UNIVERSITÉ DU QUÉBEC À MONTRÉAL

CARTOGRAPHIE GÉNÉTIQUE FINE SIMULTANÉE DE DEUX GÈNES

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

MARIE FOREST

JUIN 2010

REMERCIEMENTS

Merci à mon directeur, Fabrice Larribe, pour son soutien, sa passion pour ses recherches et son implication. Ce fut un plaisir de travailler avec lui toutes ces années.

Merci à mes correcteurs, Serge Alalouf et Jean-Philippe Boucher, pour leurs commentaires éclairants.

Merci à Malorie et Hugues, sans vous mes études à l'UQAM n'auraient pas été les mêmes. Merci pour les discussions et les échanges, j'ai appris beaucoup de ceux-ci.

Merci à Jérôme, pour son aide en informatique, pour le partage de ses ordinateurs et pour son écoute dans les moments de découragements.

Merci à Josette, pour avoir pris le temps de lire mon mémoire et d'y avoir déniché les coquilles.

Merci surtout à toi, Didier, pour ton magnifique programme Python, pour tes encouragements à me dépasser et à me questionner. J'apprends beaucoup en ta compagnie.

Merci aussi à mon père, pour son soutien et son amour inconditionnel, et aussi pour m'avoir transmis son talent à rêver...

TABLE DES MATIÈRES

LISTE DES FIGURES

x

LISTE DES TABLEAUX

RÉSUMÉ

Dans le domaine de la recherche de gènes causaux, il est maintenant connu que plusieurs caractères complexes peuvent en fait être influencés par une multitude de gènes. Dans ce mémoire, nous présentons l'adaptation d'une méthode de cartographie génétique fine à la cartographie de caractère polygénique. Nous présentons tout d'abord un aperçu de certains outils statistiques utilisés en génétique. En particulier, certaines mesures d'association généralement employées en cartographie génétique. Puis, nous présentons la méthode de cartographie que nous souhaitons adapter : méthode qui suppose que le caractère est causé par l'effet d'un seul gène. Nous supposons plutôt que le caractère est causé par la combinaison de deux gènes. Après avoir présenté notre modélisation et les aspects théoriques de l'adaptation proposée, nous utilisons des données simulées pour tester nos développements. Nous comparons aussi nos résultats avec ceux obtenus avec une mesure d'association, ainsi qu'avec la méthode de cartographie dont nous proposons une adaptation. Les résultats démontrent la nécessité de développer des méthodes de cartographie génétique adaptées aux caractères polygéniques ; avec quelques améliorations concernant l'inférence des génotypes aux gènes causaux, notre adaptation devrait offrir de meilleurs résultats que les autres méthodes présentées.

MOTS-CLÉS : statistique génétique, cartographie génétique, caractère polygénique, processus de coalescence, arbre de recombinaison ancestral

INTRODUCTION

Depuis maintenant une vingtaine d'années, la course aux gènes affectant certains caractères et maladies est ouverte ; de nouvelles technologies permettent le séquençage du génome humain ; de grands projets à l'échelle planétaire accumulent les données génétiques. Plusieurs gènes ayant un impact sur certaines maladies ont été découverts à l'aide d'outils statistiques. Mais parfois, ces méthodes échouent face à certaines maladies complexes influencées simultanément par plusieurs gènes. C'est pourquoi le développement de méthodes de cartographie génétique adaptées à cette réalité est souhaitable.

L'objectif de ce mémoire est l'adaptation de la méthode de cartographie génétique fine MapArg (Larribe, Lessard et Schork, 2002 ; Larribe et Lessard, 2008) à la cartographie de caractère polygénique (c'est-à-dire influencé par plusieurs gènes). Nous supposerons que le caractère est causé uniquement par l'effet combiné de deux gènes et qu'il n'est pas influencé par des facteurs environnementaux. Bien que ces hypothèses posées sur la maladie paraîtront possiblement peu réalistes, elles nous permettront de simplifier la modélisation et de vérifier si nous sommes sur la bonne voie.

Pour permettre au lecteur non initié de comprendre les termes spécifiques à la génétique utilisés tout au long de ce mémoire, nous présentons au premier chapitre certains concepts de base de ce domaine. Nous poursuivons ensuite par une description de certains outils statistiques employés généralement en cartographie génétique, ce qui permettra au lecteur d'avoir un aperçu d'où, dans l'univers de la statistique génétique, s'inscrivent la méthode de cartographie MapArg et l'adaptation proposée de celle-ci. MapArg sera présentée en détail au chapitre trois. Suivra la présentation de l'adaptation de cette méthode à la cartographie de deux gènes causant un caractère. Nous terminerons ce mémoire, au chapitre cinq, avec la présentation des résultats obtenus lors de l'analyse de bases de données simulées.

CHAPITRE I

INTRODUCTION À LA GÉNÉTIQUE

Nous devons la découverte des bases de la génétique à Gregor Mendel. Ce moine et botaniste autrichien a mené au début des années 1860 une vaste expérience de reproduction des pois. Il s'intéressait aux mécanismes de transmissions des caractères d'une génération à l'autre, c'est-à-dire aux mécanismes de l'hérédité. Pour ce faire, il observa des caractéristiques bien précises de différentes variétés de pois qui offraient deux choix possibles pour chaque caractéristique ; par exemple la couleur des fleurs (blanches ou violettes), la couleur des graines (jaunes ou vertes) et leurs formes (rondes ou ridés). À la suite de cette expérience (Mendel, 1866), Mendel décrivit un modèle de l'hérédité qui est, encore aujourd'hui, à la base de la génétique. Dans ce chapitre, nous expliquerons les notions de base de génétique essentielles à la compréhension du travail. La majorité des informations contenues dans ce chapitre proviennent du livre *Biologie* de Neil A. Campbell et Richard Mathieu (1995). Le lecteur familiarisé avec les bases de la génétique, telles la transmission des allèles, les recombinaisons et les mutations, pourra passer directement au chapitre 2. Un glossaire des termes techniques utilisés dans ce mémoire se trouve à la fin de celui-ci.

1.1 La cellule, source de vie

Tout être vivant est composé de cellules. En tant qu'être humain, nous sommes originaires d'une seule cellule qui s'est formée lors de la rencontre d'un spermatozoïde et d'un ovule. Ensuite, cet ovule fécondé s'est divisé un grand nombre de fois, créant une

multitude de cellules, plusieurs d'entre elles ayant des fonctions bien précises. Pour que cette magie opère, la cellule originale doit contenir toutes les informations nécessaires pour faire fonctionner le corps humain ; nous appelons l'ensemble de ces informations le *génome*. Nous allons tout d'abord regarder de plus près ce qui se passe lors de la conception d'un être humain, pour expliquer quelques termes et phénomènes génétiques qui nous seront utiles plus tard.

Le génome humain est composé de 46 *chromosomes* ; les chromosomes ressemblent à des bâtonnets et sont formés d'une longue molécule d'ADN entremêlé. Sans entrer dans les détails, notons que l'*ADN* est formé de deux brins en forme d'hélice reliés par des paires de bases azotées, notées G, C, T et A. La base G est toujours en paire avec la base C, et la base T est en paire avec la base A. Ces paires de bases agissent sur des protéines, qui elles, conditionnent le développement et le fonctionnement de l'organisme. La figure 1.1 illustre un chromosome et l'ADN qui le compose. Lors de la division cellulaire, ces 46 chromosomes sont reproduits dans les nouvelles cellules. Il en résulte que, à l'exception des cellules sexuelles, toutes les cellules humaines possèdent une copie de ces 46 chromosomes et par conséquent, une copie du génome en entier.

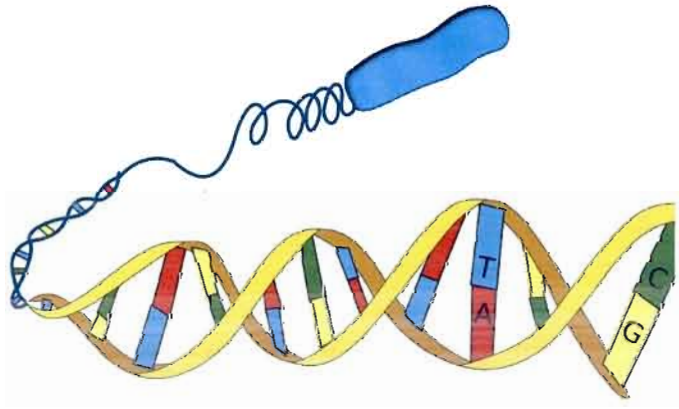
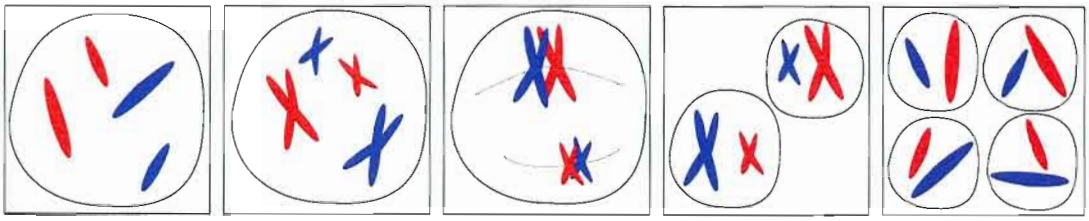


Figure 1.1 Un chromosome : l'enchevêtrement d'une molécule d'ADN formé de deux brins en forme d'hélice reliés par des paires de bases azotées (G, C, T et A).

Les cellules sexuelles sont particulières et ne sont pas créées de la même façon que les cellules traditionnelles. Si l'ovule et le spermatozoïde contenaient chacun 46 chromosomes, la cellule créée lors de leur union en contiendrait 92. Il y aurait alors deux possibilités ; soit le nombre de chromosomes chez l'être humain augmenterait de génération en génération, soit cette cellule éliminerait la moitié des chromosomes qu'elle possède. La première possibilité est peu réaliste, il faudrait avoir des cellules de plus en plus grosses pour contenir tous ces chromosomes, de plus, on peut imaginer la confusion des informations envoyées par tous ces chromosomes à la cellule. Tandis que si la cellule formée d'un ovule et d'un spermatozoïde élimine de manière aléatoire les 46 chromosomes en trop, elle pourrait, par exemple, éliminer tous les chromosomes transmis par la mère ou tous ceux transmis par le père. L'enfant ainsi créé serait alors une copie génétiquement conforme à l'un de ses parents. Cette possibilité ne favoriserait pas la diversité chez l'être humain.

En réalité, ce sont les cellules sexuelles qui contiennent moins, exactement la moitié moins, de chromosomes. Chez toutes les espèces sexuées, les chromosomes viennent par paire, chaque paire étant formée d'un chromosome venant du père et un autre venant de la mère. Les chromosomes d'une même paire sont dits *homologues*. Ils sont de même longueur et agissent sur les mêmes fonctions de l'organisme. Par exemple, on sait que l'information sur le groupe sanguin d'un individu se situe sur une certaine paire de chromosomes, en fait chacun des chromosomes de la paire possède au même endroit de l'information sur le groupe sanguin et la combinaison de ces informations détermine le groupe sanguin de l'individu. Nous allons voir plus en détail comment s'expriment et se combinent ces informations dans la prochaine partie de ce chapitre. Notons qu'une cellule composée de paires de chromosomes homologues est appelée cellule *diploïde*. De plus, les paires de chromosomes homologues sont numérotées de 1 à 22 chez l'être humain, la dernière paire de chromosomes est constituée des chromosomes sexuels, notés X et Y.

Voyons maintenant plus en détail le phénomène de division cellulaire des cellules sexuelles nommé *méiose*. Ce phénomène est illustré par la figure 1.2, où, pour simplifier, nous suivons deux paires de chromosomes de longueurs différentes ; la couleur rouge



(a) Avant la méiose (b) Réplication (c) Appariement (d) 1^e division (e) 2^e division

Figure 1.2 Étapes de la méiose. On suit ici une cellule contenant deux paires de chromosomes, ceux de même longueur sont homologues. Les chromosomes rouges sont ceux hérités de la mère et les bleus du père.

représente les chromosomes hérités de la mère, le bleu ceux hérités du père. Bien que les chromosomes viennent par paire, ils ne sont pas reliés entre eux et ils évoluent de manière indépendante à l'intérieur de la cellule (figure 1.2(a)). Avant toute division cellulaire, les chromosomes doivent être reproduits ; ce phénomène, nommé *réplication*, commence près du centre du chromosome (plus précisément au *centromètre*) et se déplace vers ses deux extrémités de manière indépendante. Le chromosome ainsi créé demeure attaché, pour l'instant, à l'original par le centromètre. Le chromosome et sa copie ressemblent alors à un X (figure 1.2(b)). Puis, il se produit un appariement des chromosomes homologues (des paires de chromosomes)(figure 1.2(c)). Cet appariement est important puisque durant celui-ci se produit le phénomène d'*enjambement* (ou de *recombinaison*) permettant de créer une plus grande variété génétique, nous y reviendrons. Suite aux enjambements se produit une première division cellulaire, les chromosomes homologues sont alors séparés en deux cellules (figure 1.2(d)). Après quoi, une deuxième division cellulaire a lieu ; celle-ci permet de diviser les chromosomes originaux de leurs copies. On obtient, après une méiose, quatre cellules filles, contenant chacune 23 chromosomes, on dit qu'il s'agit de cellules *haploïdes* (figure 1.2(e)). Nous nommons ces cellules les *gamètes*.

Revenons maintenant aux enjambements qui se produisent lors de l'appariement des chromosomes homologues. Le fonctionnement exact de ce phénomène demeure encore un mystère pour les biologistes, mais on sait que lors de la réunion des chromosomes

homologues certaines parties de ces chromosomes sont échangées entre les deux membres de la paire. Les deux chromosomes homologues résultants sont donc formés des mélanges des chromosomes paternel et maternel. La figure 1.3 illustre ce phénomène. L'endroit où se produit un enjambement est aléatoire. Tout au long de ce mémoire, nous allons parler de recombinaison génétique lorsque nous ferons référence aux enjambements.

En résumé, deux phénomènes de la méiose permettent de créer des variations génétiques : l'un d'eux est la recombinaison. De plus, un mélange des chromosomes paternels et maternels a aussi lieu lors de la première division cellulaire. Bien que durant cette division les chromosomes homologues se retrouvent à l'intérieur de différentes cellules, le choix des chromosomes allant dans chacune d'entre elles se fait aléatoirement. Par exemple, suite à la première division cellulaire, une des cellules peut contenir le chromosome noté 1 d'origine maternel et contenir le chromosome noté 2 d'origine paternel tandis que l'autre cellule contiendra le chromosome 1 paternel et le chromosome 2 maternel (la figure 1.2(d) est un exemple). Par conséquent, sans tenir compte des enjambements, il y a, chez l'être humain, 2^{23} (soit 8 388 608) cellules sexuelles (haploïdes) différentes possibles. Si de plus, nous ajoutons le phénomène de recombinaison génétique, il y a en réalité une immensité de possibilités. C'est, entre autres, ce qui explique les différences que peuvent avoir des frères et soeurs.

1.2 Gènes, allèles, transmission de l'information

Nous avons vu que toute l'information nécessaire pour faire fonctionner le corps humain d'un individu est contenue dans son génome. La moitié de son génome lui provient de son père et l'autre de sa mère. Nous allons maintenant voir la façon dont cette information s'exprime. Bien que visuellement nous pouvons observer beaucoup de différences entre deux êtres humains, ces deux personnes partagent en fait environ 99,9% de leurs génomes (Collins et Mansoura, 2001). En réalité, ce nombre n'est pas si farfelu ; on peut imaginer facilement la grande quantité d'informations nécessaires pour faire fonctionner un corps humain : faire battre un coeur, créer les vaisseaux sanguins, faire fonctionner un cerveau. Il s'agit là d'activités qui ne varient pas (ou presque) d'une

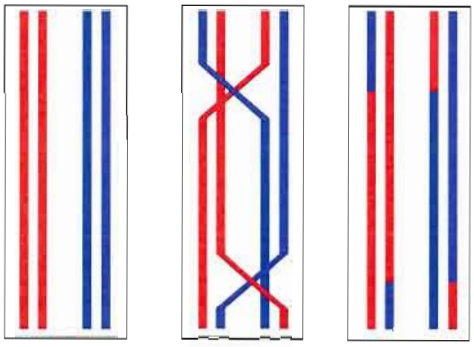


Figure 1.3 Étapes de la recombinaison. De gauche à droite, les chromosomes homologues à la suite de la réplication (non reliés par leurs centromètres), puis deux événements de recombinaisons ont lieu et finalement les chromosomes résultants.

personne à une autre. Mais des différences existent bel et bien entre individus d'une même espèce et se manifestent sur le génome. Nous allons voir de quelle façon à l'aide d'exemples.

Il est évident que certaines des différences entre deux individus peuvent être causées par des facteurs environnementaux, par exemple un excès de poids peut être le résultat d'une malnutrition et d'un manque d'exercice. Mais plusieurs particularités nous définissant sont génétiques, et dépendent de l'information contenue dans notre génome. En génétique, ces particularités biologiques se nomment *caractère*. Par exemple, la couleur naturelle de nos cheveux est un caractère, de même que notre groupe sanguin. À l'origine, un gène représentait, par définition, l'information sur un chromosome influençant un caractère. La réalité est plus complexe : certains caractères sont en fait influencés par plusieurs gènes. Un *gène* est composé d'un ensemble de paires de bases consécutives qui ensemble transmettent de l'information sur une tâche bien précise à la cellule. Un emplacement exact de paires de bases sur un chromosome se nomme *locus* (loci au pluriel). Le locus d'un gène fera référence à l'emplacement précis et physique des paires de bases le formant. Pour un même gène les informations transmises peuvent différer : nous n'avons

pas tous la même couleur de cheveux, ni le même groupe sanguin. On nomme *allèle* les différentes possibilités d'informations transmises par un gène, les allèles sont différentes versions d'un même gène. En simplifiant, on pourrait dire que pour le gène de la couleur des cheveux, il y a l'allèle brun, l'allèle blond et l'allèle roux. On se souviendra que les chromosomes viennent en paire, on en déduit que les gènes aussi et par conséquent deux allèles déterminent un caractère influencé par une paire de gènes (nous parlerons des gènes au singulier lorsqu'en réalité nous faisons référence à une paire de gènes).

Prenons l'exemple du groupe sanguin (sans tenir compte du groupe Rhésus ; Rh+ et Rh-), on observe quatre groupes sanguins : O, A, B et AB. Chacun de ces groupes est un *phénotype*, c'est-à-dire le résultat observé de l'action d'un gène. Au gène du groupe sanguin, il y a trois allèles possibles : A, B et O. Puisqu'un individu a une paire d'allèles pour le gène du groupe sanguin, on en déduit qu'il y a 6 paires distinctes possibles. On nomme la paire d'allèles qu'un individu possède pour un gène en particulier son *génotype*. Si le phénotype est ce que l'on observe d'un caractère, le génotype est plutôt ce qui est inscrit dans notre génome concernant ce caractère. La différence entre le nombre possible de phénotypes et de génotypes s'explique par le type d'allèle. Il y a des allèles *dominants* et des allèles *récessifs*. Lorsqu'un allèle récessif est en présence d'un allèle dominant, c'est l'allèle dominant qui s'exprime ; c'est comme si l'allèle dominant parle tellement plus fort que la cellule n'entend plus du tout ce que l'allèle récessif essaie de lui dire. L'allèle O du groupe sanguin est récessif et les allèles A et B sont dominants. Une personne de groupe sanguin A peut donc avoir deux génotypes différents, soit OA ou AA. Si par contre nous savons qu'une personne est de groupe sanguin O nous pouvons en déduire que ses deux parents lui ont transmis un allèle O, du fait que l'allèle O est récessif. Puisque A et B sont tous deux dominants, il se produit un phénomène de *codominance* : une personne ayant ces deux allèles sera du groupe AB. On dit qu'une personne est *homozygote* à un gène particulier si elle possède deux fois le même allèle à ce gène, sinon on dit qu'elle est *hétérozygote*.

En résumé, un individu reçoit pour un gène un allèle de sa mère et un autre de son père. Ces deux allèles forment son génotype au gène en question. En supposant

qu'un caractère précis soit déterminé que par les allèles à ce gène, si les deux allèles sont identiques, l'individu est homozygote et son phénotype est déterminé par cet allèle. Si les deux allèles sont différents, l'individu est hétérozygote et son phénotype dépendra de l'allèle dominant. Un individu transmet de façon aléatoire un de ses allèles à son enfant. En observant plusieurs individus reliés entre eux, il est possible, à l'aide de ces règles de transmission, de déduire leurs génotypes à partir de leurs phénotypes. La figure 1.4 représente un arbre généalogique (ou lignage), où une femme est représentée par un cercle et un homme par un carré, c'est la symbolisation usuelle en génétique. Un trait horizontal relie un homme et une femme ayant eu des enfants ensemble. Chaque individu est identifié par un numéro. Ici, l'exemple concerne le gène du groupe sanguin ; le phénotype d'un individu est inscrit à l'intérieur de la forme le représentant. On remarque que les individus 1 et 2 ont eu trois enfants (3, 4 et 5), dont deux de sexe féminin. Les individus 5 et 6 ont eu un enfant de sexe masculin (7). Les phénotypes sont connus pour les individus de 3 à 7 ; sommes-nous en mesure, pour cet exemple, de déduire leurs génotypes et ceux des individus 1 et 2 ? Voici ce que nous pouvons déduire :

1. Le phénotype de 7 est O $\Rightarrow$ son génotype est OO (puisque O est récessif).
2. Le phénotype de 6 est B, mais du point 1 nous savons qu'il a transmis l'allèle O à 7 $\Rightarrow$ le génotype de 6 est OB.
3. De manière analogue on déduit que le génotype de 5 est OA.
4. Puisque le phénotype de 3 est O $\Rightarrow$ son génotype est OO.
5. Le phénotype de 4 est AB $\Rightarrow$ son génotype est AB.
6. Par le point 4, on déduit que les individus 1 et 2 ont chacun un allèle O.
7. Maintenant, puisque le phénotype de 4 est AB et celui de 5 A $\Rightarrow$ les génotypes des parents sont OA et OB.

Savoir le groupe sanguin de l'un d'eux nous permettrait de connaître qui de la mère ou du père est de génotype OA et qui est OB. Nous verrons au chapitre suivant comment ces déductions nous permettent de calculer des probabilités afin d'évaluer la vraisemblance de la position d'un gène.

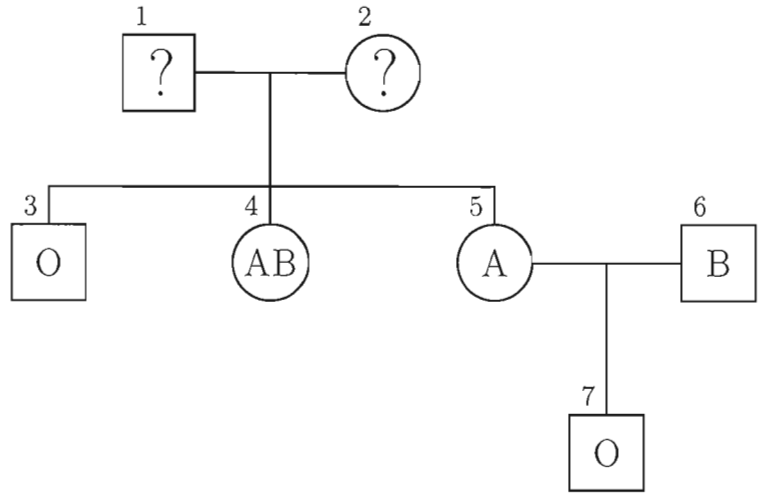


Figure 1.4 Exemple d'un arbre généalogique pour le gène du groupe sanguin. Une femme est représentée par un cercle et un homme par un carré. À l'intérieur des formes, on retrouve le phénotype de l'individu pour le gène du groupe sanguin. Les individus 1 et 2 sont les parents des individus 3 à 5, l'individu 7 est l'enfant des individus 5 et 6.

1.3 Mutations, unités de mesure et séquence de marqueurs

Fréquemment, dans les journaux, on annonce la découverte d'un gène causant une certaine maladie. Mais avoir découvert un gène causant un caractère signifie avoir réussi à estimer sa position sur le génome : on dit qu'on a réussi à le cartographier. La cartographie génétique est un domaine s'intéressant à la position des gènes sur les chromosomes ; ce domaine utilise des outils statistiques pour y arriver. Avant de décrire ces outils au chapitre suivant, il nous reste quelques termes génétiques à définir.

Lors de la méiose et de la duplication des chromosomes, il peut se produire des erreurs : une paire de bases azotées peut être oubliée, reproduite deux fois ou simplement mal reproduite. Nous appelons ces erreurs des *mutations*. Les mutations peuvent être sans conséquence, mais elles peuvent être également à l'origine de la création d'un nouvel allèle pour un gène. Les conséquences d'une mutation peuvent être, par exemple, d'empêcher la production d'une certaine protéine. Si la mutation est présente une seule fois, l'allèle

« sain » de ce gène peut produire en quantité suffisante la protéine, et l'organisme fonctionnera comme à l'habitude. Mais si deux allèles mutants sont présents, la protéine peut ne pas être produite et certaines cellules ne peuvent pas accomplir leurs tâches. L'individu peut alors être atteint d'une *maladie génétique*. Le phénotype pour un gène lié à une maladie est : atteint de la maladie (*cas*) ou non atteint de la maladie (*contrôle*). Certaines maladies ou certains caractères sont causés par plusieurs gènes, on parle alors de caractères *polygéniques*. Le sujet de ce mémoire est l'adaptation d'une méthode de cartographie génétique nommée MapArg (Larribe, Lessard et Schork, 2002 ; Larribe et Lessard, 2008), aux caractères polygéniques.

Pour arriver à positionner des gènes sur les chromosomes et éventuellement créer une carte génétique, il nous faut une unité de mesure de position. En génétique, il existe deux unités de mesure généralement utilisées. La première est simple et utilise les paires de bases azotées. Cette unité de mesure est physique, elle compte le nombre de paires de bases situées entre deux gènes. Elle est notée bp, de l'anglais « base pairs ».

La deuxième unité de mesure utilisée est basée sur les travaux de Thomas Hunt Morgan et de son équipe au début du 20^e siècle. Elle repose sur la recombinaison ; un événement de recombinaison est observé entre deux gènes s'ils ne proviennent pas du même chromosome parental. C'est-à-dire, s'il y a eu un nombre impair d'enjambements entre ces gènes lors de l'appariement des chromosomes homologues. De plus, si une recombinaison entre deux gènes par 100 méioses est observée, on dit qu'il y a un *taux de recombinaison* de 1% entre ces gènes. Morgan a étudié plusieurs caractères de la drosophile (mouche à fruits) et a réussi à cartographier certains de ses gènes. Contrairement à Mendel, qui a étudié des caractères des pois se trouvant sur des chromosomes différents, certains des caractères étudiés sur la drosophile se trouvent sur les mêmes chromosomes. Morgan a remarqué que ces caractères n'étaient pas transmis aléatoirement aux descendants, le taux de recombinaison entre les gènes concernés était inférieur à 50%. Si la transmission avait eu lieu de manière parfaitement aléatoire, il y aurait eu une chance sur deux que les allèles d'un même chromosome soient transmis ensemble (Almgren et al., 2003). En réalité, Morgan a découvert que plus deux gènes sont situés près l'un de l'autre,

moins il y a de recombinaison entre ces deux gènes. Et c'est en utilisant les taux de recombinaison entre les gènes que l'équipe de Morgan a réussi à ordonner les gènes étudiés sur les chromosomes de la drosophile. Un taux de recombinaison de 1% entre deux gènes, équivaut à une distance de 1 *centimorgan* (cM). L'équivalence entre les paires de bases et le centimorgan dépend de plusieurs facteurs, dont le sexe et l'espèce. Chez l'être humain, 1 cM équivaut approximativement à 1 mégabase (1 million de paires de bases).

Les méthodes de cartographie génétique actuelles utilisent des *marqueurs génétiques* pour obtenir de l'information sur les variations de l'ADN à certaines positions. Il s'agit de gènes ou de morceaux de séquences d'ADN ayant une position connue et possédant des variations (différents allèles) dans la population. Il existe différents types de marqueurs génétiques ; par exemple, on peut retrouver sur le génome de courtes séquences d'ADN, où une seule paire de bases varie dans la population. C'est-à-dire que pour cette courte séquence de paires de bases, tous les individus de la population ont exactement la même séquence à l'exception d'une paire de bases bien précise, on utilisera cette paire de bases comme marqueur génétique. Ces marqueurs se nomment *polymorphisme nucléotidique simple* (SNP, de l'anglais « single nucleotide polyomorphisms », prononcé « snip »). En général, il n'y a que deux allèles possibles par SNP. Nous utiliserons ce type de marqueur.

Les échantillons utilisés en cartographie génétique sont composés de séquence de marqueurs, c'est-à-dire d'une suite de plusieurs marqueurs ordonnés sur un bout de chromosome. Pour chaque individu de l'échantillon, il est possible de savoir quels allèles il possède à chacun de ces marqueurs. La figure 1.5 illustre comment, à partir d'une séquence d'ADN, il est possible d'extraire des marqueurs et de déterminer la distance entre ceux-ci. Habituellement, les séquences de marqueurs sont supposées indépendantes, et non pas en paires comme les chromosomes, ceci simplifie les méthodes de cartographie. Lorsqu'un seul chromosome est considéré, ou seulement une séquence de marqueurs extraite de ce chromosome est considérée, on parle alors d'*haplotype*. Par conséquent, plusieurs méthodes de cartographie génétique utilisent des échantillons d'haplotypes (McPeek et Strahs, 1999 ; Morris, Whittaker et Balding, 2000 ; Zöllner et Pritchard, 2005).

C'est aussi le cas de la méthode MapArg, bien que Gabrielle Boucher présente dans son mémoire de maîtrise (2009) une adaptation à la réalité diploïde.

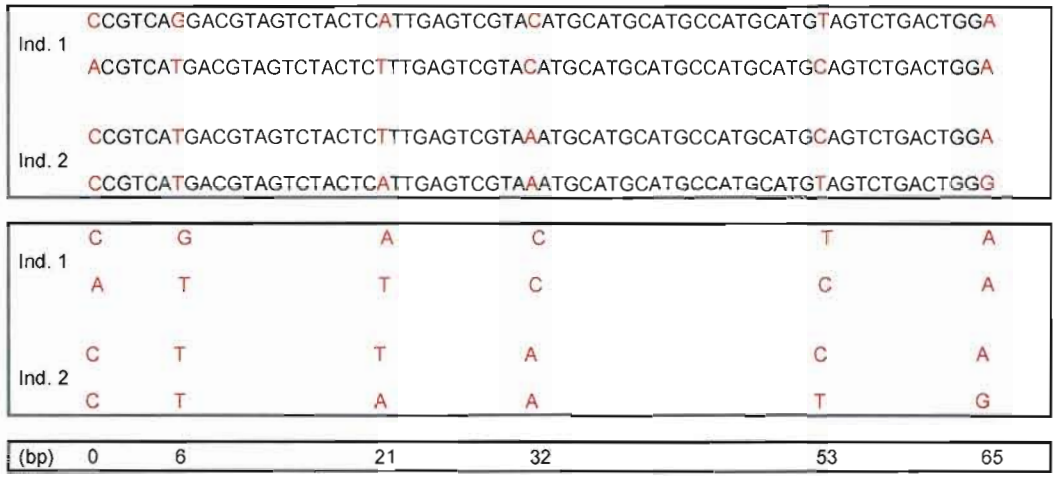


Figure 1.5 Séquences de marqueurs : courtes séquences d'ADN formées de 65 paires de bases. Les différences entre ces séquences sont utilisées comme marqueurs. La distance entre ces marqueurs est mesurée en paires de bases.

CHAPITRE II

CARTOGRAPHIE GÉNÉTIQUE

La *coségrégation* de deux allèles situés sur un même chromosome signifie qu'ils ont été transmis ensemble à un gamète lors de la méiose. On a vu au chapitre précédent que Thomas Hunt Morgan a découvert, lors de ses études sur la drosophile, que plus deux gènes sont situés près l'un de l'autre, moins il est probable qu'une recombinaison ait lieu entre ces deux gènes, et donc plus il y a de chance de coségrégation. La cartographie génétique est basée sur cette observation, et est donc fortement liée à la probabilité de recombinaison entre les gènes. Dans le cas le plus simple, si deux personnes sont affectées par un certain caractère génétique, alors non seulement elles partageront le même allèle pour ce caractère, mais elles partageront aussi, sous certaines conditions, les mêmes allèles pour des gènes avoisinant le gène associé au caractère étudié.

Il est possible de regrouper en deux grandes familles les méthodes de cartographie génétique en fonction du type d'échantillon qu'elles utilisent. En analyse de liaison (le terme anglais « linkage » est aussi employé en français), l'échantillon est composé de plusieurs groupes indépendants d'individus reliés entre eux (on parlera de famille). Tandis que les méthodes dites d'association sont basées sur le déséquilibre de liaison et utilisent plutôt un échantillon d'individus, de cas et de contrôles, non reliés entre eux.

Dans ce chapitre, nous présenterons sommairement ces deux groupes de méthodes de cartographie génétiques, leurs caractéristiques et leurs limitations. Ensuite, nous présenterons le processus de coalescence et une de ses généralisations, le graphe de recom-

binaison ancestral, qui permettent de simuler la généalogie d'un échantillon d'individus non reliés. La méthode de cartographie génétique MapArg, que nous souhaitons adapter, utilise le graphe de recombinaison ancestral ainsi que le déséquilibre de liaison pour estimer la position d'un caractère causant une maladie.

2.1 Méthodes de cartographie génétique

Nous avons choisi de ne pas présenter dans ce chapitre une méthode de cartographie génétique en particulier, mais plutôt de présenter les fondements et les outils utilisés par les deux groupes de méthodes. Ainsi, les particularités de la méthode de cartographie MapArg pourront être mises en contexte. Le lecteur intéressé pourra obtenir des informations complémentaires dans les articles de Lander et Schork (1994), de Olson *et al.* (1999) et de Robert C. Elston (2000), ainsi que dans les notes de Almgren *et al.* (2003).

2.1.1 L'analyse de liaison

Les premières études en génétique ont été effectuées sur des espèces pour lesquelles il est possible de maîtriser la reproduction. Par exemple, plusieurs études ont été faites avec la drosophile, cette espèce de mouche ayant des attributs facilitant son élevage en laboratoire : temps de gestation court et production de plusieurs descendants. On peut choisir deux mouches en fonction de leurs caractéristiques génétiques et les accoupler en vue d'obtenir plusieurs descendants, ce qui permet d'étudier la transmission de certains gènes et aussi d'estimer la distance entre ces gènes. En choisissant bien les drosophiles accouplées, il est possible d'observer s'il y a eu recombinaison entre deux gènes chez les descendants et donc de calculer le nombre de descendants pour lesquels une recombinaison a eu lieu. En répétant l'expérience plusieurs fois, on obtient une estimation assez juste du taux de recombinaison θ entre ces deux gènes et ainsi on sait s'ils sont situés près l'un de l'autre. Il est difficile de mener de telles expériences pour l'espèce humaine ; on ne peut qu'observer le résultat des reproductions ayant eu lieu. Il a donc été nécessaire de développer de nouvelles méthodes statistiques pour déterminer s'il y a évidence d'association, ou non, entre un marqueur et une maladie. Les méthodes d'analyse de

liaison sont basées sur l'utilisation de familles dont nous connaissons les généalogies, c'est-à-dire les liens reliant les individus entre eux.

