et d'autre part, les modèles non-déterministes. Pour les modèles déterministes, nous décrirons deux modèles particuliers proposés par Sulsky [Sulsky et al., 1984] et Graner et Sawada [Graner et Sawada, 1993]. Ces modèles proposent une expression fidèle de l'hypothèse DAH. Les modèles non-déterministes quant à eux, sont usuellement classés en différentes catégories selon la géométrie choisie pour modéliser une cellule. Nous trouvons tout d'abord les modèles de réseau pour lesquels chaque cellule a la même géométrie et représente un élément d'une grille régulière. Par la suite, des modèles sur géométrie continue ont été développés. Enfin des modèles hybrides, pour lesquels une cellule est représentée par un nombre aléatoire d'éléments d'une grille discrète, ont été étudiés.

### 3.3.2 Modèles déterministes

#### Modèle de Sulsky

En 1984, Sulsky et ses collaborateurs proposent un modèle étudiant l'organisation cellulaire d'un tissu [Sulsky et al., 1984]. Les auteurs se placent explicitement sous l'hypothèse DAH, et choisissent de modéliser les cellules biologiques par des cellules de Dirichlet centrées sur les noyaux des cellules biologiques. Ce choix de géométrie est motivé par une étude parallèle sur une méthode Lagrangienne pour la résolution d'équations de Navier Stokes portant sur la vitesse et la localisation spatiale de $N$ marqueurs fluides [Borgers et Peskin, 1987].

Les auteurs proposent comme Hamiltonien, noté $H_S$ et correspondant à l'énergie à minimiser, l'expression :

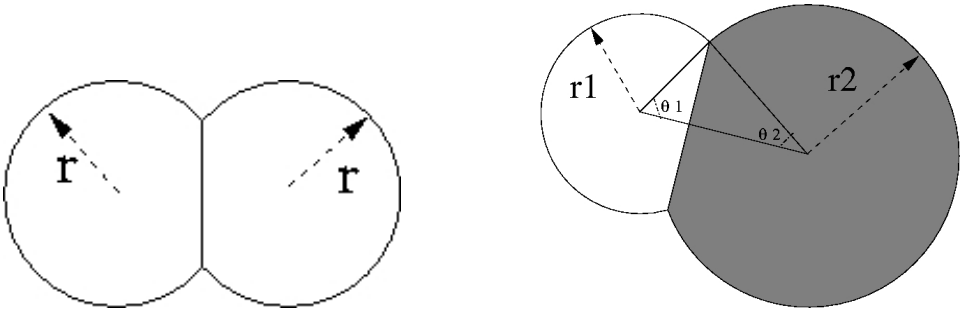
$$H_S = \sum_{i \sim j} l_{ij} e_{ij}, \tag{3.1}$$

où $i \sim j$ signifie que $i$ et $j$ sont voisins au sens de Delaunay, $e_{ij}$ est la tension de surface par unité de longueur associée à l'arête de Dirichlet commune aux cellules $i$ et $j$ et $l_{ij}$ représente la longueur de l'arête de Dirichlet associée aux cellules $i$ et $j$.

Les auteurs se proposent de formuler une dynamique qui non seulement permettrait d'atteindre l'état d'équilibre mais qui décrirait également l'évolution du tissu pour atteindre cet état d'équilibre. Pour cela, les auteurs utilisent un parallèle entre le comportement de cellules dans un tissu et le comportement d'un liquide visqueux élastique [Phillips et Steinberg, 1969]. Par la suite les auteurs se placent dans un contexte déterministe de résolution par équations aux dérivées partielles. Les auteurs ont ensuite réalisé des simulations permettant de reproduire des situations biologiques telles que la séparation ou l'engloutissement total ou partiel. Le principal intérêt de ce modèle est de vérifier que la modélisation des cellules par le diagramme de Dirichlet est satisfaisante et permet une meilleure flexibilité dans le modèle par rapport à des modèles incluant la mécanique des membranes.

#### Modèle de Graner et Sawada

Graner et Sawada ont introduit un modèle géométrique de cellules permettant de modéliser le réarrangement cellulaire par adhésion de surface [Graner et Sawada, 1993].




(a) (b)

**FIG. 3.6: Schéma de domaines de Dirichlet libres [Graner et Sawada, 1993]**. *Dans ce modèle géométrique, la distance entre la membrane et le noyau est contrainte à être inférieure à une certaine valeur $r$. La figure $(a)$ représente le cas de deux cellules de même type (ayant le même $r$). Si deux noyaux sont situés à une distance inférieure à $2r$, alors leur surface commune correspondra à la surface de Dirichlet. La figure $(b)$ représente le cas de 2 cellules de types différents (ayant 2 rayons différents, $r_1$ et $r_2$). Dans ce cas, pour garantir l'unicité des sommets des membranes, les rayons $r_1$ et $r_2$ ainsi que les angles $\theta_1$ et $\theta_2$ doivent respecter les contraintes : $r_1.sin(\theta_1) = r_2.sin(\theta_2)$ et $r_1.cos(\theta_1) + r_2.cos(\theta_2) = r$, où $r$ est la distance entre les deux centres des deux cellules.*

Les auteurs introduisent des domaines de Dirichlet libres (voir figure 3.6). Le domaine de Dirichlet libre d'un point $p$, noté $DL(p)$, correspond à son domaine de Dirichlet privé de tous les points de l'espace se trouvant à une distance supérieure à $r$ de $p$. Les auteurs définissent l'Hamiltonien de leur système par une formulation proposée par [Graner, 1993], analogue à la formule de Sulsky [Sulsky et al., 1984] (voir Equation 3.1) :

$$H_{GS} = \sum_{i \sim j} S_{ij} e_{ij} + \sum_{i} S_{iM} e_{iM} \tag{3.2}$$

où la première somme se fait sur l'ensemble des cellules $i$ et $j$ voisines, $S_{ij}$ représente l'énergie surfacique entre les cellules $i$ et $j$ tandis que $e_{ij}$ correspond à la longueur de

la surface commune entre $i$ et $j$. La deuxième somme se fait sur l'ensemble des cellules, $S_{iM}$ représente l'énergie surfacique entre la cellule $i$ et le milieu extérieur, tandis que $e_{iM}$ correspond à la longueur de la surface entre la cellule $i$ et le milieu extérieur.

En ajustant les paramètres, les auteurs parviennent à simuler différentes situations biologiques telles l'engloutissement, la configuration en damier, l'auto-arrangement partiel ou total. Le principal intérêt de ce modèle est de fournir une géométrie particulière aux cellules biologiques. Cette géométrie assure notamment que le périmètre ($\approx$ la surface de contact) de la cellule est borné, inférieur à $2\pi r$. De plus, comme le modèle travaille à nombre de cellules borné (il n'y a pas de création de cellules), l'énergie $E_{adh}$ est toujours bornée. Cependant les auteurs mettent en avant les problèmes de stabilité numérique de leur algorithme. Il propose, comme alternative, une simulation par méthode de Monte-Carlo, analogue à celle utilisée chez Graner et Glazier, 1992 [Graner et Glazier, 1992]. Ce type de simulation garantit sous certaines conditions la convergence de l'algorithme.

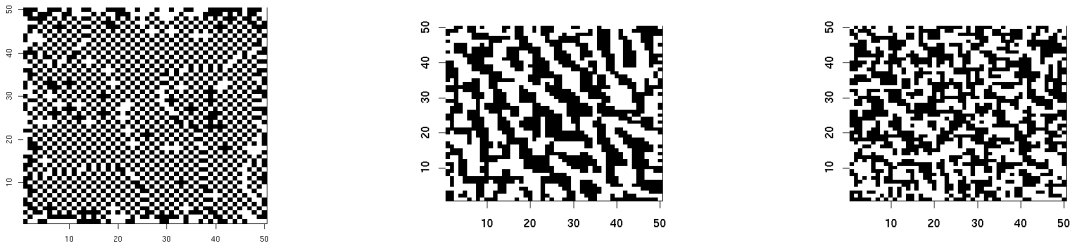
Nous allons à présent nous focaliser sur les modèles non-déterministes. Ces modèles peuvent se différencier par la géométrie choisie pour représenter un tissu.

### 3.3.3 Modèles sur grille : « Cell-lattice models »

Les modèles sur grille sont les premiers modèles développés pour étudier spécifiquement les interactions cellule-cellule [Goel et al., 1970]. Ces modèles caractérisent le tissu comme un ensemble de cellules où chaque cellule est représentée géométriquement par un polygone régulier (carré ou hexagone). Grâce à la facilité de programmation et la rapidité d'exécution des algorithmes associés, ces modèles ont remporté un franc succès. Ils permettent également de simuler un large panel de phénomènes avec une interprétation simple des paramètres.

Mochizuki et ses collaborateurs ont étendu ce modèle à un modèle stochastique [Mochizuki et al., 1996]. Dans ce cas, les cellules, de deux types histologiques différents (type I et II), sont réparties sur une grille régulière. Les auteurs se placent dans le cadre de l'hypothèse DAH et déterminent la probabilité de passer de la configuration $\pi$ à la configuration $\pi'$ en un temps $\Delta t$ par l'expression :

$$P(\pi \to \pi' \text{ en un temps } \Delta t) = \frac{2m}{1+\exp(\frac{-\Delta H}{m})}\Delta t + O(\Delta t),$$



$(a)$ $(b)$ $(c)$

**Fig. 3.7: Exemples de réalisations du modèle de Mochizuki**. *Le cas (a) représente des relations hétérotypiques, le cas (b) des relations homotypiques et le cas (c) des relations aléatoires.*

où $m$ est le taux de changement de type par cellule, $\Delta H$ correspond au changement de la force totale d'adhésion (une forme d'énergie d'adhésion) entre la configuration $\pi$ et $\pi'$. Des exemples de simulation sont proposés à la figure 3.7. Les auteurs proposent ensuite deux indices pour mesurer l'adhésion cellulaire : la fraction de cellules de type I dans le voisinage d'une cellule de type I d'une part, et le nombre de cellules de type I d'autre part.

Le modèle de Mochizuki a été adapté par Takano et ses collaborateurs pour étudier le phénomène de séparation de tissus [Takano et al., 2002]. Un troisième type de cellule correspondant à des cellules neutres a été introduit tandis que les deux autres types de cellules vont correspondre à deux types de tissus différents. La géométrie choisie pour les cellules est une géométrie hexagonale. Dans cet article, les auteurs se sont placés sous l'hypothèse DAH et ont caractérisé les forces d'adhésion par l'expression de molécules de cadhérines au sein de chaque cellule. Ils concluent qu'un changement de la quantité d'expression de cadhérines ne suffit pas pour induire la séparation de tissus. Cependant ce phénomène permet un nouveau réarrangement spatial du tissu.

Ce type de modèle, très simple, s'apparente à des automates cellulaires et permet la simulation de nombreux phénomènes biologiques. Cependant, la simplicité de cette catégorie de modèles exclut la validation d'hypothèses biologiques. En effet, chaque cellule,

peu importe son type, est modélisée soit par un carré, soit par un hexagone, de taille identique. Les échanges entre les cellules ne sont donc pas très réalistes du point de vue quantitatif puisque les surfaces sont toutes identiques et fixes.

### 3.3.4 Modèles sur géométrie continue : modèles centrés et modèles sur sommets

Afin de rendre la modélisation géométrique plus réaliste, des modèles à géométrie continue ont été développés. Ces modèles proposent de modéliser chaque cellule par un polygone aléatoire. La dynamique liée aux phénomènes d'adhésion cellulaire est intégrée aux modèles en permettant le mouvement soit du centre de la cellule (modèles centrés) soit des sommets des polygones formant les cellules (modèles sur sommets).

#### Modèles centrés

Les modèles centrés partent du constat que les cellules biologiques sont assez bien caractérisées par des cellules de Dirichlet centrées aux centres de masse des cellules biologiques, correspondant aux noyaux des cellules [Honda, 1978] et [Honda, 1983]. L'avantage de cette méthode est que la configuration du tissu est entièrement déterminée par les noyaux des cellules. Cela permet d'alléger les contraintes de modélisation concernant les interactions.

Un modèle permettant de simuler l'organisation spontanée de cellules épithéliales chez les mammifères a été étudié dès 1996 [Honda et al., 1996]. Par la suite d'autres géométries de cellules ont été introduites. En 2000, la différenciation des cellules des ailes de papillon a été modélisée grâce à un modèle centré [Honda et al., 2000]. Pour cela, les auteurs ont choisi de modéliser les cellules par un diagramme de Dirichlet pondéré et le phénomène de différenciation par un processus d'inhibition aux plus proches voisins.

Dans les modèles centrés, la géométrie des cellules est donc calculée exclusivement à partir du processus ponctuel formé par les noyaux des cellules. Par conséquent, il n'y a pas de contrôle direct sur la forme des cellules. Ainsi des cellules à géométrie fortement anisotropique ne pourront jamais être simulées.

**Modèles sur sommets**

Une autre catégorie de modèles d'organisation cellulaire à géométrie continue consiste à utiliser des polygones non centrés pour la géométrie des cellules. En effet, ces modèles déterminent la géométrie d'une cellule par un polygone. Les contraintes de modélisation vont alors s'appliquer aux sommets de ces polygones pour permettre une approximation plus fine et plus flexible de la géométrie d'une cellule.

Dans cette catégorie de modèles, les forces entrant en jeu lors des processus d'adhésion cellulaire ou de mécanique cellulaire sont reportées sur les sommets des polygones (en 2D, les sommets des polygones sont localisés aux points d'intersection de trois arêtes).

Cette classe de modèles a été introduite pour étudier notamment la dynamique des bulles de savon et des polycristaux [Nagai et al., 1988]. Certaines modifications ont été apportées à ce modèle initial pour traiter des cellules biologiques en incluant notamment un terme de potentiel pour contraindre les mouvements des sommets des polygones en 2D [Nagai et Honda, 2001] puis en 3D [Honda et al., 2004]. Plus précisemment, ce dernier modèle étudie le comportement des cellules dans un agrégat homogène. Pour cela, les auteurs utilisent une équation de mouvement pour les sommets :

$$\eta \frac{\mathrm{d}r_i}{\mathrm{d}t} = -\nabla_i U,$$

où, $\eta$ est l'analogue d'un coefficient de viscosité, $r_i$ est la position de la cellule $i$, $t$ le temps et $U$ une fonction potentielle. Cette fonction $U$ peut-être vue comme un Hamiltonien. Les auteurs décomposent cet Hamiltonien en une somme de 2 termes principaux : $U = U_s + U_v$, où $U_s$ représente l'énergie de surface, $U_v$ l'énergie de compression des polygones.

Le principal problème de ces méthodes est qu'elles ne permettent pas l'incorporation de contraintes mécaniques à l'intérieur d'une cellule. En effet, l'action du cytosquelette sur la forme de la cellule est un élément primordial de développement cellulaire qui ne peut pas être appliqué dans ce cadre.

### 3.3.5 Modèles sous-latticiels (Potts étendu)

Les modèles qui ont connu le plus de succès sont des modèles hybrides inspirés d'un modèle initial proposé par Graner et Glazier [Graner et Glazier, 1992]. Ce modèle constitue l'une des bases de notre modélisation.

**Modèle initial de Graner et Glazier ou modèle de Potts étendu**

En 1992, Graner et Glazier proposent un modèle original qui traite du problème d'auto-arrangement des cellules embryonnaires [Graner et Glazier, 1992]. Les auteurs ont choisi un compromis intéressant pour la géométrie des cellules. Partant du constat qu'un modèle sur grille est trop restrictif pour modéliser finement une cellule et qu'un modèle à géométrie continue ne permet pas, au contraire, d'inclure suffisamment de contraintes géométriques, les auteurs proposent un modèle à géométrie sous-latticielle pour lequel une cellule est composée de plusieurs pixels d'une grille. Une illustration sommaire du modèle est proposée à la figure 3.8.

Plus précisemment, à chaque point de la grille est affectée une cellule d'appartenance, notée à l'aide de la fonction $\sigma$. Par exemple, le point de coordonnées $(i, j)$ dans la grille appartient à la cellule $\sigma(i, j)$. Ainsi, une cellule est composée de plusieurs pixels (ou carrés de la grille), permettant une certaine flexibilité dans les géométries des cellules. Puis, à chaque cellule est affecté un type, noté $\tau$. Ainsi le type de la cellule à laquelle appartient le pixel $(i, j)$ vaut $\tau(\sigma(i, j))$.

Dans le modèle original, les cellules sont de trois types différents correspondant aux types histologiques : type « light », type « dark" et type « Medium ». Les cellules de type « Medium » composent le milieu extracellulaire comme par exemple la solution de culture, le substrat ou la matrice extracellulaire, tandis que les types « light » et « dark » caractérisent deux types de cellules actives.

L'Hamiltonien du système proposé par Graner et Glazier s'écrit comme une somme de deux termes : le premier terme modélise l'adhérence entre cellules par un modèle de Potts, et le second terme permet de contraindre l'aire des cellules du tissu par leur élasticité. Ce second terme s'interprète comme une contrainte de forme sur la cellule. L'Hamiltonien s'écrit :

$$H_{GG} = \sum_{(i,j)\sim(i',j')} J\left(\tau(\sigma(i,j)), \tau(\sigma(i',j'))\right)\left(1 - \delta_{\sigma(i,j),\sigma(i',j')}\right) + \lambda \sum_{\sigma} (a(\sigma) - A_{\tau(\sigma)})^2 \Gamma(A_{\tau(\sigma)}) \tag{3.3}$$

La première somme se fait sur l'ensemble des paires de pixels voisins, $(i, j) \sim (i', j')$ signifie que les pixels $(i, j)$ et $(i', j')$ sont voisins. Le voisinage choisi par les auteurs est un voisinage octogonal pour lequel les huit pixels qui entourent le pixel $(i, j)$ forment

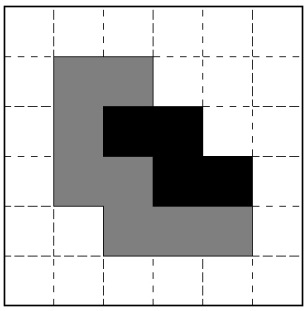


**Fig. 3.8:** *Un exemple de configuration du modèle de Graner et Glazier. Dans cette réalisation, le tissu est composé de trois cellules. Une cellule de type « Medium » en blanc, composée de 24 pixels, une cellule de type « light » en gris, composée de 8 pixels, et une cellule de type « dark » en noire, composée de 4 pixels.*

le voisinage de $(i,j)$ (voir figure 3.9). Cette première somme correspond au potentiel de paires en physique statistique. La fonction $J$ représente l'interaction entre deux pixels voisins en fonction des types des cellules auxquelles appartiennent ces deux pixels, $\delta$ représente le symbole de Kronecker. Par conséquent l'interaction entre deux pixels voisins appartenant à la même cellule est nulle.

La deuxième somme se fait sur l'ensemble des cellules, correspondant à un potentiel de singleton en physique statistique. Dans ce terme, $\lambda$ caractérise le coefficient d'élasticité du tissu, $a(\sigma)$ est l'aire de la cellule $\sigma$, $A_\tau$ l'aire cible d'une cellule de type $\tau$ et $\Gamma$ est la fonction de Heavyside : $\Gamma(x) = \mathbb{1}_{\mathbb{R}_+^*}(x)$.

Ce modèle, appelé par les auteurs, le modèle de Potts étendu, a été appliqué à un grand nombre de processus biologiques par le biais de simulations utilisant une version de l'algorithme de Metropolis [Glazier et Graner, 1993]. Une itération de l'algorithme se déroule de la façon suivante

- un pixel $(i,j)$ de la grille est choisi au hasard,
- un pixel voisin de $(i,j)$, noté $(i',j')$ est choisi au hasard parmi les huit voisins,
- si les $(i,j)$ et $(i',j')$ n'appartiennent pas à la même cellule (c'est à dire si $\sigma(i,j) \neq \sigma(i',j')$), alors la différence d'énergie, $\Delta H_{(i,j)\to(i',j')}$ pour changer la cellule d'ap-

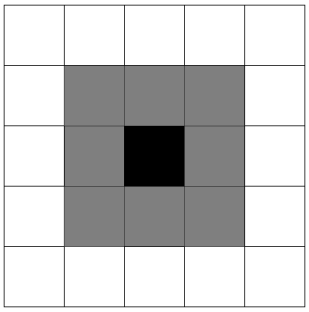


**FIG. 3.9:** *Le pixel $(i, j)$ est colorié en noir. Le voisinage du pixel $(i, j)$ est colorié en gris.*

partenance de $(i, j)$, notée $\sigma(i, j)$, en $\sigma(i', j')$ est calculée,

- si cette différence est négative ($\Delta H_{(i,j)\to(i',j')}$), le changement est accepté, sinon le changement est accepté avec une probabilité égale à $\exp(-\Delta H_{(i,j)\to(i',j')}/kT)$, où $k$ est la constante de Boltzmann et $T$ la température du système.

Pour chacun de ces processus biologiques (interactions hétérotypiques, interactions homotypiques, engloutissement, auto-arrangement cellulaire total ou partiel et dispersion cellulaire), les auteurs se sont efforcés de caractériser qualitativement les paramètres choisis et l'influence de la température [Glazier et Graner, 1993]. La figure 3.10 montre un exemple de calcul d'un Hamiltonien et la figure 3.11 explique une itération de l'algorithme de simulation.

Selon les auteurs, un intérêt majeur du modèle de Potts étendu réside dans le fait que la dynamique interdit des grands sauts énergétiques. En effet, à chaque itération seul un pixel est susceptible d'être modifié ce qui veut dire que seule une petite partie de la surface entre deux cellules peut être modifiée. Ainsi, contrairement aux modèles à géométrie continue, le modèle de Potts étendu risque de ne parcourir que le bassin énergétique auquel appartient la condition initiale. Dans leurs campagnes de simulation, les auteurs fixent les paramètres de contraintes d'aires (paramètres $A_{\tau(\sigma)}$) et font varier les paramètres liés aux forces d'adhésion intercellulaires (paramètres $J$). Une condition initiale pour laquelle le potentiel lié à la contrainte d'aire des cellules est déjà optimisé, est alors fixée pour l'ensemble des simulations. Ce faisant les auteurs se placent initialement

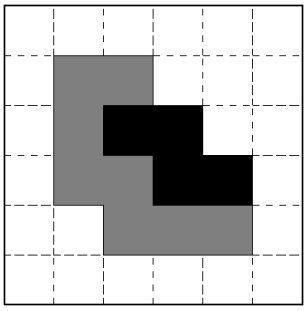


**Fig. 3.10:** *Un exemple de calcul d'Hamiltonien pour le modèle de Graner et Glazier. Dans cette réalisation, le tissu est composé de trois cellules, une cellule blanche de type « Medium », une cellule grise de type « light » et une cellule noire de type « dark ». Les 3 couleurs sont codées 1 pour noir, 2 pour gris et 3 pour blanc. En choisissant comme paramètres, d'une part la fonction $J$ symétrique telle que $J(i,i) = 0$ pour tout $i$, $J(1,2) = 1$, $J(1,3) = 4$ et $J(2,3) = 3$, et d'autre part la fonction $A$, telle que $A(1) = 3$, $A(2) = 6$ et $A(3) = -1$, où l'unité de surface est un carré, l'Hamiltonien de cet état vaut : $H_{GG} = (13\times1)+(9\times4)+(29\times3)+(4-3)^2+(8-6)^2 = 136+1+4 = 141$.*

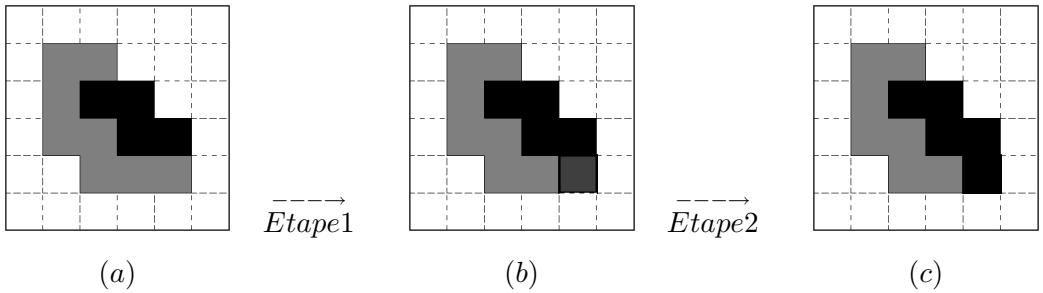



**Fig. 3.11:** *Exemple d'une itération de l'algorithme de Métropolis-Hastings pour le modèle de Graner Glazier. Les paramètres choisis sont d'une part la fonction $J$ symétrique telle que $J(i,i) = 0$ pour tout $i$, $J(1,2) = 1$, $J(1,3) = 4$ et $J(2,3) = 3$, et d'autre part la fonction $A$, telle que $A(1) = 3$, $A(2) = 6$ et $A(3) = -1$, où l'unité de surface est un carré. En $(a)$, l'Hamiltonien du système vaut $H_{GG} = 141$ (Cf figure 3.10). La première étape de l'itération consiste à choisir un site au hasard, représenté par un carré en gris foncé sur la figure $(b)$. La deuxième étape consiste à choisir une couleur pour ce site. Pour ce faire, l'algorithme choisit une couleur aléatoire parmi les couleurs des voisins du site considéré. En $(c)$, la couleur choisie est celle du voisin au dessus du site : la couleur noire. La différence des Hamiltoniens est alors calculée localement $\Delta H_{GG} = (21 - 22) + \left((5-3)^2 + (7-6)^2 - (4-3)^2 - (8-6)^2\right) = -1$. La configuration $(c)$ est alors acceptée avec une probabilité $p = e^{(-1/(kT)*q)}$, où $q = \min(0, -\Delta H_{GG})$, $k$ est la constante de Boltzmann et $T$ la température. Dans notre exemple, la transition est acceptée avec la probabilité* 1.

dans une même cuvette de potentiel, réaliste du point de vue biologique, pour la géométrie des cellules. Grâce à cette condition initiale optimisée, le modèle de Potts étendu décrit avec succès de nombreuses configurations biologiques.

Ce succès a inspiré par la suite de nombreux autres modèles de simulations et tout particulièrement des modèles appliqués à la dynamique d'un cancer.

### Modèles issus du modèle de Potts étendu appliqués à une problèmatique cancéreuse

Le modèle de Potts étendu a été récemment utilisé pour modéliser la dynamique de tissus cancéreux. Certains phénomènes, comme la croissance en milieu non vascularisé ou l'invasion, ont été particulièrement étudiés. Nous allons décrire ici deux extensions particulières du modèle de Potts étendu : le modèle de Scott et les modèles de Turner.

#### Le modèle de Scott

En 1999, Scott et ses collaborateurs ont appliqué le modèle de Potts étendu à la croissance d'une tumeur dans un milieu non vascularisé [Scott et al., 1999]. Les auteurs ont légèrement complexifié le modèle pour inclure une dynamique de réponse à des facteurs de croissance et à la diffusion de nutriments. Pour résumer, les différents types cellulaires, ainsi que les coefficients d'élasticité, sont vus comme des fonctions des concentrations de nutriments. Les auteurs supposent qu'un tissu normal contient une source homogène de nutriments. La concentration de nutriments au sein de la tumeur est alors calculée en fonction de la distance des cellules saines à la périphérie de la tumeur et du volume total de cellules actives (et donc consommatrices de nutriments) au sein de la tumeur. Le taux de croissance des cellules en prolifération dépend lui aussi de la concentration de nutriments. Les auteurs concluent que malgré d'importantes hypothèses simplificatrices, ce modèle de simulation de croissance de tumeurs en milieu non vascularisé reproduit des résultats expérimentaux sur la croissance du volume d'une tumeur.

#### Les modèles de Turner

Le processus d'invasion du cancer a été traité en utilisant le modèle de Potts étendu. Par des arguments biologiques et par souci de réalisme, les auteurs ont complexifié le

modèle en incluant le phénomène d'haptotaxie [Turner et Sherratt, 2002]. L'haptotaxie est l'action du substrat sur le déplacement des cellules. En effet, le substrat peut guider les cellules par adhésion vers leurs régions de préférence. Les auteurs ont modélisé l'haptotaxie en attachant à chaque point $(i,j)$ de la grille du modèle un paramètre $f_{ij}$, correspondant à la concentration de protéines du substrat au point $(i,j)$.

Ainsi, le potentiel Hamiltonien du modèle de Turner et Sheratt s'exprime par la formule suivante :

$$\begin{aligned} H_T &= \sum_{(i,j)\sim(i',j')} J\left(\tau(\sigma(i,j)),\tau(\sigma(i',j'))\right)\left(1-\delta_{\sigma(i,j),\sigma(i',j')}\right) \\ &\quad +\lambda\sum_{\sigma}(a(\sigma)-A_{\tau(\sigma)})^2\Gamma(A_{\tau(\sigma)})+\sum_{(i,j)}k_H f_{ij} \end{aligned}$$

Les deux premiers termes de la somme correpondent exactement à la formulation du potentiel Hamiltonien du modèle de Potts étendu. Le dernier terme représente le potentiel lié à l'haptotaxie, où $k_H$ définit la force de l'haptotaxie relativement à la force d'adhésion. La simulation de ce modèle s'effectue à l'aide d'une dynamique de Métropolis similaire à celle utilisée pour la simulation du modèle de Potts étendu.

Ce modèle a permis aux auteurs de conclure qu'une modification de la force d'adhésion entre une cellule et le substrat avait plus d'impact qu'une modification de la force d'adhésion entre cellules. De plus une modification de la force d'adhésion entre cellules n'a quasiment aucune influence sur l'invasion sauf si le taux d'expression de protéases est élevé.

Dans un article suivant, une technique permettant d'étendre la version discrète du modèle de Potts étendu à une version continue de diffusion est proposée [Turner et al., 2004]. Pour cela, les auteurs observent, à partir de simulations du modèle de Potts étendu, le comportement global d'un ensemble de cellules. Ils déduisent de ces simulations, en calculant la surface couverte par les cellules, un coefficient de diffusion. Ainsi, les auteurs peuvent déterminer une relation explicite entre les coefficients du modèle de Potts étendu et l'expression du coefficient de diffusion associé. Cette relation permet d'éclaircir les liens existant entre des lois microscopiques (l'interaction entre deux cellules voisines) et des comportements macroscopiques (l'invasion d'une tumeur). Ces liens sont d'une importance capitale pour faire le pont entre la biologie moléculaire étudiant le comportement

de la cellule et la thérapeutique cherchant à influer sur le comportement d'organismes multicellulaires.

En 2005, Turner se propose d'étudier le coefficient de diffusion des particules ainsi que la densité à l'équilibre d'un système issu du modèle de Potts étendu [Turner, 2005]. A cet effet, l'auteur se base essentiellement sur les travaux de [Cohen et Murray, 1999] et [Murray, 2002]. Dans ce modèle, les cellules biologiques sont supposées être d'une part adhésives et d'autre part des sphères élastiquement déformables. L'évaluation des coefficients d'adhésion, des coefficients mécaniques d'élasticité et des coefficients de diffusion se fait à partir de simulations du modèle de Potts étendu.

Ce modèle propose des géométries de cellules particulières et peu réalistes. Pour rendre ces cellules plus réalistes, il faut choisir une grille très fine. Cette contrainte affecte grandement la qualité des simulations, comme nous le verrons par la suite.

Le modèle proposé par Graner et Glazier est biologiquement très intéressant puisqu'il permet la simulation de nombreuses configurations classiques en biologie. Cependant, l'étude mathématique de ce modèle se résume à des études par simulation. Ce manque de cadre mathématique nuit à la fiabilité du modèle. Les modèles que nous venons de présenter proposent des complexifications du modèle de Graner et Glazier. Ces complexifications rendent leurs études théoriques encore plus compliquées.

En résumé, dans cette partie, nous avons détaillé les différentes catéogories de modèles mathématiques appliqués à l'adhésion cellulaire. Ces modèles se différencient par la géométrie utilisée pour modéliser la cellule : géométrie sur grille ou continue, mais aussi par la modélisation des interactions considérées. Certains modèles proposent des modèles de réaction-diffusion pour expliquer la dynamique cellulaire. D'autres modèles se basent sur les hypothèses formulées par Steinberg dans les années 60 qui supposait que l'auto-organisation de cellules au sein d'un tissu suivait un processus de minimisation d'énergie.

Dans la plupart des modèles énergétiques, l'énergie à minimiser s'exprime par un Hamiltonien. Au fur et à mesure des découvertes de biologie moléculaire, les auteurs ont complété et/ou amélioré l'expression de l'Hamiltonien pour coller au plus près avec le phénomène biologique qu'ils souhaitaient modéliser. Depuis plusieurs années, de nouvelles

problématiques liées à l'adhésion cellulaire ont émergé. En particulier, le cancer semble affecter fortement les mécanismes d'adhésion cellulaire. Ainsi l'adaptation de ces modèles à un tissu cancéreux apparait comme un nouveau challenge.