Il y a association entre deux gènes lorsque le taux de recombinaison entre ces gènes (noté θ en analyse de liaison) est inférieur à un demi ($\theta < 1/2$). Lorsqu'on cherche à cartographier un gène causant un certain caractère, on souhaite rejeter l'hypothèse nulle $H_0 : \theta = 1/2$ contre l'hypothèse alternative $H_1 : \theta < 1/2$. Notons qu'une valeur supérieure à $1/2$ pour θ n'a aucun sens d'un point de vue génétique. Pour tester cette hypothèse, on a recours au rapport de vraisemblance. Nous allons présenter premièrement un exemple très simple (peu réaliste) pour démontrer une version de l'analyse de liaison. Ensuite, nous ferons un survol rapide des autres versions existantes.

Prenons l'exemple de la famille illustré par la figure 2.1 (on reconnaît le style de la figure 1.4 présentée au chapitre 1 à la page 11). Cet exemple est tiré des notes de Almgren (2003) et nous permettra de comprendre comment il est possible, à l'aide de l'analyse de liaison, de déterminer si un marqueur de position connu se situe près d'un gène influençant un caractère. Ici, chacun des membres de la famille a été génotypé à un marqueur, c'est-à-dire que leurs allèles à ce marqueur ont été déterminés à l'aide de test biologique. Les allèles possibles sont notés par 1, 2, 3 et 4. Nous connaissons de plus, les phénotypes (atteint ou non) pour un certain gène influençant un caractère ; une personne affectée est représentée par une forme pleine (on se souviendra qu'un carré représente un homme et un cercle une femme). Sous certaines conditions, il est possible de savoir s'il y a eu recombinaison entre le marqueur et le gène influençant le caractère.

Nous devons tout d'abord déterminer les génotypes du caractère étudié à partir des phénotypes des individus. Pour ce faire, le nombre d'allèles possibles à ce gène doit être connu, ou du moins supposé connu, ainsi que la *fonction de pénétrance* nous indiquant la probabilité d'observer un certain phénotype en fonction du génotype. Il est usuel de présumer que le gène causant le caractère possède deux allèles ; pour notre exemple nous noterons l'allèle sain par **a** et l'allèle «mutant» par **A**. Il nous reste maintenant à déterminer la fonction de pénétrance : nous présumerons que l'allèle **A** est dominant, et

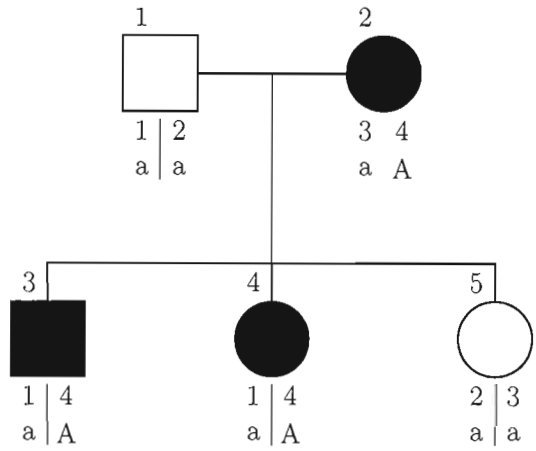


Figure 2.1 Généalogie de cinq individus reliés et génotypés à un marqueur ayant pour allèles possibles 1, 2, 3 et 4. La fondatrice et deux des enfants sont affectés par un certain caractère. En présumant que le caractère est influencé par un seul gène avec allèles possibles **a** (allèle sain) et **A** (allèle mutant), et qu'un seul allèle mutant suffit pour être affecté, on peut déduire les génotypes des individus à ce gène. Lorsque la phase est connue, un trait sépare les allèles appartenant au même chromosome. Par exemple, l'individu 3 a les allèles 1 et **a** sur le même chromosome et les allèles 4 et **A** sur le chromosome homologue.

donc qu'un seul allèle **A** suffit pour être affecté. Nous supposerons de plus qu'il n'y a pas de *phénocopie*, c'est-à-dire qu'aucun facteur autre que l'allèle **A** ne peut causer le caractère ; par exemple, une personne de génotype **aa** ne peut pas être affectée par le caractère à cause d'un facteur environnemental. Il est maintenant possible de déterminer les génotypes des individus de l'exemple de la figure 2.1.

Les individus 1 et 5 ont le génotype **aa**, puisqu'ils ne sont pas affectés. La mère affectée, doit être de génotype **aA**, puisqu'elle a transmis l'allèle **a** à sa fille 5 (non affectée). Les enfants 3 et 4 ont le génotype **aA**, puisqu'ils sont affectés et que leur père ne peut que leur avoir transmis l'allèle **a**. Pour savoir s'il y a eu recombinaison, il faut connaître la *phase*, c'est-à-dire connaître quels allèles des deux gènes étudiés se retrouvent

sur le même chromosome ; lesquels ont été hérités de la mère et lesquels du père (on séparera ces couples d'allèles par un /). Si les phases des parents sont connues, ainsi que celles d'un de leurs descendants, il est possible de savoir s'il y a eu recombinaison chez ce descendant. Par exemple (attention cet exemple n'est pas celui de la figure 2.1), si les parents ont les phases suivantes **1a/2A** et **3A/4a**, et que leur descendant a la phase **2a/3A** alors, dans ce cas particulier, il y a eu une recombinaison pour le chromosome hérité du premier parent, car celui-ci a transmis l'allèle **1** provenant d'un de ses chromosomes et l'allèle **a** provenant de son chromosome homologue. Mais il n'y a pas eu de recombinaison pour le chromosome hérité du deuxième parent, car les allèles hérités par le descendant proviennent du même chromosome appartenant au parent.

Revenons maintenant à l'exemple de la figure 2.1. La phase du père (individu 1) est simplement **1a/2a**, puisqu'il est homozygote au gène étudié, ses deux parents lui ont transmis l'allèle **a**. Les phases des enfants se trouvent assez facilement, puisque les parents sont tous deux hétérozygote au marqueur et n'ont pas les mêmes allèles, elles sont : individu 3 : **1a/4A**, individu 4 : **1a/4A** et individu 5 : **2a/3a**. Il n'y a que pour la mère que la phase est inconnue. Nous allons devoir compter le nombre de recombinaisons conditionnellement à la phase de la mère. Puisque rien n'est connu des parents (on dit qu'ils sont les *fondateurs* car ils n'ont pas de parents dans la généalogie), nous présumerons que les deux phases possibles sont équiprobables.

La fonction de vraisemblance du paramètre θ est de la forme de la fonction de masse d'une binomiale. À chaque méiose observable, il y a recombinaison avec probabilité θ. Le nombre de recombinaisons ayant eu lieu dans notre généalogie suit donc une binomiale de paramètre (n,θ), où n est le nombre de méioses informatives. Une méiose est informative s'il est possible de déterminer s'il y a eu recombinaison ou pas. Dans notre exemple, le père est homozygote au gène étudié, il est donc impossible de savoir s'il y a eu recombinaison, par conséquent, pour notre exemple, $n = 3$ (il s'agit des méioses associées à la mère). La forme générale de la fonction de vraisemblance $L(\theta)$ est :

$$L(\theta) = \binom{n}{r}\theta^r(1-\theta)^{n-r},$$

où r est le nombre de recombinaisons observées chez les méioses informatives.

La vraisemblance de notre exemple est conditionnelle à la phase de la mère, notons les différentes phases possibles ainsi : $P_1 = \mathbf{3a/4A}$ et $P_2 = \mathbf{3A/4a}$, la figure 2.2 illustre ces phases. On obtient la fonction suivante :

$$L(\theta) = \sum_{i=1}^{2} L(\theta|P_i)Pr(P_i) = \frac{1}{2}\left[\binom{3}{0}(1-\theta)^3 + \binom{3}{3}\theta^3\right].$$

Rappelons que les deux phases sont équiprobable, par conséquent $Pr(P_1) = Pr(P_2) = 1/2$. Notons que puisque $\theta \in [0{,}0.5]$, le maximum est atteint pour $\theta = 0$. Pour rejeter ou non l'hypothèse nulle $H_0 : \theta = 1/2$, on utilise le score LOD, noté $Z(\theta)$, équivalent au log du quotient de $L(\theta = \hat{\theta})$ sur $L(\theta = 1/2)$. Pour notre exemple on obtient :

$$Z(\theta = 0) = \log\left(\frac{L(\theta = 0)}{L(\theta = 0.5)}\right) = \log\left(\frac{1+0}{1/8+1/8}\right) = \log(4) = 0.6.$$

Il est généralement accepté qu'un score LOD supérieur à 3 permet de rejeter l'hypothèse nulle, et qu'un inférieur à -2 permet d'accepter l'hypothèse nulle (Lander et Schork, 1994 ; Olson, Witte et Elston, 1999). Une des propriétés du score LOD est d'être additif, c'est-à-dire, que nous pouvons utiliser plusieurs familles indépendantes et additionner leurs scores LOD pour obtenir le résultat final. Dans notre exemple, cinq autres familles ayant un même score LOD que celui calculé permettrait de rejeter l'hypothèse nulle et de conclure qu'il y a association entre le gène influençant le caractère et le marqueur, c'est-à-dire qu'ils sont situés près l'un de l'autre.

L'exemple de la figure 2.1 en est un très simple et peu réaliste. La famille était composée de peu d'individus, le modèle choisi pour le caractère était simple (la fonction de pénétrance aussi), et la phase d'un seul individu était inconnue. Très rapidement, pour des modèles plus réalistes, la fonction de vraisemblance se complique et des algorithmes de résolution à l'aide d'ordinateur s'avèrent indispensables. Il est aussi possible d'utiliser plusieurs marqueurs à la fois, dans ce cas, θ est un vecteur de longueur équivalant au nombre de marqueurs utilisé.

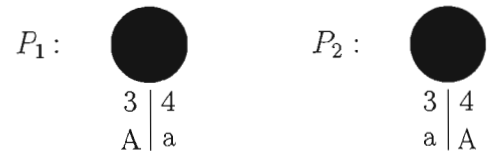


Figure 2.2 Les deux phases possibles de l'individu 2 de la généalogie de la figure 2.1 sont présentées.

Ce type d'analyse de liaison est basé sur un modèle ; le nombre d'allèles au gène étudié, ainsi que la fonction de pénétrance, ont été déterminés à l'avance. Il existe aussi des analyses de liaison pour lesquelles ces suppositions ne sont pas faites, on parle alors d'analyse de liaison sans modèle, ou non paramétrique. Ces méthodes sont basées sur le nombre d'allèles identiques par descendance.

Deux allèles sont dits *identiques par descendance* (IBD), s'ils proviennent du même ancêtre. Revenons sur l'exemple du groupe sanguin introduit au chapitre 1 en page 9, et supposons que deux parents ont comme groupe sanguin O, pour le père et AB, pour la mère, et que leurs deux enfants sont de groupe sanguin A. On se souviendra que les allèles A et B sont tous deux dominants, et que l'allèle O est récessif, on en déduit que les parents sont de génotype OO et AB. Il est alors clair que les enfants ont le génotype OA au gène du groupe sanguin et qu'ils partagent la même version de l'allèle A, alors ils ont un allèle identique par descendance. Ils ont aussi, tous deux, l'allèle O, mais on ne peut savoir s'ils possèdent la même version de cet allèle. Un des enfants peut avoir une copie de l'allèle O de son grand-père paternel et l'autre, l'allèle O de sa grand-mère paternelle.

L'idée de l'analyse de liaison non paramétrique est de regarder, chez un grand nombre de paires d'individus affectés par un caractère et ayant les mêmes parents (des frères et soeurs), s'ils ont un plus grand nombre d'allèles IBD que la normale, à un certain marqueur. Si c'est le cas, c'est que le marqueur en question est associé au gène causant le caractère.

Nous avons vu qu'un marqueur en association avec un certain gène signifie qu'il y a moins de recombinaison entre ces gènes et donc, plus de chance de coségrégation. Si on remarque qu'il y a plus d'allèles IBD à un marqueur chez des gens affectés par un caractère, c'est qu'il y a eu plus de coségrégation des allèles que de recombinaison.

Plusieurs tests utilisant les allèles IBD ont été développés. Ces méthodes s'avèrent souvent plus robustes que celles basées sur un modèle, puisqu'elles ne reposent pas sur des suppositions préalables. Biernacka et al. (2005) ont adapté la méthode de cartographie créée par Liang et al. (2001) à la cartographie simultanée de deux gènes causant un caractère, un peu comme nous souhaitons le faire avec MapArg. Cette méthode est basée sur un échantillon de paires d'individus et des marqueurs IBD. Ils font les suppositions suivantes : premièrement, ils supposent que le caractère est causé par l'effet combiné de deux gènes en association, ensuite le gène un est toujours situé à gauche du deuxième et pour leurs simulations chacun de ces gènes n'a que deux allèles possibles. En comparant leur adaptation à la méthode originale, ils arrivent à la conclusion que lorsque les résultats peuvent sembler ambigus, leur méthode peut apporter une meilleure précision dans l'estimation des positions des gènes.

Bien que l'analyse de liaison, dans ses différentes formes, ait réussi à trouver des régions de chromosome impliquées dans l'expression de certains caractères, cette méthode de cartographie génétique est tout de même limitée. Il est à noter que lors de la méiose, très peu de recombinaisons ont lieu sur un chromosome. Par conséquent, les régions trouvées avec l'analyse de liaison ne sont pas très précises. De plus, il est possible de trouver une association entre deux gènes dans une famille, mais qu'elle ne se retrouve pas dans la population. En raison de leurs limites, les analyses de liaison peuvent être utilisées pour estimer la position d'un gène sur l'ensemble du génome en utilisant très peu de marqueurs.

2.1.2 Étude d'association

Bien que l'analyse de liaison s'avère utile en cartographie génétique, il s'agit d'une méthode peu précise à cause du faible nombre de recombinaisons se produisant à chaque méiose. Elle ne permet pas de trouver le gène causant le caractère, mais trouve plutôt une région plus ou moins grande. Si un plus grand nombre de recombinaisons avait lieu lors d'une méiose, la région trouvée avec l'analyse de liaison serait plus courte, et l'estimation plus précise. Puisqu'il est impossible d'augmenter le nombre de recombinaisons, l'idée est plutôt de laisser passer le temps.

La figure 2.3 illustre la transmission, au cours des générations, d'un haplotype (en rouge) et d'un allèle causant un caractère (identifié par un astérisque). La première génération formée des enfants des fondateurs est illustrée, puis quelques individus de la n^e génération se trouvent au bas. L'haplotype en rouge a bien sûr été segmenté au

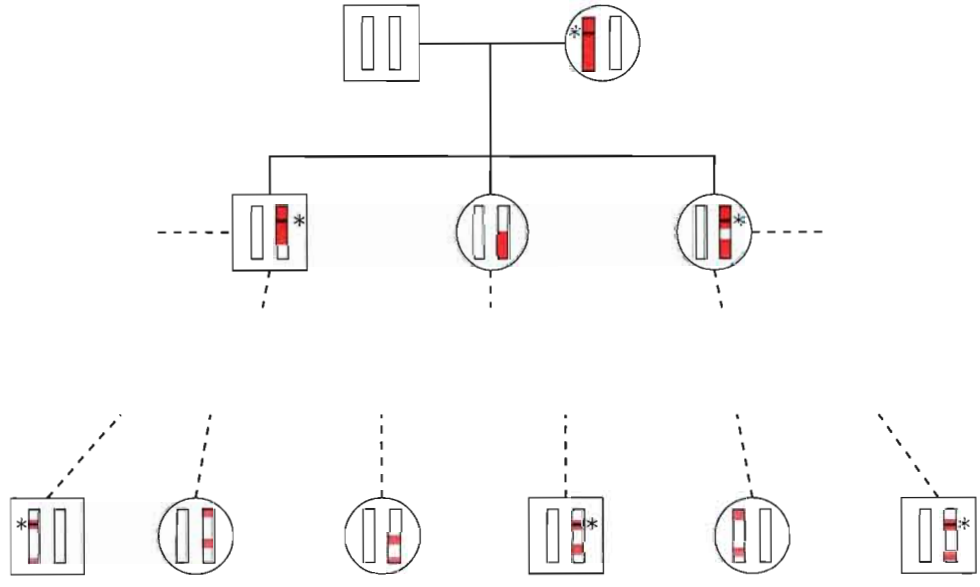


Figure 2.3 Transmission d'un haplotype sur plusieurs générations. On suit ici un haplotype en rouge transmit par la fondatrice. La première génération est illustrée, ainsi qu'une n^e au bas de la figure. L'haplotype suivit a été segmenté au gré des recombinaisons.

gré des recombinaisons, les individus de la dernière génération n'en possédant que des parcelles. Ceci entraîne que les individus de la dernière génération affectés par le caractère partagent une région beaucoup plus petite autour du gène le causant que ceux de la première génération par exemple. Si n est très grand, alors tous les haplotypes partageant encore des parcelles du fondateur apparaîtront indépendants les uns des autres.

Lorsque l'on choisit un échantillon d'individus non reliés entre eux, en réalité on choisit un échantillon de gens reliés, mais de façon lointaine. La façon dont ils sont reliés et l'histoire des recombinaisons ayant eu lieu nous sont toutefois inconnues. Si on trouve une association entre un marqueur et les gens affectés par un caractère d'un échantillon de cas et de contrôles indépendants, on sait que le marqueur sera beaucoup plus près du gène cherché que si on avait utilisé l'analyse de liaison puisqu'un plus grand nombre de recombinaisons a eu lieu. Lorsqu'une telle association est trouvée entre un marqueur et un gène qui sont effectivement situés près l'un de l'autre, on dit qu'ils sont en *déséquilibre de liaison (DL).*

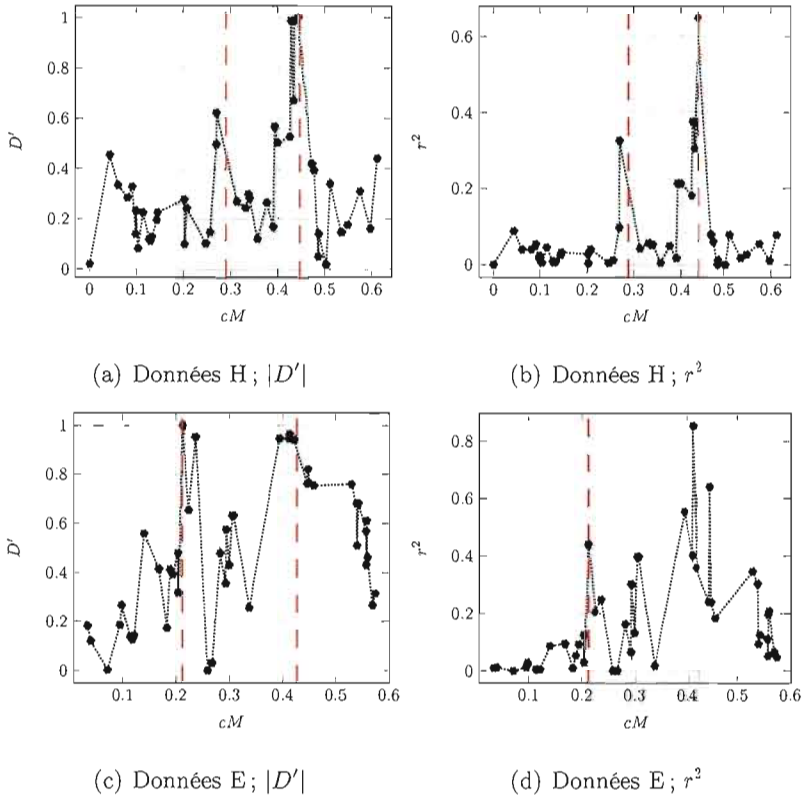
On utilise le terme déséquilibre de liaison par opposition au terme *équilibre de liaison.* En regardant seulement les individus d'une famille, deux gènes peuvent sembler en association, mais si on laisse le temps agir, au bout d'un certain nombre de générations, ces deux gènes atteindront l'équilibre de liaison, c'est-à-dire que la fréquence des différents haplotypes possibles sera seulement le produit des fréquences des allèles les composant. Par exemple, supposons que deux gènes possèdent chacun deux possibilités d'allèles : au premier gène il y a les allèles **a** et **A**, et au second **b** et **B**. Les haplotypes possibles sont alors : **ab**, **aB**, **Ab** et **AB**. Il y aura équilibre de liaison si $p_{ij} = p_i p_j$ pour tout i, j appartenant à $\{\mathbf{a}, \mathbf{A}, \mathbf{b}, \mathbf{B}\}$, où p_{ij} représente la fréquence de l'haplotype **ij** et p_i la fréquence de l'allèle **i**. Au niveau d'une famille, on peut trouver une association entre ces gènes. Ce qui peut signifier que l'on retrouvera, par exemple, plus souvent l'allèle **A** en présence de l'allèle **B**. Mais au niveau de la population, où davantage de recombinaisons ont eu lieu, cette association peut être brisée, et l'haplotype **AB** ne sera pas surreprésenté dans la population. Les deux gènes auront alors atteint l'équilibre de liaison.

Les méthodes de cartographie génétique que l'on appelle études d'association sont basées sur le déséquilibre de liaison. Elles utilisent un certain nombre de marqueurs couvrant une région d'un chromosome et testent, marqueur par marqueur, l'association entre le caractère étudié et le marqueur. Il existe différentes mesures pour tester cette association, mesures que nous utiliserons au dernier chapitre. Nous allons en présenter quelques-unes, la notation utilisée est celle de Nordborg et Tavaré (2002).

Si p_a représente la fréquence de l'allèle **a** et p_{ab} la fréquence de l'haplotype **ab**, les fréquences des allèles **A** et **B** sont alors $1 - p_a$ et $1 - p_b$. Une des mesure présentée est la valeur absolue de D normalisée (notée $|D'|$), où $D = p_{ab} - p_a p_b$ est une mesure d'écart à l'équilibre. La deuxième est $r^2 = D^2/(p_a(1-p_a)p_b(1-p_b))$, qui estime la corrélation entre les gènes A et B. Les valeurs obtenues pour ces deux mesures se situeront entre 0 et 1. Un seuil, tenant possiblement compte des tests multiples, est fixé pour rejeter ou non l'hypothèse d'équilibre (pas d'association) entre les deux marqueurs.

La figure 2.4 présente les résultats de ces deux mesures d'association pour deux ensembles de données (H et E qui seront présentés au chapitre 5). Dans ces ensembles, deux gènes causent le caractère, leurs positions sont représentées par des lignes pointillées rouges. Pour ces données, r^2 semble moins variable que $|D'|$ et estime que l'association est grande entre le caractère et un marqueur situé près de la position d'un des gènes cherchés. Certains exemples démontrent l'efficacité des études d'association, mais il n'en demeure pas moins que ces méthodes ne sont pas toujours fiables.

En effet, il a été montré (Nordborg et Tavaré, 2002) que les mesures d'association sont très variables. La figure 2.5 présente trois graphiques où l'association entre paires de marqueurs de positions connues est mesurée. Chaque point de couleur représente la valeur de l'association sur une échelle de 0 à 1 et la projection de ces points sur les deux axes représente la position des deux marqueurs pour lesquels l'association est mesurée. Les figures 2.5(a) et 2.5(b), proviennent d'une simulation utilisant le processus de coalescence avec recombinaison : processus stochastique permettant de simuler des généalogies et qui sera introduit à la prochaine section. Il y a donc eu des événements de recombinaison ;

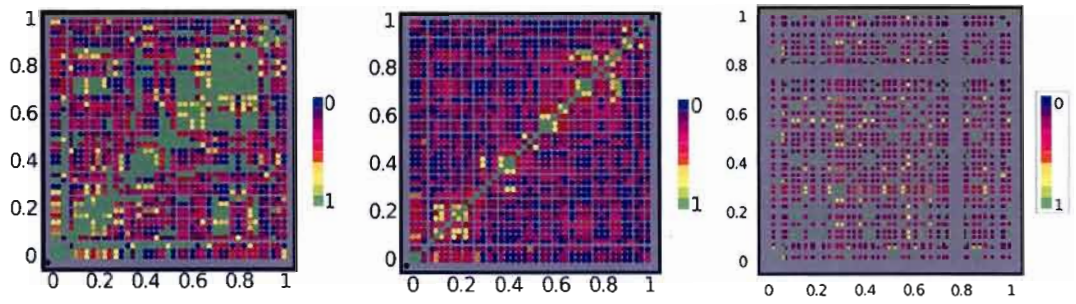


(a) Données H ; $|D'|$ (b) Données H ; r^2

(c) Données E ; $|D'|$ (d) Données E ; r^2

Figure 2.4 Exemple d'utilisation de mesure d'association à des fins de cartographie génétique. Deux bases de données ont été simulées (des détails sur ces simulations seront donnés au chapitre 5). On mesure l'association entre les marqueurs et le statut de la maladie. Sur l'axe des abscisses, on retrouve la position des marqueurs de la séquence. Chaque point correspond à l'association entre un marqueur génétique et la maladie.

la première utilise D comme mesure d'association et la seconde r. L'association due au déséquilibre de liaison entre des marqueurs voisins est détectée (la diagonale est large), mais on remarque aussi un certain bruit ; des marqueurs pourtant éloignés semblent en association. La figure 2.5(c) représente la mesure de r pour plusieurs marqueurs pris deux à deux. Les données ont été simulées dans ce dernier cas à l'aide du processus de coalescence, mais sans inclure des événements de recombinaison. Bien évidemment, chaque marqueur est en association avec lui-même (la diagonale), mais il y a encore un certain bruit. Puisqu'il n'y a pas eu de recombinaison, ces associations ne viennent



(a) Avec recombinaison $|D'|$ (b) Avec recombinaison $|r|$ (c) Sans recombinaison $|r|$

Figure 2.5 Mesure d'association entre paires de marqueurs. Les données des deux premiers graphiques (a) et (b) ont été simulées à l'aide du processus de coalescence avec recombinaison, et l'association entre les marqueurs a été estimée à l'aide de deux mesures d'association. Un deuxième ensemble de données a été simulé, sans recombinaison cette fois ; les résultats sont présentés sur le graphique (c) (tirées de Norborg et Tavaré (2002)).

pas du déséquilibre de liaison. C'est une des raisons qui expliquent pourquoi les études d'association ne sont pas toujours fiables.

Ces variabilités peuvent être dues à différents facteurs. La population utilisée n'était peut-être pas homogène. Deux mutations peuvent être apparues à peu près à la même époque et semblent alors être associées. Il ne faut pas oublier que lorsque l'on teste plusieurs marqueurs les uns après les autres, le danger de faux positif est plus grand et doit être pris en compte dans le choix du seuil de rejet.

2.2 Processus de coalescence

S'il était possible de connaître l'histoire des recombinaisons et mutations ayant eu lieu sur un bout de chromosome, pour un échantillon de cas et de contrôles, jusqu'à leur plus récent ancêtre commun (noté MRCA), et donc d'obtenir pour cet échantillon un graphique semblable à la figure 2.3 (page 23), il serait possible d'évaluer la vraisemblance de la position d'un gène causant un certain caractère. Il suffirait de comparer les sections de chromosomes que les personnes affectées partagent et que les contrôles n'ont pas, de

manière analogue au calcul de la vraisemblance pour des généalogies en analyse de liaison (voir section 2.1.1 page 19). Qu'on le veuille ou non, la généalogie d'un échantillon apporte de l'information non négligeable pour la quête des gènes affectant des caractères (Nordborg et Tavaré, 2002) ; comme cette généalogie est inconnue, nous allons la simuler. C'est pourquoi des méthodes de simulation de généalogie ont été créées. Dans cette section, nous présenterons le processus de coalescence, un processus stochastique, utilisé en génétique pour la simulation de généalogie. Pour faciliter la compréhension, nous présenterons tout d'abord le processus de coalescence simplifié, puis nous ajouterons les événements de mutation et finalement les événements de recombinaison. Plus d'informations sur le processus de coalescence et ses différentes généralisations peuvent être trouvées dans la littérature (Hein, Schierup et Wiuf, 2005 ; Nordborg, 2007).

2.2.1 Modèle de Wright-Fisher et processus de coalescence

Avant de simuler une généalogie, nous devons tout d'abord définir comment la population évolue et comment les accouplements ont lieu. Nous allons utiliser le modèle de Wright-Fisher pour l'évolution de la population. Selon ce modèle, la population est formée d'un nombre constant N d'individus non sexués, et les générations sont discrètes ; chaque individu a un seul parent dans la génération précédente. Comme un individu peut avoir plus d'un enfant, mais que la taille de la population demeure constante, intuitivement cela revient à ce que chaque enfant choisisse son parent, au hasard, parmi les N de la génération précédente. En fait, le nombre d'enfant qu'aura un individu, suit une binomiale de paramètres $(N, 1/N)$ et par conséquent un individu aura en moyenne 1 enfant.

La figure 2.6 illustre l'évolution d'une population de 10 individus sur 10 générations. Les individus d'une même génération sont représentés par les points alignés horizontalement, la génération la plus récente se situe au bas du graphique et un trait relie un individu à son parent de la génération précédente (celle au dessus). À droite, la généalogie de la génération la plus récente est surlignée. Un événement de *coalescence* se produit, du présent vers le passé, lorsque des individus d'une même génération ont le

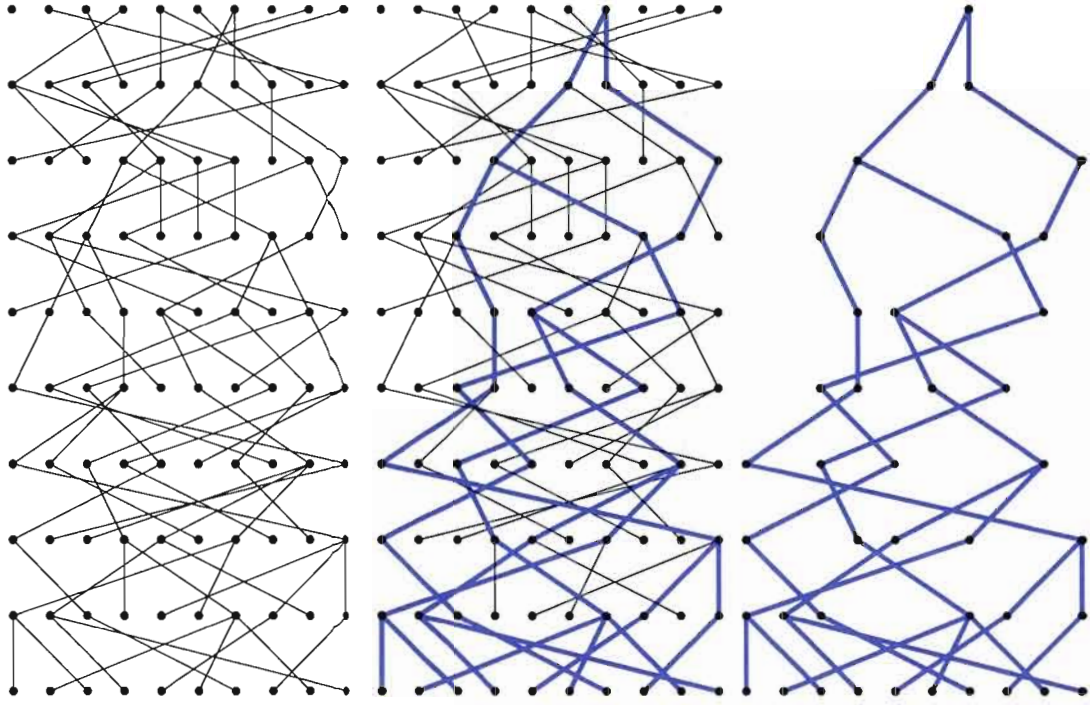
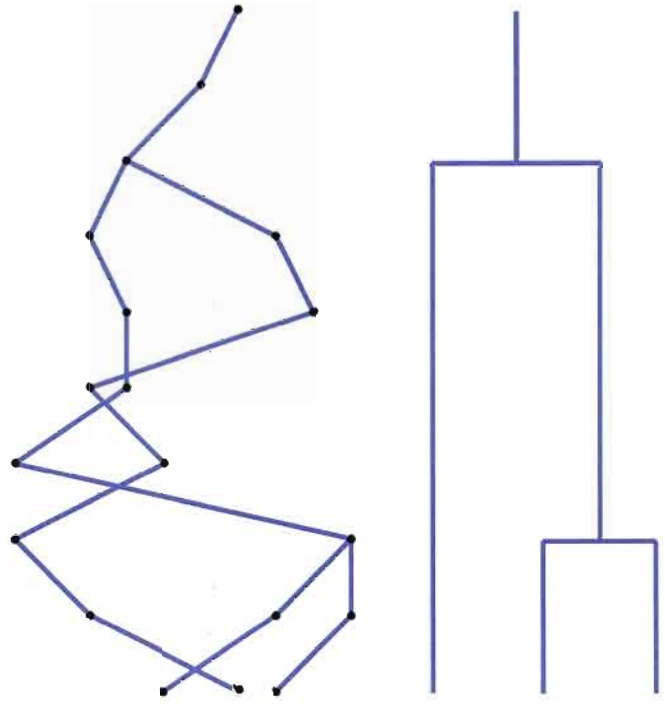


Figure 2.6 Exemple d'une généalogie selon le modèle Wright-Fisher. La figure (a) représente dix générations d'une population évoluant selon le modèle Wright-Fisher. Les individus d'une même génération sont représentés par les points alignés horizontalement, la génération la plus récente se situe au bas du graphique et un trait relie un individu à son parent de la génération précédente (celle au dessus). En (b), la généalogie de la plus récente génération est surlignée en bleue et ensuite isolée à la figure (c).

même ancêtre. Pour simuler une généalogie d'une population de taille N selon le modèle de Wright-Fisher, il est possible d'aller du passé vers le présent, en échantillonnant, à chaque génération, N parents avec remise parmi la population. Après avoir créé le nombre de générations voulu, il faut partir de la plus récente génération et remonter vers le passé pour trouver le MRCA de cette population (comme sur la figure 2.6(b)). Mais plus N est grand, plus cette façon de faire devient lourde. En réalité, nous allons utiliser un échantillon de n individus de la génération la plus récente, et nous intéresser à la généalogie de cet échantillon (comme illustré par la figure 2.7).



(a) (b)

Figure 2.7 Échantillon d'une généalogie et simulation à l'aide du processus de coalescence. À gauche, il s'agit d'un échantillon de taille 3 de la population représenté à la figure 2.6. À droite, la généalogie de l'échantillon a été démêlée et peut représenter une généalogie simulée à l'aide du processus de coalescence.

Regardons de plus près la généalogie de notre échantillon de la figure 2.6 : en faisant de l'ordre, nous obtenons le graphe présenté par la figure 2.7(b). Ce qui nous intéresse en réalité est le temps avant chaque événement de coalescence, et le temps avant l'atteinte du MRCA. Puisqu'ici, selon notre modèle, un parent transmet son matériel génétique intégralement (il n'y a pas de mutation ni de recombinaison), les étapes intermédiaires entre les événements de coalescences ne nous intéressent pas.

En regardant le problème dans le sens inverse, c'est-à-dire du présent vers le passé, il faudrait seulement, pour simuler une généalogie, estimer le temps avant les événements de coalescence et déterminer quelles lignées coalescent. C'est ce que permet de faire le processus de coalescence, découvert indépendamment par Kingman (1982), Hudson (1983)

et Tajima (1983). Cette méthode n'offre toutefois qu'une approximation d'une généalogie évoluant selon le modèle de Wright-Fisher puisqu'elle ne permet pas la coalescence de plus de deux lignées à la fois, mais elle s'avère juste lorsque N est grand.

Tout d'abord, regardons deux individus ; ils auront le même parent à la génération précédente avec probabilité $1/N$. Le premier choisi parmi les N disponibles, et le deuxième choisira le même avec probabilité $1/N$. On peut imaginer que pendant un certain temps leurs lignées resteront distinctes, mais éventuellement il y aura coalescence. Le temps (en nombre de génération) avant la coalescence de leurs lignées est distribué selon une loi géométrique de paramètre $1/N$. Le temps moyen avant que deux lignées coalescent est de N générations. En notant T_i le temps avant un événement de coalescence lorsqu'il y a i lignées, la probabilité que nos deux lignées coalescent après j générations est :

$$P(T_2 = j) = \left(1 - \frac{1}{N}\right)^{j-1} \frac{1}{N}.$$

Si on a plutôt k lignées, la probabilité que deux d'entre elles (fixées) coalescent à la génération précédente est aussi de $1/N$. Puisqu'il y a $\binom{k}{2}$ couples possibles, la probabilité que deux lignées coalescent est de $\binom{k}{2}/N$. Le temps avant d'observer un événement de coalescence suit approximativement une loi géométrique, mais cette fois-ci de paramètre $\binom{k}{2}/N$. Lorsqu'il y a k lignées, la probabilité qu'il y ait coalescence à la j^e génération est :

$$P(T_k = j) \approx \left[1 - \binom{k}{2}\frac{1}{N}\right]^{j-1} \binom{k}{2}\frac{1}{N}.$$

Il s'agit d'une approximation, puisque nous présumons que seulement deux lignées peuvent coalescer à la fois. Mais cette approximation est très près de la réalité. Revenons à notre exemple où $N = 10$ et suivons trois lignées. La probabilité qu'aucune d'entre elles ne coalesce à la génération précédente sera de 1 moins la probabilité que deux coalescent ou que les trois coalescent. Les trois coalesceront avec probabilité $1/N^2 = 1/100$, deux coalesceront avec probabilité $\binom{3}{2}1/N = 3/10$, par conséquent aucune lignée ne coalescera avec probabilité $1 - (3/10 + 1/100) = 69/100$. L'approximation de cette probabilité par le processus de coalescence est plutôt $1 - 3/10 = 70/100$. De plus, ici N est petit pour

permettre d'illustrer la méthode, en réalité N sera habituellement de l'ordre d'une dizaine de milliers. L'approximation par le processus de coalescence est donc bonne. Ce qui est logique, puisque plus la population est grande, plus les chances sont petites que plusieurs individus ($\geq$3) possèdent le même ancêtre.

Sous le modèle de Wright-Fisher, les générations sont distinctes, le temps est mesuré de façon discrète. Il est plus réaliste, et aussi plus facile, de modéliser le temps sous une échelle continue. Et pour éviter que les probabilités de coalescence dépendent de la taille de la population, il est usuel de dire qu'une unité de temps correspond à N générations. Soit $t = j/N$, où j fait référence au nombre de génération, et soit T_k^* le temps en continu avant un événement de coalescence lorsqu'il y a k lignées. On fait une approximation de la loi géométrique par la loi exponentielle. Alors $T_k^* \sim \text{Exp}(\binom{k}{2})$, et

$$P(T_k^* \leq t) = 1 - e^{-\binom{k}{2}t}.$$

Pour simuler une généalogie d'un échantillon de n individus d'une population de taille N, il suffira de simuler des temps d'attentes avant les événements de coalescence et de choisir au hasard quelle paire de lignées coalescent. Les distributions géométrique et exponentielle sont sans mémoire, par conséquent les $n-1$ temps d'attentes sont indépendants entre eux. Puisque le temps est continu, la probabilité que deux événements de coalescence aient lieu en même temps est nulle. Le plus récent ancêtre commun (MRCA) est atteint lorsqu'il ne reste qu'une lignée, la généalogie de notre échantillon est alors complète.

2.2.2 Processus de coalescence avec mutations

Nous avons montré comment il est possible de simuler les généalogies d'un échantillon à l'aide du processus de coalescence. Il s'agit maintenant d'ajouter la possibilité d'observer des mutations. Nous allons tout d'abord définir comment nous modéliserons ces mutations, puis nous verrons comment modifier le processus de coalescence pour les inclure. Finalement, nous présenterons une équation récursive utilisant le processus de

coalescence qui permet de calculer la probabilité d'observer un certain échantillon de séquence génétique.

Nous modéliserons les mutations selon le modèle des sites infinis. Ce modèle stipule que les mutations sont des événements rares qui, à chaque fois, ont lieu à de nouvelles positions sur le matériel génétique. Il est basé sur l'observation qu'une très grande partie du matériel génétique d'une espèce est identique d'un individu à l'autre. Ce qui laisse supposer que des mutations ont lieu rarement. Puisque le nombre d'endroits possibles pour une mutation est très grand et qu'il s'agit d'un événement rare, alors la probabilité qu'une mutation ait lieu plus d'une fois au même endroit est presque nulle. En cohérence avec le modèle des sites infinis, nous utiliserons maintenant des séquences génétiques pour représenter les individus de la population étudiée. De plus, nous supposerons qu'il y a absence de sélection, les mutations seront neutres, c'est-à-dire qu'elles n'influent pas sur la reproduction. Ce qui revient à ajouter des mutations sur une généalogie.

Pour ajouter les événements de mutation au processus de coalescence, nous allons revenir au modèle de Wright-Fisher. On se souviendra que selon ce modèle, la population est de taille constante et que les générations sont discrètes. Le modèle de Wright-Fisher avec mutations, selon le modèle des sites infinis, est tel que chaque fois qu'un individu a des enfants, un événement de mutation a lieu avec probabilité u. L'endroit où a lieu la mutation est choisi au hasard sur la séquence. Lorsqu'il y a un événement de mutation, tous les descendants du porteur de cette mutation la posséderont aussi.

En changeant l'orientation du temps pour aller du présent vers le passé, la probabilité d'observer un événement de mutation, pour un individu, à la génération précédente est aussi u. Le temps avant que survienne un événement de mutation pour une lignée (noté T_M) est, par conséquent, distribué selon une loi géométrique de paramètre u. La probabilité d'observer, pour la première fois, un événement de mutation à la j^e génération est :

$$P(T_M \leq j) = 1 - (1-u)^j \approx 1 - e^{-\theta t/2},$$

où $t = j/N$, et représente le temps en continu, mesuré sur une échelle telle qu'une unité

correspond à N générations. Le paramètre $\theta = 2Nu$ représente le taux de mutation pour la population (à une constante près). L'approximation en temps continu nous permet d'ajouter les mutations au processus de coalescence introduit précédemment.

Pour un échantillon de taille n, le temps avant un événement de coalescence est distribué selon une loi exponentielle de taux $\binom{n}{2}$; de plus, le temps avant un événement de mutation pour une de ces n lignées est distribué selon une loi exponentielle de taux $\theta/2$. Notons que les mutations ont lieu de manière indépendante pour chacune des lignées et sont aussi indépendantes des événements de coalescence. Par conséquent, le temps avant un événement de mutation ou de coalescence est distribué selon une exponentielle de taux

$$\binom{n}{2} + \frac{n\theta}{2} = \frac{n(n-1+\theta)}{2}.$$

L'événement sera alors une coalescence avec probabilité

$$\frac{\binom{n}{2}}{\binom{n}{2} + \frac{n\theta}{2}} = \frac{n-1}{n-1+\theta},$$

ou un événement de mutation avec probabilité

$$\frac{\frac{n\theta}{2}}{\binom{n}{2} + \frac{n\theta}{2}} = \frac{\theta}{n-1+\theta}.$$

S'il s'agit d'une mutation, on choisira laquelle mutera au hasard parmi les n lignées.

Après avoir construit la généalogie d'un échantillon du présent vers le passé, on doit redescendre la généalogie pour déterminer les séquences composant notre échantillon. Notons que la position des mutations sur les séquences n'a pas d'importance puisque nous n'utilisons pas encore les recombinaisons. Le type d'échantillon obtenu est semblable à la figure 2.8, nous savons quels individus partagent les mêmes mutations. Il est possible de déterminer quelle est la probabilité d'observer un échantillon de ce type à l'aide d'une équation récursive.

Avant de présenter cette équation, nous allons introduire quelques notations utilisées par Larribe et Lessard (2008), lesquelles nous seront utiles tout au long de ce mémoire. Premièrement, le temps évoluera du présent vers le passé, et nous noterons par t_τ le temps

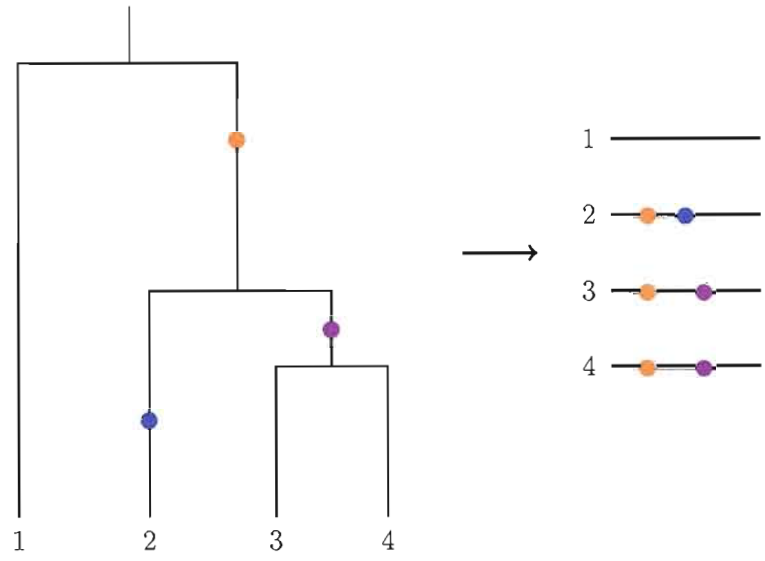


Figure 2.8 Processus de coalescence avec mutation. À gauche l'image représente la simulation de la généalogie à l'aide du processus de coalescence. Les points de couleur représentent les événements de mutation. Le premier événement, du présent vers le passé, étant une mutation (point bleu), suivi d'une coalescence (individus 3 et 4). À droite l'échantillon de séquences obtenues suite à cette simulation. Notons que l'emplacement des mutations n'a pour l'instant aucune importance.

au τ^e événement. De cette façon, l'événement $\tau = 0$ fait référence à l'échantillon de départ et nous noterons par τ^* le dernier événement conduisant au MRCA. Deux séquences sont identiques lorsqu'elles ont exactement les même mutations. Si pour un échantillon de n séquences, il y a exactement k séquences différentes, nous dirons qu'il y a k types de séquences et n_i représentera le nombre de séquences de type i $(i = 1, \ldots, k)$ présentent dans l'échantillon. Un échantillon de séquences est défini entièrement par l'ensemble des différents types de séquence le composant, ainsi que par leurs multiplicités. Par exemple, pour l'échantillon simulé à la figure 2.8, $n = 4$ car quatre séquences ont été simulées, mais les séquences 3 et 4 sont identiques, par conséquent, ici $k = 3$. Notons par H_τ, l'ensemble des types de séquences et leurs multiplicités, résultant du τ^e événement. L'équation récursive que nous présenterons est la solution de $P(H_0)$. La figure 2.9, indépendante de

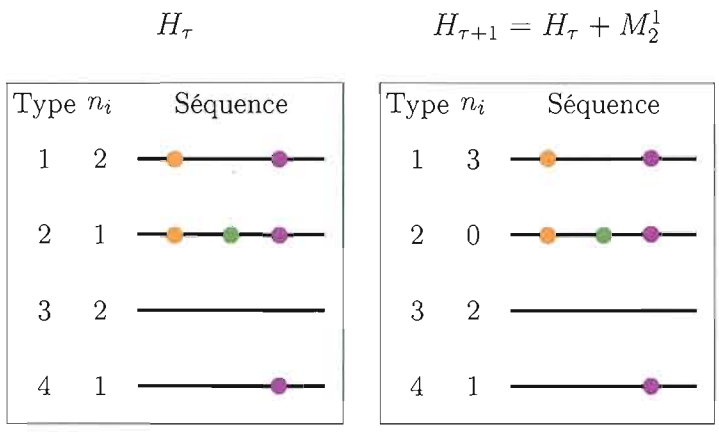


Figure 2.9 Ensemble de séquences suite à un événement de mutation. À gauche, H_τ représente l'ensemble de séquences avant la mutation. À droite, $H_{\tau+1}$ représente l'ensemble des séquences après l'événement de mutation M_2^1. Les séquences sont présentées par type et n_i représente la multiplicité du type de séquence i.

la figure 2.8, présente H_τ, un exemple d'un ensemble de séquences et leurs multiplicités et, suite à un événement de mutation d'une séquence de type 2 vers une séquence de type 1 (le $(\tau+1)^e$ événement), l'ensemble des séquences résultantes.

Regardons maintenant de plus près les conditions nécessaires pour observer un événement de coalescence ou de mutation en fonction de l'échantillon. Deux séquences peuvent coalescer si, et seulement si, elles sont identiques, c'est-à-dire si elles sont du même type. Une coalescence de deux séquences de type i sera notée C_i et sera possible si $n_i > 1$. Pour l'ensemble H_τ de la figure 2.9, seules les séquences de types 1 et 3 peuvent coalescer, car on retrouve au moins deux séquences de chacun de ces types. Selon notre modélisation, une mutation peut avoir lieu du présent vers le passé, si une seule séquence possède cette mutation, car il ne peut y avoir plus d'une mutation à une position en particulier. Supposons que la séquence de type i est la seule à posséder une certaine mutation et que $n_i = 1$, alors un événement de mutation est possible de la séquence i vers la séquence j, où la séquence de type j est identique à celle de type i la mutation en moins. Suite à celui-ci, $n_i = 0$, et la multiplicité du type de séquence j sera augmentée de

1. Un seul événement de mutation peut avoir lieu pour l'ensemble de séquence H_τ de la figure 2.9, la séquence de type 2 est la seule à avoir la mutation représenté par ●, et elle est de multiplicité 1 (*i.e.* $n_2 = 1$). Suite à cet événement de mutation, on retrouve une séquence de plus de type 1 et il n'y a plus de séquence de type 2. Une mutation d'une séquence i vers une séquence de type j sera notée M_i^j.

Si le $(\tau+1)^e$ événement est une M_i^j, nous écrirons que $H_{\tau+1} = H_\tau + M_i^j$ pour signifier que l'ensemble des séquences présentent suite au τ^e événement est modifié par une mutation d'une séquence de type i vers une séquence de type j au $(\tau+1)^e$ événement. Maintenant, sachant qu'une mutation M_i^j nous a amené à l'ensemble $H_\tau + M_i^j$, la probabilité que $H_{\tau+1}$ ait généré l'ensemble H_τ (maintenant nous regardons du passé vers le présent) est $(n_j+1)/n$. Puisqu'avec probabilité égale, chacune des n_j+1 séquences de type j présentent dans l'ensemble $H_{\tau+1}$ peuvent avoir subi cette mutation.

Si le $(\tau+1)^e$ événement est plutôt une C_i, nous écrirons que $H_{\tau+1} = H_\tau + C_i$. Maintenant, sachant qu'une coalescence C_i nous a amené à l'ensemble $H_\tau + C_i$, la probabilité que $H_{\tau+1}$ ait généré l'ensemble H_τ (du passé vers le présent) est $(n_i+1)/(n-1)$. Puisqu'avec probabilité égale, chacune des $n_i - 1$ séquences de type i parmi les $n-1$ séquences présentent dans l'ensemble $H_{\tau+1}$ peuvent avoir eu deux descendants.

La probabilité d'observer un certain échantillon H_0, peut se formuler comme suit :

$$\begin{aligned} P(H_0) = &\frac{n-1}{n-1+\theta} \sum_{n_i>1} \frac{n_i-1}{n-1} P(H_0 + C_i) \\ &+ \frac{\theta}{n-1+\theta} \sum_{\text{singleton}} \frac{n_j+1}{n} P(H_0 + M_i^j). \end{aligned} \tag{2.1}$$

On remarquera, devant les symboles de sommation, les probabilités d'observer un événement de coalescence et un événement de mutation, présentées précédemment. La deuxième sommation se fait sur les singletons, c'est-à-dire les ensembles contenant la seule séquence possédant une certaine mutation. Cette équation tient compte de tous les événements possibles pouvant avoir mener à l'échantillon de départ. Il est possible de calculer explicitement $P(H_0)$ si l'échantillon est petit, tel que présenté par Hein, Schierup

et Wiuf (2005, section 2.4). En général, il ne sera cependant pas possible de calculer cette probabilité explicitement.

2.2.3 Graphe de recombinaison ancestral

Au début de ce chapitre, nous avons vu l'importance des recombinaisons en cartographie génétique. C'est pourquoi rapidement une généralisation du processus de coalescence incluant les recombinaisons a été proposée. Hudson (1983) fut l'un des premiers à proposer l'inclusion des recombinaisons au processus de coalescence. Puis, Griffiths et Marjoram (1996) nommèrent cette généralisation le graphe de recombinaison ancestral (ARG), et présentèrent une équation récursive permettant de calculer la probabilité d'observer un certain échantillon en simulant plusieurs généalogies à l'aide de l'ARG ; cette équation est une généralisation des travaux de Griffiths et Tavaré (1994a, 1994b). Nous présenterons ici comment les recombinaisons peuvent être modélisées et incluses au modèle de Wright-Fisher, et comment il est possible de simuler un échantillon à l'aide de l'ARG. L'équation récursive de Griffiths et Marjoram (1996) ne sera pas présentée dans sa version originale, mais plutôt dans la version adaptée par Larribe, Lessard et Schork (2002) au chapitre suivant.

Lorsque nous parlons de recombinaison, il s'agit d'un événement ayant lieu chez des individus sexués. Rappelons qu'une recombinaison a eu lieu sur un haplotype d'un individu lorsque celui-ci est composé d'un mélange des haplotypes d'un de ces parents. La figure 2.10 illustre un événement de recombinaison ayant lieu du passé vers le présent, et le même événement de recombinaison, mais cette fois-ci du présent vers le passé. En regardant une recombinaison du présent vers le passé, on remarque plus facilement qu'un haplotype a deux haplotypes parentaux différents. Il a été montré que le modèle de Wright-Fisher et le processus de coalescence s'adaptent bien à plusieurs situations. Par exemple, il est possible de faire varier la taille de la population selon certains modèles. Il est aussi possible de les adapter aux individus de types sexués. Sans entrer dans les détails, ni en faire la démonstration, il est possible de considérer une population de N individus diploïdes comme une population de $2N$ haplotypes (haploïdes). Lorsque la

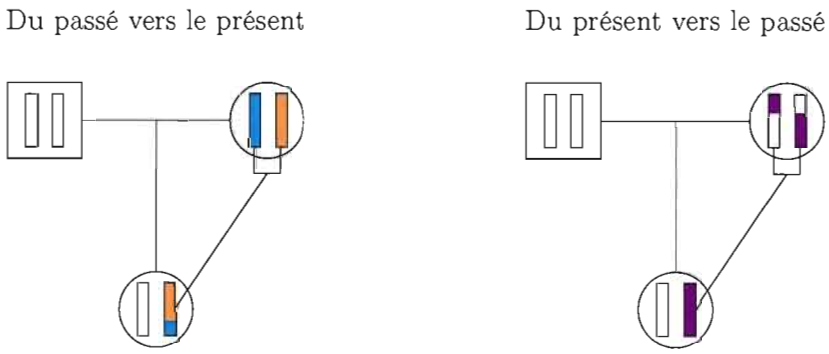


Figure 2.10 Événement de recombinaison vu dans les deux sens du temps. À gauche du passé vers le présent et, à droite, du présent vers le passé.

taille de la population est très grande en comparaison à la taille de l'échantillon, cette approximation est raisonnable et facilite grandement les calculs.

La grande différence entre une population d'individus non sexués et une population d'individus sexués, se situe dans le nombre de parents que possède une séquence. Maintenant, une séquence aura deux séquences parentales dans la génération précédente. Si une recombinaison a lieu, cette séquence sera un amalgame de ses séquences parentales, sinon elle sera identique à une des deux. Une recombinaison aura lieu avec probabilité r. En suivant l'histoire d'une lignée sous le modèle de Wright-Fisher, à chaque génération, soit un événement de recombinaison se produit, soit il n'a pas lieu. Par conséquent, le temps avant qu'une séquence en particulier subisse une recombinaison (T_R) est distribué selon une loi géométrique de paramètre r. Et la probabilité qu'un événement de recombinaison ait lieu à la génération j est :

$$P(T_R = j) = r(1-r)^{j-1}.$$

Comme pour les événements de coalescence et de mutation, nous allons utiliser une échelle de temps continue, où une unité de temps équivaut maintenant à $2N$ générations. Notons que nous remplaçons N par $2N$, puisque nous considérons notre population de N individus comme une population de $2N$ haplotypes. Par conséquent, maintenant

$\theta = 4Nu$ et nous allons utiliser $\rho = 4Nr$ comme taux de recombinaison. Sous la nouvelle échelle de temps,

$$P(T_R \leq t) = 1 - (1-r)^j = 1 - \left(1 - \frac{2Nr}{2N}\right)^{2Nt} \approx 1 - e^{-\rho t/2},$$

où $j = 2Nt$. Lorsqu'il y a k séquences, elles évoluent de manière indépendante les unes des autres, alors le temps avant un événement de recombinaison est distribué selon une loi exponentielle de paramètre $k\rho/2$. De plus, les événements de recombinaisons sont indépendants des événements de coalescences et de mutations.

En ajoutant maintenant les recombinaisons au processus de coalescence avec mutation défini précédemment, il est possible de simuler des généalogies. Le résultat obtenu ne sera plus de la forme d'un arbre, il s'agira maintenant d'un graphe, d'où le nom graphe de recombinaison ancestral. Lors de la simulation d'une généalogie d'un échantillon de n séquences, le temps avant le premier événement, quel qu'il soit, est distribué selon une exponentielle de paramètre

$$\frac{n(n-1)}{2} + \frac{n\theta}{2} + \frac{n\rho}{2} = \frac{n(n-1+\theta+\rho)}{2}.$$

Il s'agira d'une coalescence, d'une mutation ou d'une recombinaison, avec probabilités respectives :

$$\frac{n-1}{(n-1+\theta+\rho)}, \quad \frac{\theta}{(n-1+\theta+\rho)} \text{ et } \quad \frac{\rho}{(n-1+\theta+\rho)}. \tag{2.2}$$

S'il s'agit d'un événement de recombinaison, une séquence est choisie au hasard, ainsi qu'une position sur cette séquence. La position choisie représente l'endroit où aura lieu la recombinaison. La séquence subissant une recombinaison sera remplacée par deux nouvelles séquences, tel qu'illustré par la figure 2.10 (à droite). Une d'entre elles possédera le même matériel génétique du début de la séquence originelle jusqu'au point de recombinaison, le restant étant inconnu (nous parlerons de matériel génétique non ancestral). L'autre nouvelle séquence possédera le matériel génétique à gauche du point de recombinaison. Notons que nous considérons ici une séquence génétique assez courte, de telle sorte que l'on peut ignorer la possibilité qu'il y ait plus d'une recombinaison par méiose.

Il existe diverses méthodes pour choisir aléatoirement le point de recombinaison. On peut choisir de mettre plus de poids à certains endroits de la séquence pour lesquels les recombinaisons sont plus probables. On parle alors de point chaud, ou «hot spot». Nous choisirons plutôt de ne pas privilégier certaines régions de la séquence et d'utiliser une loi uniforme pour sélectionner le point de recombinaison. Aussi, lorsque les séquences sont courtes, r représentera la longueur d'une séquence en centimorgan.

La figure 2.11 illustre la simulation d'une généalogie d'un échantillon de trois séquences à l'aide du graphe de recombinaison ancestral. Lors d'une recombinaison, du matériel génétique non ancestral apparaît sur les séquences résultantes, les fragments de séquences non ancestraux sont illustrés en gris sur la figure 2.11. L'histoire et l'état (mutation présente ou non) de ces segments ne nous intéressent pas. Avec le graphe de recombinaison ancestral, on simule la généalogie du matériel ancestral, c'est-à-dire des séquences composant l'échantillon. C'est pourquoi une séquence avec un segment non ancestral peut coalescer avec une séquence ancestrale. À droite de la figure, on retrouve l'échantillon obtenu suite à cette simulation.

Il est légitime de se demander si l'atteinte du MRCA est assurée avec le graphe recombinaison ancestral. Un grand nombre de recombinaisons pourrait laisser supposer qu'il est possible de ne pas trouver d'ancêtre commun aux séquences de l'échantillon. Lorsqu'il y a k lignées, nous devons remarquer que le taux de coalescence est quadratique, il est de $k(k-1)/2$, tandis que le taux de recombinaison est linéaire, il est de $k\rho/2$. Ce qui assure l'atteinte d'un MRCA. En considérant seulement les événements de coalescence et de recombinaison, le nombre de séquences dans le graphe est un processus de naissance et de mort.

Le processus de coalescence et le graphe de recombinaison ancestral s'avèrent des outils de simulation très robustes qui s'adaptent à plusieurs situations. En plus de servir en cartographie génétique, ces méthodes sont utilisées pour estimer les paramètres θ et ρ.

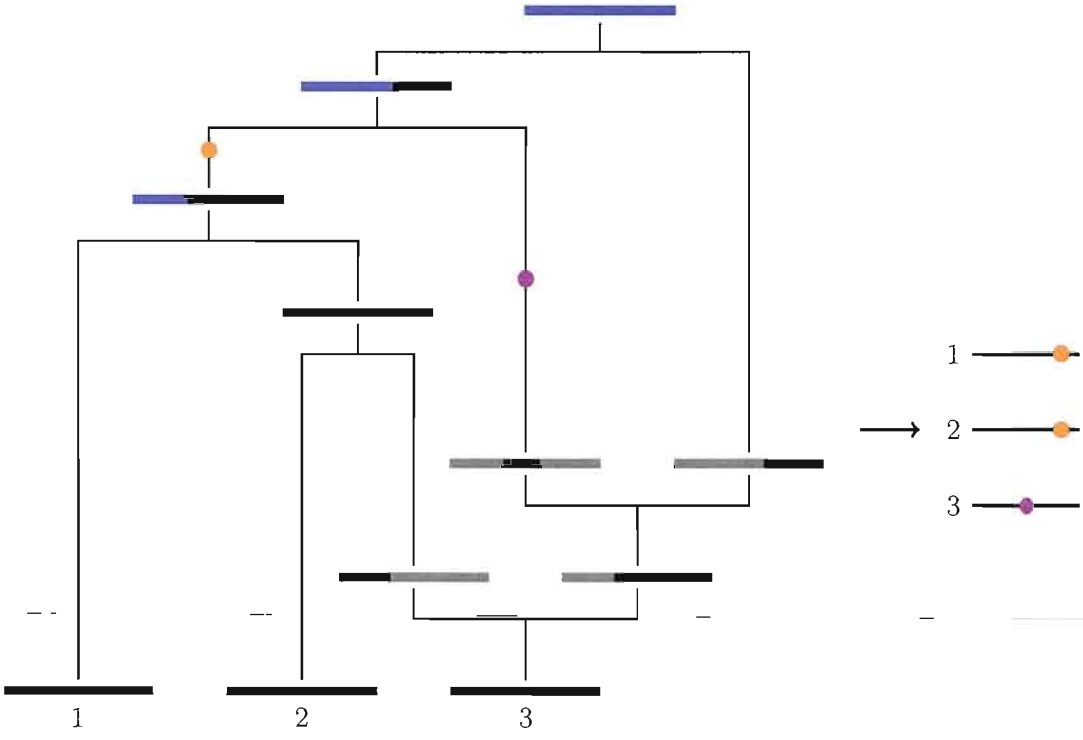


Figure 2.11 Simulation d'une généalogie à l'aide de l'ARG. Les portions de séquences grises représentent le matériel génétique non ancestral. Les portions en bleu représentent l'atteinte du MRCA pour cette portion de séquence. Remarquons que la séquence 3 n'est pas affectée par la mutation ● puisque cette mutation se trouvait sur une section de la séquence qui n'a pas été transmise suite à un événement de recombinaison. La position d'une mutation est choisie aléatoirement sur la séquence selon le modèle des sites infinis.

CHAPITRE III

CARTOGRAPHIE GÉNÉTIQUE FINE VIA LA MÉTHODE MAPARG

La méthode de cartographie génétique fine MapArg a été proposée en 2002 (Larribe, Lessard et Schork, 2002), puis améliorée en 2008 (Larribe et Lessard, 2008). Cette méthode, basée sur le déséquilibre de liaison, permet d'estimer la position d'un caractère influençant une mutation (TIM : «trait influencing a mutation»). MapArg, comme son nom l'indique, utilise le graphe de recombinaison ancestral (ARG) pour modéliser la généalogie d'un échantillon. En utilisant l'information contenue dans un échantillon et dans un grand nombre de généalogies plausibles, il est possible de déduire une fonction de vraisemblance récursive, qui avec l'aide de l'échantillonnage pondéré nous permettra d'estimer la position du TIM. Nous allons tout d'abord voir en détail le fonctionnement de cette méthode, puis nous allons donner un bref aperçu des derniers développements de MapArg. Parmi ceux-ci, on retrouve une modification de la fonction de vraisemblance, rendant celle-ci conditionnelle et composite. Cette modification permet d'accélérer les calculs informatiques en utilisant des fenêtres de marqueurs plutôt que tous les marqueurs à la fois.

3.1 Mise en situation

Supposons que nous sommes en présence d'un échantillon de personnes tirées d'une population, dont certaines sont atteintes d'une maladie génétique causée par une mutation à un gène en particulier dont nous souhaitons trouver la position. Nous savons,

suite à une étude d'association (ou une analyse de liaison), que la mutation se situe à l'intérieur d'une certaine région d'un chromosome. Nous possédons, pour cette région, les haplotypes de toutes les personnes formant notre échantillon. Si nous savions la généalogie de notre échantillon, et ce, jusqu'au plus récent ancêtre commun (MRCA), si de plus nous connaissions l'histoire des recombinaisons ayant eu lieu, nous avons vu qu'il serait possible, et même relativement facile de déterminer la position de la mutation créant la maladie en analysant quelles portions de l'haplotype du MRCA les gens affectés partagent, c'est-à-dire en utilisant le déséquilibre de liaison.

Malheureusement, nous ne sommes jamais en présence d'autant d'information concernant un échantillon. Mais nous savons qu'il est important de connaître la généalogie de notre échantillon, ou à défaut, d'être en mesure de la simuler le plus efficacement possible, car elle contient de l'information non négligeable sur la position du TIM. Nous avons vu précédemment que le graphe de recombinaison ancestral est un outil complet — puisqu'il tient compte des événements de recombinaison — pour simuler les généalogies d'un échantillon en partant du présent et en allant vers le passé, jusqu'au MRCA plus exactement. C'est pourquoi les concepteurs de MapArg on choisit d'utiliser l'ARG pour simuler différentes généalogies possibles de l'échantillon de départ.

La fonction de vraisemblance développée par Griffiths et Marjoram (1996) et adaptée par Larribe, Lessard et Schork (2002) est basée sur une fonction récursive ; elle s'amorce avec l'échantillon de départ et prend en compte tous les graphes possibles. On calcule la vraisemblance d'observer cet échantillon, vraisemblance qui dépend en fait de sa généalogie. Une des façons d'aborder la question est de commencer avec ce que l'on observe dans le présent et de reculer dans le passé, car faire le chemin inverse serait à peu près impossible. On va chercher quelle est la probabilité que le premier événement à se produire soit une recombinaison, une mutation ou une coalescence ; ensuite, si le premier événement est une coalescence, quels événements de coalescence sont plausibles avec notre échantillon. L'idée derrière MapArg est de calculer la vraisemblance d'une généalogie en même temps que celle-ci est créée grâce à l'ARG. En fait, on estime la vraisemblance de la position du TIM entre chaque marqueur de façon indépendante.

Pour chacun de ces intervalles entre marqueurs, la fonction de vraisemblance est une moyenne de la vraisemblance de chaque généalogie — ou graphe — créée. On obtient alors pour chaque intervalle, une courbe de vraisemblance, et la réunion de ces courbes forme l'estimation de la fonction de vraisemblance de notre échantillon, son maximum nous indiquant la position estimée du TIM.

3.2 Modélisation

Nous présentons ici, la modélisation choisie pour la méthode MapArg, dont certains des éléments principaux ont été présentés dans les chapitres précédents. Premièrement, comme pour le graphe de recombinaison ancestral, la population est formée de N individus que nous considèrerons comme une population de $2N$ haplotypes. Nous utiliserons le modèle Wright-Fisher pour la population ; elle sera de taille constante, les différentes générations seront discrètes et ne s'entremêleront pas et les croisements seront aléatoires. Notons qu'il est aussi possible de modéliser une population de taille non constante assez simplement (voir Larribe, Lessard et Schork (2002)), mais nous ne traiterons pas de cette situation ici. Le temps du présent vers le passé sera mesuré sur une échelle de temps continue, où une unité de temps représentera $2N$ générations.

L'échantillon utilisé par MapArg est composé d'un ensemble de séquences de marqueurs génétiques. Les marqueurs utilisés sont les polymorphismes nucléotidiques simples (SNPs) définis au chapitre un et nous les considérons comme ordonnés et distants entre eux sur les séquences. En général, il n'y a que deux allèles possibles par SNP, le contraire étant rare, nous n'en tiendrons pas compte, comme cela est usuel. Soit le marqueur est primitif (0 ou □), c'est-à-dire qu'il descend directement du MRCA sans avoir connu de mutation ; soit le marqueur est dérivé (ou mutant) (1 ou ■), c'est-à-dire qu'il y a eu une mutation à ce marqueur au cours de son histoire le menant du MRCA à l'échantillon. Qu'ils soient primitifs ou dérivés, tous les marqueurs de notre échantillon de départ sont dits ancestraux, c'est-à-dire qu'ils proviennent tous du MRCA.

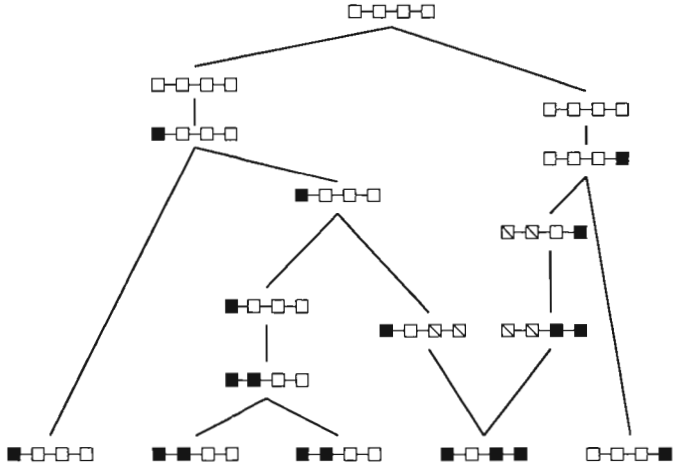
On se souviendra que suite à un événement de recombinaison, les séquences résultantes (parentes) ne partagent qu'une partie des marqueurs de leur descendant. L'une d'elles possède les mêmes marqueurs que le descendant à gauche du point de recombinaison, tandis que l'autre, ceux de droite. Nous dirons que les marqueurs non-définis, ceux que les séquences parentes ne partagent pas avec le descendant, sont des marqueurs non-ancestraux (– ou ⊘) ; ils ne proviennent pas nécessairement du MRCA. Il n'est pas nécessaire de posséder plus d'informations sur ces marqueurs puisque nous nous intéressons seulement à la généalogie de notre échantillon composé entièrement de marqueurs ancestraux.