## 3.4 Une version continue de modèle de Potts étendu : la classe CC (Cadhérine-Caténine)

Parmi les modèles non-déterministes, le modèle de Potts étendu, introduit par Graner et Glazier est l'un des modèles les plus simples permettant la simulation d'un important spectre de configurations biologiques [Graner et Glazier, 1992]. Ce modèle, basé sur l'hypothèse DAH émise par Steinberg, propose une minimisation d'un Hamiltonien par une dynamique de Métropolis-Hastings. Dans ce chapitre, nous proposons une alternative au modèle de Graner et Glazier dans le cas d'une géométrie continue. L'Hamiltonien choisi par Graner et Glazier est rappelé ci dessous :

$$H_{GG} = \sum_{(i,j)\sim(i',j')} J\left(\tau(\sigma(i,j)),\tau(\sigma(i',j'))\right)\left(1-\delta_{\sigma(i,j),\sigma(i',j')}\right)+\lambda\sum_{\sigma}(a(\sigma)-A_{\tau(\sigma)})^2\Gamma(A_{\tau(\sigma)}).$$

La dynamique de Métropolis-Hastings utilisée par Graner et Glazier est décrite en détail à la section 3.3.5. Dans la suite de cette section, nous allons montrer que l'étude approfondie de ce modèle soulève quelques limites. Ces limites sont principalement dues à la discrétisation de l'espace en pixels.

### 3.4.1 Les limites du modèle de Potts étendu

En premier lieu, Graner et Glazier ont mis en avant le fait qu'aucune contrainte n'était spécifiée quant à la connexité des cellules (voir Figure 3.12). Pour contrôler ce problème, la dynamique de Graner et Glazier utilise une configuration initiale qui satisfait déjà à une géométrie régulière. Etant donné que le paramètre de température est faible, une perte de connexité constitue dans leurs campagnes de simulation un événement rare. Néanmoins, ce problème de connexité est lié au fait que l'espace soit discrétisé : sur un modèle centré, la connexité des cellules est évidente.

En un second lieu, nous pouvons remarquer que l'algorithme utilisé par Glazier et Graner construit une chaine de Markov non reversible. En effet, leur algorithme s'autorise

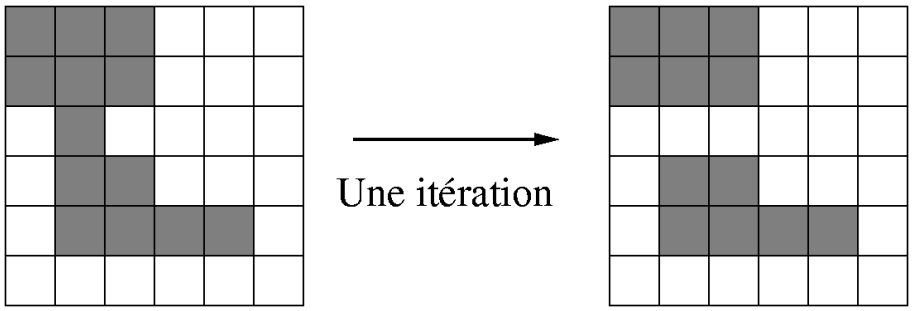
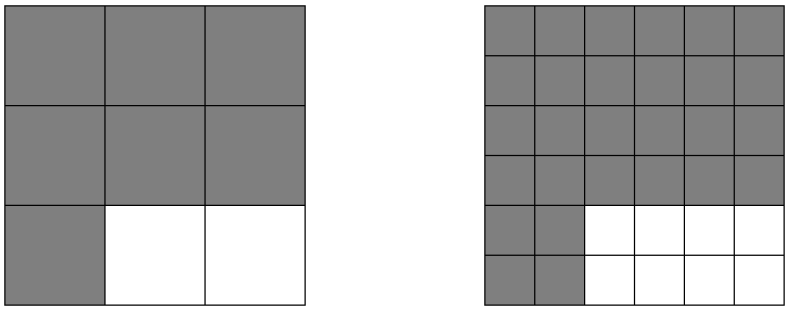



**Fig. 3.12:** *Exemple de perte de connexité d'une cellule. La cellule grise suite à l'itération schématisée ci-dessus n'est plus connexe.*

à « tuer » certaines cellules mais ne peut pas en créer. La convergence de l'algorithme n'est donc pas garantie et ce problème n'est pas traité par Graner et Glazier. Comme pour le problème de connexité des cellules, les auteurs initialisent leur algorithme en une configuration pour laquelle la contrainte de forme est déjà presque respectée. Par conséquent, l'algorithme se trouve, dès l'initialisation, proche d'un minimum local. Or, comme le précise les auteurs, leur algorithme ne s'autorise que des « petits » sauts énergétiques à chaque itération et va donc rester bloqué dans le minimum local le plus proche de la configuration initiale.

En troisième lieu, nous avons remarqué que les auteurs ne s'étaient pas préoccupés de l'invariance par changement d'échelle. En effet, Graner et Glazier se sont concentrés sur la simulation du modèle dans le cas d'un pas de discrétisation égal à 1 : chaque pixel est représenté par un carré de côté égal à 1 et donc d'aire égale à 1. Pourtant, nous pouvons constater que pour un même jeu de paramètres et une discrétisation deux fois plus fine, par exemple, l'Hamiltonien du système se trouve changé et modifie également le comportement de l'algorithme.

Premièrement, le choix de la taille de la grille de discrétisation modifie le calcul de l'Hamiltonien (voir figure 3.13). Pour une grille carrée composée de neuf pixels, chacun des carrés étant de côté 1, l'Hamiltonien de la configuration décrite dans figure 3.13(a) vaut 31. Pour la même configuration, mais sur une grille deux fois plus fine, composée de 36 pixels, chacun carrés de côté 1/2, l'Hamiltonien vaut 58, figure 3.13(b).



$(a)$ $(b)$

**Fig. 3.13:** *Exemple de calculs d'Hamiltonien pour une même configuration et un même jeu de paramètres, mais avec des grilles de discrétisation de tailles différentes. Les configurations* $(a)$ *et* $(b)$ *sont géométriquement identiques : elles sont composées de deux cellules, une grise et une blanche, dont la forme est conservée entre la configuration* $(a)$ *et* $(b)$*. Dans la figure* $(a)$ *la grille est composée de carrés de côté* $1$ *et dans la figure* $(b)$ *la grille est composée de carrés de côté* $1/2$*. En choisissant comme paramètres* $J = 3$ *pour une interaction entre un pixel d'une cellule grise et un pixel d'une cellule blanche,* $A = 4$ *comme aire cible pour une cellule grise et une cellule blanche, dans le cas* $(a)$ *l'Hamiltonien du système vaut* $31$ *et dans le cas* $(b)$ $58$*. Si l'interaction entre deux pixels est pondérée par le pas de discrétisation, l'Hamiltonien du système* $(a)$ *vaut toujours* $31$ *(car le pas vaut* $1$*), et l'Hamiltonien du système* $(b)$ *vaut* $35.5$ *(le pas vaut dans ce cas* $1/2$*).*

Nous pouvons facilement remarquer que plus la grille sera fine plus le poids sur les interactions entre pixels voisins deviendra grand par rapport à la contrainte de forme sur les cellules. Afin de conserver le même ordre de grandeur entre ces deux types de contrainte, indépendamment de la finesse de discrétisation, il est naturel de pondérer une interaction entre deux pixels par un facteur $h$, où $h$ est le pas de discrétisation. Cette pondération permet de modérer l'interaction entre deux pixels par la longueur de leur surface de contact et conserve donc la philosophie du modèle de Graner et Glazier, basée sur l'hypothèse DAH de Steinberg. Dans ce cas, les Hamiltoniens du système calculés à la figure 3.13 reste du même ordre de grandeur : $H_{GG} = 31$ contre $H_{GG} = 35.5$.

Deuxièmement, ce problème d'échelle se répercute également sur l'algorithme de simulation. En effet, la discrétisation joue un rôle important dans le calcul de la probabilité d'acceptation d'une transition. Dans l'exemple décrit à la figure 3.14, si l'interaction entre deux pixels n'est pas pondérée par le pas de discrétisation, la transition étudiée s'effectue quasi-sûrement dans le cas $(a)$ et avec une probabilité de $\exp(-4/(kT))$ dans le cas $(b)$. Comme $k \approx 1.38.10^{-23}$ et $T \approx 10$, cette probabilité dans le cas $(b)$ est très faible. Si au contraire, l'interaction entre deux pixels est pondérée par le pas de discrétisation, la transition sera acceptée presque-sûrement dans les deux cas $(a)$ et $(b)$.

La plupart des problèmes mis en avant ci-dessus sont liés à la discrétisation de l'espace. Dans notre approche nous proposons une alternative au modèle de Graner et Glazier sur un espace continu. Nous proposons ici un modèle centré, de type processus ponctuel, pour lequel chaque cellule est caractérisée par un point et une zone d'influence. L'interaction entre les cellules est également définie à l'aide d'un Hamiltonien, afin de conserver l'esprit de l'hypothèse biologique DAH.

### 3.4.2 Introduction d'une classe de fonctions Hamiltoniennes sur un espace continu : la classe CC (Cadhérine-Caténine)

#### Notations

Dans notre approche, nous considérons $n$ cellules, décrites par les coordonnées spatiales de leurs centres $(x_i)_{i=1,\dots,n}$. A chaque cellule est attaché un type, noté $(\tau_i)_{i=1,\dots,n}$. Ainsi, un ensemble de cellules, appelé configuration et noté $\varphi$, peut se formaliser par :

$$\underline{\varphi} = \{(x_i, \tau_i), \dots, (x_n, \tau_n)\}.$$

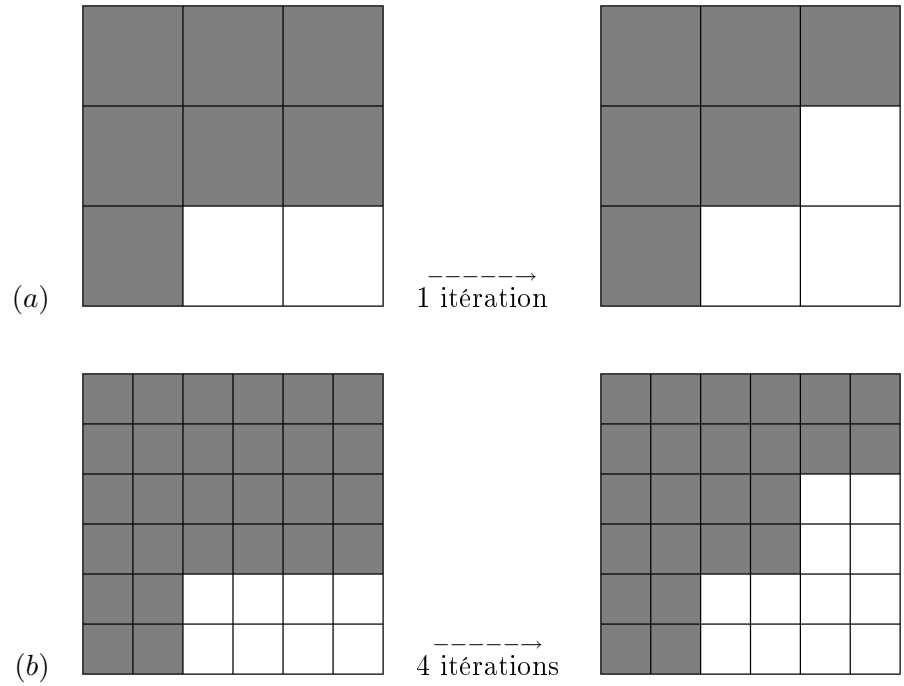



**Fig. 3.14:** *Exemple de calculs de différence d'Hamiltoniens entre une configuration initiale et une configuration terminale en fonction du pas de discrétisation pour le modèle de Graner et Glazier. Le jeu de paramètres choisi est :* $J = 3$ *pour une interaction entre un pixel d'une cellule grise et un pixel d'une cellule blanche,* $A = 4$ *comme aire cible pour une cellule grise et pour une cellule blanche. Dans le cas* $(a)$*, où le pas de discrétisation vaut* $1$*, la différence d'Hamiltonien entre la configuration de droite et la configuration de gauche vaut* $\Delta H = 26 - 31 = -5$*. Dans le cas* $(b)$*, où le pas de discrétisation vaut* $1/2$*,* $\Delta H = 62 - 58 = 4$*. Ainsi dans le cas* $(a)$*, le changement est accepté avec une probabilité de* $1$*, tandis que dans le cas* $(b)$*, le changement est accepté avec une probabilité de* $exp(-4/(kT))$*. Si l'interaction entre 2 pixels est pondérée par le pas de discrétisation, dans le cas* $(a)$*,* $\Delta H$ *reste inchangé :* $\Delta H = 5$*. Dans le cas (b), nous avons :* $\Delta H = 34,5 - 35,5 = -1$*. La probabilité d'acceptation reste différente entre le cas* $(a)$ *et* $(b)$ *mais reste du même ordre de grandeur.*

Par la suite nous noterons $\underline{x} = (x, \tau_x)$ un point marqué, où $x$ correspond à la localisation spatiale du point (ou du centre de la cellule), et $\tau_x$, la marque attachée à ce point. Ainsi une configuration de cellules, $\underline{\varphi}$, peut-être vue comme une configuration de points marqués de la manière suivante :

$$\underline{\varphi} = \{\varphi, \tau_\varphi\},$$

où $\varphi = \{x_1, \ldots, x_n\}$ correspond à un processus de points non marqués et $\tau_\varphi = \{\tau_{x_1}, \ldots, \tau_{x_n}\}$ correspond aux marques associées à chaque point.

Nous proposons un modèle continu en associant à chaque cellule biologique une zone d'influence dans $\mathbb{R}^2$ définie par la cellule de Dirichlet centrée sur chaque point $(x_i)_{i=1,\ldots,n}$. L'utilisation du diagramme de Dirichlet pour modéliser des cellules biologiques est assez répandue. Cette modélisation a été étudiée en détail par Honda [Honda, 1978] et [Honda, 1983]. Au début des années 2000, un test mesurant l'agressivité d'une tumeur a été développé à partir d'une modélisation de cellules biologiques par un diagramme de Dirichlet [Nawrocki-Raby et al., 2001]. Nous allons rappeler ici la définition du diagramme de Dirichlet, également appelé diagramme de Voronoï. Pour de plus amples détails, le lecteur est invité à se référer à [Okabe et al., 2000].

### Définition du diagramme de Dirichlet

**Définition 12** *Soit $\varphi$ un ensemble de points dans une configuration quadratique, i.e. trois points de $\varphi$ ne peuvent être sur la même ligne et quatre points de $\varphi$ ne peuvent être sur le même cercle. La cellule de Dirichlet d'un point $x \in \varphi$ est définie par :*

$$\mathrm{Dir}_\varphi(x) = \{\eta \in \mathbb{R}^2 : |x - \eta| \leq |y - \eta| \text{ pour tout } y \in \varphi \backslash x\}.$$

*La cellule de Dirichlet de $x$ est donc l'ensemble des points de $\mathbb{R}^2$ qui sont les plus proches de $x$ que des autres points de $\varphi$. Ces cellules sont des polygones convexes qui constituent une subdivision du plan appelé diagramme (ou mosaïque) de Dirichlet.*

La figure 3.15 reproduit un exemple de diagramme de Dirichlet calculé à partir de 100 points répartis selon un processus de Poisson sur le carré unité.

**Remarque :** L'aire de la cellule de Dirichlet associée au point $x$ pour la configuration $\varphi$, s'écrit $|\mathrm{Dir}_\varphi(x)|$. Lorsque deux points $x$ et $y$ possèdent des cellules de Dirichlet

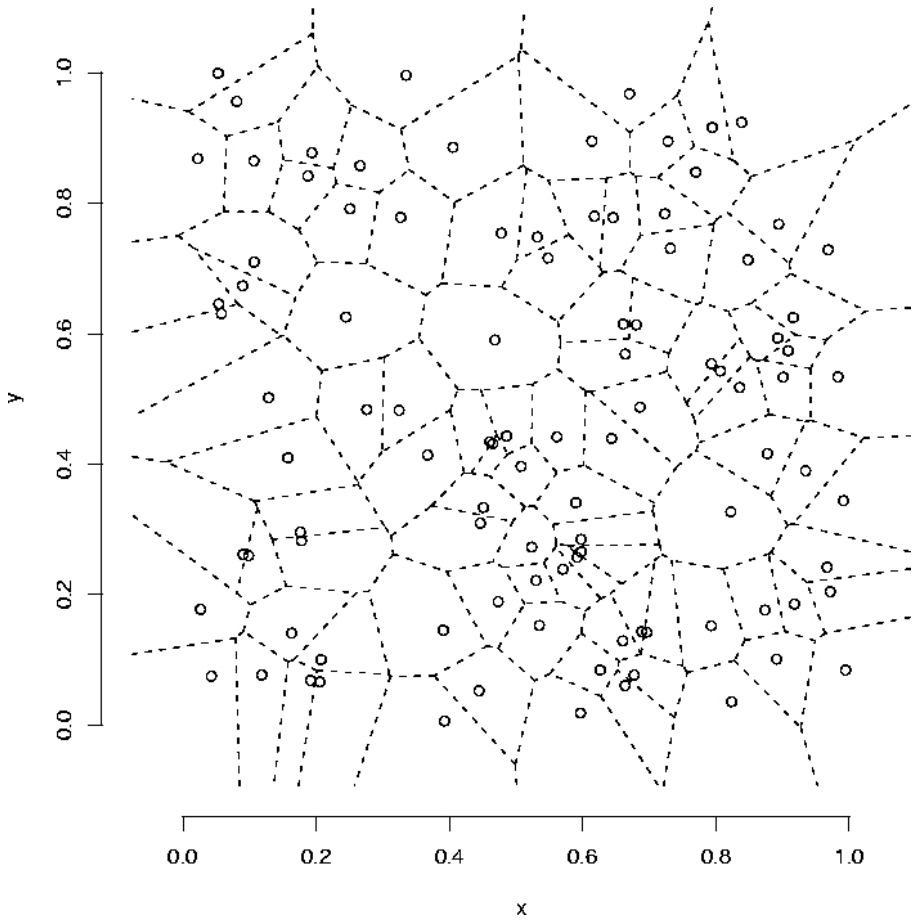



**Fig. 3.15:** *Exemple de diagramme de Dirichlet engendré par un processus de Poisson d'intensité 100 sur le carré unité.*

qui s'intersectent, nous notons l'arête de Dirichlet commune à $x$ et $y$ par le symbole $\mathrm{Dir}_\varphi(x) \cap \mathrm{Dir}_\varphi(y)$. La longueur de l'arête commune à $x$ et $y$ pour la configuration $\varphi$ se note $|\mathrm{Dir}_\varphi(x) \cap \mathrm{Dir}_\varphi(y)|$.

Lorsqu'il n'y a pas d'ambiguïté sur la configuration de points considérée et pour alléger les notations, nous omettrons la dépendance entre la cellule de Dirichlet et la configuration en notant

- $\mathrm{Dir}(x)$ la cellule de Dirichlet du point $x$,
- $|\mathrm{Dir}(x)|$ l'aire de la cellule de Dirichlet du point $x$,

- $\mathrm{Dir}(x) \cap \mathrm{Dir}(y)$ l'arête commune aux point $x$ et $y$,
- $|\mathrm{Dir}(x) \cap \mathrm{Dir}(y)|$ la longueur de l'arête commune aux point $x$ et $y$.

**Diagramme de Dirichlet et triangulation de Delaunay**

Le diagramme de Dirichlet permet la définition d'une topologie de voisinage pour les points d'une configuration quadratique. Cette topologie de voisinage, appelée triangulation de Delaunay, est définie de la manière suivante :

$$x \sim_\varphi y \Longleftrightarrow \mathrm{Dir}_\varphi(x) \cap \mathrm{Dir}_\varphi(y) \neq \emptyset,$$

où le symbole $\sim_\varphi$ représente une relation de voisinage pour la configuration $\varphi$.

L'exemple de la figure 3.15 est repris à la figure 3.16 en rajoutant les relations de voisinage, au sens de Delaunay, entre les points de la configuration. Pour une configuration donnée, $\varphi$, nous pouvons noter l'ensemble des arêtes de Delaunay, $\mathrm{Del}_2(\varphi)$ et l'ensemble des triangles de Delaunay, $\mathrm{Del}_3(\varphi)$.

De nouveau, lorsque le contexte n'introduit aucune ambiguïté sur la configuration, nous omettrons la dépendance entre la topologie de voisinage et la configuration. Dans ce cas, deux points $x$ et $y$ voisins sont notés $x \sim y$.

**Définition de la classe de fonctions CC**

Ces notations nous permettent d'introduire une nouvelle classe de fonctions Hamiltoniennes, la classe CC, définie de la manière suivante :

$$H_{CC}(\underline{\varphi}) = \sum_{x \sim_\varphi y} g(|\mathrm{Dir}_\varphi(x) \cap \mathrm{Dir}_\varphi(y)|) J(\tau_x, \tau_y) + \sum_{x \in \varphi} h(\mathrm{Dir}_\varphi(x), \tau_x), \tag{3.4}$$

où les fonctions $g$, $J$ et $h$ sont des fonctions à valeurs réelles.

Ces fonctions Hamiltoniennes se décomposent en deux termes. Le premier terme est une somme sur l'ensemble des cellules voisines au sens Delaunay. Ce terme fait intervenir une fonction d'interaction $J$, analogue de la fonction $J$ dans le modèle de Potts étendu, qui dépend des types des cellules voisines. Cette dépendance aux types modélise l'hypothèse DAH. De plus, l'interaction entre cellules voisines est pondérée par la longueur de la surface de contact entre les cellules. Cette pondération modélise, quant à elle, l'aspect fermeture éclair des liaisons adhérentes. Le second terme fait référence à la contrainte

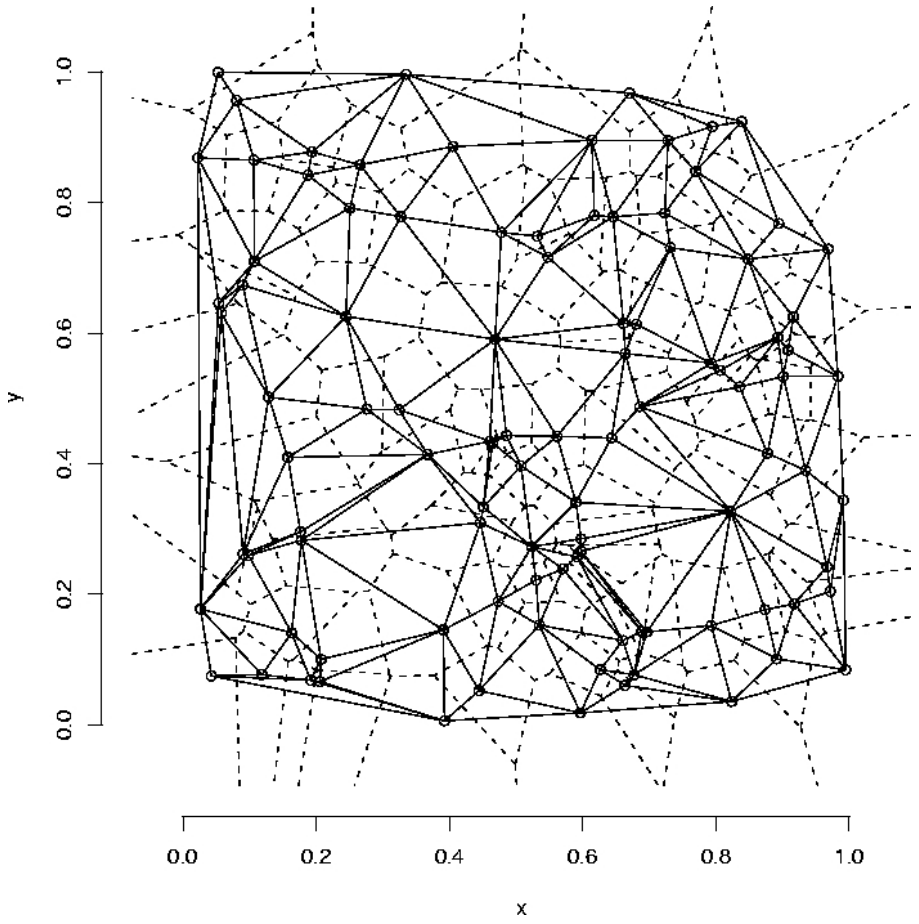



**Fig. 3.16:** *Exemple de diagramme de Dirichlet engendré à partir d'un processus de Poisson d'intensité 100 sur le carré unité. Sur cet exemple, les relations de voisinage représentées en traits pleins forment les triangles de Delaunay.*

de forme proposée par le modèle de Potts étendu. En effet, ce terme est une somme sur l'ensemble des cellules, et met en jeu la forme de la cellule en fonction de son type. De manière analogue au modèle de Potts étendu, la classe de modèles CC propose des fonctions d'énergie composées d'une contrainte d'interaction entre cellules contigües et d'une contrainte de forme sur les cellules.

### 3.4.3 Lien entre la classe CC et le modèle de Potts étendu

Dans cette section, nous allons donner quelques ingrédients heuristiques expliquant dans quelle mesure la classe de modèles CC (Equation 3.4) correspond à l'expression de l'Hamiltonien (Equation 3.3) du modèle de Potts étendu. Pour cela, nous allons réécrire l'équation 3.3, rappelée ci-dessous :

$$H_{GG} = \sum_{(i,j)\sim(i',j')} J\left(\tau(\sigma(i,j)),\tau(\sigma(i',j'))\right)\left(1-\delta_{\sigma(i,j),\sigma(i',j')}\right)+\lambda\sum_{\sigma}(a(\sigma)-A_{\tau(\sigma)})^2\Gamma(A_{\tau(\sigma)}).$$

La notion de voisinage dans le modèle de Potts étendu est définie uniquement sur les pixels de discrétisation de la façon suivante : les huit pixels qui entourent le pixel $(i,j)$ forment le voisinage du pixel $(i,j)$ (voir figure 3.9). Nous pouvons étendre la notion de voisinage dans le modèle de Potts étendu au voisinage entre cellules de la façon suivante.

Soient deux cellules, $\sigma$ et $\sigma'$, nous définissons :

$$\sigma \sim \sigma' \Longleftrightarrow \exists (i,j)\in\sigma, \exists (i',j')\in\sigma' \text{ tels que } (i,j)\sim(i',j').$$

Ainsi, deux cellules sont voisines si et seulement s'il existe une relation de voisinage entre un pixel de chaque cellule respective.

La surface de contact entre les cellules $\sigma$ et $\sigma'$ est alors notée $\sigma\cap\sigma'$. La longueur de cette surface de contact, notée $|\sigma\cap\sigma'|$, peut se définir comme suit :

$$|\sigma\cap\sigma'| = \frac{1}{2}\text{card}\left(\{(i,j),(i',j')\} \text{ tels que } (i,j)\in\sigma,(i',j')\in\sigma' \text{ et } (i,j)\sim(i',j')\right),$$

où card$(E)$ représente le cardinal de l'ensemble $E$. La longueur de la surface de contact entre les cellules $\sigma$ et $\sigma'$ correspond donc au nombre de relations de voisinage entre $\sigma$ et $\sigma'$. Le facteur $1/2$ traduit le fait que $\sim$ est symétrique. Un exemple de calcul de surface de contact est proposé à la figure 3.17.

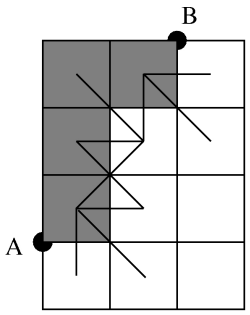


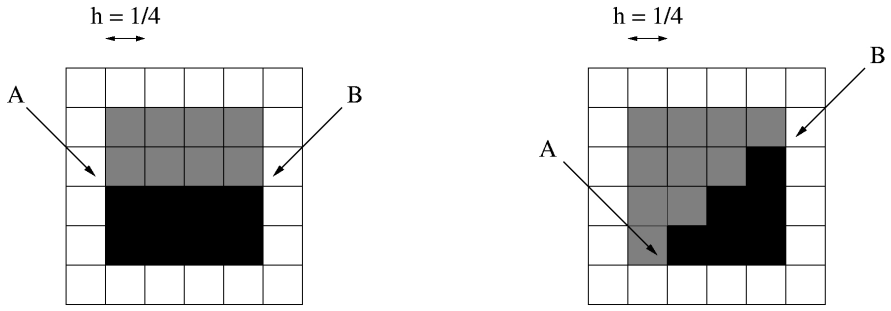

**Fig. 3.17:** *Exemple de calcul de la surface de contact entre deux cellules. La cellule $\sigma$ (en gris) et la cellule $\sigma'$ (en blanc) ont une surface de contact de longueur $|\sigma \cap \sigma'| = 10$. Chaque trait noir représente une relation de voisinage.*

L'Hamiltonien du modèle de Potts étendu peut donc se réécrire de la façon suivante :

$$H_{GG} = \sum_{\sigma \sim \sigma'} |\sigma \cap \sigma'| J\left(\tau(\sigma), \tau(\sigma')\right) + \lambda \sum_{\sigma} (a(\sigma) - A_{\tau(\sigma)})^2 \Gamma(A_{\tau(\sigma)}).$$

Lorsque les surfaces de contact entre deux cellules contigües s'apparentent à des droites (ce qui est biologiquement fondé), le terme $|\sigma \cap \sigma'|$ est de l'ordre de grandeur de $h.d(A, B)$ où h est le pas de discrétisation, $A$ et $B$ sont les extrémités de la surface de contact et $d(.,.)$ la distance euclidienne (voir figure 3.18). Ainsi $d(A, B)$ représente la longueur de la surface de contact entre les cellules $\sigma$ et $\sigma'$ dans le cas où cette surface est modélisée par une droite.

La figure 3.18 montre deux exemples représentatifs pour lesquels la longueur de la surface de contact est de l'ordre de grandeur de $h\times$(le nombre de relations de voisinage). Dans le cas d'une surface de contact horizontale entre $\sigma$ et $\sigma'$, $|\sigma \cap \sigma'| = 3/h - 2$, où $h$ est le pas de discrétisation. Dans le cas d'une surface de contact en diagonale entre $\sigma$ et $\sigma'$, $|\sigma \cap \sigma'| = 4/h - 5$, où $h$ est le pas de discrétisation. Ainsi, si nous considérons que le potentiel d'interaction pour une relation de voisinage, noté $J$ dans le modèle de Potts étendu, est pondéré par le pas de discrétisation, $|\sigma \cap \sigma'|$ est bien du même ordre de grandeur que $l_{\sigma,\sigma'}$, où $l_{\sigma,\sigma'}$ représente la longueur de la surface de contact, modélisée par une droite, entre $\sigma$ et $\sigma'$.




(a) (b)

**FIG. 3.18:** *Exemple de modélisation de la distance entre le point $A$ et le point $B$ par un voisinage latticiel. Dans les deux configurations proposées, le tissu est composé de deux cellules, une grise et une noire, et du milieu extracellulaire (en blanc). Nous nous intéressons à la surface de contact entre la cellule grise et la cellule noire. Pour un pas de discrétisation de $h = 1/4$, la distance euclidienne entre $A$ et $B$ vaut $d(A, B) = 1$ dans le cas $(a)$ et $d(A, B) = \sqrt{2}$ dans le cas $(b)$. Le nombre de relations de voisinage, pour le modèle étendu vaut $10$ dans le cas $(a)$, et $11$ dans le cas $(b)$. Nous pouvons montrer facilement que pour un pas de discrétisation de $1/n$, le nombre de relations de voisinage dans le cas $(a)$ vaut $3n-2$ et dans le cas $(b)$ vaut $4n-5$. La pondération par le pas de discrétisation permet de conserver le même ordre de grandeur pour le nombre de voisins et donc pour l'interaction de paires.*

Si nous supposons que la grille de discrétisation est infiniment fine, nous avons donc :

$$H_{GG} \approx \sum_{\sigma \sim \sigma'} l_{\sigma,\sigma'} J\left(\tau(\sigma), \tau(\sigma')\right) + \lambda \sum_{\sigma} (a(\sigma) - A_{\tau(\sigma)})^2 \Gamma(A_{\tau(\sigma)}).$$

On retrouve donc ici, pour le premier terme de la somme (le potentiel de paires), une expression analogue à la fonction Hamiltonienne formulée dans les modèles de Sulsky (voir Equation 3.1) et de Graner et Sawada (voir Equation 3.2). Nous rappelons ici que les auteurs de ces modèles proposent une résolution déterministe. Notre classe de modèles, définie par les Hamiltoniens formulés à l'équation 3.4, propose une résolution non-déterministe du problème initié par Sulsky. Nous proposons également une généralisation de ce type de modèle en intégrant une contrainte de forme sur chacune des cellules.