La figure 3.1 présente un exemple d'une généalogie pour un échantillon composé de séquences de quatre marqueurs. Ainsi que les arbres partiels de chacun des marqueurs, c'est-à-dire la généalogie d'un marqueur précis représenté cette fois par un arbre. On remarque sur la figure 3.1(a) qu'il s'agit d'un graphe puisqu'il y a eu un événement de recombinaison sur la quatrième séquence de l'échantillon au deuxième intervalle. On peut aussi représenter la généalogie d'un seul marqueur par un arbre partiel ; les figures 3.1(b) et 3.1(c) présentent respectivement, en bleu, l'arbre partiel des deux premiers marqueurs et l'arbre partiel des deux derniers marqueurs.

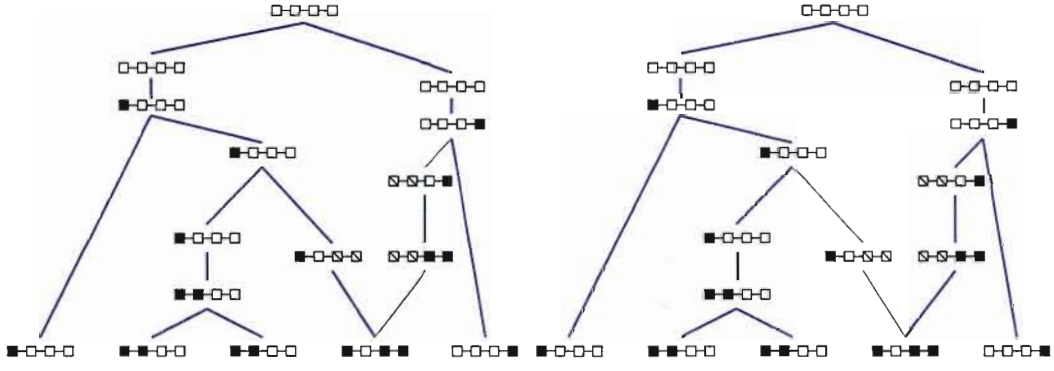
Le modèle des sites infinis est utilisé pour les mutations et aussi pour le TIM (représenté par ■). Les taux de mutation des SNPs et du TIM sont considérés petits. Les mutations sont rares, tellement rares, qu'il ne peut y avoir qu'une seule mutation par marqueur dans l'histoire de notre échantillon. Nous assumons donc que la maladie, le TIM recherché, est elle aussi causée par un seul événement de mutation dans l'histoire de notre population.

Chaque séquence de notre échantillon est composée de L marqueurs ordonnés. Le TIM fait partie de ces L marqueurs, et par conséquent la position de seulement $L-1$ marqueurs est connue (la position du TIM étant bien sûr inconnue). Nous supposons que le TIM se situe entre le premier et le dernier des marqueurs ordonnés ayant des positions connues. Les marqueurs sont ordonnés de la gauche vers la droite ; par conséquent le

marqueur m est le m^e marqueur de notre séquence en débutant par la gauche. Les distances entre les marqueurs sont mesurées en centimorgan et nous les considérons additives (pas d'interférence). Nous allons noter la distance en centimorgan entre les marqueurs m et $m+1$ par r_m, il s'agira aussi de la longueur de l'intervalle m. Il y a $L-1$ intervalles, dont $L-2$ sont de longueurs connues. De plus nous allons noter par r la longueur totale d'une séquence, donc $r = \sum_m r_m$. Notons qu'il s'agit en fait du même r correspondant au taux de recombinaison par séquence par génération, présenté au chapitre deux en page 39.



(a) Graphe



(b) Arbre partiel des deux premiers marqueurs. (c) Arbre partiel des deux derniers marqueurs

Figure 3.1 La figure (a) présente la généalogie d'un échantillon de séquences. L'arbre partiel des deux premiers marqueurs est souligné en bleu sur la figure (b). Tandis que l'arbre partiel des deux derniers marqueurs est souligné en bleu sur la figure (c).

La distance entre le premier marqueur et le TIM sera notée par r_T. Nous cherchons donc à estimer le paramètre r_T par la méthode du maximum de vraisemblance. Une notation particulière sera utilisée pour représenter la distance entre le TIM et le marqueur directement à sa gauche (r_l) et la distance entre le TIM et le marqueur directement à sa droite (r_r), tel qu'illustré sur la figure 3.2. Par exemple, si le TIM est le marqueur m, il s'ensuit alors que $r_l = r_{m-1}$ et $r_r = r_m$. La position, en centimorgan, d'un marqueur m sur la séquence sera notée par x_m ; cette notation est telle que la position du premier marqueur est 0 ($x_1 = 0$) et celle du dernier marqueur est r ($x_L = r$), on obtient ainsi que $x_m = \sum_{p=1}^{m-1} r_p$ pour $2 \leq m \leq L$.

Les distances entre les marqueurs seront évidemment identiques pour chacune des séquences. Nous pouvons par conséquent représenter une séquence s, simplement par la suite de ses L marqueurs ordonnés, $s = (s_1,...,s_L)$, où s_m sera l'allèle (0, 1 ou –) du m^e marqueur de la séquence.

On peut retrouver aux différentes étapes de notre processus deux séquences — ou plus — identiques, c'est-à-dire ayant les mêmes allèles aux mêmes marqueurs. Nous parlerons donc de type de séquences : $s^{(i)} = \left(s_1^{(i)},...,s_L^{(i)}\right)$ sera une séquence de type i ; nous dirons que deux séquences sont de même type ($s^{(i)} = s^{(j)}$) si, et seulement si

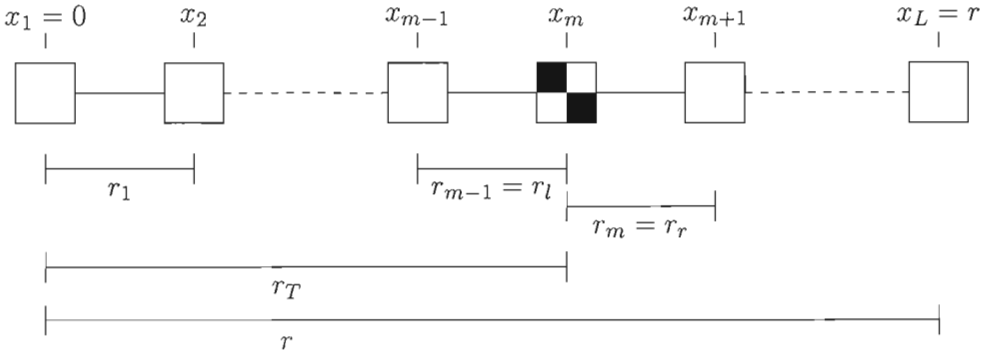


Figure 3.2 Paramètres d'une séquence selon la méthode MapArg. Une séquence de L marqueurs est représentée ; le TIM (■) est le m^e marqueur. Les différents paramètres de distances sont présentés.

$s_k^{(i)} = s_k^{(j)}$ pour tout $k \in \{1, \ldots, L\}$. La figure 3.3 présente, selon les deux représentations utilisées, six séquences de quatre types différents. La méthode MapArg n'utilise pas des séquences ordonnées, c'est-à-dire que cette méthode ne fait pas de distinction entre les différentes séquences d'un même type.

Le graphe de recombinaison ancestral est utilisé pour modéliser la généalogie de notre échantillon du présent vers le passé, c'est-à-dire de notre échantillon jusqu'à son plus récent ancêtre commun (MRCA). Comme vu précédemment, l'ARG est un processus de naissance et de mort. Les événements de transition possibles sont les coalescences, les recombinaisons et les mutations. MapArg est une méthode qui tient seulement compte des événements dans la généalogie de notre échantillon qui modifient le matériel ancestral.

En assumant que le temps évolue du présent vers le passé, il y a coalescence lorsque deux lignées (séquences) ont le même ancêtre, c'est alors la mort d'une séquence (un haplotype). Deux séquences peuvent coalescer si, et seulement si leurs marqueurs ancestraux sont identiques. Par exemple, les séquences de type 1 de la figure 3.3 peuvent coalescer (car il y en a deux), ainsi que les séquences de type 2 et 3 puisque les marqueurs ancestraux qu'elles partagent sont identiques. Nous assistons à un événement de recombinaison, toujours du présent vers le passé, lorsqu'une lignée (séquence) provient de deux ancêtres. Cette séquence partagera les mêmes marqueurs à gauche d'un certain point de recombinaison avec un de ses ancêtres et ses marqueurs à droite de ce même

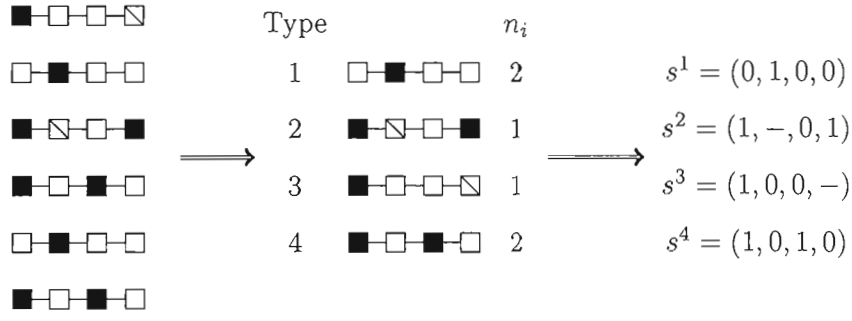


Figure 3.3 Illustration d'un ensemble de séquences, il ne s'agit pas de l'échantillon de départ. On retrouve en tout six séquences, dont quatre types de séquences différents.

point avec son autre ancêtre (voir la figure 3.1). Lorsque nous sommes en présence d'un tel événement, on assiste alors à la «naissance» d'une séquence (un haplotype). Une recombinaison introduit des marqueurs non ancestraux dans l'ensemble des séquences. Un événement de mutation est possible, du présent vers le passé, lorsqu'une des séquences de notre ensemble est la seule à posséder un marqueur mutant à une certaine position et que son ancêtre ne possède pas cette mutation. Par exemple, pour l'ensemble de séquences de la figure 3.3, un seul événement de mutation est possible : seule la séquence de type 2 est de multiplicité un et possède une mutation qu'aucune autre séquence n'a (ici au dernier marqueur). Les autres marqueurs de la séquence ancêtre seront identiques à ceux de son descendant. La séquence doit être la seule à posséder la mutation, car une seule mutation par marqueur peut avoir lieu selon notre modélisation. Un tel événement ne modifie pas le nombre de séquences présentes.

Nous avons déjà vu, au chapitre précédent, que si nous utilisons une échelle de temps continue, telle qu'une unité correspond à l'espérance du temps avant que deux séquences fixées de notre population coalescent, soit $2N$ générations, alors le temps avant le prochain événement de coalescence est approximativement distribué selon une loi exponentielle de taux $k(k-1)/2$ lorsqu'il y a k lignées (haplotypes). Si r représente le taux de recombinaison par haplotype par génération et que l'on définit $\rho := 4Nr$, alors sous notre nouvelle échelle de temps, le temps avant le prochain événement de recombinaison est approximativement distribué selon une exponentielle de moyenne $k\rho/2$, lorsque N est grand et qu'il y a k séquences. De manière analogue, si u est le taux de mutation par haplotype par génération et que l'on définit $\theta := 4Nu$, alors sous la nouvelle échelle de temps, le temps avant le prochain événement de mutation est approximativement distribué selon une exponentielle de taux $k\theta/2$, lorsque N est grand et qu'il y a k lignées. Notons que, comme nous l'avons déjà souligné, nous sommes assurés d'atteindre le plus récent ancêtre commun puisque les nouvelles séquences apparaissent selon un taux linéaire $(k\rho/2)$ et disparaissent selon un taux quadratique $(k(k-1)/2)$.

Modélisons maintenant le processus de naissance et de mort. Nous dirons que le τ^e événement modifiant le matériel ancestral se produira au temps t_τ (mesuré sous notre

nouvelle échelle : 1 unité $= 2N$ générations), et ce toujours du présent vers le passé. L'événement $\tau = 0$ fait référence à l'échantillon de départ et $\tau = \tau^*$ correspond au dernier événement conduisant au MRCA. Nous noterons H_τ, l'ensemble des séquences présentes suite au τ^e événement. Par conséquent, H_0 correspond à l'ensemble des séquences formant notre échantillon initial et H_{τ^*} correspond à la seule séquence du MRCA composée d'une suite de L marqueurs ayant tous un allèle primitif (0).

3.3 Probabilités des événements du processus

Pour chaque nouvelle étape de la création d'un graphe, il existe un nombre fini d'événements possibles, chacun ayant leur propre probabilité de survenir. Nous allons maintenant spécifier avec plus de détails les événements pouvant se produire à chacune des étapes, ainsi que leurs probabilités respectives. Définissons tout d'abord quelques autres variables qui permettront de définir ces probabilités.

Nous allons noter le nombre de séquences appartenant à H_τ par n ($|H_\tau| = n$) et le nombre de séquences de type i par n_i ; par conséquent à l'étape 0, n représente la taille de notre échantillon. Soit maintenant A^i l'ensemble des marqueurs ancestraux d'une séquence de type i, et soit B^i l'ensemble des intervalles ancestraux d'une séquence de type i. Rappelons qu'un marqueur est ancestral si son allèle est primitif ou dérivé (0 ou 1) ; un intervalle sera dit ancestral s'il est situé entre deux marqueurs ancestraux. Il n'est pas nécessaire que ces deux marqueurs soient consécutifs, c'est-à-dire qu'un intervalle m d'une séquence de type i est considéré ancestral si, et seulement si le marqueur m est $m_1 \leq m < m_2$ pour m_1, m_2 appartenant à A^i. Revenons à l'ensemble des séquences de la figure 3.3 ; pour cet exemple, A^2 est l'ensemble des marqueurs ancestraux de la séquence de type 2 soit : $\{1, 3, 4\}$ tandis que B^2 est l'ensemble de tous les intervalles car les marqueurs 1 et 4 appartiennent à A^2. Tandis que A^3 est l'ensemble formé des marqueurs 1, 2 et 3, et que B^3 est l'ensemble des deux premiers intervalles seulement, c'est-à-dire l'ensemble des intervalles compris entre les marqueurs 1 et 3.

Définissons le nombre total de marqueurs ancestraux au temps t_τ, par $a = \sum_{i=1}^{d} n_i \left|A^i\right|$, où d est le nombre de types de séquences différentes dans H_τ et $\left|A^i\right|$ est la cardinalité de l'ensemble A^i. Notons que $n \leq a \leq nL$. De plus, soit b, la longueur totale des intervalles ancestraux au temps t_τ ; $b = \sum_{i=1}^{d} n_i[\max\{x_m : \text{marqueur } m \in A^i\} - \min\{x_m : \text{marqueur } m \in A^i\}]$. Ces deux valeurs nous permettront de calculer les probabilités qu'un événement modifie le matériel ancestral. Notons aussi que $0 \leq b \leq nr$. Notons de plus que n, n_i, a et b dépendent de τ. Pour l'ensemble des séquences de la figure 3.3 en page 49, on trouve que $a = 22$, car les séquences de type 1 et 4 ont chacune quatre marqueurs ancestraux et sont de multiplicité deux (on a déjà 16 marqueurs ancestraux), et les séquences de type 2 et 3 ont respectivement trois marqueurs ancestraux et sont de multiplicité un ; on ajoute six à seize et on obtient ainsi les 22 marqueurs ancestraux de notre ensemble de séquences. Nous ne pouvons trouver la valeur exacte de b, puisque la longueur des séquences de l'ensemble et des intervalles n'a pas été définie, mais puisqu'il n'y a que le dernier intervalle de la séquence de type 3 qui n'est pas ancestral, on sait alors que b est équivalent à six fois r moins une fois la longueur du dernier intervalle.

En continuité avec l'article de Griffiths et Marjoram (1996), définissons la distribution de probabilité de l'état H_τ par $Q(H_\tau)$. Notons qu'ici, $Q(H_\tau)$ est la distribution de probabilité de H_τ conditionnellement au fait que l'événement τ ait modifié le matériel ancestral des séquences. Une mutation peut avoir lieu à un marqueur non ancestral —ou une recombinaison dans un intervalle non ancestral—, mais puisque nous nous intéressons seulement aux événements modifiant le matériel ancestral, nous n'allons pas tenir compte de ces événements. Définissons la proportion des marqueurs ancestraux de l'ensemble des séquences présentent au temps t_τ par $\alpha = a/nL$, et la proportion des longueurs de séquences ancestrales par $\beta = b/nr$. Un événement de mutation modifiera le matériel ancestral si la mutation a lieu à un marqueur ancestral. Par conséquent, le temps avant un événement de mutation modifiant le matériel ancestral est distribué selon une loi exponentielle de taux $k\alpha\theta/2$ lorsqu'il y a k séquences. De manière similaire, une recombinaison modifiera le matérial ancestral si elle a lieu dans un intervalle ancestral ; le temps

avant un tel événement est distribué selon une loi exponentielle de taux $k\beta\rho/2$, lorsqu'il y a k séquences.

Ainsi, les probabilités d'observer un événement de coalescence (C), mutation (M) ou recombinaison(R), du présent vers le passé, sont respectivement :

$$P(C) = \frac{(n-1)}{((n-1)+\alpha\theta+\beta\rho)},$$

$$P(M) = \frac{\alpha\theta}{((n-1)+\alpha\theta+\beta\rho)}$$

et

$$P(R) = \frac{\beta\rho}{((n-1)+\alpha\theta+\beta\rho)}.$$

Ensuite, pour être en mesure de construire notre fonction de vraisemblance, il faut établir les probabilités reliées aux séquences présentes aux différentes étapes du processus. La probabilité d'observer un certain ensemble de séquences à l'étape H_τ dépend en fait des séquences présentes à l'étape suivante (lorsque l'on considère les étapes du présent vers le passé). Donc H_τ dépend de $H_{\tau+1}$, puisqu'en réalité c'est de cette façon que fonctionne l'évolution (du passé vers le présent).

Premièrement, si le prochain événement est une coalescence, il nous faut déterminer quels sont les événements de coalescence possibles, en tenant compte des séquences présentes, ainsi que la probabilité de chacun d'entre eux. Il peut, bien évidemment, avoir coalescence de deux séquences de même type i, événement que nous allons noter par C_i. Un exemple est présenté à la figure 3.4(a). Si nous observons une coalescence, c'est qu'en fait deux séquences possèdent le même ancêtre ; par conséquent, si $|H_\tau| = n$ alors, $|H_{\tau+1}| = n-1$. Mais chacune des $n-1$ séquences appartenant à $H_{\tau+1}$ ont la même probabilité d'avoir engendré deux séquences (du passé vers le présent). Donc, si l'événement de coalescence est de type C_i alors il y aura $n_i - 1$ séquences de type i par la suite (dans $H_{\tau+1}$), on obtient alors que la probabilité de C_i (sachant qu'il y a coalescence) est de $(n_i - 1)/(n-1)$.

Une séquence de type i et une séquence de type j ($j \neq i$), peuvent aussi coalescer vers une séquence de type k, événement noté C_{ij}^k, si elles sont compatibles (voir figure

3.4(b)). C'est-à-dire que leurs marqueurs ancestraux communs ont des allèles identiques, ou autrement dit que $\forall m \in A^i \cap A^j$, $s_m^{(i)} = s_m^{(j)}$. Rappelons que $s_m^{(i)}$ représente le matériel génétique au marqueur m d'une séquence de type i. Alors la séquence de type k résultant de l'événement de coalescence sera tel que $A^k = A^i \cup A^j$ et que si $s_m^{(i)} = s_m^{(j)} = -$ alors $s_m^{(k)} = -$. Il est possible que la séquence k soit identique à la séquence i ou à la séquence j et nous devons en tenir compte dans le calcul des probabilités. Donc suite à un événement de coalescence de type C_{ij}^k, on se retrouve avec une séquence de plus de type k et avec une séquence en moins au total dans $H_{\tau+1}$, mais si la séquence k est identique à la séquence i (ou j) le nombre de séquences de type k demeure le même. Définissons $\delta_{ik} = 1$ si $k = i$ et égal à 0 sinon, ainsi que $\delta_{jk} = 1$ si $k = j$ et égal 0 sinon. La probabilité d'observer une coalescence C_{ij}^k, sachant que l'événement est une coalescence, est de

$$\frac{(n_k + 1 - \delta_{ik} - \delta_{jk})}{(n-1)}.$$

Maintenant, si le prochain événement est une mutation modifiant le matériel ancestral (voir figure 3.4(c)), il faut déterminer quelles mutations sont possibles et avec quelles probabilités. Une mutation d'une séquence i vers une séquence j au marqueur m, que nous noterons $M_i^j(m)$, — toujours du présent vers le passé — est possible si, et seulement si le marqueur $s_m^{(i)}$ est ancestral et dérivé ($s_m^{(i)} = 1$) et s'il s'agit de l'unique marqueur dérivé en m^e position sur nos séquences au temps de l'événement ($\forall k \neq i$, $s_m^{(k)} \neq 1$ et $n_i = 1$). En effet, comme mentionné précédemment nous supposons qu'une seule mutation peut avoir lieu par marqueur au cours de l'histoire de notre échantillon. La séquence résultante j peut déjà être présente ou non et elle sera telle que $s_m^{(j)} = 0$ et $\forall m' \neq m$ $s_{m'}^{(j)} = s_{m'}^{(i)}$. Suite à cette mutation ($M_i^j(m)$), on retrouve une séquence j de plus, tout en ayant encore le même nombre de séquences au total. De plus, nous assumons que le taux de mutation est constant pour chacun des marqueurs, par conséquent la probabilité d'observer $M_i^j(m)$ — sachant que l'événement est une mutation — est de $(n_j + 1)/n\alpha L$, car chacune des $n_j + 1$ séquences de type j a la possibilité de muter au marqueur m sur un total de $n\alpha L$ mutations possibles (modifiant le matériel ancestral).

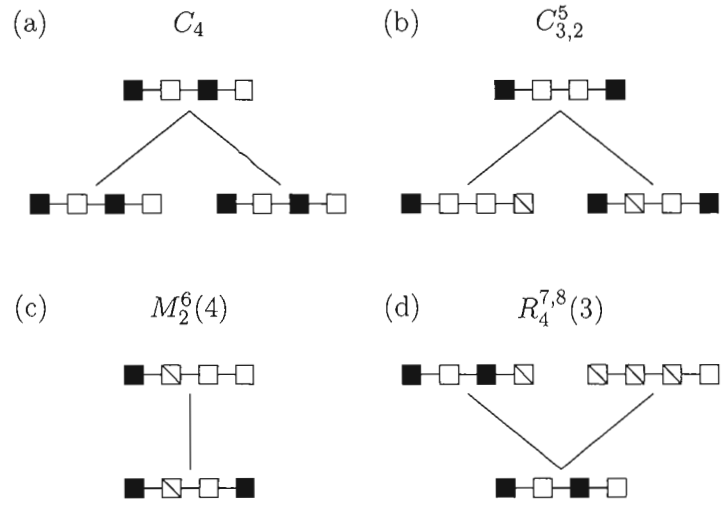


Figure 3.4 Illustrations des différents événements possibles. On représente ici quatre événements possibles à partir de l'ensemble de séquences de la figure 3.3 à la page 49 : (a) une coalescence de deux séquences de type 4 est présentée, (b) une coalescence des séquences 3 et 2 menant à une séquence de type 5, (c) une mutation du deuxième marqueur de la séquence de type 1, vers une séquence de type 6 et (d) une recombinaison à eu lieu dans le troisième intervalle d'une séquence de type 4 vers les séquences 7 et 8.

Déterminons maintenant les possibilités et les probabilités si le prochain événement est plutôt une recombinaison modifiant le matériel ancestral. Une telle recombinaison peut avoir lieu à l'intérieur de n'importe quel intervalle ancestral de nos séquences. Une recombinaison à l'intérieur de l'intervalle m d'une séquence de type i (où $m \in B^i$) vers deux séquences de types j et k sera notée par $R^{jk}_i(m)$. La séquence résultante j (k) sera telle que tous les marqueurs à gauche (droite) de l'intervalle seront identiques à ceux de la séquence de type i, tandis que tous les autres seront non-ancestraux (–)(voir la figure 3.4(d)). C'est-à-dire, si on assiste à une recombinaison $R^{jk}_i(m)$, alors $\forall m_1 \leq m$, $s^{(j)}_{m_1} = s^{(i)}_{m_1}$ et $\forall m_2 > m$, $s^{(j)}_{m_2} = -$, et inversement pour la séquence de type k. Suite à cet événement on se retrouve avec une séquence de plus de type j et une séquence de plus de type k, avec une séquence de plus au total, il y a donc $(n_j + 1)(n_k + 1)$ façons de choisir le couple (j,k) sur un total de $(n + 1)n$ possibilités. De plus, la probabilité d'observer

une recombinaison dans un certain intervalle m dépend aussi de sa longueur et donc de $r_m/\beta r$. Le dénominateur est βr, car nous devons tenir compte de la proportion des séquences qui sont ancestrales puisqu'il s'agit d'une recombinaison modifiant le matériel ancestral. On obtient alors que la probabilité d'observer une telle recombinaison de type $R_i^{jk}(m)$ — toujours sachant que le prochain événement est une recombinaison— est de

$$\frac{(n_j+1)(n_k+1)}{n(n+1)} \cdot \frac{r_m}{\beta r}.$$

Tentons maintenant d'assembler tous les morceaux. $Q(H_\tau)$ représente la probabilité d'observer H_τ et ce, conditionnellement au fait que les événements modifient le matériel ancestral des séquences. Comme mentionné plus haut, cette probabilité dépend de $H_{\tau+1}$. Si l'événement modifiant H_τ est une coalescence C_i, nous allons représenter $H_{\tau+1}$ par $H_\tau + C_i$ et ainsi de suite. Un événement de coalescence affectera toujours le matériel ancestral. Une mutation $M_i^j(m)$ affectera le matériel ancestral si, et seulement si $m \in A^i$. Tandis qu'une recombinaison $R_i^{jk}(m)$ affectera le matériel ancestral si, et seulement si l'intervalle $m \in B^i$. En unissant le tout, on obtient la fonction de récurrence suivante :

$$\begin{aligned} Q(H_\tau) = {} & P(C) \sum_{n_i>1} \frac{(n_i-1)}{n-1} Q(H_\tau + C_i) \\ & + P(C) \sum_{\substack{i \neq j \\ \text{comp.}}} \frac{(n_k+1-\delta_{ik}-\delta_{jk})}{n-1} Q(H_\tau + C_{ij}^k) \\ & + P(M) \sum_i \sum_{\substack{m \in A^i \\ \text{unique}}} \frac{1}{\alpha L} \frac{(n_j+1)}{n} Q(H_\tau + M_i^j(m)) \\ & + P(R) \sum_i \sum_{m \in B^i} \frac{r_m}{\beta r} \frac{(n_j+1)(n_k+1)}{n(n+1)} Q(H_\tau + R_i^{jk}(m)). \end{aligned} \tag{3.1}$$

On a que $Q(H_0) \equiv Q_{r_T}(H_0) = L(r_T)$ est en fait notre fonction de vraisemblance et permettra de créer nos graphes de recombinaison ancestraux.

Démontrons maintenant que l'équation 3.1 dépend du paramètre cherché r_T, représentant la position du TIM. En fait, le paramètre r_T se cache derrière les probabilités de recombinaison, plus précisément dans le quotient r_m/r. Supposons tout d'abord que le TIM est le m^e marqueur de notre séquence, lorsqu'une recombinaison a lieu dans

l'intervalle $m-1$, $r_{m-1} = r_l$ et si la recombinaison à lieu dans l'intervalle m alors $r_m = r_r$ et ces deux valeurs — r_l et r_r — dépendent de r_T. Il est possible de réécrire la dernière sommation de 3.1 pour faire ressortir les paramètres r_r et r_l du quotient r_m/r. Définissons tout d'abord $\delta_m^l = 1$ si $x_{m+1} = r_T$ et 0 sinon, ainsi que $\delta_m^r = 1$ si $x_m = r_T$ et 0 sinon. En supposant maintenant qu'il y a recombinaison dans l'intervalle m, si le TIM est le marqueur $m+1$ alors $r_m = r_l$ et si le TIM est plutôt le marqueur m alors $r_m = r_r$. La dernière sommation de 3.1 peut donc se reformuler comme suit :

$$\sum_i \sum_{m \in B^i} \frac{1}{\beta} \left[\frac{r_m}{r}(1-\delta_m^l)(1-\delta_m^r) + \frac{r_l}{r}\delta_m^l + \frac{r_r}{r}\delta_m^r \right] \frac{(n_j+1)(n_k+1)}{n(n+1)} Q(H_\tau + R_i^{jk}(m)) \tag{3.2}$$

et ainsi, notre paramètre apparaît dans la fonction de vraisemblance. Pour simplifier la notation, nous allons utiliser seulement $Q(H_0)$ sans le paramètre r_T, bien que cette fonction dépend toujours du paramètre à estimer.

3.4 Fonction de vraisemblance et échantillonnage pondéré

À partir de la distribution de probabilité 3.1 de H_τ, il est possible d'obtenir la fonction de vraisemblance. En fait, la fonction de vraisemblance est la distribution de probabilité de H_0 et donc $L(r_T) = Q_{r_T}(H_0)$. Cette équation est de la forme

$$Q(H_\tau) = \sum_{H_{\tau+1}} a(H_\tau, H_{\tau+1}) Q(H_{\tau+1}),$$

où

$$a(H_\tau, H_{\tau+1}) = \begin{cases} P(C)\frac{n_i-1}{n-1} & \text{si } H_{\tau+1} = H_\tau + C_i, \\ P(C)\frac{n_k+1-\delta_{ik}-\delta_{jk}}{(n-1)} & \text{si } H_{\tau+1} = H_\tau + C_{ij}^k, \\ P(M)\frac{1}{\alpha L}\frac{n_{j+1}}{n} & \text{si } H_{\tau+1} = H_\tau + M_i^j(m), \\ P(R)\frac{r_m}{\beta r}\frac{(n_j+1)(n_k+1)}{n(n+1)} & \text{si } H_{\tau+1} = H_\tau + R_i^{jk}(m). \end{cases}$$

Cette équation de vraisemblance ne peut pas être utilisée directement pour déterminer quelle valeur de r_T la maximise et estimer la position du TIM. Bien que cette équation ne tient compte que des généalogies plausibles avec notre échantillon et seulement des événements modifiant le matériel ancestral, il est impossible d'évaluer cette fonction de vraisemblance : il y a trop de généalogies possibles. Par conséquent, nous allons plutôt estimer cette fonction de vraisemblance en générant une grande quantité de graphes possibles à l'aide de la fonction de vraisemblance et de l'échantillonnage pondéré. Pour construire les graphes, nous devons tout d'abord définir qu'elles sont les probabilités associées aux différents événements possibles. Il faut déterminer qu'elles sont les probabilités de transitions de H_τ à $H_{\tau+1}$ de notre chaîne de Markov. Nous avons besoin d'une distribution proposée pour les probabilités conditionnelles $P(H_{\tau+1}|H_\tau)$. Plusieurs distributions proposées peuvent être utilisées, celle utilisée par la méthode MapArg est basée sur celle de Griffiths et Marjoram (1996). Nous allons définir les probabilités conditionnelles comme suit :

$$P(H_{\tau+1}|H_\tau) = \frac{a(H_\tau, H_{\tau+1})}{f(H_\tau, H_{\tau+1})},$$

où

$$f(H_\tau, H_{\tau+1}) = \sum_{H_{\tau+1}} a(H_\tau, H_{\tau+1}).$$

Nous pouvons maintenant réécrire l'équation 3.1 de cette façon :

$$Q(H_\tau) = \sum_{H_{\tau+1}} f(H_\tau, H_{\tau+1}) P(H_{\tau+1}|H_\tau) Q(H_{\tau+1}), \tag{3.3}$$

et ce pour $\tau = 0, \ldots, \tau^*$. Ce qui entraîne que

$$\begin{aligned} Q(H_0) = &\sum_{H_1} \sum_{H_2} \cdots \sum_{H_{\tau^*}} f(H_0, H_1) f(H_1, H_2) \ldots f(H_{\tau^*-1}, H_{\tau^*}) \\ &\times P(H_1|H_0) P(H_2|H_1) \ldots P(H_{\tau^*}|H_{\tau^*-1}) Q(H_{\tau^*}). \end{aligned}$$

Et puisque H_{τ^*} contient seulement la séquence du MRCA, on obtient que $Q(H_{\tau^*}) = 1$ pour cette séquence et 0 pour toutes les autres possibles. Il n'y a, en réalité, qu'une seule

séquence possible pour le MRCA. On peut alors simplifier :

$$Q(H_0) = \sum_{H_1}\sum_{H_2}\cdots\sum_{H_{\tau^*-1}}\left[\prod_{\tau=0}^{\tau^*-1} f(H_\tau, H_{\tau+1}) \times \prod_{\tau=0}^{\tau^*-1} P(H_{\tau+1}|H_\tau)\right].$$

Ce qui est en fait, une représentation d'une espérance en fonction de la distribution proposée P, c'est-à-dire

$$L(r_T) = Q(H_0) = E_P\left[\prod_{\tau=0}^{\tau^*-1} f(H_\tau, H_{\tau+1})\right], \tag{3.4}$$

où E_P signifie l'espérance en fonction de P. Cette façon de réorganiser une équation en espérance nous vient de l'échantillonnage pondéré. Il serait possible d'estimer cette espérance par une moyenne sur K graphes. Mais encore une fois, les calculs seraient énormes. Par exemple, si nous souhaitons estimer la fonction de vraisemblance pour plusieurs valeurs possibles de r_T par intervalle, il faudrait pour chacune d'elles construire K graphes en utilisant cette valeur pour la distribution proposée.

Pour contourner ce problème, nous allons substituer r_T par une valeur conductrice r_{T_0} dans notre distribution proposée qui sera pour l'occasion renommée P_0. En utilisant l'échantillonnage pondéré, nous pouvons réécrire l'équation 3.3 comme suit :

$$Q(H_\tau) = \sum_{H_{\tau+1}} \frac{f(H_\tau, H_{\tau+1})P(H_{\tau+1}|H_\tau)}{P_0(H_{\tau+1}|H_\tau)} P_0(H_{\tau+1}|H_\tau)Q(H_{\tau+1}). \tag{3.5}$$

Et donc de façon analogue à la précédente on obtient

$$L(r_T) = Q(H_0) = E_{P_0}\left[\prod_{\tau=0}^{\tau^*-1} \frac{f(H_\tau, H_{\tau+1})P(H_{\tau+1}|H_\tau)}{P_0(H_{\tau+1}|H_\tau)}\right]. \tag{3.6}$$

Il est alors possible d'estimer cette espérance en simulant K graphes à partir de la distribution P_0 et en faisant la moyenne du produit à l'intérieur de l'espérance 3.6. On obtient alors l'équation suivante :

$$\hat{Q}(H_0) = \frac{1}{K}\sum_{k=0}^{K}\left[\prod_{\tau=0}^{\tau^*-1} \frac{f(H_\tau^{(k)}, H_{\tau+1}^{(k)})P(H_{\tau+1}^{(k)}|H_\tau^{(k)})}{P_0(H_{\tau+1}^{(k)}|H_\tau^{(k)})}\right]. \tag{3.7}$$

Mais l'utilisation d'une seule valeur conductrice offrirait une bien mauvaise estimation de r_T. La méthode MapArg choisit plutôt d'en utiliser une par intervalle entre

marqueurs de position connue. Par conséquent, nous utiliserons $L-2$ valeurs conductrices. Par convention, la valeur conductrice d'un intervalle de position connue sera située au centre de celui-ci. La fonction de vraisemblance est alors estimée intervalle par intervalle de façon indépendante. Si, par exemple, nous sommes rendus à tester l'hypothèse que le TIM est le marqueur m parmi les L marqueurs, alors r_{T_0} sera le point milieu de l'intervalle $[x_{m-1}, x_{m+1}]$ que l'on nommera l'intervalle m. Nous allons par la suite estimer la fonction de vraisemblance pour les valeurs de r_T incluses dans l'intervalle m avec la fonction $\hat{L}_m(r_T)$:

$$\hat{L}_m(r_T) = \frac{1}{K}\sum_{k=0}^{K}\left[\prod_{\tau=0}^{\tau^*-1}\frac{f(H_\tau^{(k)}, H_{\tau+1}^{(k)})P(H_{\tau+1}^{(k)}|H_\tau^{(k)})}{P_0(H_{\tau+1}^{(k)}|H_\tau^{(k)})}\right]. \tag{3.8}$$

Il a été choisi d'utiliser la fonction suivante :

$$\hat{L}(r_T) = \prod_{m=2}^{L-1}\hat{L}_m(r_T), \tag{3.9}$$

comme estimation de la fonction de vraisemblance. L'estimation de r_T, $\hat{r}_T$, est le maximum de cette fonction. Suite au choix de cette fonction, nous devons convenir que $\hat{L}_m(r_T) = 1$ pour toutes les valeurs de r_T situées à l'extérieur de l'intervalle m. Cette procédure sera répétée pour $m = 2, \ldots, L-1$, en tout $K \cdot (L-2)$ graphes seront simulés à partir de l'ensemble des valeurs conductrices.

En pratique, lorsque nous estimons $L_m(r_T)$, nous insérons, pour chacune des séquences, le TIM en position r_{T_0} dans l'intervalle m de position connue. Lorsqu'on insère le TIM, on doit lui inférer son statut, mutant ou non, en fonction du phénotype de la séquence pour le caractère étudié, et de la fonction de pénétrance pour ce caractère.

3.5 Détails de l'estimation de la fonction de vraisemblance

Regardons maintenant de plus près comment la méthode MapArg estime la fonction de vraisemblance pour un intervalle de position connue. En pratique, la vraisemblance est estimée pour un nombre y de valeurs possibles pour le paramètre r_T à l'intérieur de l'intervalle. Lors de la création d'un graphe, il suffit de calculer pour l'ensemble de ses

valeurs

$$\frac{f(H_\tau, H_{\tau+1})P(H_{\tau+1}|H_\tau)}{P_0(H_{\tau+1}|H_\tau)} \tag{3.10}$$

et ce, à chaque étape de la construction du graphe, on fait ensuite le produit des résultats obtenus à chacune de ces différentes étapes, pour chacune des y valeurs considérées. Après avoir construit K graphes, on obtiendra l'estimation de la vraisemblance, pour l'intervalle de position connue, en calculant la moyenne des résultats des différents graphes pour chacune des valeurs considérées.

Explicitons maintenant l'équation 3.10, pour démontrer l'effet des différents événements possibles sur l'estimation de la vraisemblance lors de la construction d'un graphe. On a :

$$\begin{aligned}\frac{f(H_\tau, H_{\tau+1})P(H_{\tau+1}|H_\tau)}{P_0(H_{\tau+1}|H_\tau)} &= \frac{f(H_\tau, H_{\tau+1})a(H_\tau, H_{\tau+1})f_0(H_\tau, H_{\tau+1})}{f(H_\tau, H_{\tau+1})a_0(H_\tau, H_{\tau+1})} \\ &= \frac{a(H_\tau, H_{\tau+1})f_0(H_\tau, H_{\tau+1})}{a_0(H_\tau, H_{\tau+1})},\end{aligned} \tag{3.11}$$

où les fonctions f_0 et a_0 utilisent la valeur r_{T_0} comme position du TIM et la fonction a est plutôt évaluée pour chacune des y valeurs pour lesquelles on estime la vraisemblance dans l'intervalle considéré. Par conséquent $f_0(H_\tau, H_{\tau+1})$ sera constante pour ces y valeurs. Définissons

$$\phi(H_\tau, H_{\tau+1}) = \frac{a(H_\tau, H_{\tau+1})}{a_0(H_\tau, H_{\tau+1})},$$

et concentrons nous plutôt sur les valeurs que prendra cette fonction.

La fonction ϕ dépend de l'événement $\tau + 1$. Considérons chacun des événements possibles, un à la fois, pour déterminer leur influence sur la fonction ϕ. Premièrement, si $H_{\tau+1} = H_\tau + C_i$ alors

$$\phi(H_\tau, H_{\tau+1}) = \frac{P(C)\frac{(n_i-1)}{(n-1)}}{P(C)\frac{(n_i-1)}{(n-1)}} = 1$$

car en fait, la position du TIM dans notre intervalle n'affecte pas la probabilité d'observer une coalescence C_i. De manière analogue et pour des raisons semblables, on obtient lorsque $H_{\tau+1} = H_\tau + C_{ij}^k$,

$$\phi(H_\tau, H_{\tau+1}) = \frac{P(C)\frac{(n_k+1-\delta_{ik}-\delta jk)}{(n-1)}}{P(C)\frac{(n_k+1-\delta_{ik}-\delta jk)}{(n-1)}} = 1.$$

Et si $H_{\tau+1} = H_\tau + M_i^j(m)$, alors

$$\phi(H_\tau, H_{\tau+1}) = \frac{P(M)\frac{(n_j+1)}{(n\cdot\alpha\cdot L)}}{P(M)\frac{(n_j+1)}{(n\cdot\alpha\cdot L)}} = 1,$$

car la probabilité d'observer une mutation à un marqueur m ne dépend pas de la position de ce marqueur sur notre séquence.

Tandis que si $H_{\tau+1} = H_\tau + R_i^{jk}(m)$ alors tout dépend maintenant de l'intervalle de position connue dans lequel notre valeur conductrice se trouve. Et on se souviendra que lors d'un tel événement

$$a(H_\tau, H_{\tau+1}) = P(R)\frac{r_m}{\beta r}\frac{(n_j+1)(n_k+1)}{n(n+1)}.$$

Rappelons que lorsque l'on fait référence à une recombinaison dans l'intervalle m, il s'agit d'un intervalle qui n'est peut-être pas de position connue, la recombinaison peut avoir lieu entre le TIM et un de ses marqueurs voisins. Notons par r_{0l} la distance entre le TIM en position r_{T_0} et le marqueur à sa gauche, et par r_{0r} la distance entre le TIM en position r_{T_0} et le marqueur à sa droite. Considérons trois cas :

1. Si le TIM est le marqueur $\bar{m}$ dans notre suite de $\bar{L}$ marqueurs alors :

$$\phi(H_\tau, H_{\tau+1}) = \frac{P(R)\frac{r_r(n_j+1)(n_k+1)}{\beta\cdot r\cdot n\cdot(n+1)}}{P(R)\frac{r_{0r}(n_j+1)(n_k+1)}{\beta\cdot r\cdot n\cdot(n+1)}} = \frac{r_r}{r_{0r}}.$$

2. Si le TIM est le marqueur $m+1$ dans notre suite de L marqueurs alors :

$$\phi(H_\tau, H_{\tau+1}) = \frac{P(R)\frac{r_l(n_j+1)(n_k+1)}{\beta\cdot r\cdot n\cdot(n+1)}}{P(R)\frac{r_{0l}(n_j+1)(n_k+1)}{\beta\cdot r\cdot n\cdot(n+1)}} = \frac{r_l}{r_{0l}}.$$

3. Sinon

$$\phi(H_\tau, H_{\tau+1}) = \frac{P(R)\frac{r_m(n_j+1)(n_k+1)}{\beta\cdot r\cdot n\cdot(n+1)}}{P(R)\frac{r_m(n_j+1)(n_k+1)}{\beta\cdot r\cdot n\cdot(n+1)}} = \frac{r_m}{r_m} = 1.$$

Notons que puisque r_{T_0} est au centre de l'intervalle on obtient que $r_{0r} = r_{0l}$.

En utilisant les mêmes notations de l'équation 3.2, on peut résumer la situation de la façon suivante :

$$\phi(H_\tau, H_{\tau+1}) = \begin{cases} 1 & \text{si } H_{\tau+1} \text{ est une } M \text{ ou une } C, \\ \frac{r_p(1-\delta_p^l)(1-\delta_p^r)+r_l\delta_p^l+r_r\delta_p^r}{r_p(1-\delta_p^l)(1-\delta_p^r)+r_{0l}\delta_p^l+r_{0r}\delta_p^r} & \text{si } R(p). \end{cases}$$

S'il n'y avait pas d'événements de recombinaison, l'estimation de la vraisemblance serait constante pour les valeurs de r_T à l'intérieur d'un intervalle de position connue. La forme polynomiale de l'estimation proposée pour la fonction de vraisemblance est due seulement aux événements de recombinaison.

3.6 Derniers développements

3.6.1 Vraisemblance composite et conditionnelle

Une des problématiques communes à plusieurs méthodes de cartographie génétique est le temps de calcul informatique nécessaire. Pour diminuer ce temps de calcul, plusieurs de ces méthodes utilisent les vraisemblances composites (Larribe et Fearnhead, 2010). Les méthodes de vraisemblance composite permettent d'estimer la vraisemblance en calculant tout d'abord une vraisemblance dite marginale, pour des sous-ensembles d'observations généralement indépendants. On obtient ensuite l'estimation de la fonction de vraisemblance en effectuant le produit de ces vraisemblances marginales. La méthode de cartographe génétique MapArg, présentée dans ce chapitre, peut rapidement être exigeante en temps de calcul informatique. Plus le nombre de marqueurs utilisés par cette méthode est grand, plus les graphes qu'elle génère seront hauts, c'est-à-dire plus le temps avant l'atteinte du MRCA de l'échantillon sera grand. Pour réduire les temps de calcul informatique, les concepteurs de MapArg ont développé une fonction de vraisemblance composite et conditionnelle pour estimer la position du TIM (Larribe et Lessard, 2008). Cette modification utilise des sous-ensembles de marqueurs pour construire les graphes. Nous présentons brièvement ici cette modification.

L'idée consiste à évaluer la vraisemblance pour chaque intervalle, mais en utilisant un sous-ensemble de marqueurs de position connue lors de la construction des graphes. On nommera fenêtre de marqueurs un sous-ensemble de marqueurs consécutifs sur la séquence. Les fenêtres de marqueurs seront de taille d fixée. La première fenêtre est constituée des d premiers marqueurs de position connue, la deuxième fenêtre contiendra inclusivement les marqueurs situés entre les deuxième et $(d+1)^e$ marqueurs de position

connue, et ainsi de suite. La figure 3.5, inspirée de Larribe et Lessard (2008), illustre deux exemples de fenêtres pour des séquences de six marqueurs de position connue ($L = 7$). Pour chacune de ces fenêtres, la vraisemblance est évaluée indépendamment, intervalle par intervalle, de la même manière que présentée précédemment, la différence réside dans le nombre de marqueurs utilisés, avant il y en avait L maintenant, il y en a $d + 1$ (les d de positions connues et le TIM que l'on insère dans l'intervalle).

Il reste maintenant à joindre ces vraisemblances marginales. Notons que la vraisemblance peut être évaluée plus d'une fois pour un même intervalle, car cet intervalle peut appartenir à plus d'une fenêtre (par exemple l'intervalle situé entre les marqueurs 3 et 4 de la figure 3.5(b) est inclus dans les trois fenêtres). La vraisemblance pour un intervalle entre marqueurs de position connue sera estimée à l'aide de la moyenne géométrique des différentes vraisemblances marginales obtenues pour cet intervalle, comme cela est usuel dans ce contexte (Lindsay, 1988 ; Varin et Vidoni, 2005). Mais chacune de ces vraisemblances marginales a été obtenue en utilisant des fenêtres différentes, et est en fait des vraisemblances marginales conjointes de la fenêtre incluant le TIM et de la fenêtre sans le TIM. C'est pourquoi Larribe et Lessard (2008) suggèrent d'utiliser une

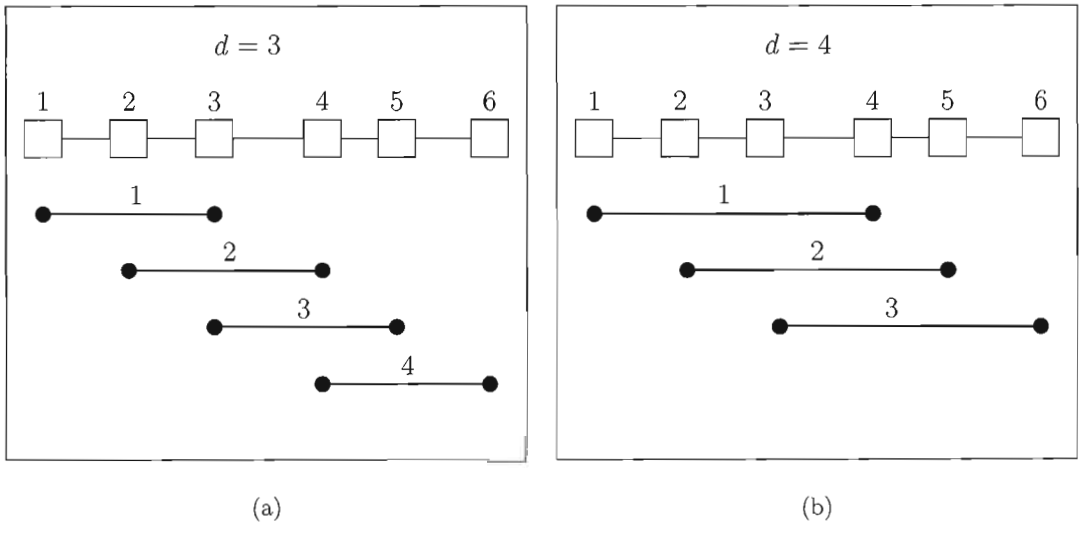


Figure 3.5 Exemples de fenêtre de marqueurs, en (a) les fenêtres sont de longueur 3, tandis qu'en (b), elles sont de longueur 4.

vraisemblance composite conditionnelle. En réalité, la vraisemblance pour un intervalle entre marqueurs de position connue sera estimée à l'aide de la moyenne géométrique des différentes vraisemblances marginales conditionnelles à la fenêtre, obtenues pour cet intervalle.

Nous pouvons formaliser le tout en se basant sur la notation de Larribe et Lessard (2008). Le nombre total de fenêtres est $G = L - d$ et les fenêtres seront numérotées de 1 à G de la gauche vers la droite. L'intervalle m est celui situé entre les $(m-1)^e$ et m^e marqueurs de position connue, il appartiendra à la fenêtre g si, et seulement si g est situé entre

$$\underline{g}(m) = \max(1, m + 1 - d)$$

et

$$\overline{g}(m) = \max(m - 1, L - d).$$

Nous noterons par $L_{m,g}(r_T)$ la vraisemblance marginale pour l'intervalle m évaluée dans la fenêtre g. La fonction de vraisemblance composite (CL) utilisant des fenêtres de d marqueurs peut s'écrire

$$CL_d(r_T) = \prod_{m=2}^{L-1} \left(\prod_{g=\underline{g}(m)}^{\overline{g}(m)} L_{m,g}(r_T) \right)^{w_m},$$

où w_m est l'inverse du nombre de fenêtres contenant l'intervalle m, c'est-à-dire :

$$w_m = \frac{1}{\overline{g}(m) - \underline{g}(m) + 1}.$$

En notant $H_0^{r_T}$, l'ensemble des séquences de l'échantillon avec le TIM à l'intérieur de l'intervalle m, et H_0^g, l'ensemble des marqueurs inclus dans la fenêtre g, alors la vraisemblance marginale se réécrit :

$$L_{m,g}(r_T) = Q(H_0^{r_T}, H_0^g)$$

et on obtient la vraisemblance marginale conditionnelle suivante :

$$L_{m,g}(r_T | H_0^g)) = \frac{Q(H_0^{r_T}, H_0^g)}{Q(H_0^g)}.$$

La fonction de vraisemblance composite conditionnelle (CCL) est alors :

$$CCL_d(r_T) = \prod_{m=2}^{L-1} \left(\prod_{g=\underline{g}(m)}^{\overline{g}(m)} L_{m,g}(r_T|H_0^g) \right)^{w_m}.$$

La figure 3.6 présente un exemple de l'estimation de la vraisemblance avec MapArg pour la base de données simulée E qui sera présentée au chapitre 5. Nous avons fait cinq simulations de 500 graphes par intervalles en utilisant des séquences de 50 marqueurs et des fenêtres de 5 marqueurs. Les courbes noires représentent les différentes répétitions ; la bleue, la combinaison de ces répétitions, soit une simulation de 2 500 graphes par intervalle. En effet, puisque les graphes construits sont indépendants les uns des autres, il est possible de réunir tous les graphes créés lors de différentes simulations pour obtenir une vraisemblance basée sur un plus grand nombre de graphes. Les différentes répétitions

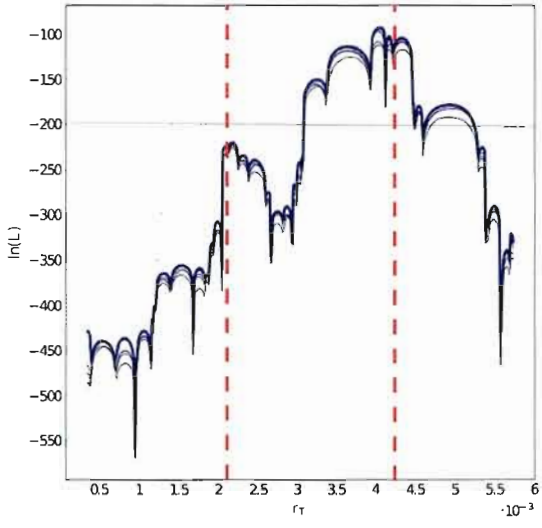


Figure 3.6 Exemple de l'estimation de la vraisemblance avec MapArg pour la base de données E, à partir de cinq répétitions de 500 graphes par intervalle en utilisant des séquences de 50 marqueurs et des fenêtres de 5 marqueurs. Une courbe noire représente une répétition, la bleue la combinaison de ces répétitions, soit une simulation de 2 500 graphes par intervalle.

nous permettent de juger de la variabilité de l'estimation obtenue. La forme arrondie de l'estimation de la vraisemblance pour un intervalle est caractéristique de l'utilisation d'une valeur conductrice.

L'utilisation des fenêtres de marqueurs permet non seulement de diminuer les temps de calculs de la méthode de cartographie MapArg, mais améliore aussi sa stabilité.

3.6.2 Autres développements

Il est aussi mentionné par Larribe et Lessard (2008) que la distribution proposée utilisée est loin d'être idéale et qu'il est difficile de trouver des généalogies très vraisemblables. Ils expliquent que la vraisemblance peut prendre de très grandes valeurs lorsqu'une généalogie très probable est trouvée et ensuite la vraisemblance diminue tranquillement avec la simulation de généalogies moins probables. Si l'on compare plusieurs estimations de la vraisemblance pour un même ensemble de données, cette particularité peut créer une variabilité dans les résultats obtenus. Par exemple, pour un même intervalle, une des simulations peut s'être arrêtée immédiatement après avoir trouvé une généalogie très probable, la vraisemblance sera alors plus haute que pour une autre simulation ayant arrêté sans récemment avoir trouvé une généalogie probable. Pour contrer cette variabilité, les auteurs suggèrent de faire plusieurs simulations pour un même échantillon et de combiner les vraisemblances obtenues en prenant leurs moyennes géométriques. Ce qui peut être vu comme le résultat d'une vraisemblance composite, où les sous-ensembles de marqueurs sont en fait tous les marqueurs.

Mentionnons aussi le travail de Gabrielle Boucher (2009), qui pour son mémoire de maîtrise a travaillé à l'adaptation de MapArg à la réalité diploïde. Elle a développé un algorithme EM conditionnel au phénotype permettant d'estimer les haplotypes d'un échantillon de génotype. Car la particularité des données relatives à des individus diploïdes est la méconnaissance des phases, on ignore quels allèles appartiennent au même haplotype.

CHAPITRE IV

CARTOGRAPHIE DE DEUX GÈNES À LA FOIS

MapArg permet d'estimer la position d'un seul caractère influençant une maladie, mais en réalité, plusieurs maladies complexes sont le résultat d'une multitude de gènes et de facteurs environnementaux. Pour cette raison, il est important de développer des méthodes de cartographie génétique tenant compte de ces difficultés. Nous avons décidé de nous intéresser à la problématique de la cartographie de caractères polygéniques. Dans ce chapitre, nous explorons une façon de modifier la méthode de cartographie génétique MapArg, qui nous permettra d'estimer la position de deux gènes causant une maladie. Nous expliquerons tout d'abord, pourquoi la fonction de vraisemblance garde la même forme. Ensuite, les détails du calcul de la fonction de vraisemblance seront présentés. Puis, la question des différentes modélisations possibles sera abordée.

4.1 Paramétrisation du problème

Nous avons vu que MapArg utilise un échantillon de séquences de marqueurs, suppose que la maladie est causée par une seule mutation sur un seul gène (TIM), et estime la position de ce TIM par le maximum d'une fonction de vraisemblance. Cette fonction de vraisemblance est estimée indépendamment pour chaque intervalle entre marqueurs de position connue. Pour ce faire, on suppose que la position du TIM est au centre de l'intervalle (il s'agit de la valeur conductrice), ensuite on simule plusieurs milliers de généalogies pour l'échantillon de séquence et on évalue la vraisemblance de la position du TIM pour cet intervalle.

Si on suppose plutôt que la maladie est causée par l'effet de deux mutations situées sur deux gènes différents (nommons-les TIM1 et TIM2), le but est maintenant d'estimer la position de ces deux TIM. De manière analogue à MapArg, l'échantillon sera aussi composé de séquences de marqueurs. Si les positions des deux TIM sont connues, la simulation des différentes généalogies possibles de l'échantillon se fait de façon similaire à MapArg, car les probabilités d'observer des événements de coalescence, mutation ou recombinaison demeurent les mêmes. Donc, la fonction de vraisemblance demeure la même, mais maintenant elle ne dépend plus seulement d'un paramètre (la position du TIM), mais plutôt de deux (les positions des TIM1 et TIM2). Avec MapArg on évaluait cette vraisemblance pour tous les intervalles où le TIM pouvait se trouver. Maintenant, nous allons devoir évaluer la vraisemblance pour tous les couples d'intervalles où peuvent se trouver le TIM1 et le TIM2. Par exemple, nous allons supposer que le TIM1 se trouve dans le i^e intervalle et le TIM2 dans le j^e et nous allons évaluer la vraisemblance pour ce couple d'intervalles. En fait, nous l'évaluerons pour tous les couples possibles. La fonction de vraisemblance peut par conséquent être représentée en trois dimensions ; une pour la position du TIM1, une autre pour la position du TIM2 et finalement, une dernière pour la valeur de la fonction de vraisemblance. Le maximum de cette surface nous indiquera les positions du TIM1 et du TIM2 les plus vraisemblables.

Mais avant d'entrer dans les détails du calcul de la vraisemblance et de décrire la forme que prendra cette surface, présentons les paramètres et les variables qui seront utilisés en se basant sur la notation définie au chapitre précédent. Les séquences génétiques utilisées formant l'échantillon sont composées d'une suite de L marqueurs (SNP), chacun étant mutant ou non-mutant. Puisque nous cherchons à estimer la position de deux gènes causant la maladie, $(L-2)$ marqueurs ont une position connue sur la séquence. Pour l'instant, nous nommerons le premier gène sur la séquence de position inconnue le TIM1 et le deuxième le TIM2. De cette façon, le TIM1 est toujours à gauche du TIM2 sur la séquence de marqueurs. Cette paramétrisation se justifie sous certaines conditions et dépend de l'action des deux TIM sur le statut malade ou non de l'individu. Nous aborderons le sujet des différentes paramétrisations possibles à la fin de ce chapitre.

Rappelons que les marqueurs sont numérotés de 1 à L, de la gauche vers la droite sur la séquence. La position en centimorgan du marqueur i est noté x_i. L'intervalle m est l'intervalle situé entre le m^e et le $(m+1)^e$ marqueur et r_m est sa longueur. La longueur d'une séquence est notée r, on suppose les distances additives ainsi que l'absence d'interférence par conséquent, $r = \sum_{i=1}^{L-1} r_i$. Il y a $L-1$ intervalles, dont $L-3$ de longueurs connues. Nous noterons par r_{T1} la distance entre le premier marqueur et le TIM1, et par r_{T2} la distance entre le premier marqueur de la séquence et le TIM2. On suppose que le TIM1 et le TIM2 ne sont ni le premier marqueur, ni le dernier. De plus, r_{l_1} (r_{l_2}) sera la distance entre le TIM1 (TIM2) et le premier marqueur à sa gauche, et r_{r_1} (r_{r_2}) sera la distance entre le TIM1 (TIM2) et le premier marqueur à sa droite. Les figures 4.1 et 4.2 illustrent cette notation dans deux situations différentes, l'une où les deux TIM ne sont pas situés à l'intérieur du même intervalle de position connue et l'autre où ils le sont. On remarque sur la figure 4.2, que r_{r_1} et r_{l_2} sont égaux lorsque le TIM1 et le TIM2 appartiennent au même intervalle de position connue.

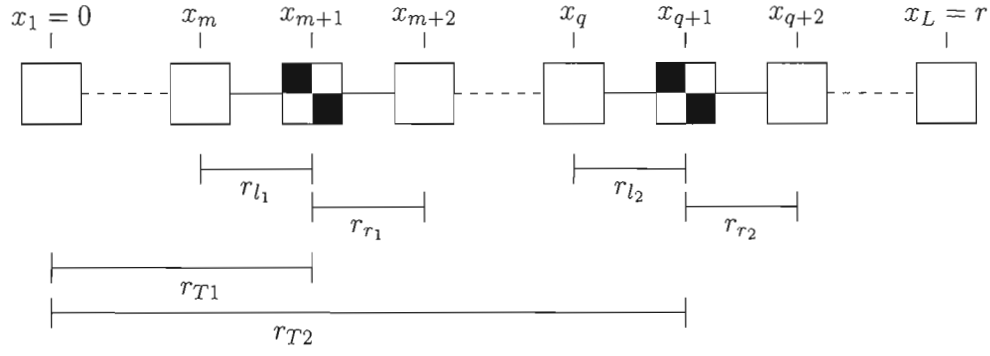


Figure 4.1 Séquence de marqueurs. Le TIM1 et le TIM2 sont représentés par un carré quadrillé (▪), et ne sont pas dans le même intervalle de longueur connue. Les positions, en centimorgan, des marqueurs sur la séquence sont représentées par les x_i. La position du TIM i est notée par r_{Ti}, et les longueurs des intervalles situés à gauche et à droite du TIM i sont notées respectivement r_{l_i}, r_{r_i}.

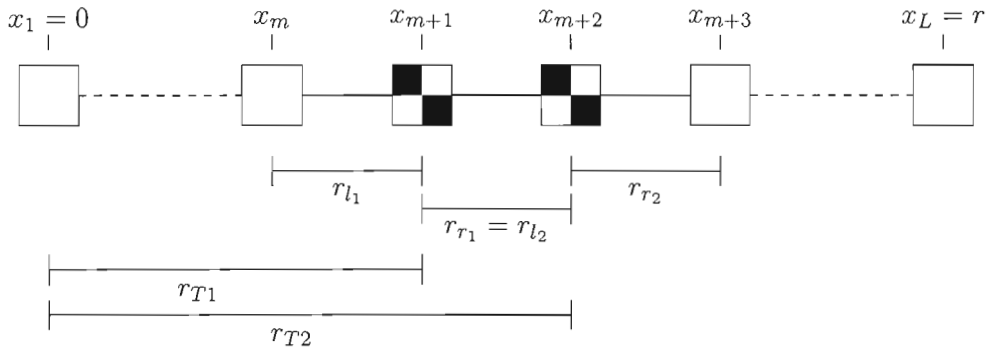


Figure 4.2 Séquence de marqueurs. Le TIM1 et le TIM2 sont représentés par un carré quadrillé (▚), et sont situés dans le même intervalle de longueur connue. La position du TIM i est notée par r_{Ti}, et les longueurs des intervalles situés à gauche et à droite du TIM i sont notées respectivement r_{l_i}, r_{r_i}. Remarquons que $r_{r_1} = r_{l_2}$, lorsque les TIM sont situés dans le même intervalle.

Les deux paramètres à estimer sont r_{T1} et r_{T2}. Pour obtenir une estimation, nous utilisons une fonction de vraisemblance similaire à celle utilisée par MapArg (l'équation 3.6), basée sur la récurrence 3.1 présentée en page 56. Les graphes sont construits de la même façon, et chaque événement (coalescence, mutation, recombinaison) a les mêmes probabilités de se produire. La différence entre MapArg et la méthode proposée ici, réside dans l'ensemble des valeurs conductrices utilisées pour évaluer la fonction de vraisemblance. Il y a maintenant deux paramètres à estimer, les valeurs conductrices seront donc tous les couples possibles (r_{0_1}, r_{0_2}) de positions pour le TIM1 et le TIM2 ; la fonction de vraisemblance sera une surface dans $\mathbb{R}^3$. Avec MapArg, la fonction de vraisemblance est évaluée indépendamment pour chaque intervalle entre marqueurs de positions connues. La valeur conductrice est située au centre de l'intervalle évalué, il y a $L - 2$ valeurs conductrices. Dans la modification proposée ici, nous évaluons la vraisemblance pour des intervalles dans $\mathbb{R}^2$. C'est-à-dire que pour chaque intervalle, il y a une valeur conductrice pour le TIM1 et une autre pour le TIM2. Par exemple, r_{0_1} peut être situé dans le premier intervalle de position connue et r_{0_2} dans le sixième, alors dans ce cas, l'intervalle de $\mathbb{R}^2$ pour lequel la vraisemblance est évaluée est : $(x_1; x_3) \times (x_7; x_9)$

(x_2 étant la position du TIM1 et x_8 celle du TIM2). Par convention, r_{0_1} et r_{0_2} sont situés au centre de leurs intervalles entre marqueurs de position connue, lorsqu'ils ne sont pas situés dans le même intervalle. S'ils le sont, r_{0_1} sera situé au premier tiers et r_{0_2} au deuxième tiers de l'intervalle. La figure 4.3 illustre, pour un exemple où $L = 7$, la région de $\mathbb{R}^2$ pour laquelle la vraisemblance est évaluée. On trouve sous les axes, une représentation des marqueurs de position connue. La valeur conductrice utilisée pour l'intervalle de $\mathbb{R}^2$ en bleu, est représentée par une étoile (★).Puisque selon notre paramétrisation le TIM1 est toujours à gauche du TIM2, la région de $\mathbb{R}^2$ pour laquelle la vraisemblance est évaluée est triangulaire.

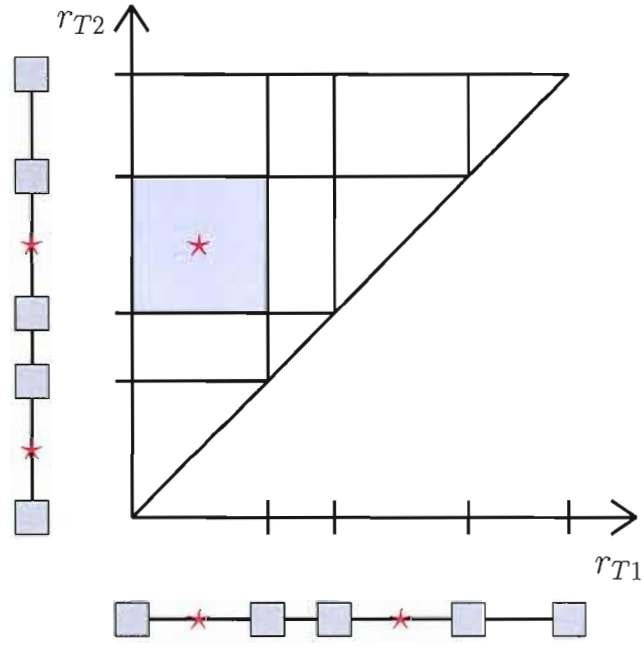


Figure 4.3 Intervalles de $\mathbb{R}^2$ pour l'évaluation de la vraisemblance, $L = 7$ et les 5 marqueurs sont représentés sous les axes. L'étoile (★) représente la valeur conductrice pour l'intervalle en bleu.

4.2 Un aperçu de la vraisemblance

Nous avons vu que l'équation de vraisemblance demeure similaire lorsque nous souhaitons estimer la position de deux TIM. Nous allons maintenant démontrer que cette vraisemblance dépend des paramètres inconnus r_{T1} et r_{T2}, car bien que ces deux paramètres n'apparaissent pas explicitement dans la fonction de vraisemblance, il est possible de les faire ressortir de l'équation. Ensuite, nous regarderons de plus près la forme que prendra la surface de vraisemblance pour différentes situations.

On se souviendra que la vraisemblance $L(r_T)$ est égale à $Q(H_0)$ (qui maintenant dépend des paramètres r_{T1} et r_{T2}) et est basée sur la récurrence suivante :

$$\begin{aligned} Q(H_\tau) = {} & P(C) \sum_{n_i>1} \frac{(n_i-1)}{n-1} Q(H_\tau + C_i) \\ & + P(C) \sum_{\substack{i\neq j \\ \text{comp.}}} \frac{(n_k+1-\delta_{ik}-\delta_{jk})}{n-1} Q(H_\tau + C_{ij}^k) \\ & + P(M) \sum_i \sum_{\substack{m\in A^i \\ \text{unique}}} \frac{1}{\alpha L} \frac{(n_j+1)}{n} Q(H_\tau + M_i^j(m)) \\ & + P(R) \sum_i \sum_{p\in B^i} \frac{r_p}{\beta r} \frac{(n_j+1)(n_k+1)}{n(n+1)} Q(H_\tau + R_i^{jk}(p)). \end{aligned} \tag{4.1}$$

En fait, $L(r_{T1},r_{T2})$ est aussi basée sur cette récurrence, puisque l'équation de vraisemblance demeure la même lorsque nous souhaitons estimer la position de deux mutations causant une maladie. Nous pouvons alors démontrer que la récurrence 4.1 dépend des paramètres r_{T1} et r_{T2}. Comme pour $L(r_T)$, les valeurs des paramètres à estimer n'influencent pas les probabilités de coalescence et de mutation, mais influencent plutôt les probabilités de recombinaison. On remarque, dans la dernière sommation, le quotient r_p/r, où r_p représente la longueur de l'intervalle p. La sommation est faite sur tous les intervalles dits ancestraux de la séquence de type i, puis faite sur tous les types de séquences. Il est important de se souvenir que pour cette récurrence et pour la vraisemblance, les deux TIM font partie des L marqueurs composant les séquences. Supposons que nous sommes dans la situation représentée par la figure 4.1 en page 71, c'est-à-dire que le TIM1 est le $(m+1)^e$ marqueur de la séquence et le TIM2 est le $(q+1)^e$ marqueur. Alors les longueurs

des intervalles m et $m+1$ dépendront de r_{T1}, car $r_m = r_{l_1}$ et $r_{m+1} = r_{r1}$ et bien sûr, r_{r_l} et r_{r_1} dépendent de la position du TIM1. Similairement, les longueurs des intervalles q et $q+1$ dépendront de r_{T2}. On arrive alors aux égalités suivantes : $r_m = r_{l_1}$, $r_{m+1} = r_{r_1}$, $r_q = r_{l_2}$ et $r_{q+1} = r_{r_2}$. Il est possible de réécrire le quotient r_p/r de la récurrence 4.1 pour faire ressortir sa dépendance avec les paramètres à estimer. Présentons tout d'abord quelques variables ; soient :

$$\delta_p^{l_1} = \begin{cases} 1 & \text{si } x_{p+1} = r_{T1}, \\ 0 & \text{sinon ;} \end{cases}$$

$$\delta_p^{r_1} = \begin{cases} 1 & \text{si } x_p = r_{T1}, \\ 0 & \text{sinon ;} \end{cases}$$

$$\delta_p^{l_2} = \begin{cases} 1 & \text{si } x_{p+1} = r_{T2} \text{ ET } x_p \neq r_{T_1}, \\ 0 & \text{sinon ;} \end{cases}$$

$$\delta_p^{r_1} = \begin{cases} 1 & \text{si } x_p = r_{T2}, \\ 0 & \text{sinon.} \end{cases}$$

Il est alors possible de réécrire r_p/r par :

$$\frac{1}{r}\left[r_p(1-\delta_p^{l_1})(1-\delta_p^{r_1})(1-\delta_p^{l_2})(1-\delta_p^{r_2}) + r_{l_1}\delta_p^{l_1} + r_{r_1}\delta_p^{r_1} + r_{l_2}\delta_p^{l_2} + r_{r_2}\delta_p^{r_2}\right]. \qquad (4.2)$$

Notons que si le TIM1 et le TIM2 sont deux marqueurs côte-à-côte, comme l'illustre la figure 4.2, alors $r_{r_1} = r_{l_2}$ et l'égalité 4.2 demeure vraie de part la définition de $\delta_p^{l_2}$. Nous venons de démontrer que la vraisemblance $L(r_{T1},r_{T2})$ dépend bien des deux paramètres à estimer (r_{T1},r_{T2}).