Dans la suite de ce chapitre, nous allons tout d'abord exhiber une classe de processus ponctuels marqués de Gibbs associée à la classe de fonction Hamiltoniennes CC, présentée dans cette section. Le cadre éprouvé des processus ponctuels de Gibbs nous permettra de déterminer certaines conditions sur les fonctions $g$, $J$ et $h$ de la classe CC, garantissant l'existence de tels modèles. Les deux ingrédients principaux caractérisant l'existence de tels modèles sont la stabilité locale de l'énergie et la quasilocalité. La stabilité locale de l'énergie nous permet de faire exister le modèle dans tout borélien borné. Cette propriété constitue également un argument fondamental pour l'étude mathématique d'un algorithme de simulation que nous proposerons par la suite. La quasilocalité permet d'étendre l'existence de notre classe de processus à $\mathbb{R}^2$ en s'appuyant sur un ensemble de spécifications locales. La propriété de quasilocalité permet également l'étude théorique d'estimateurs de pseudo-vraisemblance pour l'ensemble des paramètres du modèle. Bien que nous ne nous attacherons pas à étudier théoriquement ces propriétés, nous proposerons des estimateurs de pseudo-vraisemblance pour les paramètres du modèle.

La définition formelle de la classe de processus ponctuels de Gibbs associée à la classe de fonction Hamiltoniennes CC nécessitent quelques rappels sur la théorie des processus ponctuels.

**Remarque :** Pour notre modélisation nous avons choisi de tenir compte de l'influence des marques sur la disposition des points. Dans le modèle que nous allons proposer, nous perturbons la mesure de Poisson d'un processus ponctuel marqué. Pour le processus ainsi

défini, nous ne connaissons pas la loi marginale de la configuration des points sans les marques. Une autre approche de modélisation, qui ne sera pas traitée dans ce manuscrit, peut consister à supposer que le modèle suit deux aléas distincts. Tout d'abord la configuration des points se modélise par un processus ponctuel non marqué. Puis, conditionnellement à la réalisation de points, les marques en chaque point sont ensuite déterminées selon un autre processus. Cette deuxième hypothèse de modélisation nous a inspirée pour proposer un estimateur de pseudo-vraisemblance à la section 3.10.

## 3.5 Les processus ponctuels

Les processus ponctuels ont été introduits pour caractériser des semis de points localisés dans l'espace, $\mathbb{R}^d$ par exemple. L'étude mathématique de ce type de données a nécessité la mise en place de structures probabilistes. Cette section est dédiée aux rappels de ces structures et permet également d'introduire de nombreuses notations utiles dans la suite du chapitre.

### 3.5.1 Les processus ponctuels sur $\mathbb{R}^d$

Un processus ponctuel $\Phi$ est une variable aléatoire définie sur un espace métrique $S \subset \mathbb{R}^d$. Une réalisation du processus ponctuel $\Phi$, notée $\varphi$, est appelée une configuration. Le nombre de points d'une configuration $\varphi$ est symbolisé par $n(\varphi)$.

De plus si $B \subset S$ est un sous-ensemble borné, alors :

$$\varphi_B = \varphi \cap B$$

représente la restriction de la configuration $\varphi$ aux points inclus dans $B$.

L'espace métrique $S$ est muni de la tribu des boréliens $\mathcal{B}$, et $\mathcal{B}_0$ représente l'ensemble des boréliens bornés, $N_{lf}$ caractérise l'espace des configurations localement finies dans $S$ :

$$N_{lf} = \{\varphi \subseteq S : n(\varphi_B) < \infty, \text{ pour tout } B \subset S\}.$$

Enfin, l'espace $N_{lf}$ est naturellement muni d'une tribu, notée $\mathcal{N}_{lf}$.

**Définition 13** *Un processus ponctuel $\Phi$ sur $S$ est une variable aléatoire définie sur un espace probabilisé $(\Omega, \mathcal{F}, \mathcal{P})$ et à valeurs dans $(N_{lf}, \mathcal{N}_{lf})$. La loi de probabilité $P_\Phi$ de $\Phi$*

*est définie par :*

$$P_\Phi(F) = \mathbb{P}(\{\omega \in \Omega : \Phi(\omega) \in F\}) \text{ pour } F \in \mathcal{N}_{lf}.$$

Pour de plus amples détails sur les processus ponctuels, le lecteur peut se réferrer aux deux livres suivants [Daley et Vere-Jones, 1988] et [Stoyan et al., 1995].

Dans de nombreux problèmes, la localisation spatiale des points n'est pas la seule information accessible. Dans ce cas, à chaque point du processus ponctuel, une marque est attachée, nous parlons alors de processus ponctuels marqués.

### 3.5.2 Les processus ponctuels marqués

La classe de processus pour lesquels une marque est attachée à chaque point est appelée la classe des processus ponctuels marqués dont une définition formelle est donnée ci-dessous.

**Définition 14** *Soit $\Phi$ un processus ponctuel défini sur l'espace $T$, et $M$ un espace mesurable, appelé espace des marques. Si pour chaque point $x$ du processus $\Phi$, une marque $\tau_x \in M$ est attachée, alors :*

$$\underline{\Phi} = \{(x, \tau_x), x \in \Phi\}$$

*est appelé un processus ponctuel marqué sur $T \times M$.*

Une étude très complète sur les processus ponctuels marqués a été réalisée dans le livre suivant [Stoyan et Stoyan, 1994].

Lorsque nous nous intéressons à un processus ponctuel défini sur un espace borné, nous pouvons remarquer que les deux notions de processus ponctuels et de processus ponctuels marqués peuvent être unifiés dans un même formalisme mathématique [Møller et Waagepetersen, 2003]. La distinction entre un processus ponctuel sur $S \subset \mathbb{R}^d$ et un processus ponctuel marqué, dont les points sont dans $T \subset \mathbb{R}^d$ et les marques dans $M$, disparaît en choisissant $S = T \times M$.

Par la suite, nous considérons des processus ponctuels marqués définis sur $S = T \times M$. Une réalisation d'un processus ponctuel marqué sera notée :

$$\underline{\varphi} = \{(x, \tau_x), x \in \varphi\},$$

où $\varphi$ est le processus ponctuel non marqué sous-jacent. L'exemple le plus célèbre pour les processus ponctuels marqués ou non, est le processus de Poisson.

### 3.5.3 Un exemple : le processus de Poisson

Le processus ponctuel de Poisson joue un rôle fondamental dans l'étude des processus ponctuels. En effet, il sert de modèle de référence, mathématiquement étudiable, pour lequel il n'existe aucune interaction entre les points. De nombreux modèles, comme les processus ponctuels de Gibbs, par exemple, sont alors construits à partir du processus ponctuel de Poisson.

Dans cette section, nous nous intéressons uniquement au processus de Poisson défini sur $S \subseteq \mathbb{R}^d$ et spécifié par une fonction d'intensité $\rho : S \to [0, \infty)$. La fonction d'intensité $\rho$ est supposée localement intégrable ce qui signifie que :

$$\int_B \rho(\xi) d\xi < \infty, \quad \text{pour tout borné } B \subseteq S.$$

Cette fonction d'intensité sert à définir la mesure d'intensité d'un processus ponctuel de Poisson.

**Définition 15** *La mesure d'intensité $\mu$ d'un processus ponctuel de Poisson de fonction d'intensité $\rho$ vaut :*

$$\mu(B) = \int_B \rho(\xi) d\xi, \quad B \subseteq S.$$

.

Nous pouvons à présent définir un processus ponctuel de Poisson.

**Définition 16** *Un processus ponctuel $\Phi$ sur $S$ est un processus ponctuel de Poisson de fonction d'intensité $\rho$ si les deux propriétés suivantes sont satisfaites*

1. *Pour tout $B \subseteq S$, tel que $\mu(B) < \infty$, le nombre de points de $\Phi$ dans $B$, $n(\Phi_B)$, suit une loi de Poisson de paramètre $\mu(B)$.*
2. *Pour $n \in \mathbb{N}$ et pour tout $B \subseteq S$, tel que $0 < \mu(B) < \infty$, conditionnellement à $n(\Phi_B) = n$, le processus $\Phi_B$ est constitué de $n$ points i.i.d. de fonction de densité $f(x) = \rho(x)/\mu(B)$.*

Nous pouvons remarquer que si la fonction d'intensité $\rho$ est constante, alors le processus ponctuel de Poisson associé est *homogène*. Dans ce cas, la distribution d'un processus ponctuel est invariante par translation, *i.e.* la distribution de $\Phi + s = \{\xi + s : \xi \in \Phi\}$ est la même que celle de $\Phi$. Nous pouvons également caractériser ce processus comme un processus *stationnaire*.

La proposition suivante nous permet d'exhiber la mesure de probabilité associée à un processus ponctuel de Poisson de fonction d'intensité $\rho$.

**Proposition 13**

1. *Soit $\Phi$ un processus ponctuel. $\Phi$ est un processus ponctuel de Poisson sur $S$ et de fonction d'intensité $\rho$ ssi pour tout $B \subseteq S$, tel que $\mu(B) = \int_B \rho(\xi)d\xi$, et pour tout $F \subseteq N_{lf}$ :*

$$\begin{aligned}&\mathbb{P}(\Phi_B \in F)\\ &= \sum_{n=0}^{\infty} \frac{\exp(-\mu(B))}{n!} \int_B \dots \int_B \mathbb{1}_{\{x_1,\dots,x_n \in F\}} \prod_{i=1}^{n} \rho(x_i)dx_1, \dots dx_n.\end{aligned}$$

2. *Soit $\Phi$ un processus ponctuel de Poisson sur $S$ et de fonction d'intensité $\rho$. Pour toute fonction $h : N_{lf} \to [0, \infty)$ et tout $B \subseteq S$, tel que $\mu(B) < \infty$ :*

$$\begin{aligned}&\mathbb{E}[h(\Phi_B)]\\ &= \sum_{n=0}^{\infty} \frac{\exp(-\mu(B))}{n!} \int_B \dots \int_B \mathbb{1}_{\{x_1,\dots,x_n \in F\}} h(\{x_1, \dots, x_n\}) \prod_{i=1}^{n} \rho(x_i)dx_1, \dots dx_n.\end{aligned}$$

**Le processus ponctuel de Poisson marqué**

Nous nous intéressons à présent au processus ponctuel marqué $\underline{\Phi} = \{(x, \tau_x) : x \in \Phi\}$, où les points sont dans un espace $T$ et les marques dans un espace $M$.

**Définition 17** *Soit $\Phi$ un processus de Poisson sur $T$ et de fonction d'intensité localement intégrable $\varrho$. Supposons que conditionnellement à $\Phi$, les marques $\{\tau_x : x \in \Phi\}$ soient mutuellement indépendantes. Alors le processus $\underline{\Phi} = \{(x, \tau_x) : x \in \Phi\}$ est un processus ponctuel de Poisson marqué. De plus, si les marques sont identiquement distribuées avec une distribution $Q$, alors $Q$ est appelée la distribution des marques.*

Les marques peuvent prendre plusieurs formes : des entiers, des réels, des objets géométriques, etc...La proposition suivante nous permet de caractériser la fonction d'intensité d'un processus ponctuel de Poisson marqué.

**Proposition 14** *Soit $\underline{\Phi}$ un processus ponctuel de Poisson marqué où $M \subseteq \mathbb{R}^p$. Le processus ponctuel sous-jacent, $\Phi$, défini sur $T$ a une fonction d'intensité $\varrho$. Nous supposons que, conditionnellement à $\Phi$, chaque marque $\tau_x$ a une densité $p$, et nous posons $\rho(x, \tau_x) = \varrho(x)p(\tau_x)$. Sous ces conditions nous avons*

$\underline{\Phi}$ *est un processus ponctuel de Poisson sur $T \times M$ et de fonction d'intensité $\rho$.*

Une réalisation d'un processus ponctuel de Poisson marqué sur :

$$[0,1]^2 \times \{\text{cercle vide, cercle plein, triangle}\}$$

est représentée à la figure 3.19. Sur cet exemple, le processus est stationnaire, homogène et défini sur le carré unité avec une densité de point égale à 50 par unité d'aire.

## 3.6 Les processus ponctuels de Markov marqués de type plus proche voisin

Par la suite nous allons nous intéresser uniquement à des processus définis à partir du processus ponctuel de Poisson. Ces processus permettent d'inclure des dépendances entre les points. Ces dépendances peuvent être de plusieurs types : dépendance sur la localisation des points ou sur les types des points, par exemple. Une classe de modèle, appelée « processus ponctuels de Markov de type plus proche voisin » et introduite en 1989 [Baddeley et Møller, 1989], est particulièrement intéressante pour notre étude. En effet, cette classe de modèles permet de modéliser certaines dépendances entre des points qui sont voisins pour un graphe donné. Ainsi, la relation de voisinage entre les points dépend explicitement de la réalisation du processus. Or la modélisation d'un tissu par un diagramme de Dirichlet a pour conséquence de faire dépendre le voisinage (de Delaunay) de la configuration considérée. En effet, l'insertion d'un point dans une configuration peut, par exemple, faire disparaître certaines relations de voisinage. Par la suite, nous nous focaliserons essentiellement sur les processus ponctuels de Markov marqués de type plus proche voisin.

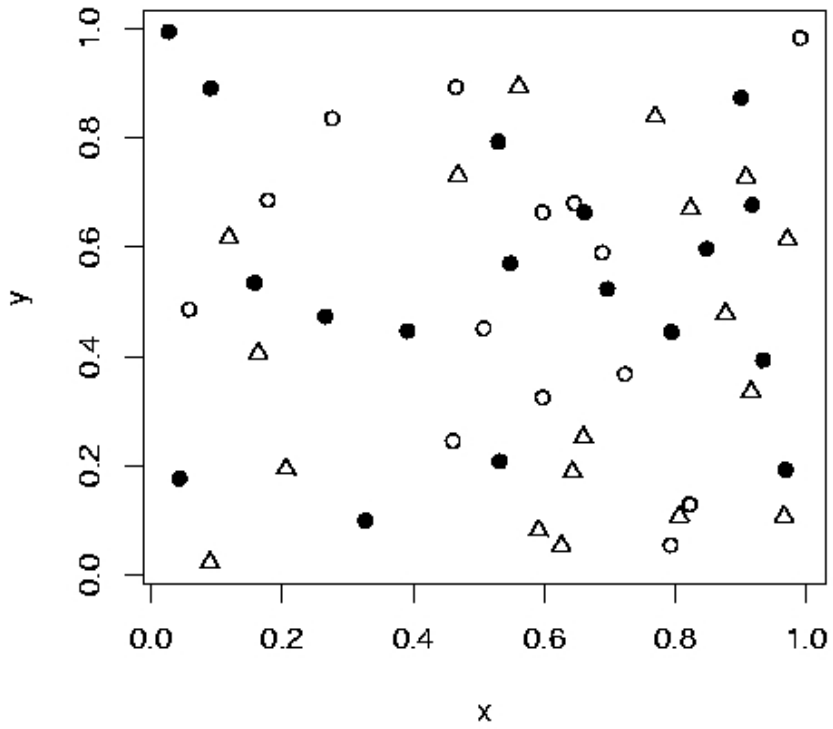



**Fig. 3.19:** *Exemple de processus ponctuel marqué de Poisson défini sur* $T \times M$*, où* $T = [0,1]^2$ *et* $M = \{$*cercle vide, cercle plein, triangle*$\}$.

### 3.6.1 Les processus ponctuels de Markov marqués

Dans cette section, nous nous intéressons aux processus ponctuels définis sur un espace $S = T \times M$, tel que $T \subseteq \mathbb{R}^d$ et $M \subseteq \mathbb{R}^p$. L'ensemble des configurations finies de $S$ se définit de la mnière suivante :

$$N_f = \{\varphi \in S : n(\varphi) < \infty\}.$$

Considérons $\underline{\Phi}$ un processus ponctuel marqué de densité $f$ par rapport au processus ponctuel de Poisson marqué standard (*i.e.* pour lequel l'intensité du processus ponctuel des points vaut $\varrho = 1$ et la densité des marques se note $p$).

Ainsi d'après la proposition 13, nous avons, pour tout $F \subseteq N_f$ :

$$\begin{aligned}\mathbb{P}(\underline{\Phi} \in F) &= \sum_{n=0}^{\infty} \frac{\exp(-|S|))}{n!} \int_T \int_M \cdots \int_T \int_M \mathbb{1}_{\{(x_1,\tau_{x_1}),\ldots,(x_n,\tau_{x_n}) \in F\}} \\ &\quad f(\{(x_1,\tau_{x_1}),\ldots,(x_n,\tau_{x_n})\}) p(\tau_{x_1}) \ldots p(\tau_{x_n}) dx_1 d\tau_{x_1} \ldots dx_n d\tau_{x_n}. \quad (3.5)\end{aligned}$$

Les processus ponctuels de Markov sont caractérisés par certaines classes de densités $f$. Ces fonctions vont s'interpréter comme des interactions entre les points du processus. Le caractère Markovien du processus ponctuel va se traduire par le fait que ces interactions ne mettront en jeu que des points voisins dans l'espace. Par conséquent, les fonctions $f$ vont représenter des fonctions d'interactions locales.

**Définition 18** *Soit $f : N_f \to [0, \infty)$, une fonction telle que :*

$$f(\varphi) > 0 \Rightarrow f(\phi) > 0 \quad \textit{pour } \phi \subseteq \varphi.$$

*$f$ est alors une fonction* **héréditaire**.

Pour définir un processus ponctuel de Markov marqué, nous devons introduire certaines notations caractérisant le voisinage définissant la propriété Markovienne spatiale utilisée. Nous notons $\sim$ une relation de voisinage, de sorte que $x \sim y$ signifie que les points $x$ et $y$ sont voisins. De plus, le voisinage de $x$ dans $T$ est noté $N_x = \{y \in S : y \sim x\}$.

**Définition 19** *Soit $\underline{\Phi}$ un processus ponctuel marqué défini sur $S$ et de densité $f$ par rapport au processus ponctuel de Poisson standard. Supposons que $f : N_f \to [0, \infty)$ soit*

*une fonction héréditaire. Supposons, de plus, que pour tout* $\underline{\varphi} \in N_f$, *tel que* $f(\underline{\varphi}) > 0$ *et pour tout* $\underline{x} \in S \backslash \underline{\varphi}$, *nous ayons :*

$$f(\underline{\varphi} \cup \underline{x})/f(\underline{\varphi}) \text{ dépend de } \underline{\varphi} \text{ uniquement par l'intermédiaire de } \underline{\varphi} \cap N_x.$$

*Sous ces conditions,* $f$ *est une fonction de Markov. Ainsi, un processus ponctuel marqué défini par une fonction de Markov par rapport au processus ponctuel de Poisson marqué est appelé* **processus ponctuel de Markov marqué**.

Le caractère markovien de ce type de processus vient du fait que le terme $f(\underline{\varphi} \cup \underline{x})/f(\underline{\varphi})$ s'interprète comme la densité du point $\underline{x}$ conditionnellement à la configuration $\underline{\varphi}$. Ainsi, par définition cette densité conditionnelle ne doit dépendre que du voisinage de $\underline{x}$ induisant une propriété markovienne dans l'espace.

Dans le domaine de la physique statistique, un processus ponctuel de Markov est appelé un processus ponctuel de Gibbs de fonction d'énergie :

$$H(\underline{\varphi}) = -\sum_{\underline{\phi} \subseteq \underline{\varphi} : \underline{\phi} \neq 0} \log(g(\underline{\phi})),$$

où $g$ est une fonction positive. Ainsi la densité conditionnelle à $n(\underline{\Phi}) = n$, notée $f_n$, d'un processus ponctuel de Gibbs, s'écrit par rapport au processus ponctuel de Poisson de la manière suivante :

$$f_n(\underline{\varphi}) \propto \exp(-\theta H(\underline{\varphi})), \quad \theta > 0.$$

Cette remarque est très importante pour la suite du chapitre. En effet, nous allons introduire une classe de modèle de Gibbs, la classe CC, dont l'interaction entre les cellules sera spécifiée par l'intermédiaire d'une fonction Hamiltonienne de la classe CC. Cependant, dans notre étude, la relation de voisinage entre les points (ou cellules) est la triangulation de Delaunay. Or cette relation de voisinage dépend explicitement de la réalisation du processus.

### 3.6.2 Définition des processus ponctuels de Markov marqués de type plus proche voisin

Les processus ponctuels de Markov peuvent être généralisés à des processus pour lesquels le voisinage dépend de la réalisation du processus. Ce type de processus est appelé « processus ponctuels de type plus proche voisin ».

Les processus ponctuels de type plus proche voisins ont été introduits et étudiés à la fin des années 80 [Baddeley et Møller, 1989]. Ils se différencient des processus ponctuels de Markov, étudiés en détail dans [Van-Lieshout, 2000], en ce sens que la relation de voisinage dépend de la réalisation du processus. Afin de marquer cette dépendance entre le voisinage et la réalisation, nous notons $\sim_\varphi$ le voisinage associé à la réalisation $\varphi$. Cette notation est à mettre en relation avec le voisinage de Delaunay introduit à la section 3.4.2. De plus, pour tout point $\underline{x} \in S\backslash\varphi$, nous définissons le voisinage de $x$ dans la configuration $\varphi \cup x$ de la manière suivante :

$$N_x(\varphi) = \{y \in \varphi : y \sim_{\varphi\cup x} x\}.$$

**Définition 20** *Soit $\underline{\Phi}$ un processus ponctuel marqué défini sur $S$ de densité $h$ par rapport au processus ponctuel de Poisson standard. Supposons que $f : N_{lf} \to [0,\infty)$ soit une fonction héréditaire. Supposons, de plus, que pour tout $\underline{\varphi} \in N_f$, tel que $f(\underline{\varphi}) > 0$, et pour tout $\underline{x} \in S\backslash\underline{\varphi}$, nous ayons :*

$f(\underline{\varphi}\cup\underline{x})/f(\underline{\varphi})$ *dépendent uniquement de $\underline{x}$, $N_x(\varphi)$ et des restrictions de $\sim_\varphi$ et $\sim_{\varphi\cup x}$ à $N_x(\varphi)$.*

*Sous ces hypothèses, $\underline{\Phi}$ est un* **processus ponctuel de Markov marqué de type plus proche voisin** *par rapport au processus de Poisson standard.*

L'existence de mesure de probabilités associées à ce type de processus nécessite certaines conditions sur $f$. Etant donné que la relation de voisinage dépend de la réalisation, nous pouvons remarquer qu'une modification locale d'une configuration donnée va modifier sensiblement le graphe de voisinage. En particulier, l'insertion d'un point marqué, $\underline{x}$, dans une configuration donnée $\underline{\varphi}$, va créer de nouvelles relations de voisinages entre $x$ et certains points de $\varphi$. Cependant cette insertion peut également modifier les relations de voisinages déjà existantes entre les points de $\varphi$, ce qui n'est pas le cas pour un processus ponctuel de Markov. La définition de processus ponctuels de type plus proche voisin va donc nécessiter certaines conditions, portant sur la densité $f$ perturbant le processus ponctuel de Poisson standard. Dans la section suivante, nous allons citer certaines de ces propriétés et énoncer un théorème nous donnant des conditions suffisantes sur $f$ pour garantir l'existence de tels processus. Nous allons nous focaliser sur les processus ponctuels de Gibbs de type plus proche voisin. Comme nous l'avons vu précédemment, ce type de

processus est défini par une fonction Hamiltonienne, $H$, et l'expression de la densité $f$, par rapport au processus de Poisson, s'écrit :

$$f(\underline{\varphi}) \propto \exp(-\theta H(\underline{\varphi})).$$

### 3.6.3 Existence de processus ponctuel de Gibbs

Dans cette section nous allons nous focaliser sur l'étude de l'existence de processus ponctuels de Gibbs. Cette étude peut se décomposer en deux cas selon que le processus ponctuel étudié soit défini sur un borélien borné $\Lambda \subset \mathbb{R}^d$, ou sur $\mathbb{R}^d$. Dans le cas d'un processus ponctuel défini sur un borélien borné $\Lambda \subset \mathbb{R}^d$, l'existence d'un tel processus est garantie par une propriété de stabilité locale. Cette propriété, que nous expliciterons par la suite, suppose que l'énergie nécessaire pour insérer un point dans une configuration donnée est bornée inférieurerement. La propriété de stabilité locale d'un processus ponctuel de Gibbs garantit que la fonction de partition associée à ce processus est finie.

L'extension d'un tel processus à un espace infini $\mathbb{R}^d$ nécessite certaines conditions supplémentaires. L'étude de tels processus présente un intérêt tout particulier dans le domaine de la physique statistique, où des systèmes constitués d'un très grand nombre de particules sont modélisés. Un des intérêts de ce type de systèmes de particules est l'étude de transition de phase caractérisant le comportement de tels systèmes. L'étude de l'existence de ce type de modèle a été développée dans de nombreux ouvrages parmi lesquels nous pouvons citer les travaux de Ruelle [Ruelle, 1969] et de Preston [Preston, 1976]. L'approche la plus courante pour étudier les propriétés de processus de Gibbs sur $\mathbb{R}^d$ consiste à utiliser des spécifications locales pour le processus. Ces spécifications locales caractérisent des densités conditionnelles du processus pour tout borélien borné $\Lambda$, en supposant que le processus est connu à l'extérieur de $\Lambda$ ($\Lambda^c$). Un ensemble de spécifications locales, $\Pi_\Lambda$, pour un processus de Gibbs de fonction d'énergie $H$ peut s'écrire de la façon suivante, pour tout $F \in \mathcal{N}_{lf}$ :

$$\begin{aligned}\Pi_\Lambda(\underline{\varphi}, F) &= \sum_{n=0}^{\infty} \frac{\exp(-|\Lambda|)}{n!} \int_\Lambda \int_M \cdots \int_\Lambda \int_M \mathbf{1}_{\{(x_1,\tau_{x_1}),\ldots,(x_n,\tau_{x_n})\}\cup\underline{\varphi_{\Lambda^c}}\in F} \\ &\quad \exp[-H((x_1,\tau_{x_1}),\ldots,(x_n,\tau_{x_n}),\varphi_{\Lambda^c})]p(\tau_{x_1})\ldots p(\tau_{x_n})dx_1 d\tau_{x_1}\ldots dx_n d\tau_{x_n}.\end{aligned}$$

Dans l'expression précédente, le terme $H((x_1,\tau_{x_1}),\ldots,(x_n,\tau_{x_n}),\underline{\varphi_{\Lambda^c}})$ représente l'éner-

gie nécéssaire pour insérer les $n$ points marqués $\underline{x_i}$ dans la configuration $\underline{\varphi_{\Lambda^c}}$. Ainsi conditionnellement au fait que $n$ points de $\underline{\varphi}$ sont localisés dans $\Lambda$, nous pouvons noter $\underline{\varphi_\Lambda} = \{(x_1, \tau_{x_1}), \ldots, (x_n, \tau_{x_n})\}$. Le terme $H(\underline{\varphi_\Lambda}, \underline{\varphi_{\Lambda^c}})$ s'interprète alors comme l'énergie nécessaire pour insérer la configuration $\underline{\varphi_\Lambda}$ sachant que la configuration est fixée à l'extérieur du domaine $\Lambda$.

La caractérisation d'un processus ponctuel de Gibbs par spécifications locales est équivalente à la résolution d'un système d'équations DLR, proposé par Dobrushin, Landford et Ruelle [Dobrushin, 1969] et [Landford et Ruelle, 1969]. La proposition suivante, formulée par Bertin, Billiot et Drouilhet, nous permet d'établir des conditions suffisantes pour vérifier un système de spécifiactions locales [Bertin et al., 1999a]. Ces conditions, portant essentiellement sur la fonction d'energie $H$, peuvent se décomposer en deux ingrédients essentiels : une condition de stabilité locale et une condition de quasilocalité. La condition de stabilité locale garantit que, conditionnellement à l'extérieur d'un domaine borné, l'énergie est stable. Cette propriété joue un rôle très important pour l'étude de la convergence d'algorithmes de simulation par chaînes de Markov. La seconde condition permet le contrôle de la portée du processus, c'est à dire l'effet de conditionnement des bords.

**Proposition 15** *Soit $H$ l'énergie locale, définie à partir d'un ensemble de spécifications locales, $\Pi_\Lambda$. Nous supposons que $H$ est invariante par translation et que*

1. **Stabilité locale.** *il existe une constante $K > 0$, telle que, pour toute configuration $\underline{\varphi} \in N_{lf}$ :*

$$H(\underline{\varphi} \cup (0, \tau_0)) - H(\underline{\varphi}) > -K \quad \textit{pour tout } \tau_0 \in M.$$

2. **Quasilocalité.** *Pour tout borélien borné $\Delta$, tel que $0 \in \Delta$, pour toute configuration $\varphi \in N_{lf}$ :*

$$|[H(\underline{\varphi} \cup (0, \tau_0)) - H(\underline{\varphi})] - [H(\underline{\varphi_\Delta} \cup (0, \tau_0)) - H(\underline{\varphi_\Delta})]| < \varepsilon(d(0, \Delta^c)) \quad \textit{pour tout } \tau_0 \in M,$$

*où $\varepsilon$ est une fonction positive qui tend vers 0 à l'infini et $d(x, B) = min_{y \in B} d(x, y)$ est la distance euclidienne entre un point $x$ et un borélien $B$.*

*Sous ces conditions, il existe une mesure de probabilité vérifiant les spécifications locales $\Pi_\Lambda$.*

La démonstration de cette propriété est donnée en détail dans [Bertin et al., 1999a] (proposition 1 page 891). L'application de ces conditions, présentées à la proposition 15, dans le cas de processus de type plus proche voisins (c'est à dire lorsque le voisinage dépend de la réalisation du processus) constitue un intérêt tout particulier de ces résultats. L'application de ces résultats à été proposée sur différents modèles : le modèle 1-plus-proche-voisins, des modèles d'interactions de Delaunay, le modèle de Ord et des modèles de type composantes connexes markoviens [Bertin et al., 1999a], [Bertin et al., 1999b].

Par la suite, nous allons pouvoir appliquer cette propriété pour montrer l'existence de mesures de probabilité, vérifiant les spécifications locales déterminées par les fonctions d'énergie de la classe CC.

## 3.7 Etude de notre classe de modèles

Dans cette section, nous allons montrer que la classe de fonctions Hamiltoniennes définies à l'équation 3.4, vérifie, sous certaines conditions, les hypothèses de la proposition 15. Tout d'abord, nous allons définir la classe de processus ponctuels de Gibbs marqués à partir de la classe de fonctions d'énergie $H_{CC}$.

### 3.7.1 Définition de la classe de processus de Gibbs CC

Nous pouvons rappeler ici que nous avions défini un modèle centré par la classe de fonctions Hamiltoniennes suivante :

$$H_{CC}(\underline{\varphi}) = \sum_{x \sim y} J(\tau_x, \tau_y) g(|\mathrm{Dir}_{\varphi}(x) \cap \mathrm{Dir}_{\varphi}(y)|) + \sum_{x \in \varphi} h(\mathrm{Dir}_{\varphi}(x), \tau_x).$$

**Définition 21 (La classe CC)** *Nous notons CC la classe de processus ponctuels marqués de Gibbs définie sur $S = T \times M$ où $T = \mathbb{R}^2$ et $M \subseteq \mathbb{R}$. Cette classe de processus est définie à partir du processus de Poisson marqué standard et nous notons $f_{CC}$, la classe des densités des processus ponctuels marqués appartenant à CC. $f_{CC}$ est l'ensemble des fonctions $f$ telles que $f(\underline{\varphi}) = \exp(-\theta H_{CC}(\underline{\varphi}))/Z$, où $H_{CC}(\underline{\varphi}) = \sum_{x \sim y} J(\tau_x, \tau_y) g(|\mathrm{Dir}_{\varphi}(x) \cap \mathrm{Dir}_{\varphi}(y)|) + \sum_{x \in \varphi} h(\mathrm{Dir}_{\varphi}(x), \tau_x)$ avec : $J : M \times M \to \mathbb{R}$, $g : \mathbb{R} \to \mathbb{R}$, $h : N_{lf} \times M \to \mathbb{R}$}.*

### 3.7.2 Existence de processus de la classe CC

Nous allons à présent déterminer certaines conditions pour que la classe CC admettent des solutions aux équations DLR. Ces conditions sont de deux types : des conditions fonctionnelles, portant sur les fonctions $J$, $g$ et $h$, et des conditions géométriques portant sur la triangulation de Delaunay. Cette étude est très largement inspirée de travaux menés sur l'existence de processus ponctuels de Gibbs dont l'interaction est fonction de la triangulation de Delaunay [Bertin et al., 1999a]. L'ingrédient essentiel pour ce type de processus est l'utilisation d'un sous-graphe, noté $\text{Delaunay}_{\beta_0}$, du graphe de Delaunay pour contrôler la stabilité de la fonction d'énergie. En effet, pour la triangulation de Delaunay associée à un processus ponctuel de Poisson n'a aucune contrainte géométrique : la distance entre deux voisins de Delaunay n'est pas bornée, deux voisins de Delaunay peuvent être proche que l'on veut, le nombre de voisins de Delaunay d'un point n'est pas borné, etc...Ceci se répercute sur le diagramme de Dirichlet associé à un processus ponctuel de Poisson. En particulier, les arêtes de Dirichlet ainsi que les aires des cellules de Dirichlet ne sont pas bornées pour un processus ponctuel de Poisson.