Au chapitre précédent, nous avons montré que la forme polynomiale de la fonction de vraisemblance (l'estimation proposée), lorsqu'elle est évaluée à l'intérieur d'un intervalle fixé de position connue, était due seulement aux recombinaisons ayant eu lieu dans cet intervalle. On se souviendra que cette vraisemblance, pour un intervalle, était estimée par une moyenne sur K graphes du produit

$$\prod_{\tau=0}^{\tau^*-1} f(H_\tau^{(k)}, H_{\tau+1}^{(k)}) P(H_{\tau+1}^{(k)} | H_\tau^{(k)}) / P_0(H_{\tau+1}^{(k)} | H_\tau^{(k)})$$

pour $k = 1, \ldots, K$. On avait un vecteur de longueur y composé de points équidistants à l'intérieur de notre intervalle de position connue, le TIM était placé au centre de cet intervalle, en position r_0 (la valeur conductrice). Il suffisait alors d'évaluer

$$\frac{f(H_\tau, H_{\tau+1})P(H_{\tau+1}|H_\tau)}{P_0(H_{\tau+1}|H_\tau)} = \frac{a(H_\tau, H_{\tau+1})f_0(H_\tau, H_{\tau+1})}{a_0(H_\tau, H_{\tau+1})} \tag{4.3}$$

à chaque étape de la construction du graphe et ce, pour chaque point du vecteur y (ou valeurs possibles pour r_T) pour obtenir une estimation de la vraisemblance. Les fonctions f_0 et a_0 utilisent la valeur conductrice r_0 comme position du TIM et la fonction a est évaluée pour chacune des valeurs du vecteur y. Puisque $f_0(H_\tau, H_{\tau+1})$, est constante pour les différentes valeurs de y, on s'était intéressé au quotient $a(H_\tau, H_{\tau+1})/a_0(H_\tau, H_{\tau+1})$ que nous avions renommé $\phi(H_\tau, H_{\tau+1})$. Rappelons que

$$a(H_\tau, H_{\tau+1}) = \begin{cases} P(C)\frac{n_i - 1}{n-1} & \text{si } H_{\tau+1} = H_\tau + C_i, \\ P(C)\frac{n_k + 1 - \delta_{ik} - \delta_{jk}}{(n-1)} & \text{si } H_{\tau+1} = H_\tau + C_{ij}^k, \\ P(M)\frac{1}{\alpha L}\frac{n_{j+1}}{n} & \text{si } H_{\tau+1} = H_\tau + M_i^j(m), \\ P(R)\frac{r_m}{\beta r}\frac{(n_j+1)(n_k+1)}{n(n+1)} & \text{si } H_{\tau+1} = H_\tau + R_i^{jk}(m). \end{cases}$$

Maintenant, pour estimer $L(r_{T1}, r_{T2})$, nous allons devoir évaluer la fonction de vraisemblance indépendamment pour chacun des intervalles de $\mathbb{R}^2$ possibles et définis plus tôt. Pour chacun de ces intervalles, nous allons avoir un couple de valeurs conductrices (r_{0_1}, r_{0_2}) et un vecteur y_1, respectivement y_2, de points équidistants situés dans l'intervalle de position connue où se trouve le TIM1, respectivement le TIM2. Pour obtenir une estimation de la vraisemblance, on évaluera l'équation 4.3 à chaque étape de la construction des graphes et le résultat pour l'intervalle m sera inscrit dans une matrice M, où l'élément (m_{ij}) de la matrice est l'estimation de la vraisemblance pour la i^e position du TIM1 (c'est-à-dire pour le i^e élément du vecteur y_1) et la j^e position du TIM2 (le j^e élément du vecteur y_2). On estime la vraisemblance pour toutes les combinaisons possibles de position pour les TIM1 et TIM2.

Tentons maintenant de comprendre quelle valeur prendra l'équation 4.3 pour différentes situations. Premièrement, notons que les fonctions f_0 et a_0 utilisent les

valeurs conductrices r_{0_1} et r_{0_2}, tandis que nous évaluerons la fonction a pour toutes les combinaisons possibles de positions pour le TIM1 et le TIM2. Par conséquent, la fonction $f_0(H_\tau, H_{\tau+1})$ est encore une fois constante pour chacune de ces combinaisons. Nous allons donc nous intéresser au quotient $a(H_\tau, H_{\tau+1})/a_0(H_\tau, H_{\tau+1})$ et aux différentes valeurs qu'il peut prendre. Remarquons tout d'abord, que ce quotient vaut 1 si on observe un événement de coalescence ou de mutation. Ceci est dû au fait que la position des marqueurs (ou la longueur des intervalles) n'entre pas dans le calcul de la fonction a lorsque l'on observe un de ces événements. Le seul endroit où la longueur des intervalles intervient dans la fonction a est lorsque l'on observe une recombinaison : si $H_{\tau+1} = H_\tau + R_i^{jk}(m)$ alors

$$a(H_\tau, H_{\tau+1}) = P(R)\frac{(r_m)}{(\beta r)}\frac{(n_j+1)(n_k+1)}{n(n+1)},$$

et la longueur de l'intervalle apparaît, il s'agit de r_m. Puisque nous nous intéressons à ce qui apporte de la variabilité dans la vraisemblance, nous allons nous concentrer sur le quotient $a(H_\tau, H_{\tau+1})/a_0(H_\tau, H_{\tau+1})$ lorsque l'on observe une recombinaison. Nous avons vu, que l'on pouvait réécrire

$$r_m = (1-\delta_m^{l_1})(1-\delta_m^{r_1})(1-\delta_m^{l_2})(1-\delta_m^{r_2}) + r_{l_1}\delta_m^{l_1} + r_{r_1}\delta_m^{r_1} + r_{l_2}\delta_m^{l_2} + r_{r_2}\delta_m^{r_2}$$

pour faire ressortir la dépendance entre la vraisemblance et les paramètre inconnus r_{T1} et r_{T2}. Ainsi, lorsque $H_{\tau+1} = H_\tau + R_i^{jk}(p)$, on obtient que

$$\frac{a(H_\tau, H_{\tau+1})}{a_0(H_\tau, H_{\tau+1})} = \frac{(1-\delta_p^{l_1})(1-\delta_p^{r_1})(1-\delta_p^{l_2})(1-\delta_p^{r_2}) + r_{l_1}\delta_p^{l_1} + r_{r_1}\delta_p^{r_1} + r_{l_2}\delta_p^{l_2} + r_{r_2}\delta_p^{r_2}}{(1-\delta_p^{l_1})(1-\delta_p^{r_1})(1-\delta_p^{l_2})(1-\delta_p^{r_2}) + r_{0l_1}\delta_p^{l_1} + r_{0r_1}\delta_p^{r_1} + r_{0l_2}\delta_p^{l_2} + r_{0r_2}\delta_p^{r_2}}, \tag{4.4}$$

où les $\delta_\prime''$ sont ceux définis plus tôt, et où r_{0l_i} (respectivement r_{0r_i}) est la distance entre r_{0_i} et le marqueur à sa gauche (respectivement le marqueur à sa droite). Puisque r_{0_i} est au centre de l'intervalle, alors $r_{0l_i} = r_{0r_i}$.

Par exemple, supposons que nous souhaitons estimer la vraisemblance pour l'intervalle de $\mathbb{R}^2$ où le TIM1 est le $(m+1)^e$ marqueur et le TIM2 le $(q+1)^e$ marqueur ; situation illustrée par la figure 4.4. Si on observe une recombinaison dans l'intervalle p (notons cet événement $Re(p)$), regardons comment se simplifie l'équation 4.4 :

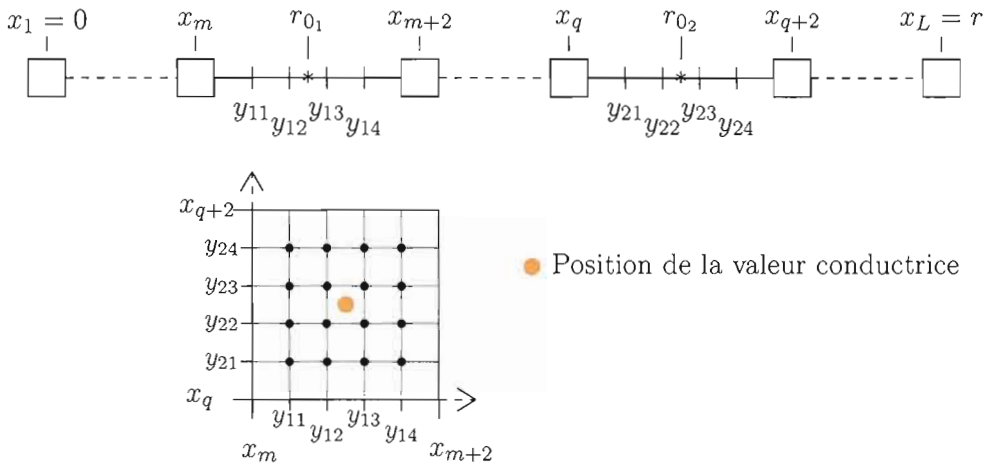


Figure 4.4 Exemple 1 d'un intervalle pour l'évaluation de la vraisemblance. Les TIM ne sont pas situés dans le même intervalle. Les vecteurs y_1 et y_2 sont de longueur 4.

- Si $Re(m)$ alors
$$\frac{a(H_\tau, H_{\tau+1})}{a_0(H_\tau, H_{\tau+1})} = \frac{r_{l_1}}{r_{0l_1}},$$
- Si $Re(m+1)$ alors
$$\frac{a(H_\tau, H_{\tau+1})}{a_0(H_\tau, H_{\tau+1})} = \frac{r_{r_1}}{r_{0r_1}},$$
- Si $Re(q)$ alors
$$\frac{a(H_\tau, H_{\tau+1})}{a_0(H_\tau, H_{\tau+1})} = \frac{r_{l_2}}{r_{0l_2}},$$
- Si $Re(q+1)$ alors
$$\frac{a(H_\tau, H_{\tau+1})}{a_0(H_\tau, H_{\tau+1})} = \frac{r_{r_2}}{r_{0r_2}},$$
- Sinon
$$\frac{a(H_\tau, H_{\tau+1})}{a_0(H_\tau, H_{\tau+1})} = 1.$$

Nommons $M_{\tau+1}$ la matrice dont l'entrée (m_{ij}) est le résultat de l'évaluation du quotient $a(H_\tau, H_{\tau+1})/a_0(H_\tau, H_{\tau+1})$ pour $r_{T_1} = y_{1i}$ et $r_{T_2} = y_{2j}$ (où y_{kl} est la l^e composante du vecteur y_k). Si on observe une $R(m)$ ou $R(m+1)$, on remarque que le quotient varie seulement par rapport à la position du TIM1, alors toutes les colonnes de la matrice $M_{\tau+1}$ seront équivalentes. De manière analogue, si on observe plutôt une $R(p)$ ou $R(p+1)$, le

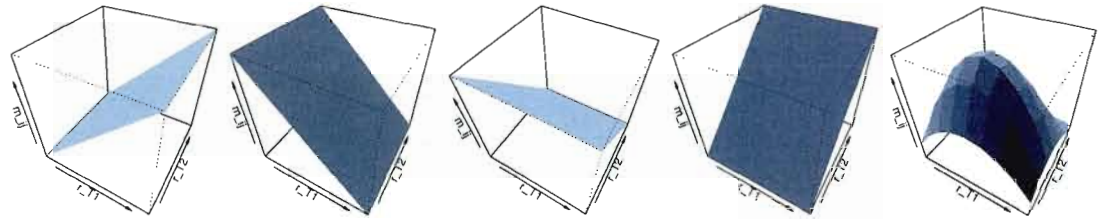
quotient variera seulement par rapport à la position du TIM2, ce qui entraîne que toutes les lignes de $M_{\tau+1}$ seront équivalentes. Pour avoir une meilleure idée de $M_{\tau+1}$, analysons un exemple numérique simple.

Exemple 4.2.1. Supposons que $x_m = 1$, $x_{m+2} = 2$, $x_q = 4$, $x_{q+2} = 5$, et que nous choisissons d'utiliser des vecteurs y_i de longueur 3 pour estimer notre vraisemblance. Alors, on peut déterminer que $r_{0_1} = 1{,}5$ et $r_{0_2} = 4{,}5$, et que $y_1' = (1{,}25; 1{,}5; 1{,}75)$ et $y_2' = (4{,}25; 4{,}5; 4{,}75)$. Si on observe une recombinaison dans l'intervalle m, alors

$$M_{\tau+1} \propto \begin{bmatrix} \frac{0{,}25}{0{,}5} & \frac{0{,}25}{0{,}5} & \frac{0{,}25}{0{,}5} \\ \frac{0{,}5}{0{,}5} & \frac{0{,}5}{0{,}5} & \frac{0{,}5}{0{,}5} \\ \frac{0{,}75}{0{,}5} & \frac{0{,}75}{0{,}5} & \frac{0{,}75}{0{,}5} \end{bmatrix}.$$

Par exemple, le premier élément de y_1 est 1,25, ce qui veut dire que pour $r_{T_1} = 1{,}25$ $r_{l_1} = 1{,}25 - 1 = 0{,}25$ et donc $r_{l_1}/r_{0l_1} = 0{,}25/0{,}5$ car $r_{0l_1} = r_{0_1} - x_m = 1{,}5 - 1$. Si on avait plutôt observé une recombinaison dans l'intervalle $m + 1$, alors $M_{\tau+1}$ aurait été proportionnelle à la même matrice à une permutation de ligne près, c'est-à-dire à la matrice ayant comme première ligne la dernière de $M_{\tau+1}$ et comme dernière ligne la première de $M_{\tau+1}$. La vraisemblance estimée pour cet intervalle de $\mathbb{R}^2$ est proportionnelle aux produits, éléments par éléments, de toutes les matrices $M_{\tau+1}$ (pour $\tau = 1 \ldots \tau^*$) et devrait ressembler à une coupole. La figure 4.5 présente cinq graphiques, les quatre premiers représentent la vraisemblance après une seule recombinaison dans des intervalles différents. Et le graphique 4.5(e) représente la vraisemblance suite à cinq recombinaisons dans les quatre différents intervalles (m, $m + 1$, q et $q + 1$). On y voit bien la forme de coupole attendue.

Maintenant, que ce passe-t-il si l'on désire estimer la vraisemblance pour un intervalle de $\mathbb{R}^2$ où le TIM1 et le TIM2 appartiennent au même intervalle entre marqueurs de positions connues ? Comme par exemple lorsque le TIM1 est le $(m + 1)^e$ marqueur et le TIM2 le $(m + 2)^e$ marqueur, situation illustrée par la figure 4.6. Premièrement, les vecteurs y_1 et y_2 seront identiques. Si on observe une recombinaison, l'équation 4.4 se réduit comme suit :



(a) $Re(m)$ (b) $Re(m+1)$ (c) $Re(q)$ (d) $Re(q+1)$ (e) Après 5 recombinaisons

Figure 4.5 Exemples de graphiques pour l'estimation de $M_{\tau+1}$ lorsque les deux TIM sont dans des intervalles différents, que les vecteurs y sont de longueur 9.

- si $Re(m)$ alors

$$\frac{a(H_\tau, H_{\tau+1})}{a_0(H_\tau, H_{\tau+1})} = \frac{r_{l_1}}{r_{0l_1}},$$

- si $Re(m+2)$ alors

$$\frac{a(H_\tau, H_{\tau+1})}{a_0(H_\tau, H_{\tau+1})} = \frac{r_{r_2}}{r_{0r_2}},$$

- si $Re(m+1)$ alors

$$\frac{a(H_\tau, H_{\tau+1})}{a_0(H_\tau, H_{\tau+1})} = \frac{r_{r_1}}{r_{0r_1}} = \frac{r_{l_2}}{r_{0l_2}} = \frac{r_{T_2} - r_{T_1}}{r_{0_2} - r_{0_1}},$$

- sinon

$$\frac{a(H_\tau, H_{\tau+1})}{a_0(H_\tau, H_{\tau+1})} = 1.$$

Ici, la matrice $M_{\tau+1}$ sera triangulaire, car le TIM1 est supposé être toujours situé à gauche du TIM2. Par conséquent, nous n'estimerons pas la vraisemblance pour les couples $(y_{1_i}; y_{2_j})$ où $i \leq j$. On remarque que la vraisemblance dépend seulement de la position du TIM1 lorsqu'une recombinaison a lieu dans l'intervalle m, par conséquent les entrées d'une même ligne de la matrice $M_{\tau+1}$ auront la même valeur (lorsqu'elles sont évaluées ($\Leftrightarrow i \leq j$)). De même, l'estimation de la vraisemblance dépend seulement de la position du TIM2 lorsqu'une recombinaison a lieu dans l'intervalle $m+2$, ce qui entraîne que les entrées d'une même colonne de la matrice $M_{\tau+1}$ auront la même valeur (lorsqu'elles sont évaluées ($\Leftrightarrow i \leq j$)). Mais si une recombinaison a lieu dans l'intervalle situé entre le TIM1 et le TIM2, alors la vraisemblance dépend de la distance entre ces deux TIM.

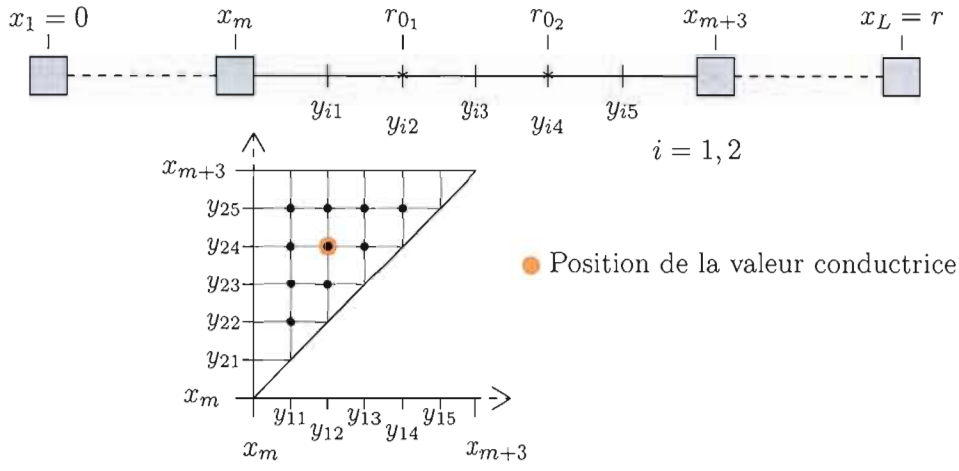


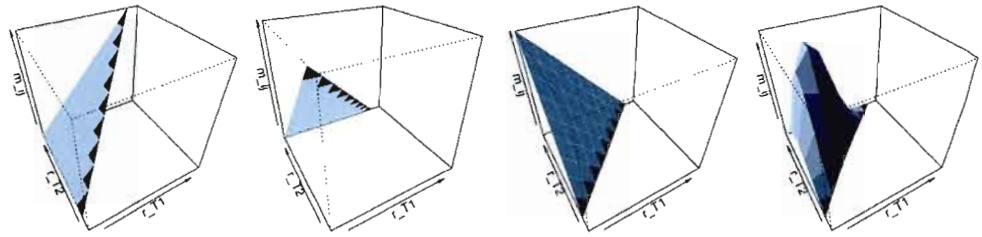
Figure 4.6 Exemple 2 d'un intervalle pour l'évaluation de la vraisemblance où le TIM1 et le TIM2 sont situés dans le même intervalle de position connue. Ici les vecteurs y_1 et y_2 sont de longueur 5 et $y_1 = y_2$.

Regardons maintenant, quelle sera la forme de la matrice $M_{\tau+1}$ dans une telle situation à l'aide d'un exemple numérique simple.

Exemple 4.2.2. Supposons que $x_m = 1$, $x_{m+3} = 4$ et que nous choisissons d'utiliser des vecteurs y_i de longueur 5 pour estimer notre vraisemblance. Alors, on peut déterminer que $r_{0_1} = 2$ et $r_{0_2} = 3$, et que $y_1' = y_2' = (1{,}5; 2; 2{,}5; 3; 3{,}5)$. Si une recombinaison à lieu dans l'intervalle entre le TIM1 et le TIM2 alors,

$$M_{\tau+1} \propto \begin{bmatrix} \nexists & 0{,}5 & 1 & 1{,}5 & 2 \\ \nexists & \nexists & 0{,}5 & 1 & 1{,}5 \\ \nexists & \nexists & \nexists & 0{,}5 & 1 \\ \nexists & \nexists & \nexists & \nexists & 0{,}5 \\ \nexists & \nexists & \nexists & \nexists & \nexists \end{bmatrix}.$$

Pour cet exemple le dénominateur est $r_{0_2} - r_{0_1} = 3 - 2 = 1$, et pour trouver le numérateur de (m_{ij}), il suffit de calculer $y_{2_j} - y_{1_i}$. Graphiquement, une telle matrice ressemblera au graphique 4.7(c). La figure 4.7 présente quatre graphiques illustrant la fonction de vraisemblance lorsque le TIM1 et le TIM2 sont situés dans le même intervalle de position



(a) $Re(m)$ (b) $Re(m+2)$ (c) $Re(m+1)$ (d) Après plusieurs recombinaisons

Figure 4.7 Exemples de graphiques pour l'estimation de $M_{\tau+1}$ lorsque les deux TIM sont dans le même intervalle de position connue et que les vecteurs y sont de longueur 11.

connue. Les trois premiers montrent le résultat après une seule recombinaison et le dernier le résultat suite à plusieurs recombinaison.

Il serait faux de penser que seulement les recombinaisons influencent la vraisemblance, elles apportent la forme polynomiale de la fonction, mais la hauteur de la vraisemblance (et par conséquent son maximum) est aussi influencée par les autres événements (coalescences et mutations) se produisant lors de la construction des graphes. La figure 4.8 est un exemple du graphique obtenu pour l'estimation de la vraisemblance pour un ensemble de données où 30 marqueurs ont été utilisés et 100 graphes ont été construits pour chacun des intervalles. On remarquera que les marqueurs ne sont pas équidistants, par conséquent les intervalles de $\mathbb{R}^2$ pour lesquelles la vraisemblance est évaluée ne sont pas de forme carrée.

4.3 Modélisation de l'interaction de deux gènes

Jusqu'à maintenant, nous avons vu quelle était la forme de la vraisemblance estimée et nous savons que cette estimation est faite indépendamment pour chaque intervalle de $\mathbb{R}^2$. Concrètement, lors de l'estimation de la vraisemblance pour un intervalle de $\mathbb{R}^2$, on insère, à chacune des séquences de l'échantillon, le TIM1 et le TIM2 au centre de leurs intervalles respectifs (ceux pour lesquels la vraisemblance est évaluée) et on construit des graphes, c'est-à-dire différentes généalogies possibles pour nos séquences. Mais on ne

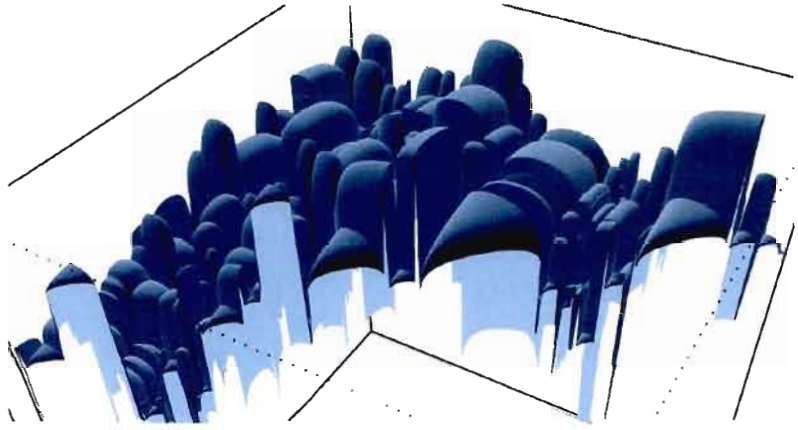


Figure 4.8 Exemple de l'estimation de la vraisemblance avec 30 marqueurs et 100 graphes par intervalle.

connait pas la situation mutant/non mutant de ces TIM. Il faut donc inférer un statut à chacun des TIM insérés, qui dépend du phénotype des séquences (cas ou contrôle), mais aussi de la façon dont ces TIM interagissent et influencent la probabilité pour les séquences d'être affectées par le caractère.

Dans cette section, nous présentons la modélisation choisie pour l'inférence des TIM, et comment cette modélisation peut rapidement se compliquer. Puis, bien que la problématique de modélisation ne soit pas tout à fait la même, nous présentons comment l'interaction entre deux gènes causant un caractère est traitée par certaines méthodes de cartographie basée sur des études d'association.

4.3.1 La modélisation choisie

Pour l'inférence des TIM, nous allons modéliser la fonction de pénétrance (notons-la F), c'est-à-dire les probabilités pour les séquences d'être affectées par le caractère en fonction des génotypes. Comme pour la méthode MapArg, nous avons choisi d'utiliser un échantillon composé d'haplotypes, ce qui simplifie la modélisation de la fonction de pénétrance. De plus, nous supposons que la maladie est causée par l'effet combiné du TIM1 et du TIM2, qu'elle est rare et récessive. Si aucun des TIM n'influence plus

que l'autre la probabilité d'être affecté, nous pouvons définir la fonction de pénétrance $F = (f_0, f_1, f_2)$ comme suit :

$$f_0 = P(\text{cas} \mid \overline{\text{TIMs}});\ f_1 = P(\text{cas}|\text{TIM1 ou TIM2}) \text{ et } f_2 = P(\text{cas}|\text{TIM1 etTIM2}).$$

Par simplicité, nous avons choisi comme modélisation $F = (0,0,1)$. Nous supposons donc, qu'une séquence affectée par le caractère possède les deux mutations, et puisque la maladie est rare, qu'une séquence contrôle possède l'allèle sain pour ces deux gènes. En supposant que le TIM1 n'influence pas plus (ou moins) que le TIM2 la probabilité d'être malade, nous pouvons affirmer, sans perte de généralité, que le TIM1 est le premier des deux sur la séquence : c'est-à-dire, le TIM1 est toujours à gauche du TIM2 ($r_{T1} < r_{T2}$).

Sans les suppositions précédentes, la modélisation se complique rapidement. Par exemple, si le fait de posséder une mutation au TIM1 augmente plus la probabilité d'être affectée par le caractère que de posséder une mutation au TIM2, on ne peut plus supposer que $r_{T1} < r_{T2}$ (la vraisemblance doit alors être évaluée pour un plus grand nombre d'intervalles de $\mathbb{R}^2$) et l'inférence des TIM n'est plus aussi directe.

De plus, l'utilisation d'un échantillon de diploïdes complexifie la fonction de pénétrance, et par le fait même sa modélisation. Le tableau 4.1, présente un exemple de fonction de pénétrance, les allèles possibles au TIM1 sont **a** et **A** (**A** étant l'allèle mutant) et les allèles du TIM2 sont **b** et **B** (**B** étant l'allèle mutant). Pour cet exemple, un individu doit posséder au moins un des allèles mutants au deux gènes pour être affecté par le caractère. En supposant que nous choisissons cette modélisation, nous devons ensuite inférer le génotype des individus. Cette étape est alors plus délicate qu'auparavant.

Notons toutefois que nous n'avons aucune connaissance préalable sur l'interaction entre les gènes et sur la façon dont elle se traduit en fonction de pénétrance. Il n'y a aucun consensus sur le sujet présentement en génétique, et la définition même de l'interaction entre les gènes et sa modélisation mathématique ne fait pas l'unanimité (Cordell, 2002).

	aa	aA	AA
bb	0	0	0
bB	0	1	1
BB	0	1	1

Tableau 4.1 Exemple de fonction de pénétrance pour un échantillon de diploïde. Les allèles du TIM1 sont **a** et **A**, tandis que les allèles du TIM2 sont **b** et **B**. Les lettres majuscules représentent les allèles mutants.

4.3.2 Interaction entre deux gènes et étude d'association

Nous avons vu précédemment que les études d'association testent la corrélation entre le caractère étudié et chacun des marqueurs un à la fois. Pour chaque individu appartenant à l'échantillon, le génotype des marqueurs utilisés, c'est-à-dire la paire d'allèles que possède l'individu au marqueur, est connu. Ces méthodes peuvent ne pas détecter un gène qui influence seulement le caractère lorsqu'il est en présence d'un autre. C'est pourquoi différentes méthodes pour cartographier des caractères polygéniques (on supposera que le caractère est influencé par deux gènes) basés sur l'association ont été présentées ces dernières années.

Marchini, Donnelly et Cardon (2005) ont proposé deux méthodes de cartographie et les ont comparées à la méthode d'association marqueur par marqueur. L'idée est, pour chaque paire possible de marqueurs, d'ajuster un modèle de régression logistique — la variable dépendante est : affectée oui ou non par le caractère — aux génotypes observés dans l'échantillon pour cette paire de marqueurs. La différence entre les deux méthodes proposées réside en la sélection des marqueurs, la première utilise tous les marqueurs disponibles, tandis que la deuxième a deux étapes : une première sélection est effectuée, ensuite on teste toutes les paires possibles de marqueurs sélectionnés. Si le modèle de régression s'avère significatif, on conclut que les deux marqueurs influencent le caractère.

D'autres méthodes testent plutôt les paramètres représentant l'interaction entre les marqueurs dans le modèle de régression logistique. Barhdadi et Dubé (2010) proposent l'utilisation d'un modèle additif des effets principaux et d'interaction multiplicative (AMMI) pour tester l'interaction entre les paires de marqueurs. Si le caractère est plutôt quantitatif, alors ces méthodes utilisent plutôt un modèle de régression linéaire.

Pour ces méthodes, la fonction de pénétrance n'a pas besoin d'être déterminée à l'avance, il s'agit plutôt de bien déterminer les paramètres utilisés dans le modèle de régression.

CHAPITRE V

SIMULATIONS ET RÉSULTATS

Afin de tester nos développements, nous n'avons d'autres choix que de simuler nous-mêmes des ensembles de données qui répondent à nos suppositions. Il est courant de faire ainsi ; on pourra voir par exemple Stephens et Donnelly (2001) et Fearnhead et Donnelly (2001). Nous expliquons tout d'abord comment nous avons construit nos ensembles de données à analyser, nous présentons ensuite le programme informatique conçu pour implanter l'adaptation proposée, puis nous terminons en présentant les résultats obtenus.

5.1 Simulation d'ensembles de séquences

Nous avons choisi de simuler dix ensembles de séquences de marqueurs pour tester notre adaptation de MapArg. La simulation de ces ensembles s'est faite en deux étapes distinctes. Premièrement, nous avons simulé dix populations à l'aide du programme *ms* de Hudson (2002), puis pour chacun de ces ensembles nous avons choisi les deux marqueurs causant le caractère, inféré le phénotype aux séquences et tiré un échantillon de cas et de contrôles, à l'aide d'un programme C++ nommé *selectM*.

Pour la première étape, nous utilisons le programme *ms* (Hudson, 2002) qui est basé sur le processus de coalescence avec recombinaison ; la généalogie des séquences est tout d'abord simulée, ensuite les mutations sont ajoutées selon le modèle des sites infinis. Ce programme est flexible, il permet entre autre de faire varier la taille de

la population selon différents modèles. Nous avons simulé dix populations de 10 000 haplotypes chacune, en considérant la taille de la population fixe à travers le temps. Les séquences sont composées de 500 marqueurs et sont de longueur $r = 0.625$ cM (soit approximativement 625 kilobases). Il s'agit de courtes séquences puisque notre méthode en est une de cartographie fine.

Pour la deuxième étape, l'utilisation d'un second programme informatique est nécessaire, car les marqueurs influençant le caractère n'ont pas encore été déterminés. Celui-ci nous permet de choisir une paire de marqueurs selon la fréquence conjointe des deux allèles mutants dans la population. Ensuite, selon la modélisation choisie, nous inférons à chacune des séquences le phénotype (affectée ou non par le caractère), et nous enlevons les deux marqueurs représentant le TIM1 et le TIM2. Nous enlevons aussi tous les marqueurs ayant un allèle de fréquence inférieure à 1% dans la population, puisque ceux-ci sont peu informatifs. Après quoi un échantillon de cas et de contrôles est tiré aléatoirement dans la population. Les positions des deux gènes influençant le caractère ont été conservées pour permettre de comparer avec l'estimation obtenue.

Selon la modélisation présentée au chapitre précédent, une séquence est influencée par le caractère seulement si elle possède les deux mutations. Il s'ensuit que certaines de nos séquences contrôles possèderont tout de même une des deux mutations. Nous avons demandé pour chacune des dix populations un échantillon de 300 cas et de 300 contrôles. La fréquence du caractère varie d'une population à l'autre, entre 2,5% et 15%. Les fréquences associées à chacun des échantillons sont représentées dans le tableau 5.1, notons que les échantillons A et B ont seulement 250 cas, puisque dans la population on en retrouve seulement 250 séquences affectées (la fréquence du caractère étant de 2,5%).