Pour étudier l'existence des processus de la classe CC, nous allons utiliser un sous-graphe, $\text{Delaunay}_{\beta_0}$, du graphe de Delaunay. Ce graphe, noté formellement $\text{Del}_{3,\beta}^{\beta_0}$, est composé de l'ensemble des triangles de Delaunay qui satisfont une contrainte de petit angle.

**Définition 22** *Soit une configuration de points $\varphi$, et soit $\beta_0 \in ]0, \pi/3]$. Le sous graphe $\text{Delaunay}_{\beta_0}$, noté $\text{Del}_{3,\beta}^{\beta_0}(\varphi)$, se définit de la manière suivante :*

$$\text{Del}_{3,\beta}^{\beta_0}(\varphi) = \{\psi \in \text{Del}_3(\varphi), \beta(\psi) > \beta_0\},$$

*où la fonction $\beta(\psi)$ désigne le plus petit angle du triangle $\psi$. Nous pouvons également définir les arêtes du sous graphe $\text{Delaunay}_{\beta_0}$ pour une configuration $\varphi$, notées $\text{Del}_{2,\beta}^{\beta_0}(\varphi)$, de la manière suivante :*

$$\text{Del}_{2,\beta}^{\beta_0} = \bigcup_{\psi \in \text{Del}_{3,\beta}^{\beta_0}(\varphi)} \mathcal{P}_2(\psi).$$

Ce sous-graphe $\text{Delaunay}_{\beta_0}$ va nous permettre en particulier de contrôler le nombre de points en interaction dans le modèle CC. Cependant, l'utilisation d'un sous graphe

Delaunay$_{\beta_0}$ ne permet pas, en l'état, de contrôler la contrainte de forme sur les cellules de Dirichlet en chaque point du processus. L'utilisation de ce sous graphe Delaunay$_{\beta_0}$ ne permet pas de définir un diagramme dual caractérisant une zone d'influence pour chaque point. Nous allons définir l'ensemble de points pour lesquels la cellule de Dirichlet peut être calculée et est identique pour le graphe de Delaunay et le sous graphe de Delaunay$_{\beta_0}$. Nous notons cet ensemble de points $\mathrm{Del}_{1,\beta}^{\beta_0}$ que nous définissons comme suit.

**Définition 23** *Soit $\varphi$ une configuration de points et soit $\beta_0 \in ]0, \pi/3]$.*

$$\mathrm{Del}_{1,\beta}^{\beta_0}(\varphi) = \{x \in \varphi : T_\varphi(x) \subset \mathrm{Del}_{3,\beta}^{\beta_0}(\varphi)\},$$

*où $T_\varphi(x)$ est l'ensemble des triangles de Delaunay de la configuration $\varphi$ contenant le point $x$.*

Afin d'étudier, en premier lieu, la stabilité locale de la classe de modèles CC, nous allons contraindre les interactions à s'opérer entre les voisins du sous graphe Delaunay$_{\beta_0}$. De plus, la contrainte de forme ne s'appliquera qu'aux points ayant une zone d'influence dans le sous graphe Delaunay$_{\beta_0}$, c'est à dire aux points de $\mathrm{Del}_{1,\beta}^{\beta_0}$. Cela revient à étudier les propriétés de la fonction d'énergie suivante :

$$H_{CC}^{\beta_0}(\underline{\varphi}) = \sum_{(x,y)\in \mathrm{Del}_{2,\beta}^{\beta_0}} J(\tau_x, \tau_y) g(|\mathrm{Dir}_\varphi(x) \cap \mathrm{Dir}_\varphi(y)|) + \sum_{x \in \mathrm{Del}_{1,\beta}^{\beta_0}} h(\mathrm{Dir}(x), \tau_x).$$

Cette énergie d'interaction permet de définir une nouvelle classe de processus, la classe $\mathrm{CC}^{\beta_0}$.

Les contraintes, utilisées par l'intermédiaire du graphe de Delaunay$_{\beta_0}$, sont essentielles pour l'étude mathématique du modèle. Cependant, l'angle minimum $\beta_0$ peut se choisir aussi petit que possible, tant qu'il reste strictement positif. Par conséquent, dans la pratique, nous pouvons considérer le modèle général, proposé à la définition 21.

**Proposition 16 (Stabilité locale de la classe $\mathrm{CC}^{\beta_0}$)**

*Soit $\beta_0 \in ]0, \pi/3]$. Soit $\underline{\varphi} = \{(x, \tau_x) : x \in \varphi\} \in N_{lf}$ une configuration et $\tau_0 \in M$. Notons $\underline{\varphi'}$ la configuration telle que :*

$$\underline{\varphi'} = \underline{\varphi} \cup (0, \tau_0).$$

*Supposons que*

– *il existe une constante* $K_J$ *telle que* $|J()| < K_j$,

– *il existe une constante* $K_g$ *telle que* $|g()| < K_g$,

– *il existe une constante* $K_h$ *telle que* $|h()| < K_h$,

*Sous ces conditions :*

$$|H_{CC}^{\beta_0}(\underline{\varphi'}) - H_{CC}^{\beta_0}(\underline{\varphi})| \leq C(K_J, K_g, K_h, \beta_0).$$

**Preuve :**

D'après la définition de $H_{CC}^{\beta_0}$, la différence entre $H_{CC}^{\beta_0}(\underline{\varphi'})$ et $H_{CC}^{\beta_0}(\underline{\varphi})$ est donnée par :

$$\begin{aligned}
&H_{CC}^{\beta_0}(\underline{\varphi'}) - H_{CC}^{\beta_0}(\underline{\varphi}) = \\
&\quad \sum_{(x,y)\in \mathrm{Del}_{2,\beta}^{\beta_0}(\varphi')} J(\tau_x,\tau_y) g(|\mathrm{Dir}_{\varphi'}(x) \cap \mathrm{Dir}_{\varphi'}(y)|) + \sum_{x\in \mathrm{Del}_{1,\beta}^{\beta_0}(\varphi')} h(\mathrm{Dir}_{\varphi'}(x),\tau_x) \\
&\quad - \sum_{(x,y)\in \mathrm{Del}_{2,\beta}^{\beta_0}(\varphi)} J(\tau_x,\tau_y) g(|\mathrm{Dir}_{\varphi}(x) \cap \mathrm{Dir}_{\varphi}(y)|) - \sum_{x\in \mathrm{Del}_{1,\beta}^{\beta_0}(\varphi)} h(\mathrm{Dir}_{\varphi}(x),\tau_x)
\end{aligned}$$

En utilisant les propriétés géométriques de la mosaïque de Dirichlet, l'expression ci-dessus se simplifie en une somme de six termes :

$$\begin{aligned}
&H_{CC}^{\beta_0}(\underline{\varphi'}) - H_{CC}^{\beta_0}(\underline{\varphi}) \\
&\quad = \sum_{(0,x)\in \mathrm{Del}_{2,\beta}^{\beta_0}(\varphi')} g(|\mathrm{Dir}_{\varphi'}(0) \cap \mathrm{Dir}_{\varphi'}(x)|) J(\tau_0,\tau_x) \\
&\quad + \sum_{(x,y,0)\in \mathrm{Del}_{3,\beta}^{\beta_0}(\varphi')} (g(|\mathrm{Dir}_{\varphi'}(x) \cap \mathrm{Dir}_{\varphi'}(y)|) - g(|\mathrm{Dir}_{\varphi}(x) \cap \mathrm{Dir}_{\varphi}(y)|)) J(\tau_x,\tau_y) \\
&\quad - \sum_{(x,y)\in \mathrm{Del}_{2,\beta}^{\beta_0}(\varphi)\setminus \mathrm{Del}_{2,\beta}^{\beta_0}(\varphi')} g(|\mathrm{Dir}_{\varphi}(x) \cap \mathrm{Dir}_{\varphi}(y)|) J(\tau_x,\tau_y) \\
&\quad + h(\mathrm{Dir}_{\varphi'}(0),\tau_0) \mathbb{1}_{0\in \mathrm{Del}_{1,\beta}^{\beta_0}(\varphi')} \\
&\quad + \sum_{x\in \mathrm{Del}_{1,\beta}^{\beta_0}(\varphi');(x,0)\in \mathrm{Del}_{2,\beta}^{\beta_0}(\varphi')} h(\mathrm{Dir}_{\varphi'}(x),\tau_x) - h(\mathrm{Dir}_{\varphi}(x),\tau_x) \\
&\quad - \sum_{x\in \mathrm{Del}_{1,\beta}^{\beta_0}(\varphi)\setminus \mathrm{Del}_{1,\beta}^{\beta_0}(\varphi')} h(\mathrm{Dir}_{\varphi}(x),\tau_x).
\end{aligned}$$

Le premier terme correspond à l'interaction entre le point 0 et ses voisins. Le second terme correspond aux différences d'interaction entre la configuration $\underline{\varphi'}$ et la configuration

$\underline{\varphi}$. La somme s'effectue sur l'ensemble des triangles de Delaunay$_{\beta_0}$ de la configuration $\underline{\varphi'}$ mettant en jeu le point 0. La propriété d'insertion locale du diagramme de Dirichlet nous indique que seules les interactions entre deux points du voisinage de 0 sont modifiées. Le troisième terme quantifie les relations de voisinage qui ont disparu suite à l'insertion du point $(0, \tau_0)$ dans la configuration $\underline{\varphi}$. Le quatrième terme est expliqué par la contrainte de forme pour la cellule associée au point 0. Le cinquième terme représente la différence dans la contrainte de forme entre la configuration $\underline{\varphi'}$ et $\underline{\varphi}$. De nouveau, la propriété d'insertion locale dans le diagramme de Dirichlet nous garantit que seules les cellules associées aux points voisins de 0 ont une contrainte de forme modifiée. Enfin, le sixième terme nous vient du fait que certaines cellules « perdent » leurs contraintes de forme à cause de la contrainte de petit angle matérialisé par le paramètre $\beta_0$.

Nous allons à présent contrôler chacun des six termes indépendamment. Nous posons tout d'abord $N = \left\lceil \frac{2\pi}{\beta_0} \right\rceil$, où $\lceil . \rceil$ représente la fonction partie entière. Le résultat suivant va nous permettre de contrôler le nombre de termes de chacune des sommes composant les six termes de l'expression précédente. En effet, nous pouvons rappeler que le nombre de triangles qui apparaissent dans le graphe de Delaunay$_{\beta_0}$ suite à l'insertion du point 0 est majoré par $N$. De même le nombre de triangles qui disparaissent du graphe de Delaunay$_{\beta_0}$ est également majoré par $N$. Ces résultats sont démontrées dans [Bertin et al., 1999a].

Premièrement, nous en déduisons que le nombre de voisins du point 0 dans le graphe de Delaunay$_{\beta_0}$ est majoré par $N$. Ainsi, puisque $g$ sont $J$ sont bornées :

$$\left| \sum_{(x,0)\in \mathrm{Del}_{2,\beta}^{\beta_0}} g(|\mathrm{Dir}_{\varphi'}(0) \cap \mathrm{Dir}_{\varphi'}(x)|) J(\tau_0, \tau_x) \right| \leq N K_g K_J. \tag{3.6}$$

Comme nous l'avons remarqué précédemment, le nombre de nouveaux triangles dans le graphe de Delaunay$_{\beta_0}$ est majoré par $N$. Ainsi en rappelant que les fonctions $g$ et $J$ sont bornées nous obtenons que :

$$\left| \sum_{(x,y,0)\in \mathrm{Del}_{3,\beta}^{\beta_0}(\varphi')} (g(|\mathrm{Dir}_{\varphi'}(x) \cap \mathrm{Dir}_{\varphi'}(y)|) - g(|\mathrm{Dir}_{\varphi}(x) \cap \mathrm{Dir}_{\varphi}(y)|)) J(\tau_x, \tau_y) \right| \leq 2N K_g K_J. \tag{3.7}$$

Le contrôle du troisième terme nécessite de remarquer que les relations de voisinages qui ont disparu entre $\underline{\varphi'}$ et $\underline{\varphi}$ mettent en jeu deux points appartenant à des triangles

ayant disparu. Comme nous avons remarquer que le nombre de triangles ayant disparus est majoré par $N$, nous pouvons donc majorer par $3N$ le nombre de relations de voisinages ayant disparu. En utilisant le fait que les fonctions $g$ et $J$ sont bornées, nous obtenons que :

$$\left| \sum_{(x,y)\in \mathrm{Del}_{2,\beta}^{\beta_0}(\varphi)\setminus \mathrm{Del}_{2,\beta}^{\beta_0}(\varphi')} g(|\mathrm{Dir}_{\varphi}(x)\cap \mathrm{Dir}_{\varphi}(y)|)J(\tau_x,\tau_y) \right| \leq 3NK_gK_J. \tag{3.8}$$

Ensuite, par hypothèse nous avons directement :

$$\left| h(\mathrm{Dir}_{\varphi'}(0),\tau_0)\mathbb{1}_{0\in \mathrm{Del}_{1,\beta}^{\beta_0}(\varphi')} \right| \leq K_h. \tag{3.9}$$

Pour le cinquième terme, nous pouvons constater que la somme se fait sur l'ensemble des voisins de Delaunay du point 0. De plus, la somme ne tient compte que des points de $\mathrm{Del}_{1,\beta}^{\beta_0}$, ce qui implique que ces points sont voisins de 0 dans le graphe de $\mathrm{Delaunay}_{\beta_0}$. Le nombre de terme de la somme peut donc se majorer par $2N$. En rappelant que la fonction $h$ est bornée, nous obtenons que :

$$\left| \sum_{x\in \mathrm{Del}_{1,\beta}^{\beta_0}(\varphi');(x,0)\in \mathrm{Del}_{2,\beta}^{\beta_0}(\varphi')} h(\mathrm{Dir}_{\varphi'}(x),\tau_x) - h(\mathrm{Dir}_{\varphi}(x),\tau_x) \right| \leq 2NK_h. \tag{3.10}$$

Enfin, pour le sixième et dernier terme, nous pouvons constater que l'ensemble des points de la somme appartiennent au voisinage de Delaunay du point 0. En effet, seuls ces points appartiennent à des triangles ayant été modifiés par l'insertion du point 0. De plus, ces points appartiennent nécessairement à des triangles ayant disparu du graphe de $\mathrm{Delaunay}_{\beta_0}$. Or nous avons constaté que le nombre de triangles ayant disparu du graphe de $\mathrm{Delaunay}_{\beta_0}$ est majoré par $N$. Ainsi, puisque la fonction $h$ est bornée, nous obtenons que :

$$\left| \sum_{x\in \mathrm{Del}_{1,\beta}^{\beta_0}(\varphi)\setminus \mathrm{Del}_{1,\beta}^{\beta_0}(\varphi')} h(\mathrm{Dir}_{\varphi}(x),\tau_x) \right| \leq 3NK_h \tag{3.11}$$

Nous pouvons à présent combiner les équations 3.6-3.11, pour obtenir que :

$$\left| H_{CC}^{\beta_0}(\underline{\varphi'}) - H_{CC}^{\beta_0}(\underline{\varphi}) \right| \leq 6N(K_gK_J + K_h).$$

■

Cette propriété nous sera particulièrement utile pour l'étude théorique de l'algorithme de simulation dans la section suivante. En effet, la dynamique utilisée est une dynamique de Métropolis-Hastings par insertion-délétion. Ainsi, à chaque itération de l'algorithme, la proposition précédente nous permet le contrôle du saut d'énergie.

L'étude de la quasilocalité de la classe de fonctions d'énergie CC, nécessite de rajouter des contraintes de portée sur les interactions. En tout premier lieu, nous allons supposer que l'interaction entre deux points voisins, situés à une distance supérieure à un entier $2R$, est nulle. De plus, nous allons considérer que la contrainte de forme ne porte plus sur l'aire des cellules de Dirichlet, mais sur l'aire des cellules de Dirichlet intersectant une boule centrée sur les points et de rayon $R$. Ce type de diagramme est très ressemblant de domaines de Dirichlet libres introduits par Graner et Sawada [Graner et Sawada, 1993] (voir figure 3.6). Pour étudier la stabilité, nous allons donc nous intéresser à la classe de fonctions suivante :

$$H_{CC}^{\beta_0,R}(\underline{\varphi}) = \sum_{(x,y)\in \mathrm{Del}_{2,\beta}^{\beta_0}} J(\tau_x,\tau_y) g(|\mathrm{Dir}_{\varphi}(x_i)\cap \mathrm{Dir}_{\varphi}(y)|)\mathbb{1}_{||x-y||<2R} + \sum_{x\in \mathrm{Del}_{1,\beta}^{\beta_0}} h(\mathrm{Dir}(x)\cap B(x,R),\tau_x),$$

où $B(x,R)$ représente la boule fermée de centre $x$ et de rayon $R$.

**Proposition 17 (Quasilocalité)** *Soient $R>0$ et $\beta_0 \in ]0,\pi/3]$. Soit $\Delta$ un borélien borné tel que $0\in\Delta$. Soit $\underline{\varphi}=\{(x,\tau_x) : x\in\varphi\}\in N_{lf}$ une configuration et $\tau_0\in M$. Notons $\underline{\varphi'}$ la configuration telle que :*

$$\underline{\varphi'} = \underline{\varphi}\cup(0,\tau_0).$$

*Sous ces conditions, nous avons :*

$$\lim_{d(0,\Delta^c)\to+\infty} \left| [H_{CC}^{\beta_0,R}(\underline{\varphi'}) - H_{CC}^{\beta_0,R}(\underline{\varphi})] - [H_{CC}^{\beta_0,R}(\underline{\varphi'}_{\Delta}) - H_{CC}^{\beta_0,R}(\underline{\varphi}_{\Delta})] \right| = 0.$$

**Preuve :**

Comme nous l'avons vu au cours de la preuve précédente nous avons :

$$
\begin{aligned}
&H_{CC}^{\beta_0,R}(\underline{\varphi'}) - H_{CC}^{\beta_0,R}(\underline{\varphi}) \\
&\quad = \sum_{(0,x)\in \mathrm{Del}_{2,\beta}^{\beta_0}(\underline{\varphi'})} g(|\mathrm{Dir}_{\varphi'}(0)\cap \mathrm{Dir}_{\varphi'}(x)|)J(\tau_0,\tau_x)\mathbb{1}_{||0-x||<2R} \\
&\quad + \sum_{(x,y,0)\in \mathrm{Del}_{3,\beta}^{\beta_0}(\varphi')} (g(|\mathrm{Dir}_{\varphi'}(x)\cap \mathrm{Dir}_{\varphi'}(y)|) - g(|\mathrm{Dir}_{\varphi}(x)\cap \mathrm{Dir}_{\varphi}(y)|))J(\tau_x,\tau_y)\mathbb{1}_{||x-y||<2R} \\
&\quad - \sum_{(x,y)\in \mathrm{Del}_{2,\beta}^{\beta_0}(\varphi)\setminus \mathrm{Del}_{2,\beta}^{\beta_0}(\varphi')} g(|\mathrm{Dir}_{\varphi}(x)\cap \mathrm{Dir}_{\varphi}(y)|)J(\tau_x,\tau_y)\mathbb{1}_{||x-y||<2R} \\
&\quad + h(\mathrm{Dir}_{\varphi'}(0)\cap B(0,R),\tau_0)\mathbb{1}_{0\in \mathrm{Del}_{1,\beta}^{\beta_0}(\varphi')} \\
&\quad + \sum_{xs\in \mathrm{Del}_{1,\beta}^{\beta_0}(\varphi');(x,0)\in \mathrm{Del}_{2,\beta}^{\beta_0}(\varphi')} h(\mathrm{Dir}_{\varphi'}(x)\cap B(x,R),\tau_x) - h(\mathrm{Dir}_{\varphi}(x)\cap B(x,R),\tau_x) \\
&\quad - \sum_{x\in \mathrm{Del}_{1,\beta}^{\beta_0}(\varphi)\setminus \mathrm{Del}_{1,\beta}^{\beta_0}(\varphi')} h(\mathrm{Dir}_{\varphi}(x)\cap B(x,R),\tau_x).
\end{aligned}
$$

La contrainte de forme, s'apparente à un modèle de Ord. Nous pouvons donc utiliser la propriété de Markoviannité à l'ordre 1, établie par [Baddeley et Møller, 1989], pour constater que cette contrainte a une portée incluse dans la boule fermée de centre 0 et de rayon $2R$. De plus la portée sur les relations de voisinages nous permet d'établir que :

$$H_{CC}^{\beta_0,R}(\underline{\varphi'}) - H_{CC}^{\beta_0,R}(\underline{\varphi}) = H_{CC}^{\beta_0,R}(\underline{\varphi'}_{B(0,2R)}) - H_{CC}^{\beta_0,R}(\underline{\varphi}_{B(0,2R)}).$$

Nous en déduisons donc que :

$$\lim_{d(0,\Delta^c)\to+\infty} \left| [H_{CC}^{\beta_0,R}(\underline{\varphi'}) - H_{CC}^{\beta_0,R}(\underline{\varphi})] - [H_{CC}^{\beta_0,R}(\underline{\varphi'}_{\Delta}) - H_{CC}^{\beta_0,R}(\underline{\varphi}_{\Delta})] \right| = 0.$$

■

Les deux propriétés précédentes nous permettent donc d'établir le théorème suivant garantissant l'existence de mesures de probabilité vérifiant les spécifications locales de la classe CC.

**Théorème 4** *Soit $\underline{\varphi}$ un processus ponctuel marqué de Gibbs de la classe CC, ayant comme fonction d'énergie $H_{CC}^{\beta_0,R}$. Supposons de plus que*

- *il existe une constante $K_J$ telle que $|J()| < K_j$,*
- *il existe une constante $K_g$ telle que $|g()| < K_g$,*

– *il existe une constante* $K_h$ *telle que* $|h()| < K_h$,

*Sous ces conditions, il existe une mesure de probabilité satisfaisant les spécifications locales liées à la fonction d'interaction* $H_{CC}^{\beta_0,R}$.

**Preuve :** Les propriétés 16 et 17 nous permettent de vérifier la propriété 15.

■

Il est important de remarquer que les contraintes supplémentaires, contraintes de petit angle et portée, pour passer de la fonction $H_{CC}$ à la fonction $H_{CC}^{\beta_0,R}$ n'ont aucun sens biologique. Elles ne sont utiles que dans un but mathématique. Cependant, nous pouvons remarquer que le paramètre $\beta_0$ peut être choisi aussi petit que souhaité (à condition qu'il reste strictement positif). De même, le paramètre $R$ peut-être fixé aussi grand que possible (à condition qu'il reste fini). Ainsi, en pratique il est raisonnable de confondre la classe de modèle définie à partir de la fonction d'énergie $H_{CC}^{\beta_0,R}$, de celle définie de manière générale par la fonction d'énergie $H_{CC}$.

Nous allons à présent nous focaliser sur la simulation de cette classe de modèle. La simulation s'effectue la plupart du temps dans un borné conditionnellement aux bords. Cette remarque nous permet de comprendre que l'étude théorique de l'algorithme de simulation va s'appuyer essentiellement sur la propriété de stabilité locale du processus.

## 3.8 Simulation de la classe de processus de Gibbs CC

La classe de modèle étant complètement définie, nous allons à présent nous intéresser à la simulation des processus de la classe CC. La simulation nous permettra en premier lieu d'étudier le comportement de notre modèle et de vérifier que celui-ci est conforme à la construction du modèle (c'est à dire à nos attentes). Puis, dans une section suivante, nous nous intéresserons à l'inférence des paramètres du modèle notamment à l'aide du simulateur.

Dans cette section, nous allons décrire, dans un premier temps, les algorithmes développés pour simuler des processus Gibbsiens, puis nous présenterons l'algorithme de Métropolis-Hastings que nous avons implémenté. Nous étudierons ensuite théoriquement notre algorithme.

### 3.8.1 Description de l'algorithme utilisé pour simuler notre modèle

La simulation de quelques processus spatiaux, comme le processus de Poisson, peut se faire directement. En effet pour le processus de Poisson, la forme analytique de la densité est connue ce qui permet une simulation simple (voir par exemple [Stoyan et al., 1995] et [Møller et Waagepetersen, 2003]). Malheureusement, la densité de processus spatiaux n'est généralement pas connue analytiquement. Par exemple, pour les processus ponctuels de Gibbs, la constante de normalisation (également appélée fonction de partition) est généralement incalculable.

L'outil de base pour simuler ce type de processus repose sur la construction d'une chaîne de Markov ergodique dont la loi invariante est la loi du processus. Ce type de simulation est communément appelé simulation MCMC (*Markov Chain Monte-Carlo*). La simulation du processus conditionnellement au nombre de points fixé est assez simple et a été étudiée en détail dans [Metropolis et al., 1953] et [Ripley, 1979].

Dans le cas général, la communauté de la statistique spatiale s'est longtemps focalisée sur l'utilisation d'algorithmes de naissance et de mort. Certains résultats théoriques et quelques utilisations pratiques de ces algorithmes sont présents dans différents ouvrages [Kelly et Ripley, 1976], [Preston, 1977], [Ripley, 1977] et [Baddeley et Møller, 1989]. Au cours des dernières années, d'importants développements ont été apportés à la simulation de processus stochastiques par l'intermédiaire de la simulation parfaite (ou exacte). Ces algorithmes font suite aux travaux originaux de Prop et Wilson [Propp et Wilson, 1996]. De nombreux algorithmes de simulation exacte ont, depuis, été développés pour différents processus spatiaux.

Cependant, une autre méthode, basée sur les travaux de [Metropolis et al., 1953] et [Hastings, 1970], a été développée par Geyer et Møller [Geyer et Møller, 1994]. Cette méthode reste à ce jour la plus utilisée pour la simulation de processus ponctuels grâce à sa simplicité de mise en œuvre et son efficacité. La simulation de processus spatiaux par l'algorithme de Metropolis-Hastings consiste à construire une chaîne de Markov ergodique, dont la distribution invariante est la loi du processus.

**Algorithme 1** : Algorithme de Metropolis-Hastings pour la simulation de processus de la classe CC

*Paramètres*

$\lambda > 0$ : intensité du processus de poisson initial;

$H_{CC}$ la fonction d'interaction du processus de Gibbs;

*Initialisation par un processus de Poisson homogène d'intensité* $\lambda$

**pour chaque** *i in n.iter* **faire**
  $U_1 \longleftarrow$ RANDOM;
  $U_2 \longleftarrow$ RANDOM;
  **si** $U_1 < 1/2$ **alors**
    Insertion : Generation uniforme d'un point;
    $\xi \longleftarrow$ Choix uniforme d'un point;
    **si** $(\exp(-\theta(H_{CC}(\varphi \cup (x,\tau)) - H_{CC}(\varphi))) > U_2)$ **alors**
      $\varphi \longleftarrow \varphi \cup (x,\tau)$;
    **fin**
  **fin**
  **sinon**
    Délétion : Choix uniforme d'un point de $\varphi$;
    $(\eta,\tau_\eta) \longleftarrow$ Choix uniforme d'un point de $\varphi$ ;
    **si** $(\exp(-\theta(H_{CC}(\varphi \backslash (\eta,\tau_\eta)) - H_{CC}(\varphi))) > U_2)$ **alors**
      $\varphi \longleftarrow \varphi \backslash (\eta,\tau_\eta)$;
    **fin**
  **fin**
**fin**

Nous avons choisi de simuler notre modèle par l'intermédiaire d'un algorithme de Metropolis-Hastings à temps discret. Cet algorithme est présenté en détails page 160. Nous allons ensuite étudier la convergence de la chaîne de Markov construite à partir de cet algorithme.

### 3.8.2 Etude théorique de l'algorithme

Afin d'étudier théoriquement la convergence de notre algorithme de simulation, nous allons effectuer un certain nombre de rappels concernant les chaînes de Markov. Puis nous formaliserons l'algorithme proposé en décrivant le noyau de transition de la chaîne de Markov associée. Enfin, nous montrerons que la chaîne de Markov décrite par notre algorithme de simulation est ergodique ce qui garantit la convergence de l'algorithme en nous appuyant sur des résultats de [Geyer et Møller, 1994].

#### Généralités sur les chaînes de Markov

Les résultats que nous allons présenter sont principalement issus des ouvrages de Nummelin et de Meyn et Tweedie [Nummelin, 1984] , [Meyn et Tweedie, 1993].

Considérons un espace $\Omega$ sur lequel est défini une distribution de probabilité $\pi$, et notons $\mathcal{B}(\Omega)$ la plus petite tribu engendrée par $\Omega$. Le principe général d'un algorithme MCMC est de construire une chaîne de Markov homogène $Y_0, Y_1, \ldots, Y_n, \ldots$ d'espace d'états $\Omega$ et telle que :

$$P^m(x, F) \to \pi(F) \quad \text{quand } m \to +\infty,$$

pour $F \subseteq \mathcal{B}(\Omega)$ et $x \in \Omega$, où :

$$P^m(x, F) = \mathbb{P}(Y_m \in F | Y_0 = x).$$

Nous allons rappeler dans la suite de cette section certaines caractéristiques générales sur les chaînes de Markov.

**Homogénéité** L'homogénéité d'une chaîne de Markov signifie que le noyau de transition, $\mathbb{P}(x, F) = \mathbb{P}(Y_{m+1} \in F | Y_m = x)$, est indépendant de $m$.

**Invariance** Une chaîne de Markov est $\pi$-invariante, c'est à dire que $\pi$ est une loi invariante pour la chaîne, si $Y_m \rightsquigarrow \pi \Rightarrow Y_{m+1} \rightsquigarrow \pi$, où le symbole $\rightsquigarrow$ signifie que $Y_m$ suit la loi de $\pi$.

**Réversibilité** Une chaîne de Markov est dite réversible si les couples $(Y_m, Y_{m+1})$ et $(Y_{m+1}, Y_m)$ sont identiquement distribués, pour tout $m$, c'est à dire, $\forall F, G \in \mathcal{B}(\Omega)$ :

$$\mathbb{P}(Y_m \in F, Y_{m+1} \in G, Y_m \neq Y_{m+1}) = \mathbb{P}(Y_m \in G, Y_{m+1} \in F, Y_m \neq Y_{m+1}).$$

**Irréductibilité** Une chaîne de Markov est dite $\Psi$-irréductible s'il existe une mesure non nulle $\Psi$, telle que pour tout $x \in \Omega$ et $F \in \mathcal{B}(\Omega)$ avec $\Psi(F) > 0$, $P^m(x, F) > 0$ pour $m \in \mathbb{N}$.

**Harris-récurrence** Une chaîne de Markov Harris-récurrente est une chaîne pour laquelle, il existe une mesure $\Psi$, telle que la chaîne soit $\Psi$-irréductible et $\forall x \in \Omega$ et $F \in \mathcal{B}(\Omega)$ avec $\Psi(F) > 0$ :

$$\mathbb{P}(Y_m \in F \text{ pour } m \in \mathbb{N} | Y_0 = x) = 1$$

La proposition suivante nous permet de faire le lien entre ces différentes notions sur les chaînes de Markov.

**Proposition 18** *Supposons qu'une loi invariante $\pi$ existe pour une chaîne de Markov. La $\Psi$-irréductibilité de la chaîne implique que*

1. *la chaîne est $\pi$-irréductible,*
2. *$\pi$ est l'unique distribution invariante pour la chaîne,*
3. *$\pi$ domine $\Psi$, c'est à dire $\pi(F) = 0 \Rightarrow \Psi(F) = 0$,*
4. *il existe $A \subset \Omega$, avec $\pi(A) = 0$ tel que la chaîne restreinte à $\Omega \backslash_A$ soit Harris récurrente et $\Omega \backslash_A$ est absorbant, c'est à dire si $Y_0 \in \Omega \backslash_A$ alors $Y_m \in \Omega \backslash_A$, pour $m \in \mathbb{N}$.*

**Définition 24** *Un espace $C$ est dit petit s'il existe $m \in \mathbb{N}^*$ et une mesure sur $\mathcal{B}(\Omega)$ $\nu > 0$ tels que :*

$$P^m(F, x) \geq \nu(F) \quad \forall x \in C, \ \forall F \in \mathcal{B}(\Omega).$$

Le caractère Harris-Recurrent pour une chaîne de Markov est bien souvent assez délicat à montrer. La proposition suivante permet de faciliter ce point.

**Proposition 19** *Soit une chaîne de Markov $\Psi$-irredutible. Supposons qu'il existe un petit espace $C \subseteq \Omega$ et une fonction $V : \Omega \to [1, \infty)$ tels que $\{x \in \Omega : V(x) \leq \alpha\}$ est petit pour $\alpha > 1$ et :*

$$\mathbb{E}[V(Y_1)|Y_0 = x] \leq V(x) \quad \forall x \in \Omega \backslash C,$$

*alors la chaîne est Harris-récurrente.*

**Apériodicité** Soit $Y_0, Y_1, \ldots, Y_m, \ldots$ une chaîne de Markov définie sur $\Omega$ et $\Psi$-irréductible. Si $\Omega$ peut être partitionné en $\Omega_0, \Omega_1, \ldots, \Omega_d$ et $A$, tel que $\mathbb{P}(x, \Omega_j) = 1$, $\forall x \in \Omega_i$, $j = i + 1 \bmod d$ et $\Psi(A) = 0$, avec $d > 1$, alors la chaîne est dite *périodique*, sinon elle est *apériodique*.

De même, la définition d'une chaîne apériodique est délicate à utiliser et il est assez usuel de passer par la proposition suivante pour montrer le caractère périodique ou apériodique d'une chaîne de Markov.

**Proposition 20** *Une chaîne de Markov irréductible admettant $\pi$ comme distribution invariante est apériodique si et seulement si il existe un petit espace $C$ avec $\pi(C) > 0$ et $n \in \mathbb{N}$ :*

$$P^m(x, C) > 0 \quad \forall x \in C \text{ et } m \geq n.$$

**Ergodicité** Une chaîne de Markov est dite ergodique si elle est Harris récurrente et apériodique.

**Proposition 21** *Si une chaîne de Markov est ergodique alors :*

$$P^m(x, F) \to \pi(F) \quad \forall x \in \Omega.$$

L'ensemble des caractéristiques introduites dans cette section vont nous permettre d'étudier avec précision la chaîne de Markov générée par l'algorithme proposé à la page 160. Cette étude nécessite tout d'abord l'écriture formelle du noyau de transition de la chaîne de Markov.

**Description formelle de notre algorithme**

Nous allons tout d'abord introduire quelques notations qui seront dans la mesure du possible conformes aux notations utilisées par Geyer et Møller [Geyer et Møller, 1994].

Soit $(S, \mathcal{B}, \lambda)$ un espace mesurable tel que $S = T \times M, 0 < \varrho(T) < \infty, 0 < p(M) < \infty$ et pour lequel $\mathcal{B}$ contient tous les singletons. $(\Omega, \mathcal{F}, \mu)$ correspond à l'espace exponentiel associé à $(S, \mathcal{B}, \lambda)$ [Carter et Prenter, 1972]. Plus précisémment, $\Omega = \bigcup_{n=0}^{\infty} \Omega_n$ où :

$$\Omega_n = \{(x_1, \tau_1), \ldots, (x_n, \tau_n) \subset S\}$$

est l'espace des configurations à $n$ points inclus dans $S$. $\mathcal{F}$ représente la tribu engendrée par $\Omega$ et $\mu$ est le processus de Poisson d'intensité $\rho = \varrho \times p$ sur $S$.

Nous pouvons également rappeler que l'hérédité d'un processus ponctuel de densité $f$ par rapport au processus de Poisson se définit de la manière suivante :

$$f(\underline{\varphi}) > 0 \Rightarrow f(\underline{\psi}) > 0 \quad \forall \varphi \subseteq \psi,$$

où $\varphi \subseteq \psi$ signifie que $\psi$ est une sous-configuration de $\varphi$.

Nous pouvons également introduire les espaces $K_n$, pour $n = 1, \ldots, +\infty$ définis par :

$$K_n = \Omega_n \cap K \text{ où } K = \{f > 0\}.$$

Les espaces $K_n$ correspondent donc aux configurations de mesure strictement positive et composées d'exactement $n$ points. Chaque espace $K_n$ engendre une tribu d'événements que nous notons $\mathcal{F}_{\setminus}$, pour $n = 1, \ldots, +\infty$.

Nous pouvons à présent décrire la chaîne de Markov construite par l'algorithme de la page 160. En supposant que $\underline{\varphi} \in K_n$ avec $n > 0$ (resp. $K_0$), le noyau $Q(\underline{\varphi}, .)$ est concentré sur $K_{n-1} \cup K_n \cup K_{n+1}$ (resp. $K_0 \cup K_1$). Plus précisémment, avec une probabilité $q(\underline{\varphi})$, un nouveau point (marqué), $\underline{x} = (x, \tau_x)$, est généré à partir d'une densité $b(\underline{\varphi}, \underline{x})$ et avec une probabilité $1 - q(\underline{\varphi})$, un point (marqué) $\underline{y} = (y, \tau_y) \in \underline{\varphi}$, sélectionné selon une probabilité $d(\underline{\varphi} \setminus \underline{y}, \underline{y})$, est enlevé au processus. La probabilité d'acceptation est notée $A$.

Pour l'algorithme décrit à la page 160, les fonctions utilisées sont les suivantes :

$$q(.) \equiv \frac{1}{2} \quad b(.,.) \equiv \frac{1}{\varrho(T)p(M)} \quad d(\varphi, .) \equiv \frac{1}{n+1} \text{ si } \varphi \in \Omega_n.$$

La probabilité d'acceptation $A$ s'écrit, quant à elle :

$$A(\underline{\varphi}|\underline{\varphi}\cup\underline{x}) = \begin{cases} \min\{1, 1/r(\underline{\varphi},\underline{x})\} & \text{si } \underline{\varphi}\cup\underline{x}\in K \\ 0 & \text{sinon} \end{cases}$$

$$A(\underline{\varphi}\cup\underline{x}|\underline{\varphi}) = \begin{cases} \min\{1, r(\underline{\varphi},\underline{x})\} & \text{si } \underline{\varphi}\cup\underline{x}\in K \\ 0 & \text{sinon} \end{cases}$$

où :

$$r(\underline{\varphi},\underline{x}) = \frac{f(\underline{\varphi}\cup\underline{x})}{f(\underline{\varphi})}\frac{\varrho(T)p(M)}{n+1} \quad \forall(\underline{\varphi}\cup\underline{x})\in K_{n+1}.$$

Le noyau de transition de la chaîne de Markov construite par l'algorithme de la page 160 s'écrit donc, pour $\underline{\varphi}\in\Omega_n$, $n>0$ :

$$Q(\underline{\varphi}, F_m) = 0 \quad \forall F_m\in\mathcal{F}_m \text{ tel que } |m-n|>1,$$

$$Q(\underline{\varphi}, F_{n+1}) = q(\underline{\varphi})\int_{\underline{\varphi}\cup\underline{x}\in F_{n+1}} b(\underline{\varphi},\underline{x})A(\underline{\varphi}\cup\underline{x}|\underline{\varphi})\varrho(dx)p(d\tau_x) \quad \forall F_{n+1}\in\mathcal{F}_{n+1},$$

$$Q(\underline{\varphi}, F_{n-1}) = (1-q(\underline{\varphi}))\sum_{\underline{y}\in\underline{\varphi}} \mathbb{1}_{F_{n-1}}(\underline{\varphi}\backslash\underline{y})d(\underline{\varphi}\backslash\underline{y},\underline{y})A(\underline{\varphi}\backslash\underline{y}|\underline{\varphi}) \quad \forall F_{n-1}\in\mathcal{F}_{n-1},$$

$$\begin{aligned} Q(\underline{\varphi}, F_n) = \mathbb{1}_{F_n}(\underline{\varphi}) \quad & \left\{ q(\underline{\varphi})\int_S b(\underline{\varphi},\underline{x})(1-A(\underline{\varphi}\cup\underline{x}|\underline{\varphi}))\varrho(dx)p(d\tau_x) \right. \\ & \left. + \quad (1-q(\underline{\varphi}))\sum_{\underline{y}\in\underline{\varphi}} d(\underline{\varphi}\backslash\underline{y},\underline{y})(1-A(\underline{\varphi}\backslash\underline{y}|\underline{\varphi})) \right\} \quad \forall F_{n+1}\in\mathcal{F}_{n+1}, \end{aligned}$$

$$Q(\emptyset, F_1) = q(\emptyset)\int_{\underline{x}\in F_1} b(\emptyset,\underline{x})A(\underline{x}|\emptyset)\varrho(dx)p(d\tau_x) \quad \forall F_1\in\mathcal{F}_1,$$

$$Q(\emptyset, F_0) = \left\{ q(\emptyset)\int_S b(\emptyset,\underline{x})(1-A(\underline{x}|\emptyset))\varrho(dx)p(d\tau_x) + (1-q(\emptyset)) \right\}.$$

En nous inspirant de l'étude menée par Geyer et Møller, nous allons nous intéresser à la convergence de la chaîne de Markov décrite ci-dessus.

**Convergence de l'algorithme**

Dans cette section nous allons montrer que la chaîne de Markov, générée par l'algorithme de la page 160, est ergodique.

**Proposition 22** *Soit $\underline{\Phi}$ un processus ponctuel marqué de Gibbs de la classe CC, ayant comme fonction d'énergie $H_{CC}^{\beta_0}$. Nous rappelons que $H_{CC}^{\beta_0}$ caractérise les fonctions d'énergie $H_{CC}$ satisfaisant une contrainte de petite angle. Supposons de plus que*

- *il existe une constante $K_J$ telle que $|J()| < K_j$,*
- *il existe une constante $K_g$ telle que $|g()| < K_g$,*
- *il existe une constante $K_h$ telle que $|h()| < K_h$,*

*Sous ces conditions nous obtenons que la densité du processus ponctuel par rapport au processus de Poisson est héréditaire.*

**Preuve.** D'après la propriété de stabilité du processus (propriété 16), nous avons pour tout $\underline{\varphi} \in \Omega$ et pour tout $\underline{x} \in S$ :

$$H_{CC}^{\beta_0}(\underline{\varphi} \cup \underline{x}) - H_{CC}^{\beta_0}(\underline{\varphi}) > -K \quad K > 0$$

Soit $\underline{\varphi} \in \Omega$ et $\underline{\psi} \subseteq \underline{\varphi}$. Nous pouvons écrire que $\underline{\varphi} = \underline{\psi} \cup (x_1, \tau_1) \cup (x_2, \tau_2) \cdots \cup (x_n, \tau_n)$. Comme $f(\underline{\varphi}) = \frac{\exp(-H_{CC}^{\beta_0}(\underline{\varphi}))}{Z} > 0$, nous avons :

$$\begin{aligned}
f(\underline{\psi}) &= \frac{\exp\left(-\theta H_{CC}^{\beta_0}(\underline{\psi})\right)}{Z} \\
&= \frac{\exp\left(-H_{CC}(\underline{\varphi}) + \sum_{i=1}^{n}\left(H_{CC}^{\beta_0}\left(\underline{\psi} \bigcup_{j=1}^{i}(x_i,\tau_i)\right) - H_{CC}^{\beta_0}\left(\underline{\psi} \bigcup_{j=1}^{i-1}(x_i,\tau_i)\right)\right)\right)}{Z} \\
&= f(x)\prod_{i=1}^{n}\exp\left(H_{CC}\left(\psi \bigcup_{j=1}^{i}(x_i,\tau_i)\right) - H_{CC}\left(\psi \bigcup_{j=1}^{i-1}(x_i,\tau_i)\right)\right) \\
&> \prod_{i=1}^{n}\exp(-K) \\
&> 0
\end{aligned}$$

■

La propriété précédente se retrouve également dans les travaux de Geyer [Geyer, 1998]

**Proposition 23** *La chaîne de Markov $X_1, X_2, ...$ générée par l'algorithme 1 est $f$-invariante.*

**Preuve.**

Par construction de l'algorithme, nous remarquons facilement que $f$ est invariante pour la dynamique de Métropolis Hastings proposée.

■

**Proposition 24** *La chaîne de Markov $X_1, X_2, ...$ générée par l'algorithme de la page 160 est $f$-réversible.*

**Preuve.**

On rappelle que $\Omega = \bigcup_{n=1}^{+\infty} \Omega_n$. Si $X_i \in \Omega_n$ $(n > 0)$ et $X_{i+1} \neq X_i$ alors, $X_{i+1} \in \Omega_{n-1} \cup \Omega_{n+1}$. Pour montrer la réversibilité $X_i$, il suffit de vérifier que $\forall n \in \mathbb{N}$, $F_n \in \mathcal{F}_n$ et $F_{n+1} \in \mathcal{F}_{n+1}$ :

$$\begin{aligned}
&\int_T \int_M \dots \int_T \int_M \int_T \int_M \mathbb{1}_{F_n}(\underline{\varphi}) \mathbb{1}_{F_{n+1}}(\underline{\varphi} \cup \underline{x}) q(\underline{\varphi}) b(\underline{\varphi}, \underline{x}) A(\underline{\varphi} \cup \underline{x} | \underline{\varphi}) \\
&f(\underline{\varphi}) \exp(-|S|)/n! p(\tau_x) \prod_{i=1}^{n} p(\tau_{x_i}) dx d\tau_x, dx_1 d\tau_1 \dots dx_n d\tau_n \\
= &\int_T \int_M \dots \int_T \int_M \sum_{i=1}^{n+1} \mathbb{1}_{F_n}(\underline{\psi} \backslash (y_i, \kappa_i)) \mathbb{1}_{F_{n+1}}(\underline{\psi}) (1 - q(\underline{\psi})) d(\underline{\psi} \backslash (y_i, \kappa_i)) A(\underline{\psi} \backslash (y_i, \kappa_i) | \underline{\psi}) \\
&f(\underline{\psi}) \exp(-|S|)/(n+1)! \prod_{i=1}^{n+1} p(\kappa_i) dy_1 d\kappa_1, \dots dy_{n+1} d\kappa_{n+1},
\end{aligned}$$

où $\underline{\varphi} = \{(x_1, \tau_1), \dots, (x_n, \tau_n)\}$ et $\underline{\psi} = \{(y_1, \kappa_1), \dots, (y_{n+1}, \kappa_{n+1})\}$.

Le second terme dans l'égalité précédente est égal à :

$$\begin{aligned}
&\int_T \int_M \dots \int_T \int_M \mathbb{1}_{F_n}(\underline{\psi} \backslash (y_1, \kappa_1)) \mathbb{1}_{F_{n+1}}(\underline{\psi}) (1 - q(\underline{\psi})) d(\underline{\psi} \backslash (y_1, \kappa_1), (y_1, \kappa_1)) A(\underline{\psi} \backslash (y_i, \kappa_i) | \underline{\psi}) \\
&f(\underline{\psi}) \exp(-|S|)/n! \prod_{i=1}^{n+1} p(\kappa_i) dy_1 d\kappa_1, \dots dy_{n+1} d\kappa_{n+1}.
\end{aligned}$$

Et en choisissant $\underline{\varphi} = \underline{\psi} \backslash (y_1, \kappa_1)$ et $\underline{x} = (y_1, \kappa_1)$, nous obtenons pour le premier terme :

$$\begin{aligned}
&\int_T \int_M \dots \int_T \int_M \mathbb{1}_{F_n}(\underline{\varphi}) \mathbb{1}_{F_{n+1}}(\underline{\psi}) q(\underline{\varphi}) b(\underline{\varphi}, \underline{x}) A(\underline{\psi} | \underline{\varphi}) \\
&f(\underline{\varphi}) \exp(-|S|)/n! \prod_{i=1}^{n+1} p(\kappa_i) dy_1 d\kappa_1, \dots dy_{n+1} d\kappa_{n+1},
\end{aligned}$$

et pour le second terme :

$$\int_T \int_M \dots \int_T \int_M \mathbb{1}_{F_n}(\underline{\varphi}) \mathbb{1}_{F_{n+1}}(\underline{\psi})(1 - q(\underline{\psi}))d(\underline{\varphi}, \underline{x})A(\underline{\varphi}|\underline{\psi})$$
$$f(\underline{\psi}) \exp(-|S|)/n! \prod_{i=1}^{n+1} p(\kappa_i) dy_1 d\kappa_1, \dots dy_{n+1} d\kappa_{n+1}.$$

La condition de réversibilité se réécrit donc :

$$q(\underline{\varphi})b(\underline{\varphi}, \underline{x})A(\underline{\psi}|\underline{\varphi})f(\underline{\varphi}) = (1 - q(\underline{\psi}))d(\underline{\varphi}, \underline{x})A(\underline{\varphi}|\underline{\psi})f(\underline{\psi}).$$

Lorsque $\underline{\psi} \notin K$, cette égalité est directement vérifiée car $A(\underline{\psi}|\underline{\varphi}) = A(\underline{\varphi}|\underline{\psi}) = 0$.
Dans le cas où $\underline{\psi} \in K$ et $r(\underline{\varphi}, \underline{x}) \geq 1$, nous avons :

$$\begin{aligned}
(1 - q(\underline{\psi}))d(\underline{\varphi}, \underline{x})A(\underline{\varphi}|\underline{\psi})f(\underline{\psi}) &= (1 - q(\underline{\psi}))d(\underline{\varphi}, \underline{x}) \frac{1}{r(\underline{\varphi}, \underline{x})} f(\underline{\psi}) \\
&= (1 - q(\underline{\psi}))d(\underline{\varphi}, \underline{x}) \frac{f(\underline{\varphi})q(\underline{\varphi})b(\underline{\varphi}, \underline{x})}{f(\underline{\psi})(1 - q(\underline{\psi}))d(\underline{\varphi}, \underline{x})} f(\underline{\psi}) \\
&= q(\underline{\varphi})b(\underline{\varphi}, \underline{x})f(\underline{\varphi}) \\
&= q(\underline{\varphi})b(\underline{\varphi}, \underline{x})A(\underline{\psi}|\underline{\varphi})f(\underline{\varphi})
\end{aligned}$$

De même, lorsque $\underline{\psi} \in K$ et $r(\underline{\varphi}, \underline{x}) \leq 1$ :

$$\begin{aligned}
q(\underline{\varphi})b(\underline{\varphi}, \underline{x})A(\underline{\psi}|\underline{\varphi})f(\underline{\varphi}) &= q(\underline{\varphi})b(\underline{\varphi}, \underline{x})r(\underline{\varphi}, \underline{x})f(\underline{\varphi}) \\
&= q(\underline{\varphi})b(\underline{\varphi}, \underline{x}) \frac{f(\underline{\psi})(1 - q(\underline{\psi}))d(\underline{\varphi}, \underline{x})}{f(\underline{\varphi})q(\underline{\varphi})b(\underline{\varphi}, \underline{x})} f(\underline{\varphi}) \\
&= f(\underline{\psi})(1 - q(\underline{\psi}))d(\underline{\varphi}, \underline{x}) \\
&= f(\underline{\psi})(1 - q(\underline{\psi}))A(\underline{\varphi}|\underline{\psi})(d(\underline{\varphi}, \underline{x})).
\end{aligned}$$

.

La chaîne $X_i$, $i = 1, 2, \dots$ est donc $f$-réversible.

■

**Proposition 25** *La chaîne de Markov $X_1, X_2, \dots$ générée par l'algorithme de la page 160 est $f$-irréductible.*

**Preuve**

Définissons $\underline{\Psi}$ la mesure de probabilité telle que $\underline{\Psi}(F) = \mathbb{1}_F(\emptyset)\ \forall F \in K$.

Soit $\underline{\varphi} \in K_n\ n \geq 1$. Soit $\underline{y} \in \underline{\varphi}$. Nous avons :

$$(1 - q(\underline{\varphi}))d(\underline{\varphi}\backslash\underline{y}, \underline{y}) > 0$$

et

$$f(\underline{\varphi}\backslash\underline{y}q(\underline{\varphi}\backslash\underline{y})b(\underline{\varphi}, \underline{y}) > 0 \quad \text{car } f \text{ est héréditaire.}$$

Ainsi :

$$\begin{aligned} Q(\underline{\varphi}\backslash\underline{y}|\underline{\varphi}) &= (1 - q(\underline{\varphi}))d(\underline{\varphi}\backslash\underline{y}, \underline{y})A(\underline{\varphi}\backslash\underline{y}|\underline{\varphi}) \\ &\geq (1 - q(\underline{\varphi}))d(\underline{\varphi}\backslash\underline{y}, \underline{y})\frac{f(\underline{\varphi}\backslash\underline{y})q(\underline{\varphi}\backslash\underline{y})b(\underline{\varphi}\backslash\underline{y}, \underline{y})}{f(\underline{\varphi})(1 - q(\underline{\varphi}))d(\underline{\varphi}\backslash\underline{y}, \underline{y})} \\ &\geq f(\underline{\varphi}\backslash\underline{y})q(\underline{\varphi}\backslash\underline{y}b(\underline{\varphi}, \underline{y})\frac{1}{f(\underline{\varphi})} \\ &> 0. \end{aligned}$$

Nous notons $P^k(\underline{\psi}|\underline{\varphi}) = \mathbb{P}(X_{m+k} = \underline{\psi}|X_m = \underline{\varphi})\ \forall m \in \mathbb{N}$ avec $P^1(\underline{\psi}|\underline{\varphi}) = P(\underline{\psi}|\underline{\varphi})$.

Par conséquent, si $n = 1$ :

$$P(\emptyset|\underline{\varphi}) > 0 \quad \forall m \in \mathbb{N}.$$

De plus, si $n > 1$ :

$$P^n(\emptyset|\underline{\varphi}) \geq P(\{\underline{\varphi}\backslash\underline{y}\}|\underline{\varphi})P^{n-1}(\emptyset|\{\underline{\varphi}\backslash\underline{y}\}).$$

Par induction sur $n$, nous obtenons donc que $\forall n > 0$ :

$$P^n(\emptyset|\underline{\varphi}) > 0.$$

Nous obtenons donc que $\forall F \in \Omega$ tel que $\underline{\Psi}(F) > 0$ et $\forall \underline{\varphi} \in K_n$ avec $n > 0$ :

$$\exists m \in \mathbb{N} \text{ tel que } P^m(F, \underline{\varphi}) \geq P^m(\emptyset, \underline{\varphi}) > 0.$$

Si à présent $\underline{\varphi} = \emptyset$ :

$$\begin{aligned} P(F, \emptyset) &\geq P(\{\emptyset\}, \emptyset) \\ &= \left\{ q(\emptyset) \int_T \int_M b(\emptyset, \underline{x})(1 - A(\underline{x}|\emptyset))p(d\tau_x)\varrho(dx) + (1 - q(\emptyset)) \right\} \\ &\geq (1 - q(\emptyset)) \\ &> 0. \end{aligned}$$

Nous en déduisons donc que la chaîne de Markov $X_1, X_2, \ldots$ générée par l'algorithme de la page 160 est $\underline{\Psi}$-irréductible.

D'après la proposition 18, et en se rappelant que $X_1, X_2, \ldots$ est $f$-invariante, nous obtenons que $X_1, X_2, \ldots$ est $f$-irréductible.

■

**Proposition 26** *La chaîne de Markov $X_1, X_2, ...$ générée par l'algorithme de la page 160 est apériodique.*

**Preuve**

D'après la démonstration précédente, nous avons que $P(\{\emptyset\}, \emptyset) > 0$. Or $\emptyset$ est un *petit* ensemble, donc nous en déduisons que $X_1, X_2, \ldots$ est apériodique.

■

**Proposition 27** *La chaîne de Markov $X_1, X_2, ...$ générée par l'algorithme de la page 160 est Harris-récurrente.*

**preuve**

Cette démonstration est largement inspirée du corollaire 2 de Tierney ([Tierney, 1994], section 3.1). Nous allons donner les ingrédients principaux de cette preuve. Le lecteur peut se référer à [Tierney, 1994] pour de plus amples détails.

Quelques définitions sont nécessaires pour cette démonstration

**Définition 25** *Une fonction mesurable $h$ est harmonique pour la chaîne $X_1, X_2, \ldots$ si :*

$$\mathbb{E}[h(X_{n+1}|X_n = x_n)] = h(x_n)$$

**Proposition 28** *Si $X_1, X_2, \ldots$ est une chaîne de Markov récurrente, alors, $X_1, X_2, \ldots$ est Harris récurrente si et seulement si toutes les fonctions harmoniques bornées de $X_1, X_2, \ldots$ sont constantes.*

Nous pouvons donc utiliser cette propriété pour étudier notre chaîne de Markov. Comme $X_1, X_2, \ldots$ est $f$-irréductible et $f$-invariante alors $X_1, X_2, \ldots$ est récurrente. Soit $h$ une fonction harmonique bornée pour $X_1, X_2, \ldots$. Donc d'après Nummelin [Nummelin, 1984], $h = f.h$ $f$ presque partout.

Supposons que $\underline{\varphi} \in K$, nous obtenons alors, d'après Tierney [Tierney, 1994], que :

$$(1 - r(\underline{\varphi}))(h(\underline{\varphi}) - \pi) = 0,$$

où $r(\underline{\varphi})$ est la probabilité que la chaîne commence en $\underline{\varphi}$ et revienne en $\underline{\varphi}$.

Puisque $\pi$ n'est pas concentrée en un point, la $\pi$-irreductibilité implique que $r(\underline{\varphi}) < 1$, pour tout $\underline{\varphi}$. Ainsi, pour tout $\underline{\varphi} \in K$, $h(\underline{\varphi}) = \pi h$. Nous pouvons en déduisons donc que $h \equiv \pi h$, ce qui équivaut à $h$ constante.

Nous obtenons donc que la chaîne est Harris-Récurrente.

■

En combinant les deux proprositions précédentes (Proposition 26 et 27), nous obtenons que la chaîne de Markov, générée par l'algorithme proposé à la page 160 est Harris-récurrente et apériodique. Or par définition, une chaîne de Markov Harris-récurrente et apériodique est ergodique. Nous en déduisons donc que la chaîne de Markov étudiée est ergodique, ce qui garantit la convergence de l'algorithme de simulation.

Nous allons à présent étudier un modèle particulier de la classe CC. Ce modèle constitue l'extension continue du modèle de Graner et Glazier.

## 3.9 Exemple d'un modèle de la classe CC

Dans cette section, nous allons nous intéresser à un modèle particulier de la classe CC. Ce modèle correspond à l'extension continue du modèle de Potts étendu proposé par Graner et Glazier. Dans un premier temps, nous allons préciser les fonctions $g$, $J$ et $h$ du modèle présenté, et nous allons montrer que ces fonctions vérifient les conditions d'existence de ce modèle. Puis, nous présenterons certaines simulations de ce modèle expliquant les phénomènes biologiques tels que les configurations en damier, les configurations en agrégats et l'engloutissement d'un tissu par un autre.

### 3.9.1 Définition et propriété du modèle

En s'inspirant du modèle de Graner et Glazier, nous définissons un processus ponctuel de Gibbs de la classe CC défini sur $S = T \times M$ où :

$$\begin{aligned} T &= \mathbb{R}^2, \\ M &= \{\tau_1, \tau_2, \tau_E\}. \end{aligned}$$

L'espace des marques est donc composé de trois éléments correspondant à trois types cellulaires précis. Deux de ces types, $\tau_1$ et $\tau_2$ représentent des types de cellules actives tandis que $\tau_E$ caractérise le milieu extracellulaire. Ainsi, les cellules de type $\tau_E$ ne sont pas sujettes à une contrainte de forme.

En s'inspirant, à nouveau, du modèle de Potts étendu, nous pouvons définir la fonction $g$, quantifiant la pondération des interactions par l'arête de Dirichlet, et la fonction $h$, décrivant la contrainte de forme sur les cellules de Dirichlet de la manière suivante :

$$g(x) = x\mathbb{1}_{[0,K_g[}(x) \quad \text{pour tout } x \in \mathbb{R}^2, \text{ avec } K_g > 0$$

et

$$h(\mathrm{Dir}_\varphi(x), \tau_x) = (|\mathrm{Dir}_\varphi(x)| - A_{\tau_x})^2 \Gamma(\tau_x) \mathbb{1}_{[0,\sqrt{K_h}[}(|\mathrm{Dir}_\varphi(x)|), \text{ avec } K_h > 0$$

où, $\Gamma$ est la fonction de Heavyside, $0 < A_{\tau_x} < \infty$, pour tout $\tau \in M$ et $K_h > 0$. Supposons enfin que $-\infty < J(\tau, \tau') < +\infty$ pour tout couple $(\tau, \tau') \in M^2$.

Ainsi, nous allons étudier le processus ponctuel marqué de Gibbs, défini par la fonction d'énergie suivante :

$$\begin{aligned} H_{CC}(\underline{\varphi}) &= \sum_{x \sim_\varphi y} |\mathrm{Dir}_\varphi(x) \cap \mathrm{Dir}_\varphi(y)| \mathbb{1}_{|\mathrm{Dir}_\varphi(x) \cap \mathrm{Dir}_\varphi(y)| < K_g} J(\tau_x, \tau_y) \\ &+ \sum_{x \in \varphi} (|\mathrm{Dir}(x)| - A_{\tau_x})^2 \Gamma(\tau_x) \mathbb{1}_{|\mathrm{Dir}(x)| < \sqrt{K_h}}, \end{aligned}$$

où $\underline{\varphi} = \{(x, \tau_x) : x \in \varphi\}$ est une configuration marquée.

Nous allons à présent montrer que ces deux fonctions $g$ et $h$ satisfont aux hypothèses du théorème 4 page 157, garantissant l'existence de mesures de probabilité.

Nous remarquons aisément que La fonction $g$ est bornée par $K_g$. De même la fonction $h$ est bornée par $K_h$ Nous en déduisons donc que le modèle, ainsi défini, satisfait

aux hypothèses du théorème 4, donc il existe une mesure de probabilité satisfaisant les spécifications locales formulées par $H_{CC}$.

D'après l'étude menée à la section 3.8, nous en déduisons que la chaîne de Markov, générée par l'algorithme proposé à la page 160, est ergodique. Nous allons pouvoir étudier des simulations obtenues en proposant des jeux de paramètres caractéristiques pour la fonction $J$. Ces paramètres seront appelés les paramètres d'interaction.

### 3.9.2 Exemples de simulations du modèle

Dans cette section, nous allons proposer trois types de simulations qui correspondent à des situations biologiques précises : configurations en damier, en agrégats et l'engloutissement d'un tissu par un autre. Pour chacune de ces situations, les coefficients de contraintes de forme ont été choisis de la manière suivante :

$$\begin{aligned} A_{\tau_1} &= 5.10^{-2}, \\ A_{\tau_2} &= 5.10^{-2}, \\ A_{\tau_E} &= -1. \end{aligned}$$

Ce choix de coefficients de contrainte de forme nous permet d'évaluer le nombre de cellules actives dans le cercle unité à environ 630 cellules à l'équilibre.

Pour les trois situations, la configuration initiale est identique. Cette configuration initiale a été obtenue en répartissant aléatoirement 1000 points dans le cercle unité. Pour chacun des ces 1000 points, une marque choisie dans l'espace $\{\tau_1, \tau_2\}$ est attachée, de sorte que ces points modélisent des cellules actives. Ces 1000 cellules sont entourées par des cellules de types $\tau_E$, modélisant le milieu extracellulaire. La configuration initiale est décrite dans la figure 3.20.

Afin de caractériser les trois types de configurations (Damier, Agrégats et Engloutissement), nous allons introduire le coefficient de tension de surface, $\gamma_{12}$ entre les types de cellules actives, donné par [Glazier et Graner, 1993] :

$$\gamma_{12} = J(\tau_1, \tau_2) - \frac{J(\tau_1, \tau_1) + J(\tau_2, \tau_2)}{2}.$$

Une configuration en damier est caractérisée par un coefficient de tension de surface négatif, tandis qu'une configuration en agrégats est caractérisée par un coefficient de

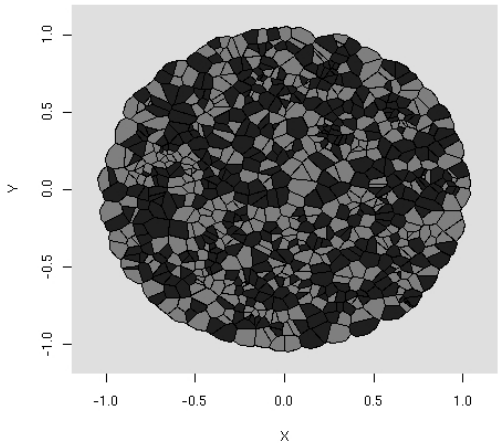



**Fig. 3.20:** *Configuration initiale.*

tension de surface positif.

Nous allons à présent présenter des exemples de simulation obtenus pour chacune des trois configurations caractéristiques.

### Configuration en Damier

Les configurations de cellules biologiques en damier ont été particulièrement étudiées dans des travaux portant sur la transformation de l'épithélium au cours de la maturation sexuelle [Honda et al., 1986].

Pour une configuration en damier, nous choisissons comme jeu de paramètres d'interaction le jeu suivant :

$$\begin{aligned} J(\tau_1,\tau_1) &= 1, \\ J(\tau_2,\tau_2) &= 1, \\ J(\tau_1,\tau_2) &= J(\tau_2,\tau_1) = 0, \\ J(\tau_1,\tau_E) &= J(\tau_E,\tau_1) = 0, \\ J(\tau_2,\tau_E) &= J(\tau_E,\tau_2) = 0, \\ J(\tau_E,\tau_E) &= 0. \end{aligned}$$

Ce jeu de paramètres conduit à un coefficient de tension de surface $\gamma_{12} = -1$, caractéristique des configurations en damier. De plus, ces paramètres peuvent s'interpréter énergétiquement. Les valeurs $J(\tau_1, \tau_1) = 1$ et $J(\tau_2, \tau_2) = 1$ signifient que si deux cellules actives de même type sont voisines, le coup énergétique est de 1, pondéré par la longueur de l'arête commune aux cellules. Les autres valeurs, égales à zéro, nous indiquent que le coup énergétique pour ce type de voisinage est nul. Ainsi, étant donné que la densité de notre processus est inversement proportionnelle à l'energie, le processus va donc favoriser des configurations ayant des cellules voisines de types actifs différents. Les configurations à forte densité ressembleront donc à des damiers. Un exemple de données simulées est décrit dans la figure 3.21.