5.2 Programme Python pour le calcul en parallèle

La méthode MapArg est actuellement implantée dans un programme C++. Notre première étape a consisté à modifier ce programme à la cartographie de deux gènes influençant le caractère, mais nous nous sommes rendus à l'évidence : les temps de

Échantillon	Cas	Contrôles	Fréquence
A	250	300	0.025
B	250	300	0.025
C	300	300	0.05
D	300	300	0.05
E	300	300	0.075
F	300	300	0.075
G	300	300	0.10
H	300	300	0.10
I	300	300	0.15
J	300	300	0.15

Tableau 5.1 Dix échantillons ont été simulés (de A à J), les colonnes cas et contrôles indiquent respectivement le nombre de cas et de contrôles de chacun des échantillons. La dernière colonne indique la fréquence du caractère étudié pour l'échantillon.

calcul étaient trop longs, et ne nous permettaient pas de véritablement tester notre adaptation. Nous avons présenté au chapitre trois l'utilisation de fenêtres de marqueurs permettant d'accélérer le temps de calcul pour la méthode de cartographie MapArg. Bien que l'implantation des fenêtres de marqueurs à l'adaptation proposée soit possible, elle s'avère complexe, c'est pourquoi nous avons choisi pour l'instant de laisser de côté cette possibilité. De plus, contrairement à MapArg, le nombre d'intervalles pour lesquels nous devons estimer la vraisemblance est énorme ; l'utilisation de séquences composée de L marqueurs entraîne l'évaluation de la vraisemblance pour $(L-3)(L-2)/2$ intervalles ($L-2$ seulement pour MapArg).

Nous nous sommes plutôt tournés vers la programmation en parallèle ; c'est-à-dire la création d'un programme informatique permettant de distribuer le calcul sur plusieurs processeurs à la fois. Avec la collaboration de Didier Amyot, étudiant en mathématiques, nous avons conçu un programme informatique, utilisant le langage Python, nommé

PyArg permettant le calcul en parallèle. Le choix du langage de programmation Python a été motivé par l'existence du module *parallel python* qui nous permettait facilement de distribuer les calculs sur plusieurs processeurs. Toutefois, Python est un langage beaucoup plus lent que le C++ ; un grand travail d'optimisation de l'algorithme a été fait pour améliorer les performances de PyArg.

Présentons tout d'abord, les grandes lignes de l'algorithme PyArg, ainsi que quelques particularités de ce programme. Ensuite, nous verrons deux exemples de l'optimisation de l'algorithme.

5.2.1 Aperçu de l'algorithme de PyArg

1. Établir l'ensemble des valeurs de r_{T1} et de r_{T2} pour lesquelles la vraisemblance sera évaluée.
2. Pour chaque intervalle de $\mathbb{R}^2$ pour lesquels la vraisemblance est évaluée :
   (a) Ajouter à chaque séquence, en fonction de son phénotype et de la modélisation choisie, le TIM1 et le TIM2, selon l'intervalle de $\mathbb{R}^2$ pour lequel la vraisemblance est évaluée.
   (b) Construire K graphes :
   - À chaque étape de la construction du graphe, jusqu'à l'atteinte du MRCA :
     i. Déterminer tous les événements possibles menant à l'étape suivante de la construction du graphe.
     ii. Attribuer une probabilité à chacun de ces événements.
     iii. Choisir, selon les probabilités attribuées, quel événement aura lieu.
     iv. Pour chaque paire de points (r_{T1}, r_{T2}) contenue dans l'intervalle, évaluer la vraisemblance de cet événement.
   - Pour chaque paire de points (r_{T1}, r_{T2}) contenue dans l'intervalle, évaluer le produit des vraisemblances de chacune des étapes de la construction du graphe.

(c) Pour chaque paire de points (r_{T1}, r_{T2}) contenue dans l'intervalle, évaluer la moyenne de la vraisemblance de chacun des graphes.

3. L'estimateur de la position du TIM1 et de la position du TIM2 est le maximum de l'estimation de la fonction de vraisemblance.

La parallélisation de cet algorithme est relativement facile et se base sur l'indépendance des généalogies construites. Des tâches à effectuer sont créées et distribuées aux processeurs disponibles. Une tâche ne doit pas être trop longue à effectuer, sinon la perte d'un processeur augmenterait le temps de calcul ; ni trop courte, car cela demanderait trop de communication entre les processeurs. Chaque tâche distribuée indique dans quel intervalle insérer le TIM1 et le TIM2 ainsi que le nombre de graphes à générer. Par exemple, si nous désirons construire 1 000 graphes par intervalle, nous pouvons décider d'envoyer vingt tâches de cinquante graphes par intervalle. Si les séquences sont composées de 30 marqueurs de positions connues ($L = 32$), la vraisemblance est alors estimée pour 435 intervalles et les processeurs utilisés se partageront 8 700 tâches. Après avoir terminé une tâche, le processeur renvoie le résultat de ses calculs à l'ordinateur s'occupant de gérer les tâches ; celui-ci, après avoir reçu toutes les tâches, compile les informations et retourne les graphiques représentant l'estimation de la vraisemblance ainsi que le maximum de celle-ci.

5.2.2 Deux exemples d'optimisation pour PyArg

En regardant l'algorithme précédent, on remarque qu'à chaque étape de la création d'un graphe, on doit recalculer plusieurs fois la même chose. Par exemple, à chaque étape, on doit refaire la liste de tous les événements possibles. Mais il est aussi possible, après chaque événement, de mettre à jour cette liste ; c'est ce qui a permis, en partie, d'optimiser PyArg. Nous avons construit une table de coalescence, indiquant quelles séquences peuvent coalescer entre elles. Cette table de coalescence est mise à jour après chaque événement de coalescence, mutation ou recombinaison. Dans cette section, nous

présentons la façon dont la mise à jour est faite. De plus, nous présentons comment la sélection des marqueurs utilisés est faite.

Avant de commencer la construction d'un graphe, une table de coalescence est faite ; on associe à chaque type de séquences présent dans l'échantillon, les types de séquences pouvant coalescer avec celui-ci. Voyons comment cette table est mise à jour en fonction de l'événement ayant eu lieu.

1. Si l'événement est une coalescence de deux séquences de type i (C_i) :
   - On retrouve une séquence en moins de type i, si $n_i = 1$ alors un nouvel événement C_i est impossible.
   - Tous les autres événements possibles restent les mêmes.
2. Si l'événement est une coalescence de deux séquences différentes (C_{ij}^k) :
   - On retrouve une séquence en moins de type i, si $n_i = 1$ alors un événement C_i est impossible. S'il n'en reste plus, on doit enlever le type i de la table.
   - On retrouve une séquence en moins de type j, si $n_j = 1$ alors un événement C_j est impossible. S'il n'en reste plus, on doit enlever le type j de la table.
   - Si la séquence de type k existait dans la table, on ajoutera l'événement C_k s'il y avait précédemment qu'une séquence de type k. Si elle n'existait pas, on l'ajoute à la table et elle pourra coalescer qu'avec les séquences pouvant coalescer avec les séquences i et les séquences j.
   - Tous les autres événements possibles restent les mêmes.
3. Si l'événement est une mutation d'une séquence i vers une séquence j (M_i^j) :
   - Il ne reste plus de séquence de type i, on enlève le type de séquence i de la table.
   - Si la séquence de type j existait dans la table, on ajoutera l'événement C_j s'il y avait précédemment qu'une séquence de type j. Si elle n'existait pas, on l'ajoute à la table et on vérifie avec quelles autres séquences elle peut coalescer.
   - Tous les autres événements possibles restent les mêmes.
4. Si l'événement est plutôt une recombinaison (R_i^{jk}) :
   - On retrouve une séquence en moins de type i, si $n_i = 1$ alors un événement C_i est impossible. S'il n'en reste plus, on doit enlever le type i de la table.

- Si la séquence de type j existait dans la table, on ajoutera l'événement C_j s'il y avait précédemment qu'une seule séquence de type j. Si elle n'existait pas, on l'ajoute à la table et on vérifie avec quelles autres séquences elle peut coalescer.
- Si la séquence de type k existait dans la table, on ajoutera l'événement C_k s'il y avait précédemment qu'une séquence de type k. Si elle n'existait pas, on l'ajoute à la table et on vérifie avec quelles autres séquences elle peut coalescer.
- Tous les autres événements possibles restent les mêmes.

La mise à jour de la table de coalescence permet d'accélérer considérablement le programme PyArg, puisqu'il n'est plus nécessaire de vérifier, à chaque étape, quels types de séquences peuvent coalescer deux à deux.

Nous avons généré des populations de séquences de 500 marqueurs chacune. Pour chaque population, certains marqueurs ont été retirés, car un trop grand nombre de séquences possédaient le même allèle à ces marqueurs. Toutefois, il reste un grand nombre de marqueurs par séquence dans chaque échantillon ; une sélection des marqueurs utilisés par MapArg et PyArg doit être effectuée sinon les temps de calcul seraient énormes. MapArg sélectionne les marqueurs les plus polymorphiques, c'est-à-dire ceux ayant une proportion d'allèles mutants le plus près de 50%. Ce choix peut être discutable, puisque les intervalles entre marqueurs peuvent s'avérer grands, ce qui peut entraîner une moins bonne estimation de la fonction de vraisemblance. Pour le programme PyArg, nous avons opté pour la sélection d'un sous-ensemble de marqueurs les plus polymorphiques, puis une sélection dans le sous-ensemble des marqueurs les plus équidistants. Nous présentons ici la façon dont la sélection des marqueurs est effectuée avec PyArg.

La méthode utilisée est itérative, les marqueurs sont enlevés les uns après les autres, jusqu'à l'obtention d'une séquence de la bonne longueur. Par exemple, si nous voulons sélectionner 30 marqueurs, nous pouvons choisir les 60 marqueurs les plus polymorphiques, puis choisir parmi ces 60 marqueurs les 30 plus équidistants en enlevant en tout 30 marqueurs les uns après les autres. Cette méthode ne retourne pas les 30 marqueurs les plus équidistants, mais plutôt une approximation.

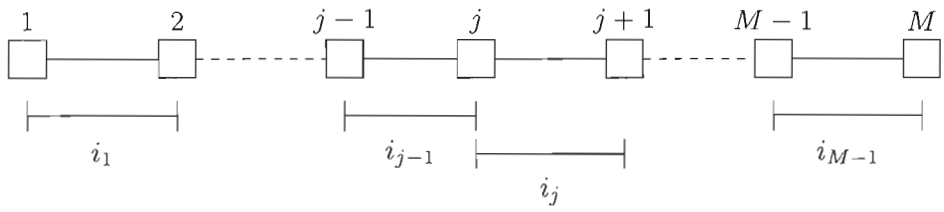


Figure 5.1 Séquence de M marqueurs. La longueur de l'intervalle j, situé entre les marqueurs j et $j+1$, est notée par i_j.

Supposons que nos séquences sont composées de M marqueurs, notés de 1 à M, et soit i_j, la longueur de l'intervalle situé entre les marqueurs j et $j+1$, tel qu'illustré par la figure 5.1. Notons par I, l'ensemble des i_j de cette séquence. Nous connaissons la variance de I, et nous souhaitons enlever un marqueur à I — notons le nouvel ensemble I_- — tel que la variance de I_- soit minimale, une variance de 0 indiquerait que les marqueurs sont équidistants. En supposant que le marqueur enlevé est le $(j+1)^e$, alors l'intervalle entre les marqueurs j et $j+1$ de la nouvelle séquence sera de longueur $i_j + i_{j+1}$. Alors la variance de l'ensemble I_- est :

$$
\begin{aligned}
\mathrm{Var}(I_-) &= \frac{1}{M-1}\sum_{m\in I_-} m^2 - \left[\sum_{m\in I_-}\frac{m}{M-1}\right]^2 \\
&= \frac{1}{M-1}\left[\sum_{\substack{m\in I_- \\ m\neq i_j+i_{j+1}}} m^2 + (i_j+i_{j+1})^2\right] - \left[\sum_{\substack{m\in I_- \\ m\neq i_j+i_{j+1}}} \frac{m}{M-1} + \frac{(i_j+i_{j+1})}{M-1}\right]^2 \\
&= \frac{1}{M-1}\left[\sum_{\substack{m\in I_- \\ m\neq i_j+i_{j+1}}} m^2 + (i_j) + (i_{j+1})^2 + 2(i_j i_{j+1})\right] - \left[\frac{M}{M}\sum_{m\in I}\frac{m}{M-1}\right]^2 \\
&= \frac{M}{M-1}\sum_{m\in I}\frac{m^2}{M} - \left[\frac{M}{M-1}\sum_{m\in I}\frac{m}{M}\right]^2 + \frac{2(i_j i_{j+1})}{M-1} \\
&= C\cdot\mathrm{Var}(I) + \frac{2(i_j i_{j+1})}{M-1},
\end{aligned}
$$

où C est une constante. Pour minimiser $\mathrm{Var}(I_{-})$, il suffit d'enlever le marqueur $j+1$ tel que le produit des longueurs des intervalles l'entourant, $i_j \cdot i_{j+1}$ est minimale. En utilisant ce résultat, il est rapide de choisir un sous-ensemble de marqueurs les plus équidistants.

5.3 Résultats

Nous présentons dans cette section les résultats obtenus avec la mesure d'association r^2, la méthode de cartographie MapArg, ainsi qu'avec l'adaptation PyArg. Les résultats pour les dix échantillons sont présentés seulement pour r^2 et MapArg. Au lieu d'analyser les dix échantillons (ce qui aurait pris beaucoup de temps), nous avons préféré faire plusieurs simulations avec les mêmes ensembles de données pour tenter de mieux comprendre les résultats obtenus avec l'adaptation proposée.

5.3.1 Résultats avec r^2 et MapArg

Pour l'analyse des données avec MapArg et r^2, nous avons choisi d'utiliser les 50 marqueurs les plus polymorphiques de chacun de nos échantillons. Nous avons de plus utilisé des fenêtres de 5 marqueurs et fait 5 répétitions de 500 graphes par intervalles pour la méthode MapArg. Rappelons que ces répétitions peuvent être combinées pour former une estimation de la vraisemblance de 2 500 graphes, puisque chaque graphe est indépendant. La combinaison des répétitions est représentée en bleu sur les graphiques.

Les figures 5.2 à 5.6 illustrent les résultats. Toutes ces figures présentent quatre graphiques ; chaque ligne représente les résultats d'une des bases de données de A à J, la première colonne correspond à la mesure d'association r^2 et la deuxième à l'estimation de la vraisemblance avec MapArg. Nous savons que MapArg suppose que le caractère est causé que par un seul gène, tandis qu'aucune supposition n'est faite par la mesure r^2. Nous souhaitons voir si un signal fort sera détecté aux positions des TIM1 et TIM2 par ces deux méthodes de cartographie.

On remarque tout d'abord que le TIM1 et le TIM2 des bases de données D et **F** (figures 5.3 et 5.4) sont situés très près l'un de l'autre. Pour ces deux bases de données,

l'estimation par maximum de vraisemblance de MapArg se situe justement près de ces deux marqueurs. La mesure r^2 a un maximum près de leurs positions, mais n'est pas aussi précise que MapArg.

Pour les bases de données A et J (figures 5.2 et 5.6), le TIM1 est assez distancé du TIM2. Les deux méthodes de cartographie trouvent un des deux marqueurs cherchés (le même), mais il n'y a aucun signal pour le deuxième marqueur.

Le TIM1 et le TIM2 des données B et I (figures 5.2 et 5.6) sont situés assez près l'un de l'autre. Pour les données I, il y a manifestement un signal dans la région où ils se trouvent, et ce, pour les deux méthodes. Tandis que pour les données B, le maximum des deux méthodes est situé un peu à droite des TIM1 et TIM2.

Maintenant, pour les données C et G (figures 5.3 et 5.5), les deux marqueurs cherchés sont assez éloignés l'un de l'autre, et le maximum obtenu par les deux méthodes de cartographie se situe entre ces deux marqueurs.

Terminons avec les bases de données E et H (figures 5.4 et 5.5). Elles sont intéressantes, puisque pour les deux méthodes le maximum se situe à un des marqueurs cherchés et que l'on observe un signal clair au deuxième marqueur. Mais pour les données E, la région entourant le maximum est plutôt large, et offre une estimation moins précise de la position du marqueur situé dans cette région.

Notons qu'en général, la forme obtenue avec la mesure r^2 est similaire à la forme de la vraisemblance estimée par MapArg. Ce qui peut surprendre puisque MapArg utilise toute l'information disponible contrairement à une analyse avec r^2 qui n'utilise qu'un marqueur à la fois.

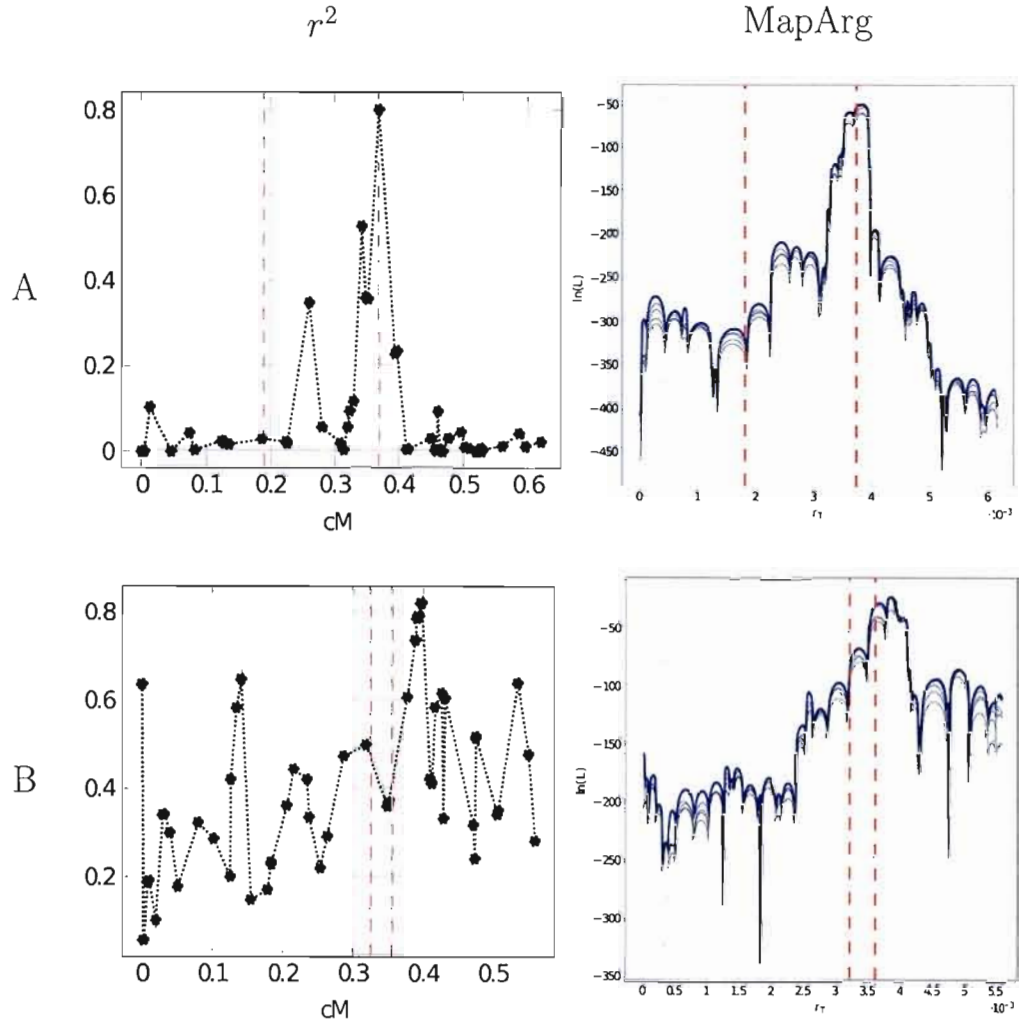


Figure 5.2 Résultats de r^2 et MapArg pour les données A et B avec 50 marqueurs. Les lignes pointillées rouges représentent la position du TIM1 et du TIM2. Les résultats pour MapArg ont été obtenus en utilisant des fenêtres de 5 marqueurs et en effectuant 5 répétitions de 500 graphes par intervalle. Les différentes courbes représentent les répétitions, la courbe bleue la combinaison des répétitions.

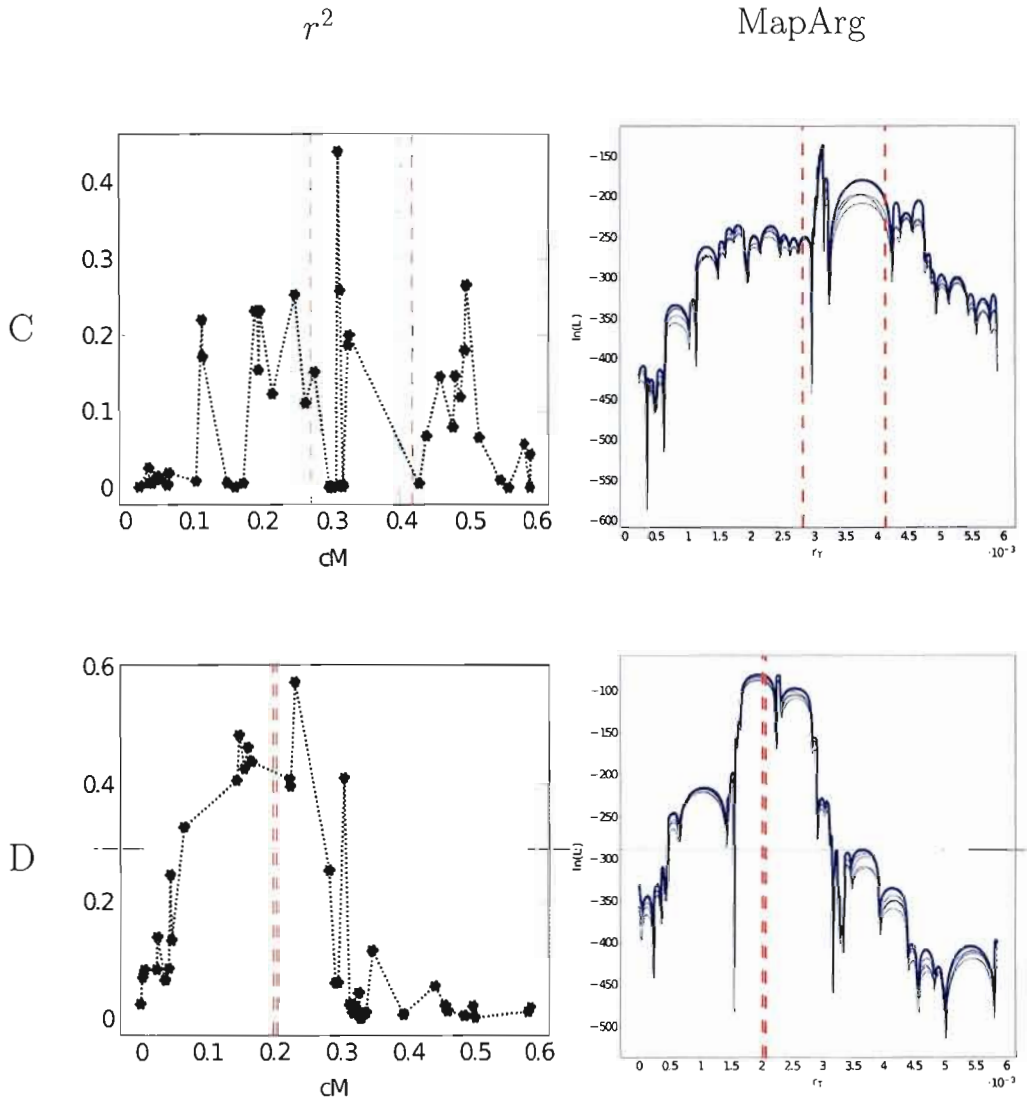


Figure 5.3 Résultats de r^2 et MapArg pour les données C et D avec 50 marqueurs. Les lignes pointillées rouges représentent la position du TIM1 et du TIM2. Les résultats pour MapArg ont été obtenus en utilisant des fenêtres de 5 marqueurs et en effectuant 5 répétitions de 500 graphes par intervalle. Les différentes courbes représentent les répétitions, la courbe bleue la combinaison des répétitions.

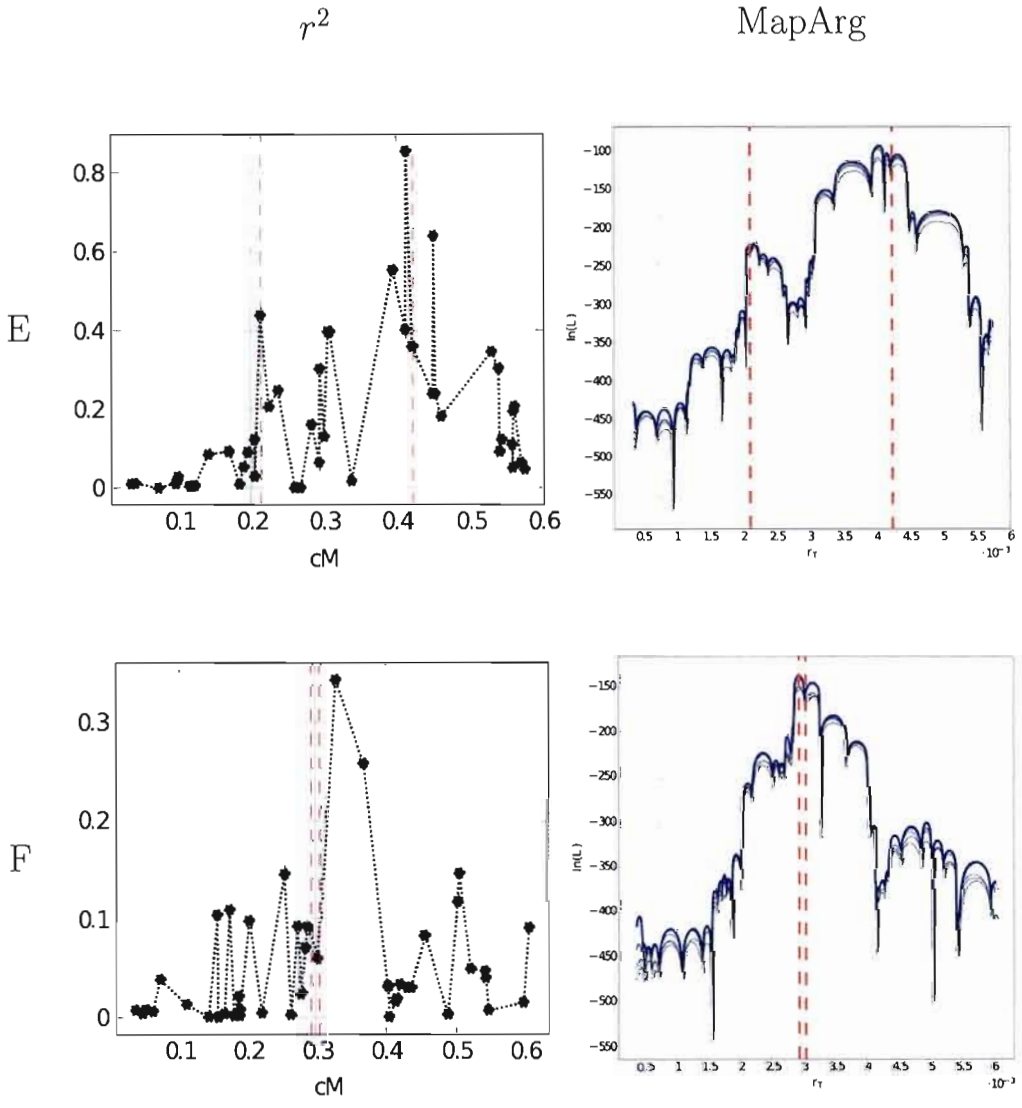


Figure 5.4 Résultats de r^2 et MapArg pour les données E et F avec 50 marqueurs. Les lignes pointillées rouges représentent la position du TIM1 et du TIM2. Les résultats pour MapArg ont été obtenus en utilisant des fenêtres de 5 marqueurs et en effectuant 5 répétitions de 500 graphes par intervalle. Les différentes courbes représentent les répétitions, la courbe bleue la combinaison des répétitions.

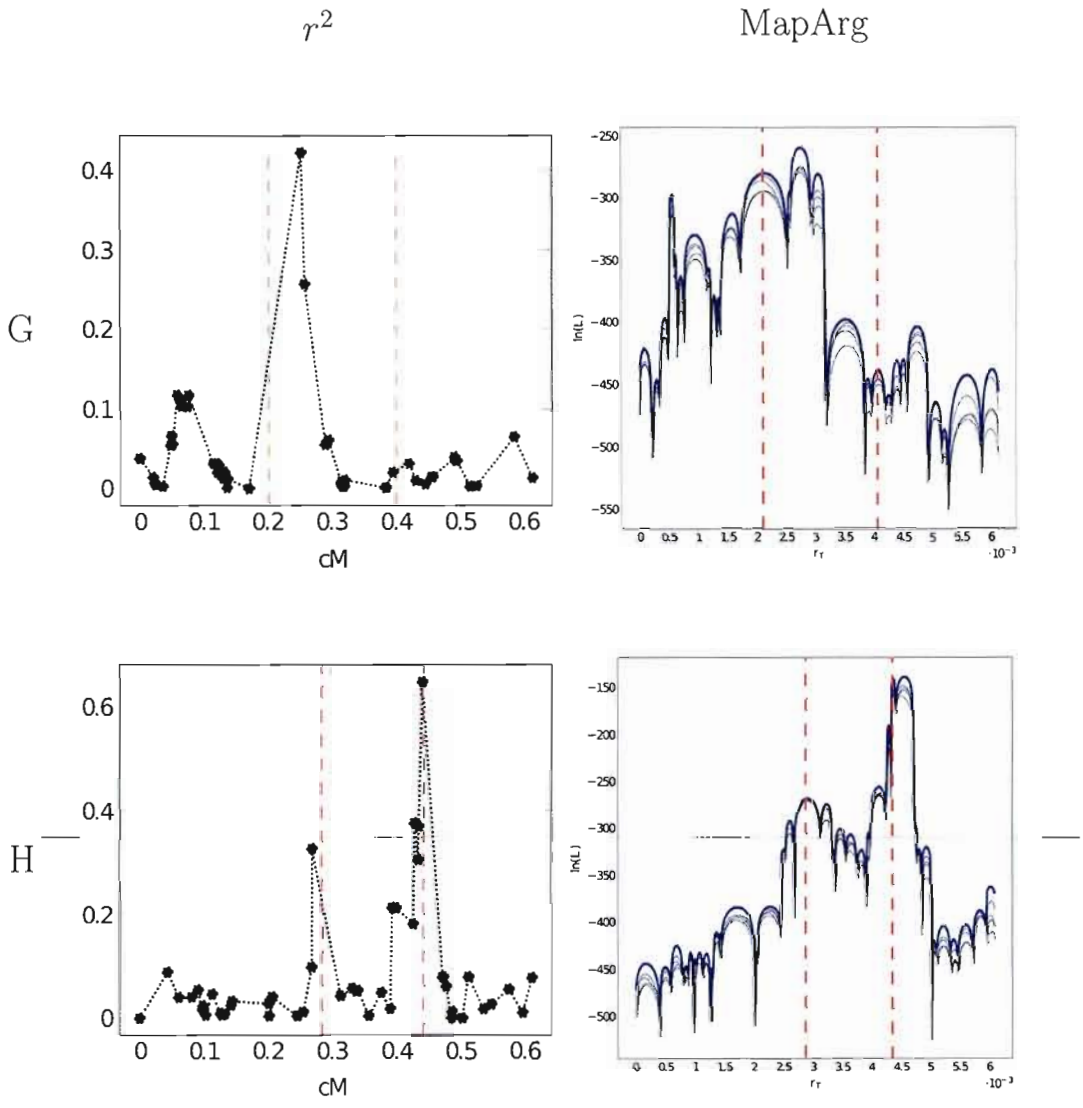


Figure 5.5 Résultats de r^2 et MapArg pour les donnécs G et H avec 50 marqueurs. Les lignes pointillées rouges représentent la position du TIM1 et du TIM2. Les résultats pour MapArg ont été obtenus en utilisant des fenêtres de 5 marqueurs et en effectuant 5 répétitions de 500 graphes par intervalle. Les différentes courbes représentent les répétitions, la courbe bleue la combinaison des répétitions.

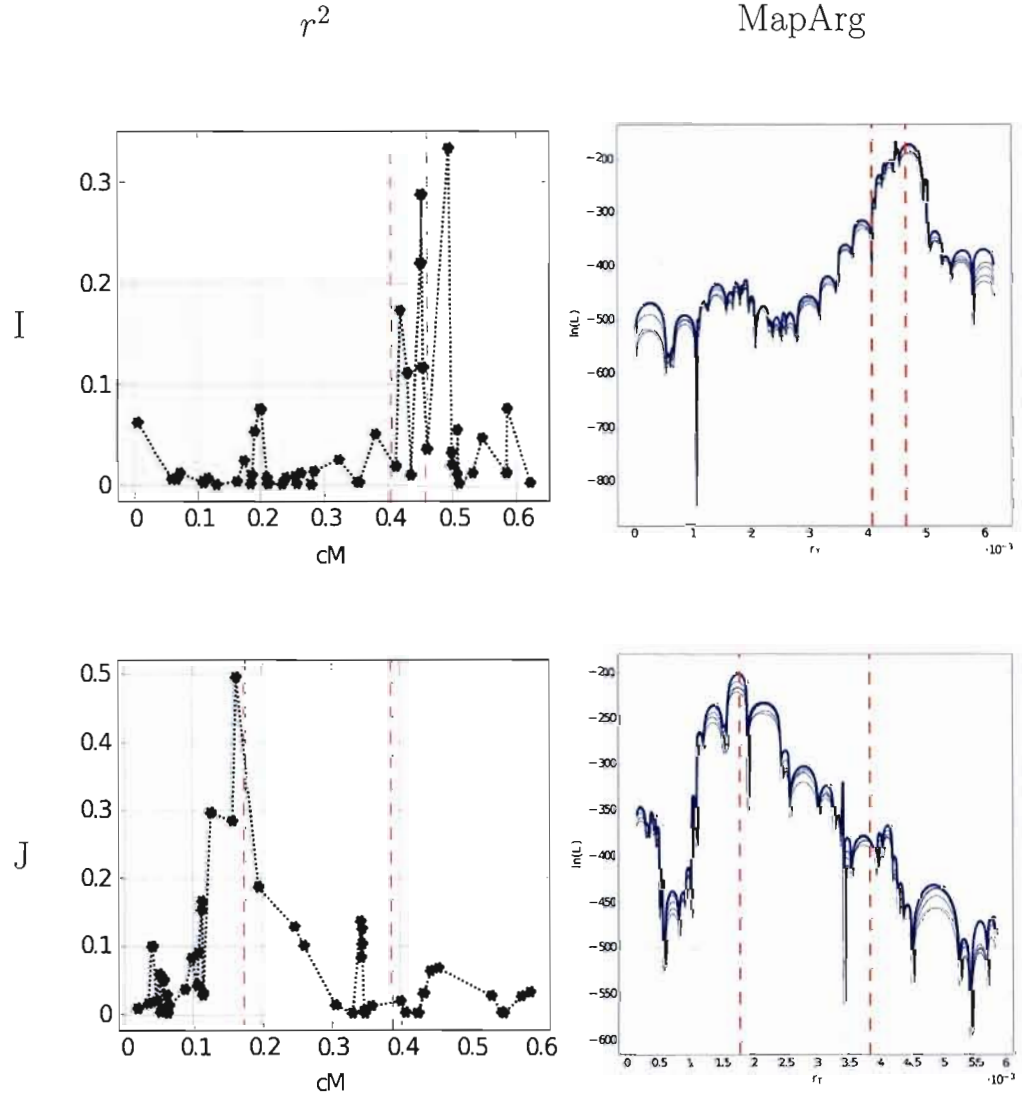


Figure 5.6 Résultats de r^2 et MapArg pour les données I et J avec 50 marqueurs. Les lignes pointillées rouges représentent la position du TIM1 et du TIM2. Les résultats pour MapArg ont été obtenus en utilisant des fenêtres de 5 marqueurs et en effectuant 5 répétitions de 500 graphes par intervalle. Les différentes courbes représentent les répétitions, la courbe bleue la combinaison des répétitions.