**Configuration en Agrégats**

Les configurations en agrégats ont été observées expérimentalement dans de nombreuses situations. Un des exemples les plus célèbres est l'arrangement spontané en agrégat de cellules homotypiques chez l'organisme vivant appelé Hydra [Gierer et al., 1972]. D'autres observations expérimentales peuvent être trouvées dans les deux articles suivants [Armstrong, 1989] et [Takeuchi et al., 1988].

Pour une configuration en damier, nous choisissons comme jeu de paramètres d'interaction le jeu suivant :

$$\begin{aligned} J(\tau_1, \tau_1) &= 0, \\ J(\tau_2, \tau_2) &= 0, \\ J(\tau_1, \tau_2) &= J(\tau_2, \tau_1) = 1, \\ J(\tau_1, \tau_E) &= J(\tau_E, \tau_1) = 0, \\ J(\tau_2, \tau_E) &= J(\tau_E, \tau_2) = 0, \\ J(\tau_E, \tau_E) &= 0. \end{aligned}$$

Ce jeu de paramètres conduit à un coefficient de tension de surface $\gamma_{12} = 1$. Ce coefficient positif est caractéristique des configurations en agrégats. De manière analogue aux configurations en damier, les paramètres $J(\tau, \tau')$ peuvent s'interpréter énergétiquement. En effet ce jeu de paramètres défavorise des configurations ayant des cellules voisines de

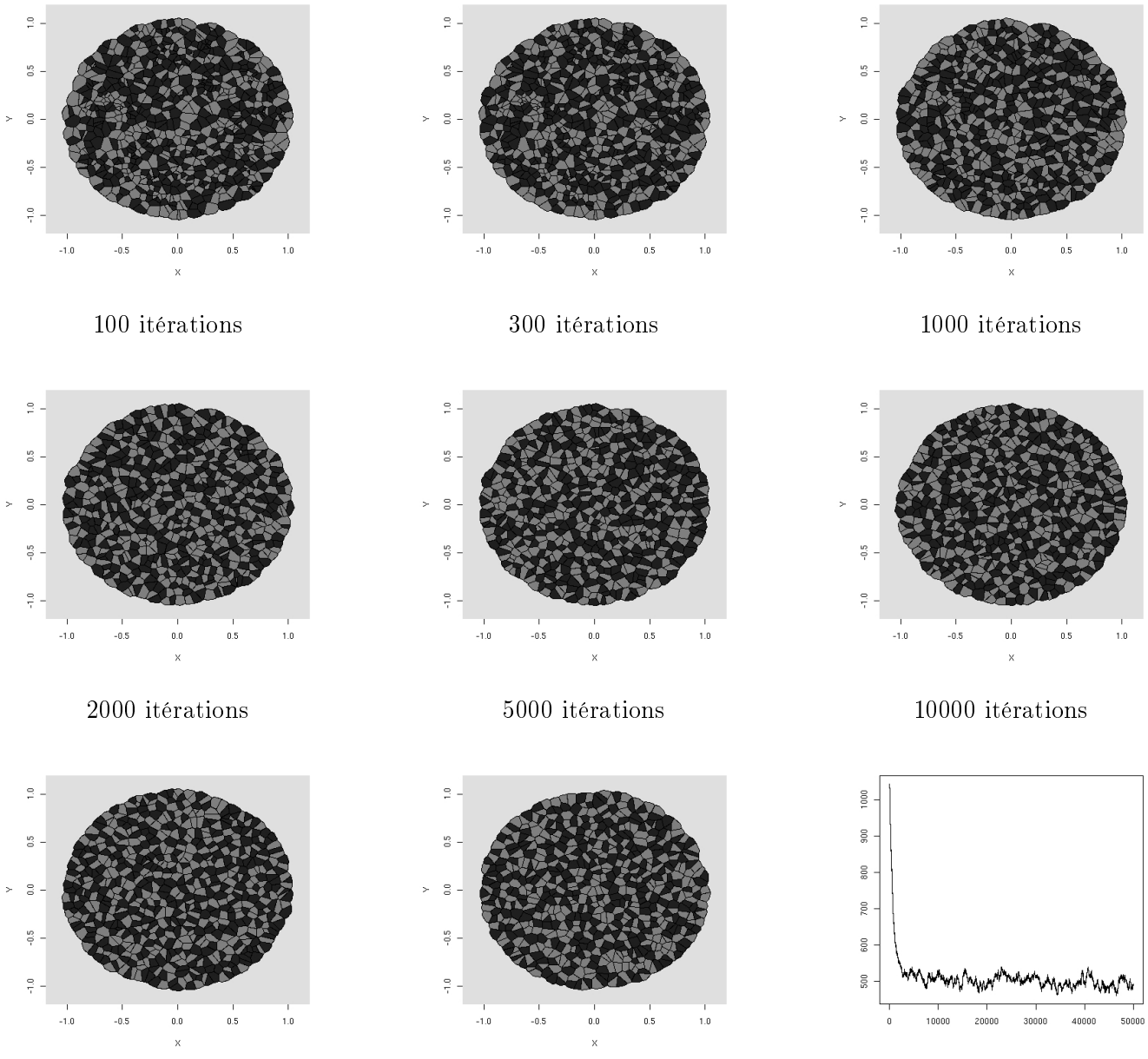



**Fig. 3.21:** *Simulations de configurations en Damier*

types actifs différents pour favoriser des cellules voisines de même type. Ainsi, les configurations à forte densité feront apparaître des agrégats (ou clusters) de cellules de même type. Un résultat de simulation est proposé dans la figure 3.22.

**Engloutissement d'un tissu par un autre**

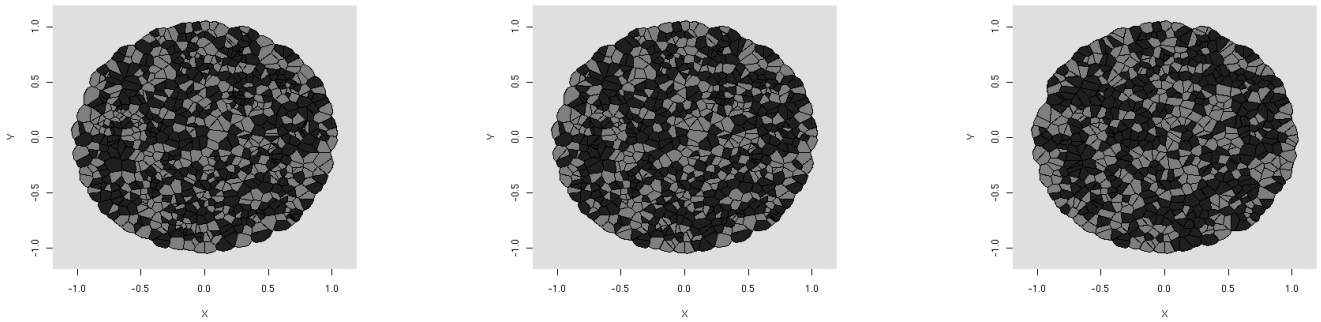
L'engloutissement d'un tissu par un autre tissu est un phénomère biologique largement étudié dans les deux articles suivants [Armstrong, 1989] et [Foty et al., 1996]. Lorsque deux tissus sont en configurations d'agrégats, et que l'un des deux tissus possède une affinité plus forte avec le milieu extracellulaire que l'autre tissu. Le premier tissu a tendance à entourer le second.

Ce phénomène peut être obtenu à l'aide du jeu de paramètres d'interaction suivant :

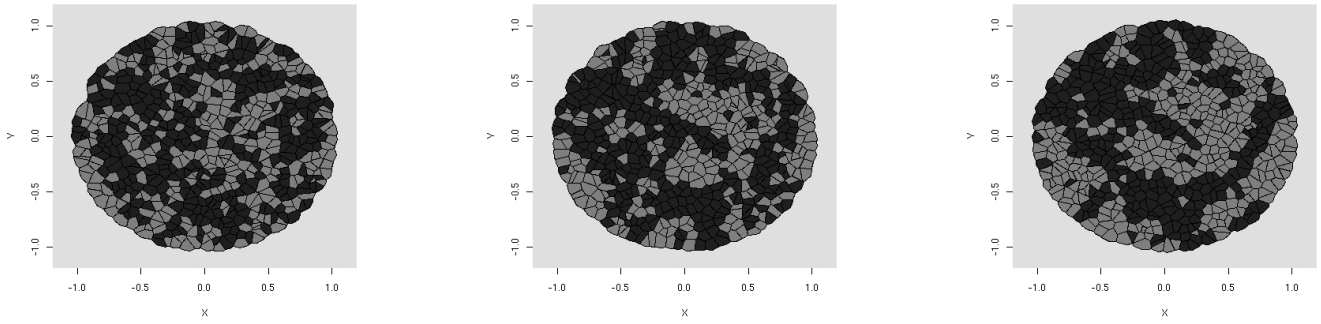
$$\begin{aligned} J(\tau_1,\tau_1) &= 0, \\ J(\tau_2,\tau_2) &= 0, \\ J(\tau_1,\tau_2) &= J(\tau_2,\tau_1) = 1, \\ J(\tau_1,\tau_E) &= J(\tau_E,\tau_1) = 0, \\ J(\tau_2,\tau_E) &= J(\tau_E,\tau_2) = 1, \\ J(\tau_E,\tau_E) &= 0. \end{aligned}$$

L'interprétation énergétique de ces paramètres est la même que pour les configurations en agrégats. Les paramètres $J(\tau_1,\tau_E) = 0$ et $J(\tau_2,\tau_E) = 1$, modélisent le fait que les cellules de type $\tau_1$ ont une affinité plus forte avec le milieu extracellulaire par rapport aux cellules de type $\tau_2$. Ainsi, pour ce jeu de paramètres, les configurations de forte densité vont avoir tendance minimiser les contacts entre les cellules de type $\tau_2$ et le milieu extracellulaire. Ainsi, les cellules de type $\tau_1$ vont « entourer » les cellules de type $\tau_2$ afin de minimiser ces contacts et donc matérialiser un engloutissement. La figure 3.23 représente une simulation du phénomène d'engloutissement à l'aide du jeu de paramètres décrit ci-dessus.

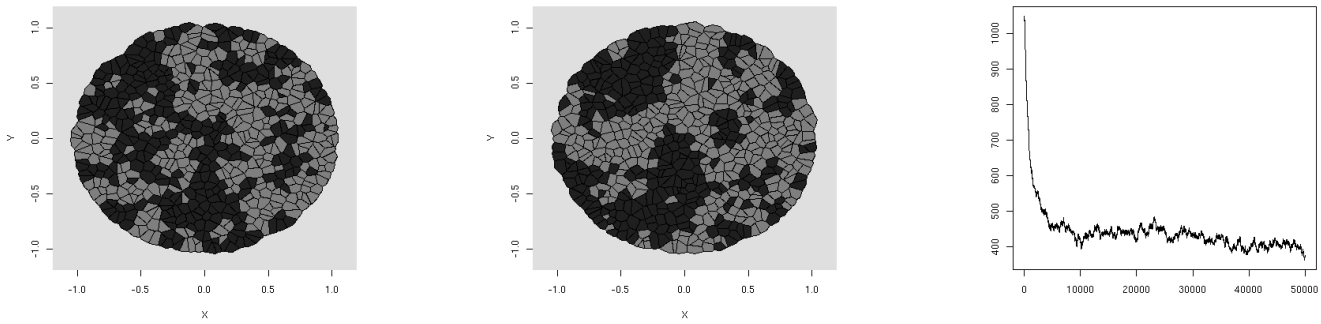
L'ensemble des résultats présentés ci-dessus (damier, agrégats et engloutissement) ont été obtenus à $\theta$ fixé : $\theta = 10$. Cependant, ce paramètre, inversement proportionnel dans



100 itérations 300 itérations 1000 itérations



2000 itérations 5000 itérations 10000 itérations



20000 itérations 50000 itérations

**Fig. 3.22:** *Simulations de configurations en Agrégats*

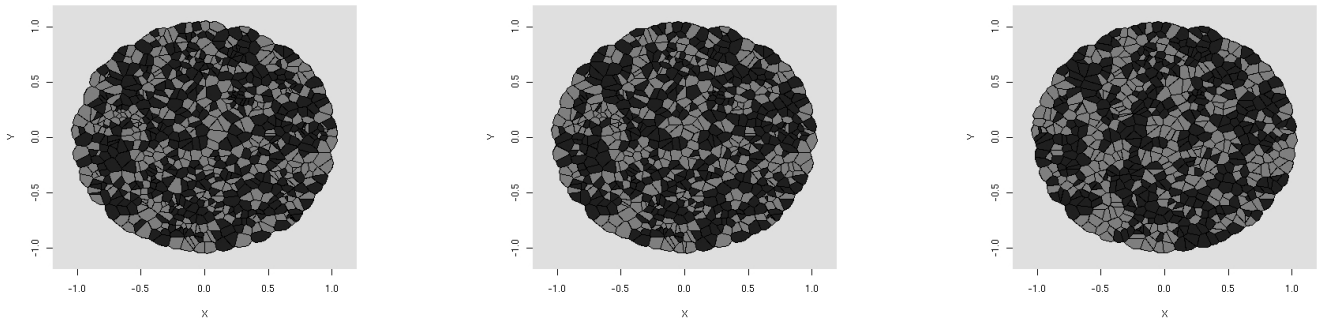
le modèle d'Ising et le modèle de Potts, joue un rôle fondamental dans la modélisation. La section suivante nous montre, sur des simulations, l'influence de ce paramètre pour l'allure des configurations obtenues. Cette étude par simulation a été menée sur des modèles de type « damier » et « agrégats » uniquement.

### 3.9.3 Influence du paramètre $\theta$

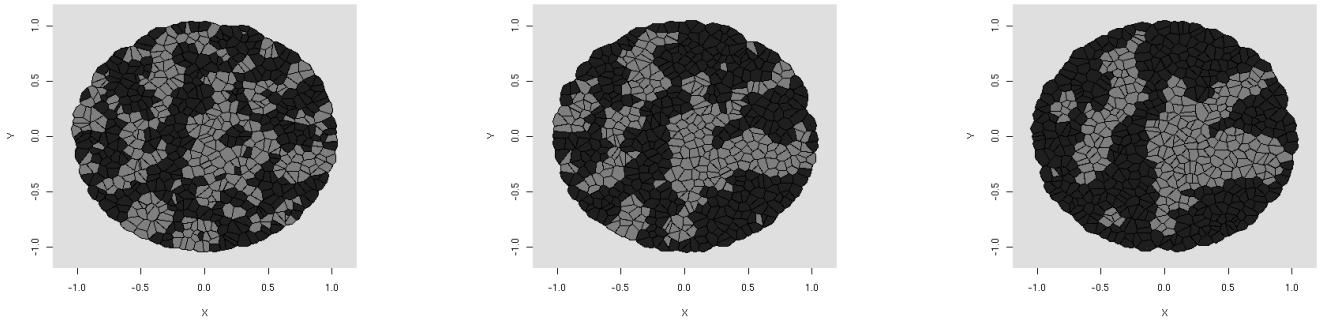
Pour étudier, par simulations, l'influence du paramètre $\theta$ sur le modèle, nous avons utilisé les paramètres d'interaction présentés lors des sections précédentes pour le cas du damier (section 3.9.2) et le cas agrégats (section 3.9.2). Nous avons alors simulé ces modèles pour les valeurs de $\theta$ suivantes : $\theta = 1, 5,$ et 10. Les configurations, présentées à la figure 3.24, sont le résultat de 50000 itérations de l'algorithme de Métropolis initialisées par la configuration de la figure 3.20.

Les simulations obtenues à la figure 3.24 nous montrent que pour des valeurs de $\theta$ faibles ($\theta = 1$ par exemple), les marques restent réparties presque aléatoirement. Il est donc très difficile de distinguer les configurations de type damier des configurations de type agrégats. Dès que $\theta$ augmente ($\theta = 5$), des regroupements homotypiques se forment pour la configuration de type agrégats. Parallèlement, nous pouvons remarquer visuellement que le nombre de relations de voisinage entre cellules de même type diminue sensiblement. Enfin, lorsque $\theta$ atteint la valeur de 10, nous remarquons que la configuration de type damier s'apparente à un échiquier tandis de larges « clusters » se sont formés pour la configuration de type agrégats.

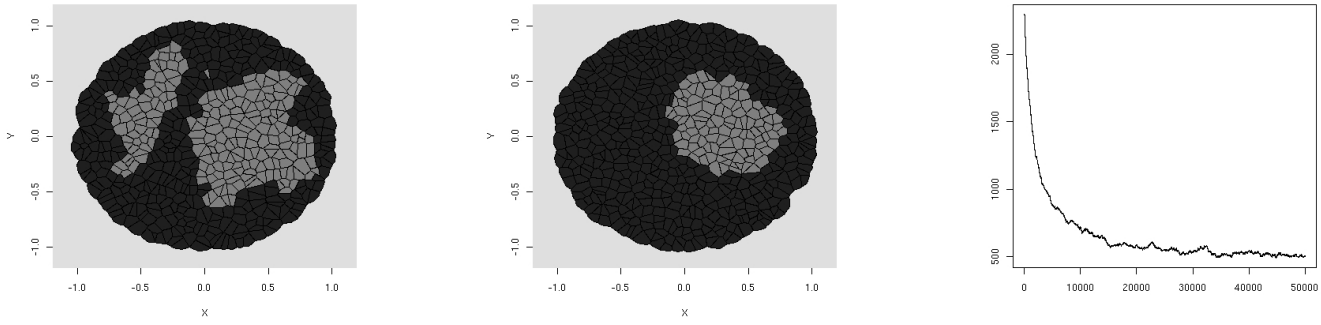
Ces résultats montrent l'importance du paramètre $\theta$ dans le modèle. En effet, ce paramètre détermine l'intensité donnée à la fonction d'énergie et peut être interprété comme l'écart entre le modèle CC et le processus de Poisson (processus de référence). Cette étude par simulations nous permet de comprendre que l'estimation du paramètre $\theta$ peut être utile pour discriminer les tissus cancéreux. Cette estimation pourrait permettre de détecter une modifications des propriétés adhérentes de façon précoce.



100 itérations 300 itérations 1000 itérations



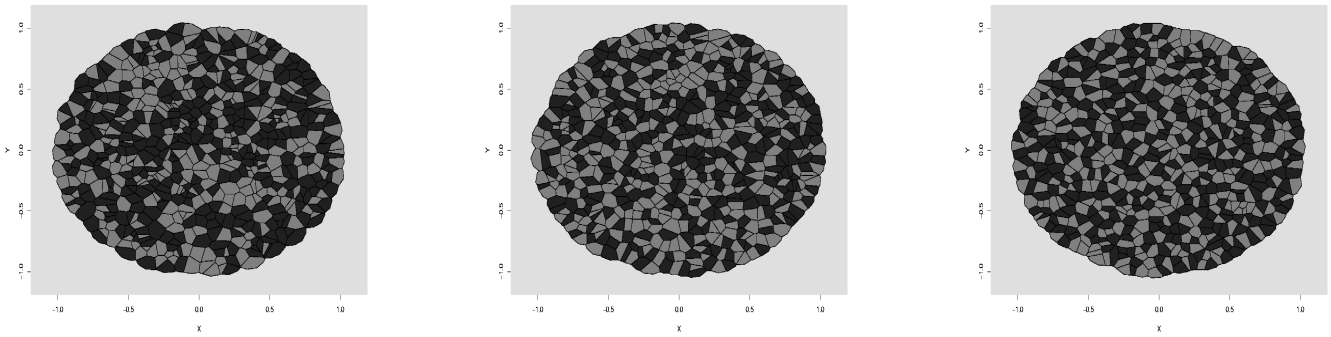
2000 itérations 5000 itérations 10000 itérations



20000 itérations 50000 itérations

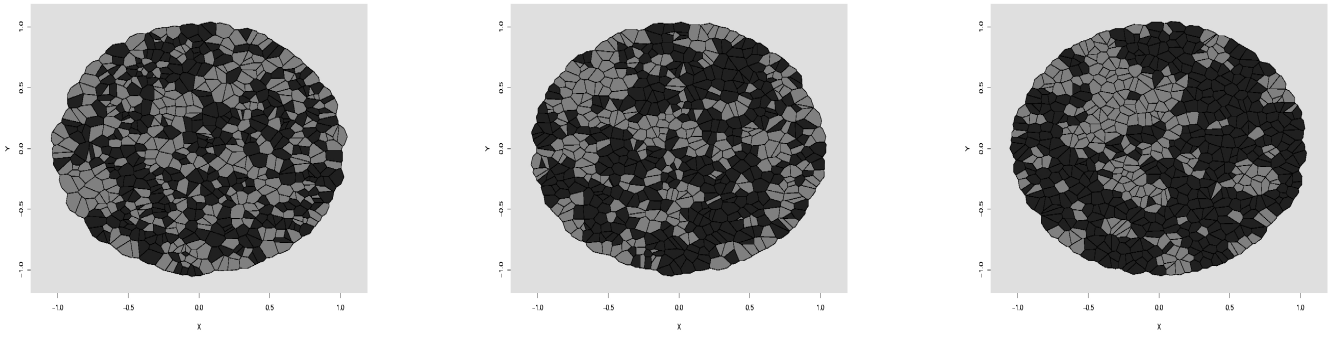
**Fig. 3.23:** *Simulations de l'engloutissement d'un tissu par un autre*

Configurations en Damier



$\theta = 1$ $\theta = 5$ $\theta = 10$

Configurations en Agrégats



$\theta = 1$ $\theta = 5$ $\theta = 10$

**Fig. 3.24:** *Influence du paramètre $\theta$ sur des simulations de type Damier et de type Agrégat. Nous présentons ici les configurations finales utilisant des valeurs pour $\theta$ : $\theta = 1, 5, 10$.*

## 3.10 Estimation du paramètre d'adhésion $\theta$

Comme nous l'avons déjà vu pour l'algorithme de simulation, les densités des processus de la classe CC sont incalculables. Ainsi, la vraisemblance du modèle proposée à la section précédente est incalculable. Dans cette section nous proposons un estimateur du paramètre d'adhésion $\theta$ en utilisant des techniques de pseudo-vraisemblance. Ces techniques ont été introduites par Besag pour l'analyse d'images bruitées [Besag, 1975]. Puis en 1982, la pseudo-vraisemblance a été introduite dans l'étude de processus d'interaction [Besag et al., 1982]. Dans ces travaux pionniers, l'étude s'est concentrée sur des processus définis sur des grilles régulières. Dans cette section nous allons proposer deux estimateurs construits sur le concept de la pseudo-vraisemblance, mais qui s'appliquent premièrement à un processus pour lequel les points sont aléatoires mais fixes et deuxièmement à un processus ponctuel de Gibbs. D'autres techniques d'estimation ont été développées dans le cadre des processus ponctuels. Nous pouvons citer ici des techniques basées sur la maximisation de la vraisemblance par Chaîne de Markov Monte Carlo [Geyer et Thompson, 1992] ou la technique de Takacs-Fiksel introduite dans les années 80 [Takacs, 1986] [Fiksel, 1988]. Ces techniques ne sont pas l'objet de notre étude, cependant, le lecteur peut se référer à une récente revue d'ensemble du problème de l'inférence statistique pour les processus ponctuels [Møller et Waagepetersen, 2003].

### 3.10.1 Définition d'estimateurs de $\theta$

Considérons une configuration finie marquée, $\varphi$ :

$$\underline{\varphi} = \{(x, \tau_x) : x \in \varphi\}.$$

De plus, nous notons la configuration des points par :

$$\varphi = \{x_1, \ldots, x_n\},$$

et la réalisation des marques par :

$$\tau_x = \{\tau_{x_1}, \ldots, \tau_{x_n}\}.$$

Nous allons proposer deux estimateurs du paramètre d'adhésion $\theta$, construits par pseudo-vraisemblance. Le premier estimateur, $\widetilde{\theta}$, suppose que la configuration est fixe, le

second estimateur, $\widehat{\theta}$, s'inscrit dans un modèle de processus ponctuels et considère que le nombre de points est aléatoire. Le premier estimateur se situe dans une philosophie de modèle de Potts. Nous considérons que les marques sont conditionnelles à la réalisation des points et n'ont aucun influence sur la configuration non marquée. Nous avons fait référence à ce type de modélisation dans la remarque qui clôt la section 3.4. Le deuxième estimateur est, quant à lui, fidèle à la classe de modèle CC que nous avons proposé dans les sections précédentes.

**Définition de $\widetilde{\theta}$**

Dans un premier temps, nous allons nous intéresser à la pseudo-vraisemblance calculée à partir des marques, en considérant que la configuration spatiale est fixe. En notant que l'espace des marques est $M$, la pseudo-vraisemblance du modèle s'écrit comme le produit des probabilités conditionnelles. Ce produit se fait sur l'ensemble des points marqués de $\underline{\varphi}$. La pseudo-vraisemblance, $\widetilde{\mathrm{PL}}(\theta)$, s'écrit alors de la manière suivante :

$$\widetilde{\mathrm{PL}}(\theta) = \prod_{\tau_{x_i} \in \tau_x} \mathbb{P}(\tau_{x_i} | \varphi, \tau_x \backslash \tau_{x_i}, \theta).$$

Dans la formule ci-dessus, la probabilité conditionnelle d'observer $\tau_{x_i}$ au point $x_i$, sachant la configuration des points $\varphi$ et la réalisation des marques sur l'ensemble des sites sauf $x_i$ est décrite par :

$$\mathbb{P}(\tau_{x_i} | \varphi, \tau_x \backslash \tau_{x_i}, \theta) = \frac{\exp(-\theta H_{CC}^{\varphi}(\tau_{x_i}))}{\sum_{\tau \in M} \exp(-\theta H_{CC}^{\varphi}(\tau))}.$$

Dans cette formule, $H_{CC}^{\varphi}(\tau_i)$ correspond à la contribution de la marque du point $x_i$ dans l'expression de la fonction Hamiltonienne $H_{CC}$, i.e. :

$$H_{CC}^{\varphi}(\tau_{x_i}) = \sum_{x \sim x_i} |\mathrm{Dir}(x_i) \cap \mathrm{Dir}(x)| J(\tau_x, \tau_{x_i}) + (|\mathrm{Dir}(x_i)| - A_{\tau_{x_i}})^2 \Gamma(A_{\tau_{x_i}}).$$

Nous pouvons à présent définir un estimateur du paramètre d'adhésion $\theta$ à partir de l'expression de la pseudo-vraisemblance.

**Définition 26** *Soit $\underline{\varphi} = \{(x, \tau_x) : x \in \varphi\}$ une configuration finie marquée, telle que $\varphi = \{x_1, \ldots, x_n\}$ est la configuration des points et $\tau = \{\tau_{x_i}, \ldots, \tau_{x_n}\}$. Sous le modèle*

*de Gibbs défini par une fonction d'énergie $H_{CC}$, conditionnellement à la réalisation des points, le logarithme de la pseudo-vraisemblance $\widetilde{PL}$ s'écrit :*

$$\widetilde{\mathrm{L}}(\theta) = -\sum_{\tau_{x_i}\in\tau}\left(\theta H_{CC}^{\varphi}(\tau_{x_i}) + \log\sum_{\tau\in M}\exp(-\theta H_{CC}^{\varphi}(\tau))\right).$$

*Nous définissons $\widetilde{\theta}$, un estimateur de $\theta$, par :*

$$\widetilde{\theta}(\tau|\varphi) = \mathrm{argmax}_{\theta}\widetilde{\mathrm{L}}(\theta),$$

**Définition de $\widehat{\theta}$**

Le second estimateur de nous proposons correspond à l'estimateur de pseudo-vraisemblance du modèle CC. L'introduction de la pseudo-vraisemblance pour l'étude de processus ponctuels de Gibbs a été effectuée par [Jensen et Møller, 1991]. Plus récemment, certains résultats ont été proposé concernant la consistance et la normalité asymptotique de processus ponctuel de Gibbs de type plus proche voisin, invariant par translation [Billiot et al., 2006].

Les travaux de Jensen et Møller ont permis de déterminer l'expression générale de la pesudo-vraisemblance pour un processus ponctuel de Gibbs. En adaptant cette expression à nos notations, et en empruntant certaines notations à [Billiot et al., 2006], la pseudo-vraisemblance, définie pour une configuration marquée $\underline{\varphi} = \{(x,\tau_x) : x \in \varphi\}$ et un domaine d'observation $\Lambda$ est donnée par :

$$PL(\underline{\varphi},\theta) = \exp\left(-\int_{\Lambda}\int_{M}\exp(-H_{CC}(\underline{x}|\underline{\varphi}))d\tau_x dx\right)\prod_{\underline{x}\in\underline{\varphi}_{\Lambda}}\exp\left(-H_{CC}(\underline{x}|\underline{\varphi}\backslash\underline{x})\right),$$

où $H_{CC}(\underline{x}|\underline{\varphi}) = H_{CC}(\underline{\varphi}\cup\underline{x}) - H_{CC}(\underline{\varphi})$, représente l'énergie nécessaire pour intégrer le point marqué $\underline{x}$ dans la configuration $\underline{\varphi}$. Nous pouvons également rappeler que dans nos notations un point marqué se note $\underline{x} = (x,\tau_x)$, où $x$ est la localisation saptiale du point et $\tau_x$ la marque associé à ce point.

En utilisant l'expression ci-dessus, nous pouvons définir un estimateur du paramètre d'adhésion $\theta$.

**Définition 27** *Soit $\underline{\varphi} = \{(x,\tau_x) : x \in \varphi\}$ une configuration telle que $\varphi = \{x_1,\ldots,x_n\}$ est la configuration des points et $\tau = \{\tau_{x_i},\ldots,\tau_{x_n}\}$. Soit $\Lambda$ un domaine d'étude. Sous*

*le modèle de Gibbs défini par une fonction d'énergie $H_{CC}$, le logarithme de la pseudo-vraisemblance $PL$ s'écrit :*

$$\widetilde{\mathrm{LPL}}(\underline{\varphi},\theta) = -\int_\Lambda \int_M \exp(-H_{CC}(\underline{x}|\underline{\varphi}))d\tau_x dx - \sum_{\underline{x}\in\underline{\varphi}_\Lambda} H_{CC}(\underline{x}|\underline{\varphi}\backslash\underline{x},\theta).$$

*Nous définissons $\widehat{\theta}$, un estimateur de $\theta$, par :*

$$\widehat{\theta}(\tau|\varphi) = \mathrm{argmax}_\theta \mathrm{LPL}(\underline{\varphi},\theta),$$

Nous allons à présent étudier, par simulations les performances de ces estimateurs $\widetilde{\theta}$ et $\widehat{\theta}$.

### 3.10.2 Performances des estimateurs de $\theta$

Les performances des estimateurs $\theta$ ont été étudiées sur des simulations utilisant l'algorithme proposé à la page 160. Nous avons effectué 100 simulations pour des valeurs de $\theta$ entre 1 et 20. Ces simulations ont été réalisées en utilisant des paramètres d'interaction de type damier (voir section 3.9.2) et de type agrégats (voir section 3.9.2). Les valeurs de la moyenne empirique et de la variance empirique ont été reportées dans les tableaux 3.1 et 3.2.

Les résultats que nous avons obtenus dans le tableau 3.1 nous montrent que l'estimateur $\widetilde{\theta}$ est très faiblement biaisé pour les deux types de configurations, damier et agrégat. Nous pouvons également remarquer que la variance de $\widetilde{\theta}$ est relativement faible (Var $< 1$) pour des valeurs de $\theta$ faibles ($\theta < 10$). Nous constatons également que la variance augmente sensiblement lorsque $\theta$ augmente pour atteindre, pour $\theta = 20$, la valeur de 3.55 dans le cas damier et 2.98 dans le cas agrégats. Nous constatons enfin que la variance est légèrement plus élevée dans le cas de configurations en damier que de configurations en agrégats.

Le tableau 3.2 récapitule les résultats obtenus pour l'estimateur $\widehat{\theta}$. Nous constatons que les performances sont assez similaires à celles de $\widetilde{\theta}$ en terme de biais. Cependant, la variance de $\widehat{\theta}$ semble légèrement plus élevée. Nous pouvons constater que nous retrouvons le fait que la variance augmente lorsque $\theta$ augmente.

Il est assez surprenant de trouver une variance plus élevé pour $\widehat{\theta}$ que pour $\widetilde{\theta}$. Ce constat semble venir du fait que l'intégrale dans la formule de la pseudo-vraisemblance

| | Damier | | Aggrégats | |
|---|---|---|---|---|
| | Moyenne | Variance | Moyenne | Variance |
| $\theta = 1$ | 0.98 | 0.70 | 1.03 | 0.4 |
| $\theta = 3$ | 3.14 | 0.66 | 3.01 | 0.51 |
| $\theta = 5$ | 5.01 | 0.57 | 4.94 | 0.94 |
| $\theta = 8$ | 8.20 | 1.07 | 8.01 | 0.81 |
| $\theta = 10$ | 10.47 | 1.20 | 9.80 | 1.00 |
| $\theta = 12$ | 12.28 | 1.81 | 12.05 | 1.09 |
| $\theta = 15$ | 14.58 | 2.22 | 15.03 | 1.20 |
| $\theta = 20$ | 20.44 | 3.55 | 20.08 | 2.98 |

**Tab. 3.1: Performances de $\widetilde{\theta}$.** *Les moyennes et variances empiriques ont été calculées à partir de* 100 *simulations pour des valeurs de* $\theta = 1, 3, 5, 8, 10, 12, 15, 20$. *Ces simulations ont utilisé les paramètres d'interaction typique de configurations en Damier et en Agrégats.*

| | Damier | | Aggrégats | |
|---|---|---|---|---|
| | Moyenne | Variance | Moyenne | Variance |
| $\theta = 1$ | 1.01 | 0.91 | 0.97 | 1.3 |
| $\theta = 3$ | 2.92 | 0.82 | 2.99 | 0.94 |
| $\theta = 5$ | 5.17 | 1.12 | 4.93 | 1.05 |
| $\theta = 8$ | 7.83 | 1.43 | 7.91 | 1.17 |
| $\theta = 10$ | 10.24 | 2.24 | 10.30 | 1.38 |
| $\theta = 12$ | 12.37 | 2.81 | 11.88 | 1.85 |
| $\theta = 15$ | 15.43 | 3.87 | 15.58 | 2.55 |
| $\theta = 20$ | 20.22 | 5.29 | 20.18 | 6.65 |

**Tab. 3.2: Performances de $\widehat{\theta}$.** *Les moyennes et variances empiriques ont été calculées à partir de* 100 *simulations pour des valeurs de* $\theta = 1, 3, 5, 8, 10, 12, 15, 20$. *Ces simulations ont utilisé les paramètres d'interaction typique de configurations en Damier et en Agrégats.*

$PL$ est approchée par une somme. En effet, le calcul de cette intégrale pose un problème d'implémentation algorithmique et surtout de temps de calcul. Pour réduire sensiblement la variance, il faut rendre la grille de discrétisation, utilisée pour le calcul de l'intégrale, suffisamment fine. Associé au second terme de la pseudo-vraisemblance, ce calcul nécessite un temps de calcul assez important. Au contraire, le calcul de $\widetilde{\theta}$ est fonction du nombre de points étudiés. En conséquence, le coût algorithmique de ce calcul est moins important et $\widetilde{\theta}$ semble capter suffisamment d'information pour être performant et discriminer entre les configurations.

## 3.11 Conclusion

Dans ce chapitre, nous avons proposé un modèle stochastique d'organisation spatiale d'un tissu biologique. Ce modèle a été construit à partir d'une hypothèse biologique, l'hypothèse DAH, formulée par Steinberg et ses collaborateurs, qui postule qu'un tissu est une configuration de cellules, minimisant une fonction d'énergie associée aux forces d'adhésion entre cellules voisines [Steinberg, 1970]. En étudiant plus en détail les avancées biologiques sur les processus chimiques gérant l'adhésion entre les cellules, nous avons constaté qu'une liaison adhérente se développe telle une fermeture éclair. L'intégration de cette propriété dans l'hypothèse DAH consiste à supposer que l'énergie d'adhésion entre deux cellules, dépend cannoniquement du type des deux cellules voisines et de la longueur de l'adhésion.

Nous avons alors étudié une classe de modèle de processus ponctuels marqués de type Gibbs. Pour ces modèles, nous avons supposé que les informations disponibles sont la localisation spatiale des centres de chacune des cellules et qu'un type a été associé à chaque cellule. En s'appuyant sur des études de Honda, nous avons modélisé géométriquement un tissu par le diagramme de Dirichlet calculé à partir du centre de chaque cellule [Honda, 1978], [Honda, 1983]. Ainsi, chaque cellule biologique est modélisée par sa cellule de Dirichlet. Cette modélisation géométrique nous a permis de définir une topologie de voisinage entre les cellules du tissu. Grâce à ce voisinage, nous avons proposé une fonction d'énergie caractérisée par un potentiel de paire et un potentiel de singleton. Le potentiel de paire mesure l'énergie utilisée par l'ensemble des paires de cellules voisines.

Cette énergie de paire est fonction des types des deux cellules voisines et de longueur d'interaction, modélisée par la longueur de l'arête commune de Dirichlet entre les deux cellules voisines. Le potentiel de singleton permet de contrôler la forme de chacune des cellules, modélisées par les cellules de Dirichlet.

Nous avons ensuite donné certaines conditions pour garantir l'existence des modèles proposés. Ces conditions, essentiellement géométriques, consistent à étudier le processus sur un sous-graphe, le graphe Delaunay$_{\beta_0}$ du graphe de Delaunay. Ce sous-graphe conserve l'ensemble des triangles de Delaunay satisfaisant une contrainte de petit angle. Cette condition a été utilisée pour garantir la stabilité locale de l'Hamiltonien $H_{CC}^{\beta_0}$. De plus, afin d'étudier la quasilocalité, une portée finie a été introduite à la fonction d'énergie. L'intérêt de ces conditions est essentiellement mathématique. En pratique, il est tout à fait raisonnable de supposer que ces conditions sont vérifiées.

Nous avons ensuite proposé un simulateur pour notre classe de modèle. Ce simulateur suit une dynamique de Métropolis-Hastings de type insertion-délétion. L'étude mathématique de la convergence de cet algorithme a été réalisée en couplant certains résultats montrés pour l'existence du processus et certaines études menées notamment par Geyer et Møller [Geyer et Møller, 1994].

La dynamique de Métropolis a déjà été utilisée dans le cadre de l'adhésion cellulaire au début des années 90 [Graner et Glazier, 1992]. Cependant, aucun contrôle mathématique n'avait été formulé : les auteurs ont initialisé leur processus à une configuration déjà satisfaisante. Etant donné que l'algorithme proposé par Graner et Glazier ne s'autorise que de « petits sauts » énergétiques à chaque itération, les résultats obtenus par les auteurs sont interprétables biologiquement et mathématiquement.

Cependant, ce manque de contrôle théorique pose une limite importante au modèle de Graner et Glazier. En effet, depuis plusieurs années, l'adhésion cellulaire a été identifiée comme un facteur important pour de nombreux processus de développement d'un organisme, comme l'embryogénèse ou la morphogénèse. Ces processus de développement sont connus pour leur caractère dynamique. Leur modélisation nécessite la prise en compte de phénomènes cellulaires tels que le cycle cellulaire, l'apoptose ainsi que la prolifération cellulaire [Steinberg, 1996]. L'étude de ces phénomènes morphologiques se fait essentiellement par des simulations intégrant une composante spatiale d'interaction

ou de diffusion couplée, à une composante temporelle. Ainsi, l'étude de ce type de phénomènes nécessite d'une part l'intégration d'une composante temporelle dans le modèle et d'autre part un contrôle mathématique de l'algorithme de simulation. Les modèles de Gibbs spatio-temporel ont déjà été introduits dans différents domaines, l'épidémiologie [Chadoeuf et al., 1992], la segmentation d'images [Kamijo et al., 2001], par exemple. L'extension de notre classe de modèle à des modèles spatio-temporels nécessite donc l'intégrétation de différents phénomènes de régulation cellulaire. Les travaux que nous avons produits sur l'existence de mesure de Gibbs et de convergence de l'algorithme de Métropolis peuvent être adaptés à des modèles de Gibbs spatio-temporels.

L'intégration de notre modèle dans le cadre mathématique des processus de Gibbs nous a permis d'utiliser une technique classique d'estimation de paramètres : la pseudo-vraisemblance. Notre modélisation nous a permis d'isoler le paramètre $\theta$ pour contrôler la force d'adhésion au sein d'un tissu biologique. Nous avons proposé deux procédures d'estimation en supposant tout d'abord le diagramme de Dirichlet fixé et en s'intéressant à la loi de la marque en chaque site puis en s'intéressant à la pseudo-vraisemblance d'un processus ponctuel de Gibbs. La première procédure est largement inspirée de l'estimation dans le modèle de Potts classique. Les résultats que nous avons obtenus nous montrent que l'estimateur proposé est faiblement biaisé. Pour de faibles valeurs de $\theta$, la variance de l'estimateur est également faible. Cependant, nous constatons que la variance augmente sensiblement lorsque $\theta$ augmente.

Pour continuer le parallèle avec le modèle de Potts, dont notre modèle est une extension, nous pouvons rappeler que le paramètre $\theta$, sous le modèle de Potts, est inversement proportionnel à la température du système. Une étude mathématique de ce modèle a montré l'existence d'une transition de phase portant sur le paramètre $\theta$. L'étude de la transition de phase dans notre classe de modèle, ce qui revient à étudier l'unicité d'une mesure de Gibbs vérifiant les spécifications du problème DLR associé, nécessite de plus amples investigations mathématiques. Certaines études ont déjà été menées sur la transition de phase dans des processus ponctuels à géométrie aléatoire [Häggström, 2000], [Bertin et al., 2004]. Ces résultats semblent pouvoir être adaptés à notre classe de modèle.

Pour la seconde procédure d'estimation, nous nous sommes reposés sur les travaux

de [Jensen et Møller, 1991] et plus récemment de [Billiot et al., 2006]. Dans cette dernière étude, les hypothèses de stabilité locale et de quasilocalité sont nécessaires pour obtenir la consistance de l'estimateur tandis que la normalité asymptotique s'appuie sur la quasilocalité de l'énergie locale. Etant donné que nous avons montré sous certaines conditions les propriétés de stabilité locale et de quasilocalité de notre classe de modèle, il est envisageable d'étudier théoriquement la consistance et la normalité asymptotique de l'estimateur $\hat{\theta}$.

Du point de vue des modèles de Gibbs, le principal intérêt de notre approche à été de pondérer l'interaction entre deux cellules par la longueur de l'arête de Dirichlet. L'étude de l'existence de cette classe de modèle, permet donc son utilisation dans un contexte plus large que le contexte de la biologie. En effet, le comportement des forêts est couramment modélisé par un processus ponctuel d'interaction. Ainsi en supposant que la localisation des arbres soit connue et qu'un type puisse être affecté à chaque arbre, notre modèle peut s'appliquer à ce type de problème.

Le modèle que nous avons décrit dans ce chapitre est issu d'un problème biologique concret, le cancer. La grande diversité liée aux processus de cancérisation et tout particulièrement aux cellules cancéreuses rend son étude difficile. Cette diversité se traduit également au niveau de la répartition spatiale des cellules au sein d'un tissu. L'adhésion cellulaire est un composant essentiel de la progression d'une tumeur. En particulier, la progression d'un clone tumoral est associée à une modification des propriétés adhérentes des cellules cancéreuses avec les cellules et le substrat environnants. Cette modification des propriétés adhérentes nécessite une évolution assez lente au début de cette progression. Une détection précoce de cette modification permet d'optimiser le traitement pour le malade. La procédure d'estimation que nous avons mise en place constitue un outil pouvant répondre à une détection des modifications d'adhésion au sein d'un tissu.

La qualité de la procédure d'estimation peut suivre deux axes de développement. Premièrement, la détection de cellules cancéreuses au sein d'un échantillon histologique se fait actuellement par un pathologiste. Cet expert discrimine les cellules en se basant essentiellement sur la forme des cellules. Par conséquent, le modèle que nous avons proposé peut être amélioré en modifiant la contrainte de forme, c'est à dire la fonction $h$ présente dans la seconde somme de l'Hamiltonien $H_{CC}$. Dans un second temps, les progrès dans le

domaine de la biologie moléculaire permettent le développement de nouveaux marqueurs de protéines. La spécificité de ces marqueurs s'est particulièrement accrue, en particulier pour le marquage de molécules d'adhésion. L'amélioration des données peut donc avoir un impact important sur la procédure d'estimation et également sur la classification des tissus.

Enfin, notre modèle peut également aider aux tests de médicaments. En effet, en plus de leurs rôles dans l'adhésion entre cellules, les molécules Cadhérines sont largement impliquées dans plusieurs signaux de régulation intra-cellulaire. Ces signaux ont été récemment caractérisés comme suppresseurs de tumeurs. De plus, de récentes expérimentations biologiques ont montré que l'injection de molécules de Cadhérines dans un tissu permet de modifier *in vitro* les propriétés adhérentes des cellules [Foty et Steinberg, 2004]. Ainsi, certaines recherches thérapeutiques se focalisent sur le développement de médications permettant de modifier les propriétés d'adhésion des cellules cancéreuses. Cette voie thérapeutique inhiberait alors le développement du clone tumoral et son invasion du milieu environnant. Notre modèle pourrait aider à tester les effets de ce type de médicament en évaluant la modification du paramètre d'adhésion et ainsi mesurer l'effet sur le ralentissement de la propagation spatiale d'un cancer.

## Chapitre 4

# Conclusion générale

Ce travail de thèse représente une contribution à la modélisation mathématique de la génèse d'un cancer. Nous avons en effet proposé deux modèles stochastiques construits à partir d'hypothèses biologiques reconnues. Pour chacun des deux modèles, nous avons présenté une procédure statistique pour inférer les paramètres des modèles et proposer une interprétation biologique de ces paramètres.

Bien que le cancer soit connu pour son extrême diversité, certains schémas généraux concernant le développement de cette maladie ont été élaborés. Il est communément admis que le développement d'un cancer suit trois phases principales : l'initiation, la promotion et la progression. Les deux premières phases, l'initiation et la promotion, constituent des modifications intra-cellulaires, tandis que la phase de progression représente le développement extra-cellulaire de la tumeur qui aboutit à l'invasion d'autres organismes par voie sanguine notamment. Au cours de cette thèse nous avons proposé un modèle correspondant à la phase d'initiation d'un cancer et un modèle s'intéressant à la progression spatiale d'une tumeur.

Pour le premier modèle, nous avons considéré que l'initiation d'un cancer est due à une altération génétique de gènes codant pour le contrôle de la fidélité de la réplication de l'ADN et pour la réparation de l'ADN. Cette altération a pour conséquence la perte du système « MisMatch Repair » (MMR), entrainant une hausse brutale du taux de mutation pour l'ensemble de la lignée issue de la cellule ancestrale. Cette théorie biologique

a été développée par Loeb et ses collaborateurs depuis les années 70 [Loeb et al., 1974]. Une vingtaine d'années sont nécessaires pour pouvoir observer phénotypiquement cet événement initiateur (la perte de MMR). Nous avons proposé un modèle stochastique intégrant l'hypothèse de perte de MMR. Ce modèle permet de tester l'occurrence de l'événement perte de MMR à partir d'un échantillon de séquences d'ADN, caractérisant les cellules d'un tissu (pré-)cancéreux. Nous avons développé une approche de coalescence pour modéliser la généalogie de cet echantillon conditionnellement à la perte de MMR à un instant donné de l'histoire de l'échantillon. L'étude théorique du modèle, axée principalement sur la topologie des arbres coalescents, nous a permis de fournir deux estimateurs sans biais du taux de mutation élevé, conséquence de l'occurrence de la perte de MMR. Ces statistiques nous ont ensuite aider à construire des tests pour détecter, l'occurrence ou l'absence de l'événement perte de MMR dans l'échantillon d'étude. Les résultats que nous avons obtenus ont été testés à partir de simulations, produites par un simulateur d'arbre conditionnel coalescent, étudié théoriquement.

Pour le second modèle, décrivant le développement spatial d'une tumeur, nous nous sommes placer sous l'hypothèse DAH, formulée par Steinberg [Steinberg, 1970]. Nous pouvons rappeler que l'hypothèse DAH, postule que l'adhésion entre les cellules, responsable de la configuration spatiale d'un tissu, est fonction des types cellulaires des cellules adhérentes. Nous avons intégré cette hypothèse biologique dans un modèle stochastique de processus ponctuels. Pour prendre en compte de l'hypothèse DAH, nous avons formulé une fonction d'énergie composée d'un potentiel de paire, modélisant les contraintes d'adhésion entre cellules voisines et un potentiel de singleton, décrivant une contrainte de forme sur les cellules. Cette fonction d'énergie, ou Hamiltonien, nous a permis de décrire une classe de processus ponctuels de Gibbs dédiée à l'organisation spatiale de cellules dans un tissu. Par la suite, nous avons proposé un algorithme de simulation inspiré d'une dynamique de Métropolis-Hastings. La convergence de cet algorithme a été étudiée en détail. Puis nous avons proposé une procédure d'estimation pour un paramètre, s'interprétant comme la force d'adhésion du tissu. Cet estimateur a été testé à partir de simulations produites par notre algorithme. Les résultats nous montrent que notre estimateur est faiblement biaisé et permet de discriminer les interactions homotypiques et les interactions hétérotypiques, par exemple.

Les deux modèles proposés dans cette thèse s'intéressent à l'étude de deux phases du développement d'un cancer : l'initiation et la progression. Ce type de modélisation rejoint l'avis de certains spécialistes quant à l'étude des processus de cancérisation [Hanahan et Weinberg, 2000]. En effet, Hanahan et Weinberg affirme que la recherche contre le cancer doit se développer de la manière suivante : la compréhension du cancer doit aujourd'hui essayer de décomposer cette maladie selon des principes fondamentaux sous-jacents. Ces auteurs proposent d'expliquer la croissance d'une tumeur par six modifications physiologiques principales. Ces modifications mettent en jeu des altérations génétiques ou moléculaires. Une cellule, pour devenir cancéreuse, doit donc acquérir les six propriétés suivantes

1. le développement d'un réseau de signalisation de croissance propre à la cellule,
2. le développement d'une insensibilité aux signaux de non-croissance produits par le milieu environnant,
3. la capacité à échapper à l'apoptose,
4. le développement d'un potentiel de croissance illimité,
5. la capacité à soutenir une angiogénèse forte,
6. la capacité à envahir les tissus environnants et à métastaser.

Nous pouvons constater que la perte de MisMatch ne fait pas partie de cette liste de six événements. Les auteurs précisent toutefois que l'acquisition d'un phénotype de mutation ne peut pas être considérée comme un événement indépendant. Au contraire, l'augmentation du taux de mutation doit être vue comme une combinaison des propriétés précédemment citées. Le travail proposé dans cette thèse fait écho à une conclusion formulée dans cet article de synthèse

« We will then be able to apply the tools of mathematical modeling to explain how specific genetic lesions serve to manifest cancer » [Hanahan et Weinberg, 2000].

En marge de cette conclusion, certains travaux ont essayé d'intégrer l'ensemble des six propriétés dans un unique modèle. Un modèle à base d'Equations Différentielles Ordinaires a été récemment proposé [Spencer et al., 2004]. Ce modèle est caractérisé par un schéma d'évolution très complexe : chaque cellule peut se trouver dans 17 états caractéristiques différents. Ainsi, le jeu de paramètres nécessaires pour modéliser chacune

des six propriétés proposées par Hanahan et Weinberg est de taille considérable. L'étude mathématique du système d'Equations Différentielles Ordinaires est rendu difficile par ce surparamétrage. Afin d'étudier la sensibilité du modèle aux paramètres, les auteurs se proposent de faire varier chacun des paramètres indépendamment, en laissant le reste des paramètres fixes. Une première difficulté de cette approche est de fixer correctement les paramètres du modèle. Ce choix est particulièrement délicat dans certains cas, pour lesquels aucune données réelles de référence sont présentes dans la littérature. En conséquence, l'interprétation biologique de cette étude de sensitivité est peu fiable. A cet effet, les auteurs précisent en conclusion que de nombreuses expériences biologiques sont nécessaires pour que leur modèle puisse mimer le développement complet d'un cancer. Il apparaît cependant que l'intégration dans un seul modèle de plusieurs phénomènes interagissant les uns avec les autres rend son interprétation quasi-impossible. Ce constat est renforcé par le fait qu'aucune étude mathématique ne permet le contrôle théorique du modèle et de sa simulation.

D'autre part, un modèle multi-agents, appelé CancerSim, a également été développé pour les dynamiques et interactions proposées par Hanahan et Weinberg [Abbott et al., 2006]. Ce modèle de simulation décrit l'évolution en trois dimensions d'un tissu sous l'influence de ces six propriétés. Les résultats expérimentaux obtenus par le modèle CancerSim, montrent un certain désaccord avec les résultats expérimentaux décrits chez Hanahan et Weinberg. De plus, les auteurs ont des difficultés à expliquer ces désaccords. De nouveau, l'interprétation biologique des résultats est très difficile du fait du grand nombre de paramètres régissant la dynamique du système. Les caractéristiques mathématiques du modèle ne sont pas non plus complètement étudiées et posent la question de la fiabilité des simulations proposées.

Les deux modèles qui viennent d'être présentés, le modèle aux Equations Différentielles Ordinaires et le modèle CancerSim, pose le problème de la modélisation et de la simulation en biologie. Ces modèles, en voulant intégrer le plus de phénomènes biologiques possibles, semblent paradoxalement perdre de leur efficacité. Cette perte d'efficacité peut s'expliquer par le fait que la biologie du cancer reste encore partiellement élucidée. Même si chaque phénomène, pris séparément, est relativement bien compris, les interactions entre ces phénomènes et les conséquences quant au développement tumoral constitue

une limite forte pour ce type d'approche. Nous rejoignons ici l'article de May qui précise que dans de nombreux cas, une modélisation moins ambitieuse donne des résultats plus satisfaisants quant à leur interprétation et aux prédictions fournies [May, 2004]. En effet, dans les deux modèles précédents, aucune hypothèse biologique n'est clairement testée. Par conséquent, ces modèles, essentiellement intéressés par la simulation du développement d'un cancer, restent difficile à utiliser pour vérifier certaines hypothèses biologiques. De plus le manque d'étude théorique de ces modèles limite leur utilisation. Il semble en effet difficile de formuler de nouvelles hypothèses biologiques à partir de simulations peu fiables.

Le travail produit dans cette thèse s'est concentré sur une modélisation mathématique de deux hypothèses biologique précises : la perte de MisMatch Repair et l'hypothèse DAH. Nous nous sommes focalisés sur la production et l'étude mathématique d'estimateurs permettant d'évaluer l'impact de ces hypothèses dans la cancérisation. Actuellement, l'évaluation d'hypothèses biologiques utilise beaucoup la simulation informatique. Grâce à l'explosion des moyens informatiques, les simulateurs ont pu intégrer de plus en plus de phénomènes. Pourtant, la prise en compte de nombreux phénomènes rend l'interprétation et la fiabilité des simulations d'autant plus délicates. En particulier, dans le cas de la simulation du cancer, l'impact des phénomènes biologiques étant partiellement connu, l'utilisation de ce type de simulateurs est peu utile pour le développement de nouvelles voies thérapeutiques.

Le modèle coalescent à deux taux de mutation, présenté dans cette thèse, trouve son utilité dans l'évaluation directe de l'instabilité génétique au cours de la cancérogénèse. L'étude mathématique des estimateurs du taux de mutation élévé fournit un outil rigoureux pour le développement de nouvelles approches thérapeutiques. En effet, nous pouvons envisager le développement de substances chimiques qui se substitueraient aux molécules associées à la réparation de l'ADN, lorsque ces dernières sont déficientes. Les statistiques que nous avons proposées pourraient permettre d'évaluer le bénéfice apporté par ces substances.

D'autre part, le modèle d'interaction spatiale proposée pour étudier l'adhésion entre les cellules d'un tissu peut également s'inclure dans un processus de validation thérapeutique. Afin d'endiguer le développement spatial d'un clone tumoral, certaines recherches

s'intéressent à modifier les propriétés adhésives des cellules cancéreuses. L'estimateur de pseudo-vraisemblance proposé pourrait permettre une évaluation rigoureuse de la modification de ces propriétés adhérentes.

# Liste des travaux publiés au cours de cette thèse

**Revues Internationales**

- **M. Emily** and O.Francois. Conditional coalescent trees with two mutation rates and their application to genomic instability, *Genetics*, Vol. 172, Mars 2006, pages 1809-1820.
- **M. Emily**, D. Morel, R. Marcelpoil and O.Francois. Spatial correlation of gene expression measures in Tissue Microarray core analysis, *Journal of theoretical Medicine*, Vol. 6, No. 1, Mars 2005, pages 33-39.

**Conférences Internationales**

- **M. Emily** and O.Francois. Estimating the raised mutation rate in a sample of gene with mutators, *JOBIM'05*, Poster 74, Lyon France, 6-8 Juillet 2005.
- **M. Emily** and O. Francois. Number of segregating sites in a sample of genes under the genetic instability hypothesis, *XIth International Symposium on Applied Stochastic Models and Data Analysis*, ASMDA 2005, Brest, France, 17-20 Mai 2005.
- **M. Emily** and O. Francois. Estimating the raised mutation rate from a sample of genes with mutators, *33rd European Mathematical Genetics Meeting*, EMGM 2005, Le Kremlin-Bicetre, Paris, 1-2 Avril 2005, pp.26. Résumé publié dans *Annals of Human Genetics* Vol. 69, page 767 Part 6, 2005.
- **M. Emily**, D. Morel, R. Marcelpoil, O.Francois. Spatial correlation of gene expression measures in Tissue Microarray core analysis, *JOBIM'04*, Page 50, Montreal Canada, 28-30 Juin 2004.

- **M. Emily** and O.Francois. Estimateur de pseudo-vraisemblance pour un processus ponctuel de Markov : application à l'histologie, *38ème Journée de Statistique*, Clamart France, 29 Mai - 2 Juin 2006.

# Bibliographie


[Abbott et al., 2006] Abbott, R., Forrest, S., et Pienta, K. (2006). Simulating the Hallmarks of Cancer. *Artificial Life*, **Sous Presse**.

[Ammari et al., 2004] Ammari, H., Seo, J. K., Kwon, O., et Woo, E. J. (2004). Mathematical framework and anomaly estimation for breast cancer detection : Electrical impedance technique using T2000 configuration. *IEEE Trans. Biomedical Engineering*, **51**(11) : 1898–1906.

[Anderson et Chaplain, 1998] Anderson, A. R. et Chaplain, M. A. (1998). Continuous and discrete mathematical models of tumor-induced angiogenesis. *Bulletin of Mathematical Biology*, **60** : 857–899.

[Anderson, 2001] Anderson, G. R. (2001). Genomic instability in cancer. *Current Science*, **81** : 501–507.

[Armitage et Doll, 1954] Armitage, P. et Doll, R. (1954). The age distribution of cancer and a multi-stage theory of cancinogenesis. *British Journal of Cancer*, **8** : 1–12.

[Armitage et Doll, 1957] Armitage, P. et Doll, R. (1957). A two-stage theory of carcinogenesis in relation to the age distribution of human cancer. *British Journal of Cancer*, **11** : 161–169.

[Armstrong, 1989] Armstrong, P. B. (1989). Cell sorting out : the self assembly of tissues in vitro. *Critical Review in Biochemistry and Molecular Biology*, **24** : 119–149.

[Ashley, 1969] Ashley, D. J. (1969). The two “hit" and multiple “hit" theories of carcinogenesis. *British Journal of Cancer*, **23**(2) : 313–328.

[Baddeley et Møller, 1989] Baddeley, A. et Møller, J. (1989). Nearest-Neighbour Markov point processes and random sets. *International Statistical Review*, **2** : 89–121.

[Baish et Jain, 2000] Baish, J. W. et Jain, R. K. (2000). Fractals and Cancer. *Cancer Research*, **60** : 3683–3688.

[Beckman et Loeb, 2005] Beckman, R. A. et Loeb, L. A. (2005). Genetic instability in cancer : Theory and experiment. *Seminars in Cancer Biology*, **6** : 423-435.

[Bertin et al., 1999a] Bertin, E., Billiot, J. M., et Drouilhet, R. (1999a). Existence of 'nearest-neighbour' spatial Gibbs models. *Advances in Applied Probability*, **2** : 89–121.

[Bertin et al., 1999b] Bertin, E., Billiot, J. M., et Drouilhet, R. (1999b). Spatial Delaunay Point Processes. *Stochastic Models*, **15** : 181–199.

[Bertin et al., 2004] Bertin, E., Billiot, J. M., et Drouilhet, R. (2004). Phase Transition in the Nearest-Neighbor Continuum Potts Model. *Journal of Statistical Physics*, **114** : 79–100.

[Berx et VanRoy, 2001] Berx, G. et VanRoy, F. M. (2001). The E-Cadherin/catenin complex : an important gatekeeper in breast cancer tumorigenesis and malignant progression. *Breast Cancer Research*, **3** : 289–293.

[Besag, 1975] Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician*, **24** : 192–236.

[Besag et al., 1982] Besag, J., Milne, R., et Zachary, S. (1982). Point process limits of lattice processes. *Annals of Applied Probability*, **19** : 210–216.

[Bielas et Loeb, 2005] Bielas, J. H. et Loeb, L. A. (2005). Mutator Phenotype in Cancer : Timing and Perspectives. *Environnemental and Molecular Mutagenesis*, **45** : 206–213.

[Bigras et al., 1996] Bigras, G., Marcelpoil, R., Brambilla, E., et Brugal, G. (1996). Cellular sociology applied to neuroendocrine tumors of the lung : quantitative model of neoplastic architecture. *Cytometry*, **24**(1) : 74–82.

[Billiot et al., 2006] Billiot, J. M., Coeurjolly, J. F., et Drouilhet, R. (2006). Maximum Pseudo-Likelihood Estimator for Nearest-Neighbour Gibbs Point Processes. *Preprint Arxiv*.

[Birchmeier et Behrens, 1994] Birchmeier, W. et Behrens, J. (1994). Cadherins expression in carcinomas : role in the formation of cell junctions and the prevention of invasiveness. *Biochimica Biophysica Acta*, **1198** : 11–26.

[Boland et al., 1998] Boland, C. R., Thibodeau, S. N., Hamilton, S. R., Sidransky, D., Eshleman, J. R., Burt, R. W., Meltzer, S. J., Rodriguez-Bigas, M. A., Fodde, R., Ranzani, G. N., et Srivastava, S. (1998). A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition : development of international criteria for the determination of microsatellite instability in colorectal cancer. *Cancer Research*, **58** : 5248–5257.

[Borgers et Peskin, 1987] Borgers, C. et Peskin, C. S. (1987). A lagrangian fractional step method for the incompressible Navier-Stokes equations on a periodic domain. *Journal of Computational Physics*, **70** : 397–438.

[Boveri, 1929] Boveri, T. (1929). *Origin of the Malignant Tumors*. Williams & Williams Publising Co.

[Bracke et al., 1996] Bracke, M. E., VanRoy, F. M., et Mareel, M. M. (1996). The E-cadherin/catenin complex in invasion and metastasis. *Current Topics Microbiology and Immunology*, pages 123–161.

[Bremnes et al., 2002] Bremnes, R. M., Veve, R., Hirsh, F. R., et A.Franklin, W. (2002). Hereditary predisposition for cancer of the breast and the ovary. *Lung Cancer*, **36** : 115–124.

[Brodland, 2002] Brodland, G. W. (2002). The differential interfacial tension hypothesis (DITH) : a comprehensive theory for the self-rearrangement of embryonic cells and tissues. *Journal of Biomechanical Engineering*, **124** : 188–197.

[Brodland et Chen, 2000] Brodland, G. W. et Chen, H. H. (2000). The mechanics of cell sorting and envelopment. *Journal of Biomechanics*, **33** : 845–851.

[Cairns, 1975] Cairns, J. (1975). Mutation selection and the natural history of cancer. *Nature*, **255** : 197–200.

[Cairns, 2002] Cairns, J. (2002). Somatic stem cells and the kinetics of mutagenesis and carcinogenesis. *Proceedings of the National Academy of Science*, **99**(16) : 10567–10570.

[Calabrese et al., 2004] Calabrese, P., Tavaré, S., et Shibata, D. (2004). Pretumor progression : Clonal evolution of human stem cell populations. *American Journal of Pathology*, **164**(4) : 1337–1346.

[Cannings, 1974] Cannings, C. (1974). The latent roots of certain Markov chains arising in genetics : A new approach. I. Haploid models. *Advances of Applied Probability*, **6** : 260–290.

[Carter et Prenter, 1972] Carter, D. S. et Prenter, P. M. (1972). Exponential spaces and couting processes. *Probability Theory and Related Fields*, **21** : 1–19.

[Chadoeuf et al., 1992] Chadoeuf, J., Nandris, D., Geiger, J. P., Nicole, M., et Pierrat, J. C. (1992). Modélisation spatio-temporelle d'une épidémie par un processus de Gibbs : Estimation et Tests. *Biometrics*, **48** : 1165–1175.

[Chandebois, 1977] Chandebois, R. (1977). Cell sociology and the problem of position effect : Pattern formation, origin and role of gradients. *Acta Biotheoretica*, **26**(4) : 203–238.

[Chen et al., 2003] Chen, C. L., Liu, S. S., Ip, S. M., Wong, L. C., Ty, N., et Hy, N. (2003). E-cadherin expression is silenced by DNA methylation in cervical cancer cell lines and tumours. *Eur. J. Cancer*, **39** : 517–523.

[Christofori et Semb, 1999] Christofori, G. et Semb, H. (1999). The role of cell-adhesion molecule E-cadherin as a tumor-suppressor gene. *Trends in Biochemical Sciences*, **24** : 73–76.

[Cohen et Murray, 1999] Cohen, D. S. et Murray, J. D. (1999). A generalized diffusion model for growth and dispersal in a population. *Journal of Mathematical Biology*, **12**(2) : 73–76.

[Conacci-Sorrell et al., 2002] Conacci-Sorrell, M., Zhurinsky, J., et Ben-Ze'ev, A. (2002). The cadherin-catenin adhesion system in signaling and cancer. *Journal of Clinical Investigation*, **109** : 987–991.

[Coop et Griffiths, 2004] Coop, G. et Griffiths, R. C. (2004). Ancestral inference on gene trees under selection. *Theoretical Population Biology*, **66** : 219–232.

[Cowin, 2000] Cowin, S. C. (2000). How is a tissue built ? *Journal of Biomechanical Engineering*, **122** : 553–569.

[Daley et Vere-Jones, 1988] Daley, D. J. et Vere-Jones, D. (1988). *An introduction to the theory of point processes.* Springer-Verlag, New-York.

[De-Wever et al., 2001] De-Wever, E. V.-A. O., da Rocha, A. S. C., et Marcel, M. (2001). Defective E-cadherin/catenin complexes in human cancer. *Virchows Archiv*, **439**(6) : 725–751.

[Dickerson, 1989] Dickerson, R. E. (1989). Definitions and nomenclature of nucleic acid structure components. *Nucleic Acids Reseach*, **17**(5) : 1797–1803.

[Dobrushin, 1969] Dobrushin, R. (1969). Gibbsian random fields. The general case. *Func. Anal. Appl.*, **3** : 22–28.

[Doki et al., 1993] Doki, Y., Shiozaki, H., Tahara, H., Inoue, M., Oka, H., Iihara, K., Kadowaki, T., Takeichi, M., et Mori, T. (1993). Correlation between E-cadherin expression and invasiveness in vitro in a human esophageal cancer cell line. *Cancer Research*, **53**(14) : 3421–3426.

[Dorudi et al., 1993] Dorudi, S., Sheffield, J. P., Poulsom, R., Northover, J. M., et Hart, I. R. (1993). E-cadherin expression in colorectal cancer. An immunocytochemical and in situ hybridization study. *The American journal of pathology*, **142** : 981–986.

[Düchting et al., 1996] Düchting, W., Ulmer, W., et Ginsberg, T. (1996). Cancer : a challenge for control theory and computer modelling. *European Journal of Cancer*, **32** : 1283–1292.

[Faeron et Vogelstein, 1990] Faeron, E. R. et Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. *Cell*, **61**(5) : 759–767.

[Fiksel, 1988] Fiksel, T. (1988). Estimation of interaction potentials of gibbsian point processes. *Statistics*, **19** : 77–86.

[Fishel et al., 1993] Fishel, R., Lescoe, M. K., Rao, M. R., Copeland, N. G., N. A. Jenkins, J. G., Kane, M., et Kolodner, R. (1993). The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colorectal cancer. *Cell*, **75** : 1027–1038.

[Fisher, 1958] Fisher, J. C. (1958). Multiple-mutation theory of carcinogenesis. *Nature*, **181** : 651–652.

[Fisher, 1922] Fisher, R. A. (1922). On the dominance ratio. *Proceedings of the Royal Society*, **42** : 321–431.

[Foty et al., 1996] Foty, R. A., Pflerger, C. M., Forgacs, G., et Steinberg, M. S. (1996). Surface tensions of embryonic tissues predict their mutual envelopment behavior. *Development*, **122** : 1611–1620.

[Foty et Steinberg, 2004] Foty, R. A. et Steinberg, M. A. (2004). Cadherin-mediated cell-cell adhesion and tissue segregation in relation to malignancy. *International Journal of Development Biology*, **48** : 397–409.

[Friedlander et al., 1989] Friedlander, D. R., Mège, R. M., Cunningham, B. A., et Edelman, G. M. (1989). Cell sorting-out is modulated by both the specificity and amount of different cell adhesion molecules (CAMs) expressed on cell surfaces. *Proceedings of the National Academy of Science*, **86** : 7043–7047.

[Gatenby et Gawlinski, 1996] Gatenby, R. A. et Gawlinski, E. T. (1996). A reaction-diffusion model of cancer invasion. *Cancer Research*, **56** : 5745–5753.

[Geiger et Ayalon, 1992] Geiger, B. et Ayalon, O. (1992). Cadherins. *Annual Review of Cell Biology*, **8** : 307–332.

[Geyer, 1998] Geyer, C. J. (1998). Likelihood inference for spatial point processes. In Barndoff-Nielsen, O. E., Kendall, W. S., et van Lieshout, M. N. M., editors, *Stochastic Geometry, Likelihood and Computation*. Chapman and Hall.

[Geyer et Møller, 1994] Geyer, C. J. et Møller, J. (1994). Simulation Procedures and Likelihood Inference for Spatial Point Processes. *Scandinavian Journal of Statistics*, **21** : 359–373.

[Geyer et Thompson, 1992] Geyer, C. J. et Thompson, E. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *Journal of the Royal Statistical Society B*, **54** : 657–699.

[Gierer et al., 1972] Gierer, A., Berking, S., Bode, H., David, C. N., Flick, K., Hansmann, G., Schaller, H., et Trenkner, E. (1972). Regeneration of hydra from reaggregated cells. *Nature : New Biology*, **239** : 98–101.

[Gierer et Meinhardt, 1972] Gierer, A. et Meinhardt, H. (1972). A theory of biological pattern formation. *Kybernetik*, **12** : 30–39.

[Giroldi et al., 1999] Giroldi, L. A., Bringuier, P. P., Shimazui, T., Jansen, K., et Schalken, J. A. (1999). Changes in cadherin-catenin complexes in the progression of human bladder carcinoma. *International Journal of Cancer*, **2** : 70–76.

[Glazier et Graner, 1993] Glazier, J. A. et Graner, F. (1993). Simulation of differential adhesion driven rearrangement of biological cells. *Physical Review E*, **47**(3) : 2128–2154.

[Goel et al., 1970] Goel, N. S., Campbell, R. D., Gordon, R., Rosen, R., Martinez, H., et Ycas, M. (1970). Self-sorting of isotropic cells. *Journal of Theoretical Biology*, **28** : 423–268.

[Graner, 1993] Graner, F. (1993). Can surface adhesion drive cell-rearrangment ? Part I : biological cell-sorting. *Journal of Theoretical Biology*, **164**(4) : 455–476.

[Graner et Glazier, 1992] Graner, F. et Glazier, J. A. (1992). Simulation of biological cell sorting using a two-dimensional extended Potts model. *Physical Review Letters*, **69**(13) : 2013–2016.

[Graner et Sawada, 1993] Graner, F. et Sawada, Y. (1993). Can surface adhesion drive cell-rearrangment ? Part II : a geometrical model. *Journal of Theoretical Biology*, **164**(4) : 477–506.

[Gray et Scott, 1984] Gray, P. et Scott, S. K. (1984). Autocatalytic reactions in the isothermal, continuous stirred tank reactor : oscillations and instabilities in the system $A + 2B \rightarrow 3B$, $B \rightarrow C$. *Chemical Engineering Science*, **39** : 1087–1097.

[Griffiths, 1980] Griffiths, R. C. (1980). Lines of descent in the diffusion approximation of neutral Wright-Fisher models. *Theoretical Population Biology*, **17** : 37–50.

[Griffiths et Tavaré, 1998] Griffiths, R. C. et Tavaré, S. (1998). The age of a mutation in a general coalescent tree. *Stochastic Models*, **14** : 273–295.

[Griffiths et Tavaré, 2003] Griffiths, R. C. et Tavaré, S. (2003). The genealogy of a neutral mutation. In Green, P., Hjort, N., et Richardson, S., editors, *Highly Structured Stochastic Systems*, pages 393–413. Oxford University Press.

[Gumbiner, 2000] Gumbiner, B. M. (2000). Regulation of cadherin adhesive activity. *The Journal of Cell Biology*, **148** : 399–403.

[Hahn et Weinberg, 2002] Hahn, W. C. et Weinberg, R. A. (2002). Modelling the molecular circuicity of cancer. *Nature Reviews Cancer*, **2** : 331–341.

[Hanahan et Weinberg, 2000] Hanahan, D. et Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, **100** : 57–70.

[Harris, 1976] Harris, A. K. (1976). Is cell sorting caused by differences in the work of intercellular adhesion ? A critique of the Steinberg hypothesis. *Journal of Theoretical Biology*, **61** : 267–285.

[Hastings, 1970] Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57** : 97–109.

[Honda, 1978] Honda, H. (1978). Description of cellular patterns by Dirichlet domains : The two-dimensional case. *Journal of Theoretical Biology*, **72** : 523–543.

[Honda, 1983] Honda, H. (1983). Geometrical models for cells in tissues. *International Review of Cytology*, **81** : 191–248.

[Honda et al., 1996] Honda, H., Tanemura, M., et Imayama, S. (1996). Spontaneous architectural organization of mammalian epiderms from random cell packing. *The Journal of Inverstigative Dermatology*, **106**(2) : 312–315.

[Honda et al., 2004] Honda, H., Tanemura, M., et Nagai, T. (2004). A three-dimensional vertex dynamics cell model of space-filling polyhedra simulating cell behavior in a cell aggregate. *Journal of Theoretical Biology*, **226** : 439–453.

[Honda et al., 2000] Honda, H., Tanemura, M., et Yoshida, A. (2000). Differentiation of Wing Epidermal Scale Cells in a Butterfly under the lateral inhibition model - Appearance of large cells in a polygonal pattern. *Acta Biotheoretica*, **48**(2) : 121–136.

[Honda et al., 1986] Honda, H., Yamanaka, H., et Eguchi, G. (1986). Transformation of a polygonal cellular pattern during sexual maturation of the avian oviduct epithelium : Computer simulation. *Journal of Embryology and Experimental Morphology*, **98** : 1–19.

[Häggström, 2000] Häggström, O. (2000). Markov Random Fields and Percolation on General Graphs. *Advances in Applied Probability*, **32** : 39–66.

[Jackson et Loeb, 1998a] Jackson, A. L. et Loeb, L. A. (1998a). The mutation rate and cancer. *Genetics*, **148** : 1483-1490.

[Jackson et Loeb, 1998b] Jackson, A. L. et Loeb, L. A. (1998b). On the origin of multiple mutations in human cancer. *Cancer Biology*, **8** : 421-429.

[Jensen et Møller, 1991] Jensen, J. L. et Møller, J. (1991). Pseudolikelihood for exponential family models of spatial point processes. *Annals of Applied Probability*, **1** : 445–461.

[Joo et al., 2002a] Joo, Y. E., Rew, J. S., Bomp, H. S., Park, C. S., et Kim, S. J. (2002a). Expression of E-cadherin, alpha- and beta-catenins in patients with pancreatic adenocarcinoma. *Pancreatology*, **2** : 129–137.

[Joo et al., 2002b] Joo, Y. E., Rew, J. S., Choi, S. K., Bomp, H. S., Park, C. S., et Kim, S. J. (2002b). Expression of E-cadherin and catenins in early gastric cancer. *Journal of Clinical Gastroenterology*, **35** : 35–42.

[Kamijo et al., 2001] Kamijo, S., K., I., et M., S. (2001). Segmentations of Spatio-Temporal Images by Spatio-Temporal Markov Random Field Model. In Figueiredo, M., Zerubia, J., et Jain, A. K., editors, *Energy Minimization Methods in Computer Vision and Pattern Recognition : Third International Workshop*, pages 300–313.

[Karlin et McGregor, 1972] Karlin, S. et McGregor, J. (1972). Addendum to a paper of W. Ewens. *Theoretical Population Biology*, **3** : 113–116.

[Kelly et Ripley, 1976] Kelly, F. et Ripley, B. D. (1976). A note on Strauss model for clustering. *Biometrika*, **63** : 357–360.

[Kemler, 1993] Kemler, R. (1993). From cadherins to catenins : cytoplasmic protein interactions and regulation of cell adhesion. *Trends in Genetics*, **9** : 317–321.

[Kimura, 1969] Kimura, M. (1969). The number of heterozygous nuclotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, **61**(4) : 893–903.

[Kimura et Ohta, 1973] Kimura, M. et Ohta, T. (1973). The age of a neutral mutant persisting in a finite population. *Genetics*, **75** : 199-212.

[Kingman, 1982a] Kingman, J. F. C. (1982a). The Coalescent. *Stochastic Processes and their Applications*, **13** : 235–248.

[Kingman, 1982b] Kingman, J. F. C. (1982b). On the genealogy of large populations. *Journal of Applied Probability*, **19** : 27–43.

[Knudson, 1971] Knudson, A. G. (1971). Mutation and Cancer : Statistical study of Retinoblastoma. *Proceedings of the National Academy of Science*, **68**(4) : 820–823.

[Komarova, 2005] Komarova, N. (2005). Mathematical modeling of tumorigenesis : mission possible. *Current Opinion in Oncology*, **17**(1) : 39–43.

[Krone et Neuhauser, 1997] Krone, S. M. et Neuhauser, C. (1997). Ancestral process with selection. *Theoretical Population Biology*, **51** : 210–237.

[Kuroda et al., 1998] Kuroda, S., Futaka, M., Nakagawa, M., Fujii, K., Nakamura, T., Ookubo, T., Izawa, I., Nagase, T., Nomura, N., Tani, H., Shoji, I., Matsuura, Y., Yonehara, S., et Kaibuchi, K. (1998). Role of IQGAP1, a target of the small GTPases Cdc42 and Rac1, in regulation of E-Cadherin-Mediated cell-cell adhesion. *Science*, **281** : 832–835.

[Landford et Ruelle, 1969] Landford, O. et Ruelle, D. (1969). Observables at infinity and states with short range correlations in statistical mechanics. *Communications in Mathematical Physics*, **13** : 194–215.

[Lengyel et Epstein, 1991] Lengyel, I. et Epstein, I. R. (1991). Modelling of Turing structures in the chlorite-iodide-malonic acid-strach reaction system. *Science*, **251** : 650–652.

[Lindblom et al., 1993] Lindblom, A., Tannergard, P., Werelius, B., et Nordenskjold, M. (1993). Genetic mapping of a second locus predisposing to hereditary non-polyposis colon cancer. *Nature Genetics*, **5** : 279–282.

[Little et Wright, 2003] Little, M. P. et Wright, E. G. (2003). A stochastic model incorporating genomic instability to colon cancer data. *Mathematical Biosciences*, **183**(2) : 111-134.

[Loeb, 2001] Loeb, L. (2001). A Mutator Phenotype in Cancer. *Cancer Research*, **61** : 3230-3239.

[Loeb, 1991] Loeb, L. A. (1991). Mutator phenotype may be required for multistage carcinogenesis. *Cancer Research*, **51** : 3075-3079.

[Loeb et al., 1974] Loeb, L. A., Springgate, C. F., et Battula, N. (1974). Errors in DNA replication as a basis of malignant changes. *Cancer Research*, **34** : 2311-2321.

[Marcelpoil, 1993] Marcelpoil, R. (1993). *Méthodologie pour l'étude de la sociologie cellulaire : application à l'étude du tissu prostatique normal et pathologique.* PhD thesis, Université Joseph Fourier - Grenoble I.

[May, 2004] May, R. M. (2004). Uses and Abuses of Mathematics in Biology. *Science*, **303** : 790–793.

[McGary et al., 2002] McGary, E. C., Lev, D. C., et Bar-Eli, M. (2002). Cellular adhesion pathways and metastatic potential of human melanoma. *Cancer Biology and Therapy*, **1** : 459–465.

[Metropolis et al., 1953] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., et Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21** : 1087–1092.

[Meyn et Tweedie, 1993] Meyn, S. P. et Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability.* Springer-Verlag, London.

[Michor et al., 2005] Michor, F., Iwasa, Y., Vogelstein, B., Lengauer, C., et Nowak, M. A. (2005). Can Chromosomal instability initiate tumorigenesis ? *Seminars in Cancer Biology*, **15** : 43–49.

[Møller et Waagepetersen, 2003] Møller, J. et Waagepetersen, R. P. (2003). *Statistical Inference and Simulation for Spatial Point Processes.* Chapman and Hall/CRC, Boca Raton.

[Mochizuki et al., 1996] Mochizuki, A., Isawa, Y., et Takeda, Y. (1996). A stochastic model for cell sorting and measuring cell-cell-adhesion. *Journal of Theoretical Biololgy*, **179** : 129–146.

[Moolgavkar et Knudson, 1981] Moolgavkar, S. H. et Knudson, A. G. (1981). Mutation and cancer : a model for human carcinogenesis. *Journal of the National Cancer Institute*, **66**(6) : 1037–1052.

[Murray, 2002] Murray, J. D. (2002). *Mathematical Biology, 3ème Ed.* Springer-Verlag, Berlin.

[Nagai et Honda, 2001] Nagai, T. et Honda, H. (2001). A dynamic cell model for the formation of epithelial tissue. *Philosophical Magazine B*, **81** : 699–719.

[Nagai et al., 1988] Nagai, T., Kawasaki, K., et Nakamura, K. (1988). Vertex dynamics of two-dimensional cellular pattern. *Journal of the Physical Society of Japan*, **57**(7) : 2221–2224.

[Nawrocki-Raby et al., 2001] Nawrocki-Raby, B., Polette, M., Gilles, C., Clavel, C., Strumane, K., Matos, M., Zahm, J. M., Roy, F. V., Bonnet, N., et Birembaut, P. (2001). Quantitative cell dispersion analysis : new test to measure tumor cell aggressiveness. *International Journal of Cancer*, **93** : 644-652.

[Neuhauser et Tavaré, 2001] Neuhauser, C. et Tavaré, S. (2001). The Coalescent. In Brenner, S. et Miller, J., editors, *Encyclopedia of Genetics*, pages 392-397. Academic Press.

[Ngwa et Maini, 1995] Ngwa, G. A. et Maini, P. K. (1995). Spatio-temporal patterns in a mechanical model for mesenchymal morphogenesis. *Journal of Mathematical Biology*, **33** : 489–520.

[Nordling, 1953] Nordling, C. O. (1953). A new theory on cancer-inducing mechanism. *British Journal of Cancer*, **7** : 68–72.

[Nose et al., 1988] Nose, A., Nagafuchi, A., et Takeichi, M. (1988). Expressed recombinant cadherins mediate cell sorting in model systems. *Cell*, **54** : 993–1001.

[Nowak et al., 2002] Nowak, M. A., Komarova, N. L., Sengupta, A., Jallepalli, P. V., Shih, I. M., Vogelstein, B., et Lengauer, C. (2002). The role of chromosomal instability in tumor initiation. *Proceedings of the National Academy of Science*, **99**(25) : 16226–16231.

[Nowell, 1976] Nowell, P. C. (1976). The clonal evolution of tumor cell populations. *Science*, **194** : 23–28.

[Nummelin, 1984] Nummelin, E. (1984). *General irreducible Markov Chains and non-negative Operators*. Cambridge University Press, Cambridge.

[Okabe et al., 2000] Okabe, A., Boots, B., Sugihara, K., et Chiu, S. N. (2000). *Spatial tessellations. Concepts and Applications of Voronoi diagrams*. Wiley, Chichester.

[Orme et Chaplain, 1996] Orme, M. E. et Chaplain, M. A. (1996). A mathematical model of the first steps of tumor-related angiogenesis : capillary sprout formation and

secondary branching. *IMA Journal of Mathematics Applied in Medicine and Biology*, **13** : 73–98.

[Oster et al., 1983] Oster, G. F., Murray, J. D., et Harris, A. K. (1983). Mechanical aspects of mesenchymal morphogenesis. *Journal of Embryology ans Experimental Morphology*, **78** : 83–125.

[Owen et Sherratt, 2004] Owen, M. R. et Sherratt, J. A. (2004). Mathematical modeling of macrophage dynamics in tumors. *Mathematical Models and Methods in Applied Sciences*, **377** : 675–684.

[Oyama et al., 1994] Oyama, T., Kanai, Y., Ochiai, A., Oda, T., Yanagihara, A., Tsukita, S., Shibamoto, S., et Ito, F. (1994). A truncated $\beta$-catenin disrupts the interaction between E-Cadherin and $\alpha$-catenin : a cause of loss intercellular adhesiveness in human cancer cell lines. *Cancer Research*, **54** : 6282–6287.

[Ozawa et al., 1990] Ozawa, M., Ringwald, M., et Kemler, R. (1990). Uvomorulin-catenin complex formation is regulated by a specific domain in the cytoplasmic region of the cell adhesion molecule. *Proceedings of the National Academy of Science*, **87** : 4246–4250.

[Painter et al., 2000] Painter, K. J., Maini, P. K., et Othmer, H. G. (2000). Chemotactic response to multiple signalling cues. *Journal of Mathematical Biology*, **41** : 285–314.

[Phillips et Steinberg, 1969] Phillips, H. M. et Steinberg, M. S. (1969). Equilibrium Measurements of Embryonic Chick Cell Adhesiveness, I. Shape Equilibrium in Centrifugal Fields. *Proceedings of the National Academy of Science*, **64** : 121–127.

[Preston, 1976] Preston, C. J. (1976). Random Fields. *Lecture Notes in Mathematics*, **534** : 1–200.

[Preston, 1977] Preston, C. J. (1977). Spatial birth-and-death preocesses. *Bulletin of the International Statistical Institute*, **46** : 371–391.

[Propp et Wilson, 1996] Propp, J. G. et Wilson, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithm*, **9** : 223–252.

[Qi et al., 1993] Qi, A., Zheng, X., Du, C., et An, B. (1993). A cellular automaton model of cancerous growth. *Journal of Theoretical Biology*, **161**(1) : 1–12.

[Quaranta et al., 2005] Quaranta, V., Weaver, A. M., Cummings, P. T., et Anderson, A. R. A. (2005). Mathematical modeling of cancer : The future of prognosis and treatment. *Clinica Chimica Acta*, **357** : 173–179.

[Ripley, 1977] Ripley, B. D. (1977). Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society Series B*, **39** : 172–212.

[Ripley, 1979] Ripley, B. D. (1979). Simulating spatial patterns : dependent samples from a multivariate density, Algorithms. *Applied Statistics*, **28** : 109–112.

[Root-Bernstein et Bernstein, 1999] Root-Bernstein, R. S. et Bernstein, M. I. (1999). Game-theory models of interactions between tumor cells. *Anticancer Research*, **19** : 4869–4876.

[Rubin et al., 2001] Rubin, M. A., Mucci, N. R., Figurski, J., Fecko, A., Pienta, K. J., et Day, M. L. (2001). E-Cadherin expression in prostate cancer : a broad survey using high-density tissue microarray technology. *Human Pathology*, **32** : 690–697.

[Ruelle, 1969] Ruelle, D. (1969). *Statistical Mechanics*. Benjamin, New-York.

[Saito et al., 2003] Saito, T., Nishimura, M., Yamasaki, H., et Kudo, R. (2003). Hypermethylation in promoter region of E-cadherin gene is associated with tumor dedifferentiation and myometrial invasion in endometrial carcinoma. *Cancer*, **97** : 1002–1009.

[Scheumman et al., 1995] Scheumman, G. F. W., Hoang-Vu, C., Cetin, Y., Gimm, O., Behrends, J., Wasielewski, R. V., Georgii, A., Birchmeier, W., zur Mulhen, A. V., Dralle, H., et Brabant, G. (1995). Clinical significance if E-Cadherin as a prognostic marker in thyroid carcinomas. *Journal of Clinical Endoctrinology and Metabolism*, **80**(7) : 2168–2172.

[Schweichheimer et al., 1998] Schweichheimer, K., Zhou, L., et Birchmeier, W. (1998). E-Cadherin in human brain tumours : loss of immunoreactivity in malignant meningiomas. *Virchows Archiv*, **432** : 163–167.

[Scott et al., 1999] Scott, E. L., Britton, N. F., Glazier, J. A., et Zajac, M. (1999). Stochastic simulation of benign avascular tumor growth using the Potts model. *Mathematical and Computer Modelling*, **30** : 183–198.

[Shapiro et al., 1995] Shapiro, L., Fannon, A. M., Kwong, P. D., Thompson, A., Lehmann, M. S., Grubel, G., Legrand, J. F., Als-Nielsen, J., Colman, D. R., et Hendrick-

son, W. A. (1995). Structural basis of cell-cell adhesion by cadherins. *Nature*, **374** : 327–337.

[Sheratt et Nowak, 1992] Sheratt, J. A. et Nowak, M. A. (1992). Oncogenes, anti-oncogenes and the immune response to cancer : a mathematical model. *Proceedings of the Royal Society - Biological sciences*, **248**(1323) : 261–271.

[Shimazui et al., 1995] Shimazui, T., Giroldi, L. A., Bringuier, P. P., Oosterwijk, E., et Schalken, J. A. (1995). Complex cadherin expression in real cell carcinoma. *Cancer research*, **56**(14) : 3234–3237.

[Slatkin et Rannala, 2000] Slatkin, M. et Rannala, B. (2000). Estimating Allele Age. *Annual Review of Genomics and Human Genetics*, **1** : 225–249.

[Spencer et al., 2004] Spencer, S. L., Berryman, M. J., Garcia, J. A., et Abbott, D. (2004). An ordinary differential equation model for the multistep transformation to cancer. *Journal of Theoretical Biology*, **231** : 515–524.

[Steinberg, 1962a] Steinberg, M. S. (1962a). On the mechanism of tissue reconstruction by dissociated cells, I. population kinetics, differential adhesiveness, and the absence of directed migration. *Proceedings of the National Academy of Science*, **48** : 1577-1582.

[Steinberg, 1962b] Steinberg, M. S. (1962b). On the mechanism of tissue reconstruction by dissociated cells, II. time-course of events. *Science*, **137** : 762–763.

[Steinberg, 1962c] Steinberg, M. S. (1962c). On the mechanism of tissue reconstruction by dissociated cells, III. free energy relations and the reorganization of fused, heteronomic tissue fragments. *Proceedings of the National Academy of Science*, **48** : 1769–1776.

[Steinberg, 1963] Steinberg, M. S. (1963). Reconstruction of tissues by dissociated cells. Some morphogenetic tissue movements and the sorting out of embryonic cells may have a common explanation. *Science*, **141** : 401–408.

[Steinberg, 1970] Steinberg, M. S. (1970). Does differential adhesion govern self-assembly processes in histogenesis ? Equilibrium configurations and the emergence of a hierarchy among populations of embryonic cells. *Journal of Experimental Zoology*, **173** : 395–433.

[Steinberg, 1996] Steinberg, M. S. (1996). Adhesion in Development : An Historical Overview. *Developmental Biology*, **180** : 377–388.

[Steinberg et Foty, 1997] Steinberg, M. S. et Foty, R. A. (1997). Intercellular adhesion as determinants of tissue assembly and malignant invasion. *Journal of Cellular Physiology*, **173** : 135–139.

[Steinberg et Takeichi, 1994] Steinberg, M. S. et Takeichi, M. (1994). Experimental specification of cell sorting, tissue spreading, and specific spatial patterning by quantitative differences in cadherin expression. *Proceedings of the National Academy of Science*, **91** : 206–209.

[Stephens, 2000] Stephens, M. (2000). Times on Trees, and the Age of an Allele. *Theoretical Population Biology*, **57** : 109–119.

[Stephens et Donnelly, 2003] Stephens, M. et Donnelly, P. (2003). Ancestral inference in population genetics with selection. *Australian and New Zealand Journal of Statistics*, **45** : 395–430.

[Stoyan et Stoyan, 1994] Stoyan, D. et Stoyan, H. (1994). *Fractals, Random Shapes and Point Fields*. Wiley, Chichester.

[Stoyan et al., 1995] Stoyan, D., W. Kendall, W., et Mecke, J. (1995). *Stochastic Geometry and its Applications*. Wiley, Chichester.

[Sulsky et al., 1984] Sulsky, D., Childress, S., et Percus, J. K. (1984). A model of cell sorting. *Journal of Theoretical Biology*, **106**(3) : 275–301.

[Tajima, 1983] Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*, **105** : 437–460.

[Tajima, 1989] Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123** : 585–595.

[Takacs, 1986] Takacs, R. (1986). Estimator for pair-potential of a gibbsian point process. *Statistics*, **17** : 429–433.

[Takano et al., 2002] Takano, R., Mochizuki, A., et Iwasa, Y. (2002). Possibility of tissue separation caused by cell adhesion. *Journal of Theoretical Biology*, **221** : 459–474.

[Takayama et al., 1996] Takayama, T., Shiozaki, H., Shibamoto, S., Oka, H., Kimura, Y., Tamura, S., Inoue, M., Monden, T., Ito, F., et Monden, M. (1996). $\beta$-catenin expression in human cancers. *American Journal of Pathology*, **148** : 39–46.

[Takeuchi et al., 1988] Takeuchi, I., Kakutani, T., et Tasaka, M. (1988). Cell behavior during formation of prestalk/prespore pattern in submerged agglomerates of *Dictyosteltum*. *Differentiation*, **18** : 191–196.

[Tavaré, 1984] Tavaré, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoretical Population Biology*, **26** : 119–164.

[Tavaré, 2004] Tavaré, S. (2004). Ancestral inference in population genetics. In Cantoni, O., Tavaré, S., et Zeitouni, O., editors, *Ecole d'été de Probabilités de Saint-Flour XXXI - 2001*, pages 1–188. Springer-Verlag.

[Thom, 1975] Thom, R. (1975). *Structural stability and morphogenesis*. Benjamin Addison Wesley, New York.

[Tierney, 1994] Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, **22** : 1701–1762.

[Tomlinson et al., 1996] Tomlinson, I., Novelli, M., et Bodmer, W. (1996). The mutation rate and cancer. *Proceedings of the National Academy of Sciences*, **93** : 14800–14803.

[Tomlinson, 1997] Tomlinson, I. P. M. (1997). Game-theory models of interactions between tumor cells. *European Journal of Cancer*, **33** : 1495–1500.

[Tomlinson et Bodmer, 1995] Tomlinson, I. P. M. et Bodmer, W. F. (1995). Failure of programmed cell death and differentiation as causes of tumors : some simple mathematical models. *Proceedings of the National Academy of Science*, **92** : 11130–11134.

[Turing, 1952] Turing, A. (1952). The Chemical Basis of Morphogenesis. *Philosophical Transactions of The Royal Society (B)*, **237**(641) : 37–72.

[Turner, 2005] Turner, S. (2005). Using cell potential energy to model the dynamics of adhesive biological cells. *Physical Review E*, **71** : 041903 (12 pages).

[Turner et Sherratt, 2002] Turner, S. et Sherratt, J. A. (2002). Intercellular Adhesion and Cancer Invasion : A discrete Simulation Using the Extended Potts Model. *Journal of Theoretical Biology*, **216** : 85–100.

[Turner et al., 2004] Turner, S., Sherratt, J. A., Painter, K. J., et Savill, N. J. (2004). From a discrete to a continuous model of biological cell movement. *Physical Review E*, **69** : 021910 (10 pages).

[Van-Lieshout, 2000] Van-Lieshout, M. N. M. (2000). *Markov Point Processes and their Applications*. Imperial College Press.

[Waddington, 1957] Waddington, C. H. (1957). *The Strategy of Genes*. MacMillan Co., New York.

[Ward et King, 1999] Ward, J. P. et King, J. R. (1999). Mathematical modelling of avascular-tumor growth : II. Modelling growth saturation. *IMA Journal of Mathematics Applied in Medicine and Biology*, **16** : 171–211.

[Watterson, 1975] Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, **2** : 256–276.

[Wheelock et Johnson, 1989] Wheelock, M. J. et Johnson, K. R. (1989). Cadherins as modulators of cellular phenotype. *Annual Reviews of Cell and Developmental Biology*, **2** : 89–121.

[Wijnhoven et al., 2000] Wijnhoven, B. P. L., Dinjens, W. N. M., et Pignatelli, M. (2000). E-Cadherin-catenin cell-cell adhesion complex and human cancer. *British Journal of Surgery*, **87**(8) : 992–1005.

[Wiuf et Donnelly, 1999] Wiuf, C. et Donnelly, P. (1999). Conditional Genealogies and the Age of a Neutral Mutant. *Theoretical Population Biology*, **56** : 183–201.

[Wodarz et Komarova, 2005] Wodarz, D. et Komarova, N. (2005). *Computational Biology of Cancer : Lecture notes and mathematical modeling*. World Scientific Publishing.

[Wolpert, 1969] Wolpert, L. (1969). Positional information and the spatial pattern of cellular differentiation. *Journal of Theoretical Biology*, **25** : 1–47.

[Wooster et Weber, 2003] Wooster, R. et Weber, B. L. (2003). Breast and Ovarian Cancer. *The New England Journal of Medicine*, **348** : 2339–2347.

[Wright, 1931] Wright, S. (1931). Evolution in the Mendelian populations. *Genetics*, **16** : 97–159.

[Yap et Goodwin, 2004] Yap, A. S. et Goodwin, M. (2004). Cell adhesion by cadherins receptors : A brief Primer. *Austrialian Biochemist*, **35** : 13–16.