5.3.2 Résultats avec PyArg

Nous avons commencé nos analyses avec PyArg en utilisant les données A, G, H et J. Ces ensembles sont tels que le TIM1 est assez distancé du TIM2, et offrent une certaine variété dans les résultats obtenus avec r^2 et MapArg. Nous avons obtenu avec les ensembles A et J une bonne estimation de la position pour seulement un des marqueurs cherchés. Pour l'ensemble H, nous observions plutôt un signal clair aux positions des deux marqueurs cherchés. Tandis que le maximum obtenu pour les données de l'ensemble G se situe entre les positions des deux marqueurs cherchés.

Nous avons choisi d'utiliser des séquences de 30 marqueurs (les 30 plus équidistants parmi les 60 plus polymorphiques), ce qui entraîne l'évaluation de la vraisemblance pour 435 intervalles. Nous voulions utiliser assez de marqueurs, mais en même temps ne pas avoir trop d'intervalles à évaluer. Chaque simulation consiste en la création de 1 000 graphes par intervalle, ce qui équivaut à la création de 435 000 graphes par simulation. Une simulation de ce type, nous prend en moyenne 20 heures de calculs en utilisant environ 134 processeurs, ce qui est équivalant à environ 112 jours de calculs sur un processeur. Nous avons effectué trois simulations de 1000 graphes par intervalle pour les quatre ensembles de données mentionnés, afin de voir la variabilité des résultats. Puis nous avons combiné ces trois simulations pour en obtenir une de 3 000 graphes par intervalle.

Les figures 5.7 à 5.10 présentent les résultats ; chaque figure est associée à un ensemble de données (A, G, H et J) et présente la vraisemblance estimée par les trois simulations de 1000 graphes par intervalle ((a), (b) et (c)), ainsi l'estimation obtenue en combinant ces trois simulations (d). On se souviendra que la position du TIM1 se trouve en abscisse et que la position du TIM2 est en ordonnée ; les lignes pointillées rouges représentent leurs vraies positions. Une représentation en couleur du logarithme de la vraisemblance est faite sur $\mathbb{R}^2$. L'échelle de couleur utilisée passe du bleu (le minimum) au rouge foncé (le maximum).

Nous observons que l'estimation de la vraisemblance est semblable entre les différentes simulations d'un même ensemble de données, mais diffère d'un ensemble à l'autre. Regardons de plus près les résultats obtenus.

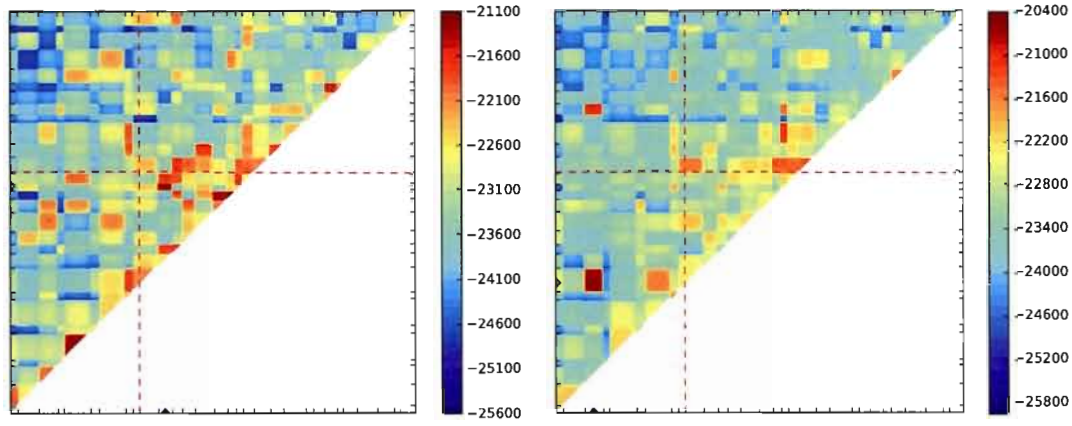
On observe pour l'ensemble A (figure 5.7), que le maximum de l'estimation de la vraisemblance de la simulation 1 (figure 5.7(a)) se trouve près de la vraie position des TIM1 et TIM2. Le maximum de la simulation 2 (figure 5.7(b)) est situé loin de la vraie position, mais celle-ci est située dans un intervalle où la vraisemblance est haute (c'est-à-dire de couleur orange). Pour la dernière simulation (figure 5.7(c)), le maximum de la vraisemblance estimée se situe lorsque le TIM1 et le TIM2 sont situés dans le même intervalle ; celui tout juste à côté de l'intervalle où se trouve le TIM2 (le même trouvé par r^2 et MapArg). On remarque aussi sur les trois simulations un certain bruit, c'est-à-dire des régions où la vraisemblance est haute mais seulement sur une simulation. Par exemple le maximum de la simulation deux est dans un intervalle où la vraisemblance n'est pas haute pour les deux autres simulations. Ces bruits ressortent aussi sur la combinaison des trois simulations (figure 5.7(d)) ; le maximum de la vraisemblance estimée est le même que celui de la simulation 3, qui était beaucoup plus grand que les maximums des simulations 1 et 2.

Pour les simulations 1 et 2 de l'ensemble de données G (figures 5.8(a) et 5.8(b)), on remarque que les maximums se situent lorsque nous plaçons les TIM1 et TIM2 entre leurs deux vraies positions, comme avec r^2 et MapArg (voir figure 5.5 page 100). Tandis que le maximum de la vraisemblance estimée lors de la simulation 3 se situe lorsque les TIM1 et TIM2 sont placés dans le même intervalle, celui où se trouve le TIM1. Puisque c'est ce maximum qui est le plus grand des maximums des trois simulations, il est aussi le maximum de la vraisemblance estimée par la combinaison de ces trois simulations.

Pour l'ensemble de données H (figure 5.9), il y a encore du bruit pour les trois simulations effectuées. Le maximum de la simulation 1 (figure 5.9(a)) se trouve lorsque le TIM1 et le TIM2 sont situés dans le même intervalle, celui contenant la vraie position du TIM2. On se souviendra que r^2 et MapArg trouvait aussi le TIM2. Le maximum de la

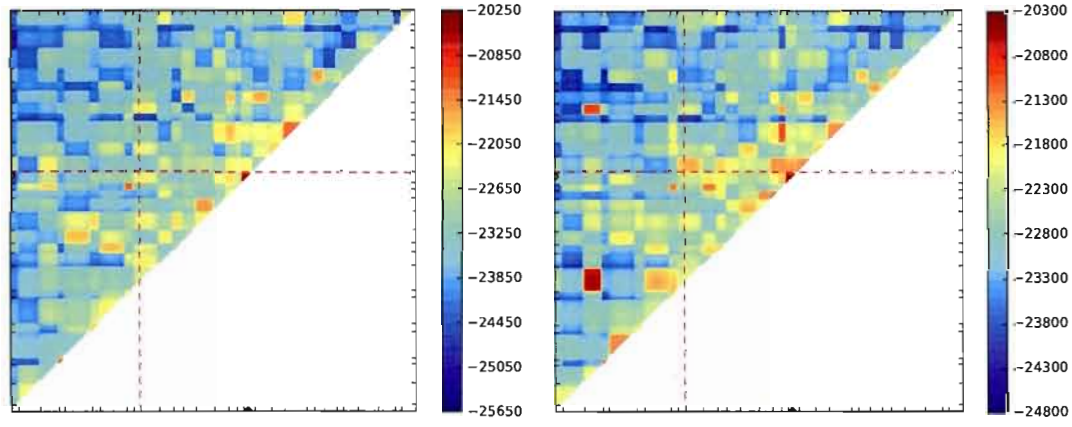
vraisemblance de la combinaison des trois simulations (figure 5.9(d)) est situé au même endroit que celui de la simulation 1. Les maximums des simulations 2 et 3 (figures 5.9(b) et 5.9(c)) sont situés loin de la vraie position.

Les maximums des vraisemblances estimées des simulations 1 à 3 de l'ensemble de données J (figure 5.10), sont tous situés dans des intervalles où les TIM1 et TIM2 sont placés entre leurs vraies positions. Seul le maximum de la simulation 3 place le TIM1 près de sa vraie position. Le TIM1 étant celui trouver par MapArg et r^2.



(a) Donnée A, simulation 1

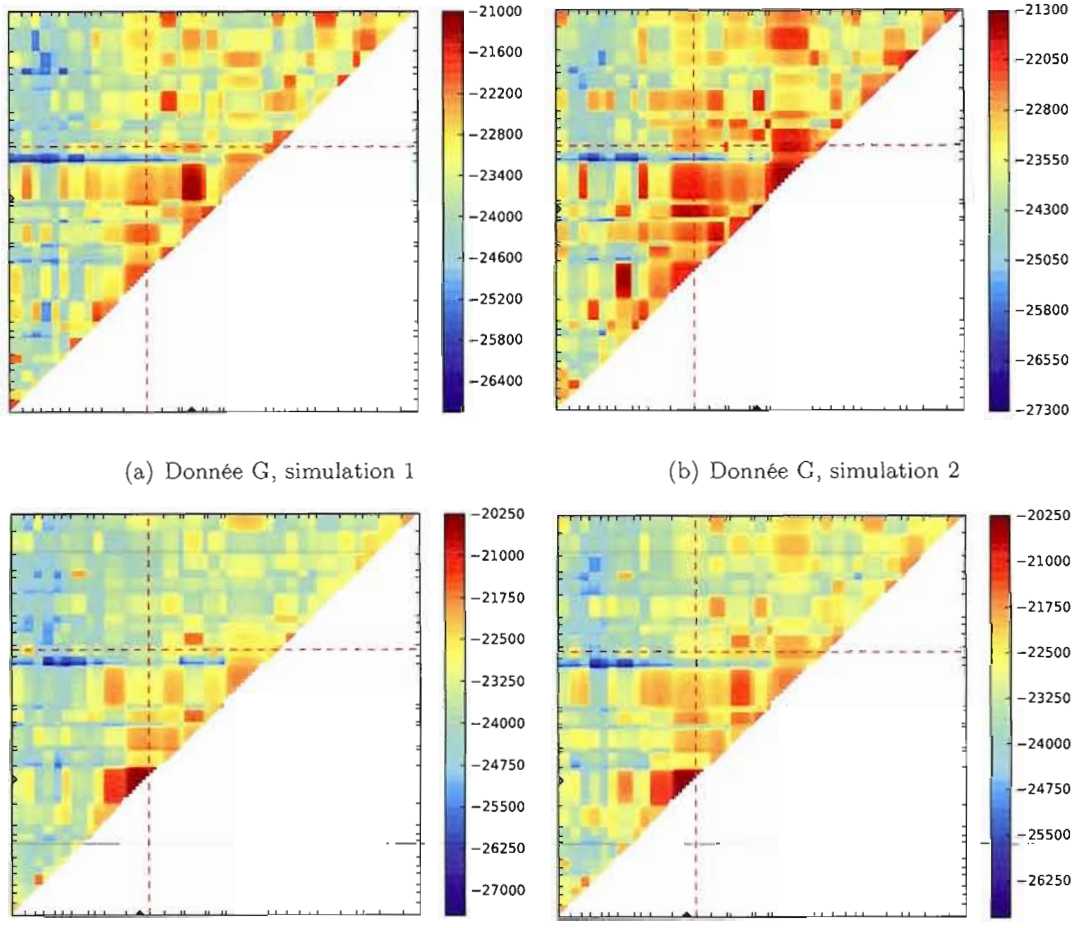
(b) Donnée A, simulation 2



(c) Donnée A, simulation 3

(d) Donnée A, combinaison des 3 simulations

Figure 5.7 Résultats avec PyArg pour les données A. Les marqueurs utilisés sont les 30 plus équidistants parmi les 60 plus polymorphiques. Trois simulations de 1 000 graphes par intervalle ont été effectuées. La figure (d) représente la combinaison de ces trois simulations, soit une simulation de 3 000 graphes par intervalle.



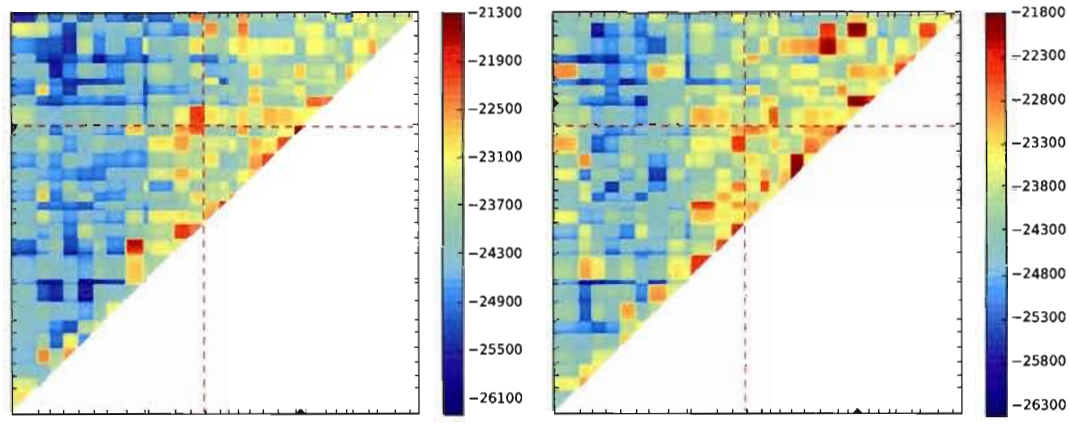
(a) Donnée G, simulation 1

(b) Donnée G, simulation 2

(c) Donnée G, simulation 3

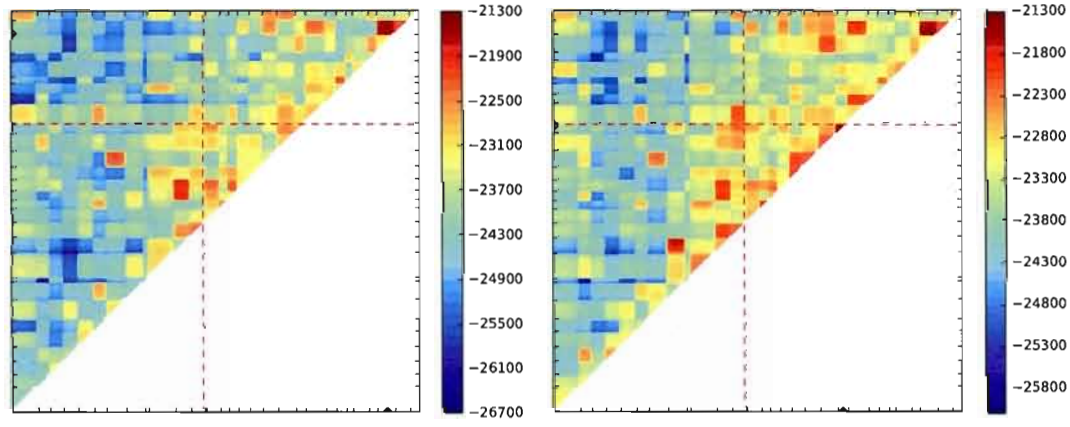
(d) Donnée G, combinaison des 3 simulations

Figure 5.8 Résultats avec PyArg pour les données G. Les marqueurs utilisés sont les 30 plus équidistants parmi les 60 plus polymorphiques. Trois simulations de 1 000 graphes par intervalle ont été effectuées. La figure (d) représente la combinaison de ces trois simulations, soit une simulation de 3 000 graphes par intervalle.



(a) Donnée H, simulation 1

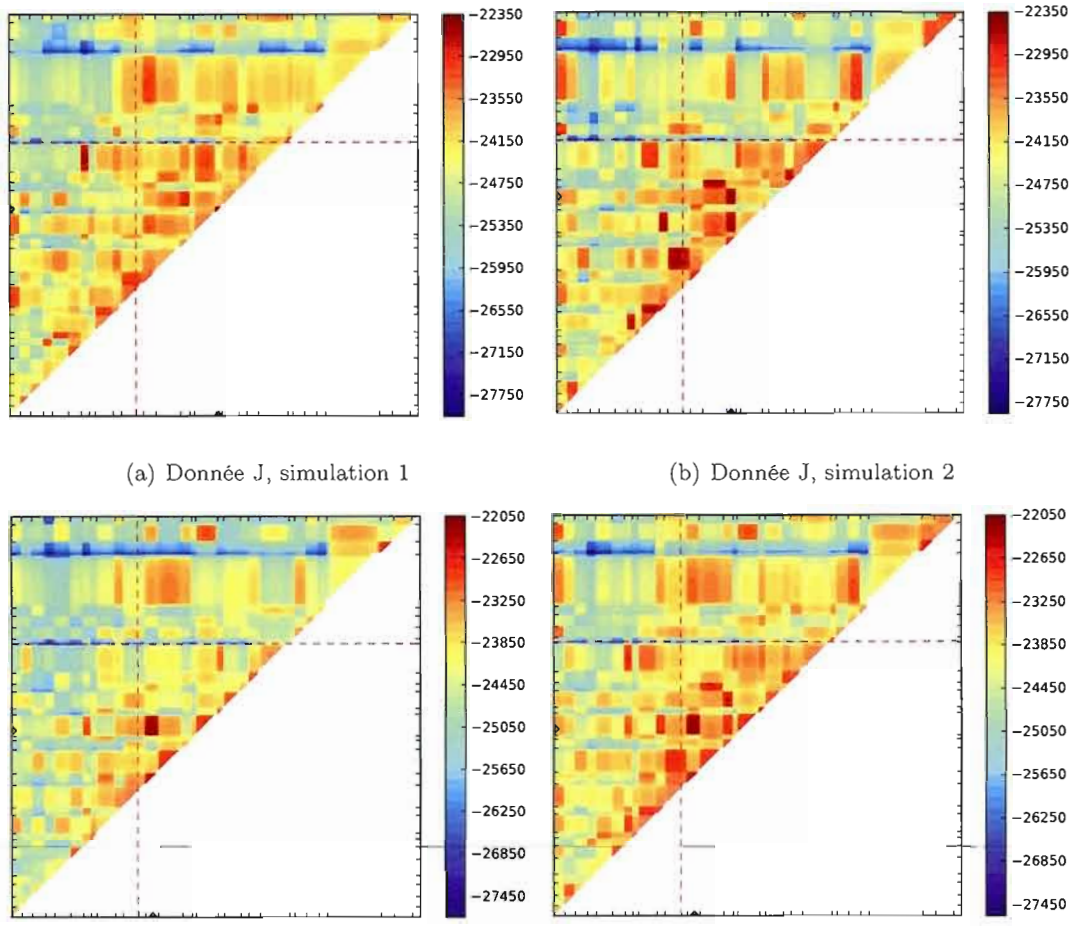
(b) Donnée H, simulation 2



(c) Donnée H, simulation 3

(d) Donnée H, combinaison des 3 simulations

Figure 5.9 Résultats avec PyArg pour les données H. Les marqueurs utilisés sont les 30 plus équidistants parmi les 60 plus polymorphiques. Trois simulations de 1 000 graphes par intervalle ont été effectuées. La figure (d) représente la combinaison de ces trois simulations, soit une simulation de 3 000 graphes par intervalle.



(a) Donnée J, simulation 1

(b) Donnée J, simulation 2

(c) Donnée J, simulation 3

(d) Donnée J, combinaison des 3 simulations

Figure 5.10 Résultats avec PyArg pour les données J. Les marqueurs utilisés sont les 30 plus équidistants parmi les 60 plus polymorphiques. Trois simulations de 1 000 graphes par intervalle ont été effectuées. La figure (d) représente la combinaison de ces trois simulations, soit une simulation de 3 000 graphes par intervalle.

5.3.3 Diverses possibilités d'amélioration de l'estimation avec PyArg

Les premiers résultats nous ont permis de constater qu'il y avait une variabilité dans les résultats obtenus et que les paramètres utilisés n'offraient peut-être pas la meilleure estimation de la vraisemblance puisqu'ils ne nous permettaient pas de détecter les mêmes signaux qu'avec MapArg.

Pour tenter de réduire la variabilité des résultats, nous avons tenté d'aller vers la limite de la vraisemblance composite utilisée par MapArg avec les fenêtres de marqueurs. Nous parlons de limite puisque nous avons fait la vraisemblance composite des trois simulations effectuées pour chaque ensemble de données. C'est comme si nous avions utilisé trois fenêtres des 30 mêmes marqueurs. Il s'agit en fait de faire la moyenne des logarithmes des vraisemblances des simulations. La figure 5.11 présente les résultats obtenus pour les quatre ensembles de données A, G, H et J. On remarque que pour tous les ensembles de données, les maximums de vraisemblances se situent lorsque les TIM1 et TIM2 sont placés dans le même intervalle, sur la diagonale, (figures 5.11(b), 5.11(c) et 5.11(d)), ou presque (figure 5.11(a)) et ces intervalles sont ceux où se trouve un des TIM (figures 5.11(a), 5.11(b) et 5.11(c)), ou presque (figure 5.11(d)).

Mais nous n'arrivons pas encore à trouver les vraies positions des TIM1 et TIM2. Nous avons voulu explorer la possibilité d'augmenter le nombre de graphes construits par intervalle. Nous avons décidé de nous concentrer sur l'ensemble de données H, puisque pour cet ensemble nous observions des signaux clairs aux TIM1 et TIM2 avec MapArg.

Pour augmenter le nombre de graphes construits par intervalle, sans augmenter les temps de calcul, nous avons réduit le nombre de marqueurs utilisés. Nous avons utilisé les 15 marqueurs les plus équidistants parmi les 30 plus polymorphiques. Nous avons fait quatre simulations de 5 000 graphes par intervalle, ce qui correspond à un total de deux millions de graphes. La figure 5.12 présente les résultats, et les figures 5.15(a) et 5.15(b) respectivement la combinaison et la vraisemblance composite de ces quatre simulations. Ces simulations n'offrent pas de meilleure estimation de la position du TIM1 et du TIM2,

même que le maximum de la vraisemblance composite n'est pas près du TIM1 ou du TIM2. Nous avons probablement enlevé trop de marqueurs aux séquences, ce qui réduit considérablement l'information utilisée.

C'est pourquoi nous avons décidé de faire plus de simulations avec les mêmes 30 marqueurs utilisés précédemment pour les données H. Nous avons effectué 7 autres simulations de 1 000 graphes par intervalle. Les résultats sont présentés sur les figures 5.13 et 5.14. Les figures 5.14(d) et 5.15(c) présentent respectivement la combinaison et la vraisemblance composite des dix simulations de 1 000 graphes par intervalle de l'ensemble H.

Encore une fois, on remarque une variabilité dans les estimateurs de maximum de vraisemblance parmi les simulations, ainsi que du bruit pour les vraisemblances estimées. Malgré tout, les vraisemblances estimées par ces dix simulations sont similaires. La combinaison des dix simulations, ou l'estimation de la vraisemblance à l'aide de 10 000 graphes par intervalle n'offre pas une meilleure estimation de la vraie position des TIM1 et TIM2.

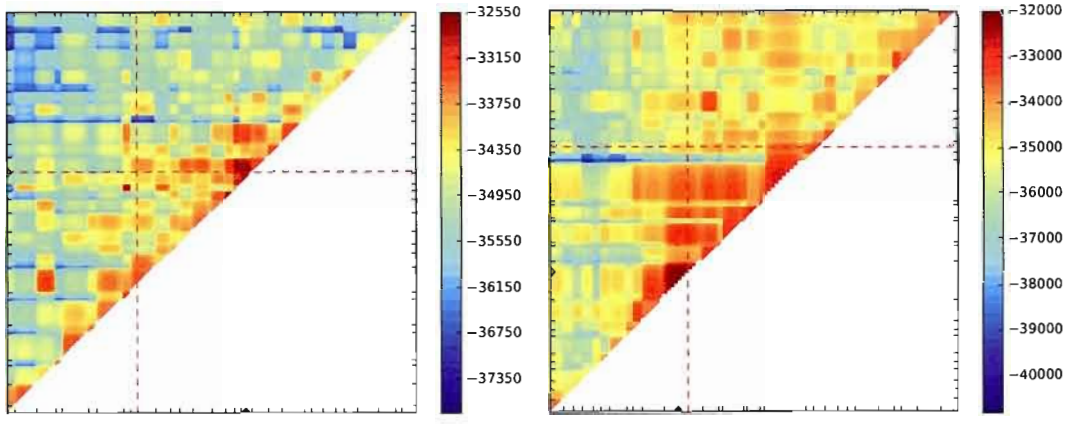
La vraisemblance composite de ces dix simulations (figure 5.15(c)) offre la même estimation que la vraisemblance composite de nos trois premières simulations (figure 5.11(c)). La position estimée est celle où les TIM1 et TIM2 sont situés dans le même intervalle, celui où se trouve le TIM2. Mais on remarque une zone triangulaire où la vraisemblance est haute, et qui correspond à la région située entre les vraies positions des TIM1 et TIM2.

On se souvient que certaines de nos séquences non affectées par le caractère possèdent tout de même une des deux mutations, et que dans notre modélisation nous leur inférons deux allèles ancestraux primitifs. Nous avons décidé de faire de nouvelles simulations en utilisant les vrais génotypes de nos séquences pour l'inférence, car ceci nous permettra de voir si cette erreur d'inférence nuit à notre estimation de la vraisemblance. Les figures 5.16 et 5.17 présentent les résultats obtenus. Nous avons effectué deux répétitions d'une simulation de 500 graphes par intervalle pour les données H et G (figure

5.16), et seulement une simulation de 500 graphes par intervalle pour les données A et J (figure 5.17), en utilisant les 30 marqueurs les plus équidistants parmi les 60plus polymorphiques.

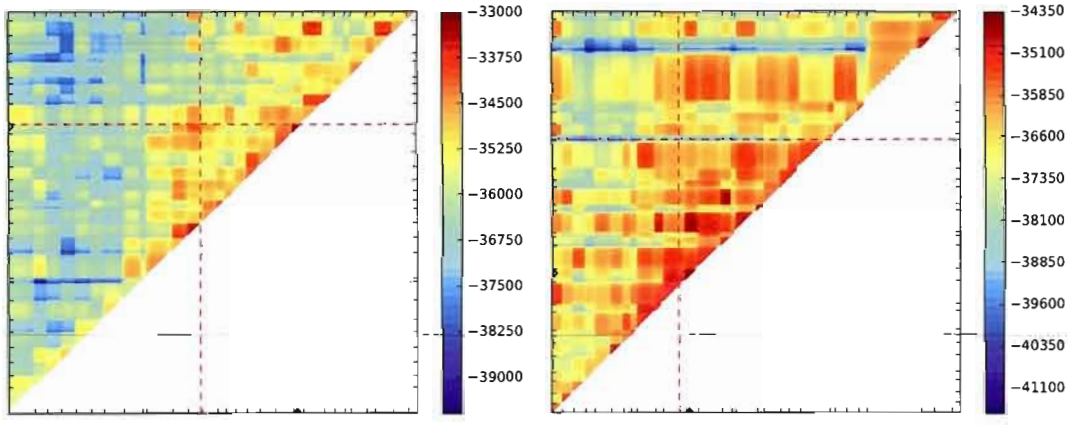
Ce qui frappe avec les résultats obtenus, en utilisant les vrais génotypes de nos séquences aux marqueurs cherchés, est que la vraisemblance pour les intervalles situés près de la diagonale est beaucoup plus basse qu'avant. Pour les données H, le maximum de vraisemblance de la première simulation (figure 5.16(a)) est situé dans le bon intervalle, tandis que pour la deuxième simulation (figure 5.16(b)), il est situé tout près de la vraie position des TIM1 et TIM2.

Pour l'ensemble de données G (figures 5.16(c) et 5.16(d)), on remarque que en moyenne la région où la vraisemblance est maximale est située assez près de l'intervalle contenant la vraie position des TIM1 et TIM2. Une contrainte de temps ne nous a pas permis d'effectuer plusieurs simulations pour les données A et J. Il est par conséquent difficile de savoir si certaines régions où la vraisemblance est haute sont seulement dues à de la variabilité ou non. Malgré tout, on remarque que la vraisemblance estimée est haute dans les régions où se trouvent les TIM1 et les TIM2 pour les deux ensembles de données.



(a) Donnée A, vraisemblance composite

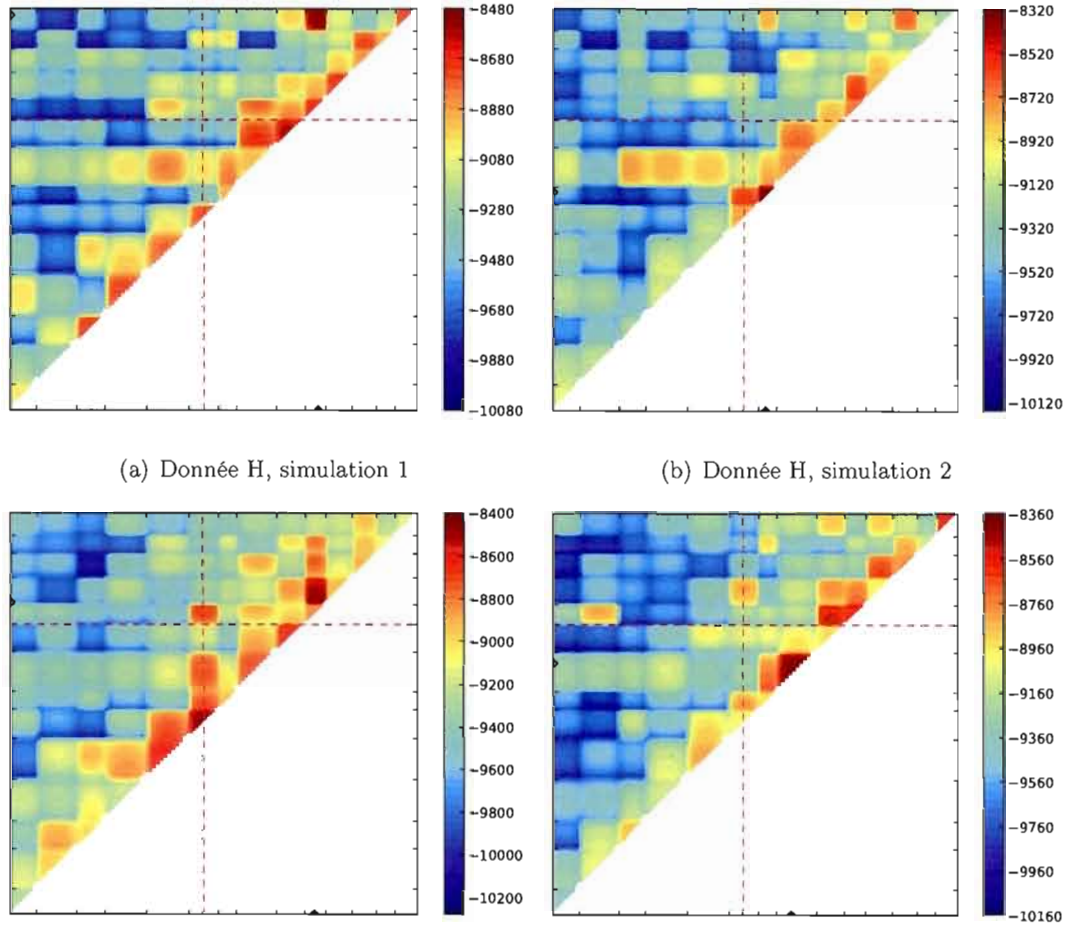
(b) Donnée G, vraisemblance composite



(c) Donnée H, vraisemblance composite

(d) Donnée J, vraisemblance composite

Figure 5.11 On présente ici les résultats de vraisemblance composite pour les données A, G, H et J, obtenus en combinant leurs trois simulations de 1 000 graphes par intervalle (les trois premières simulations pour H). Il s'agit en fait de la moyenne des logarithmes des trois estimations de la fonction de vraisemblance pour ces données.



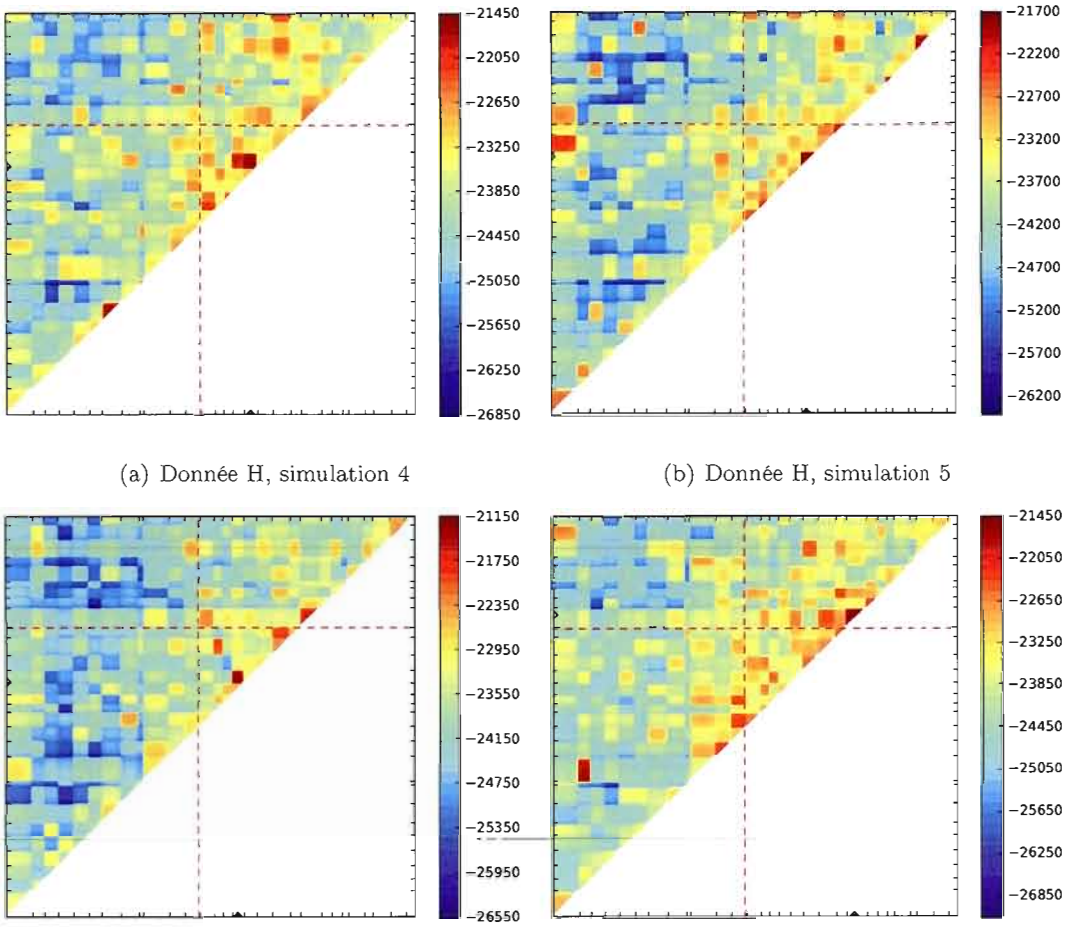
(a) Donnée H, simulation 1

(b) Donnée H, simulation 2

(c) Donnée H, simulation 3

(d) Donnée H, simulation 4

Figure 5.12 Résultats avec PyArg pour les données H. Les marqueurs utilisés sont les 15 plus équidistants parmi les 30 plus polymorphiques. Quatre simulations de 5 000 graphes par intervalle ont été effectuées.



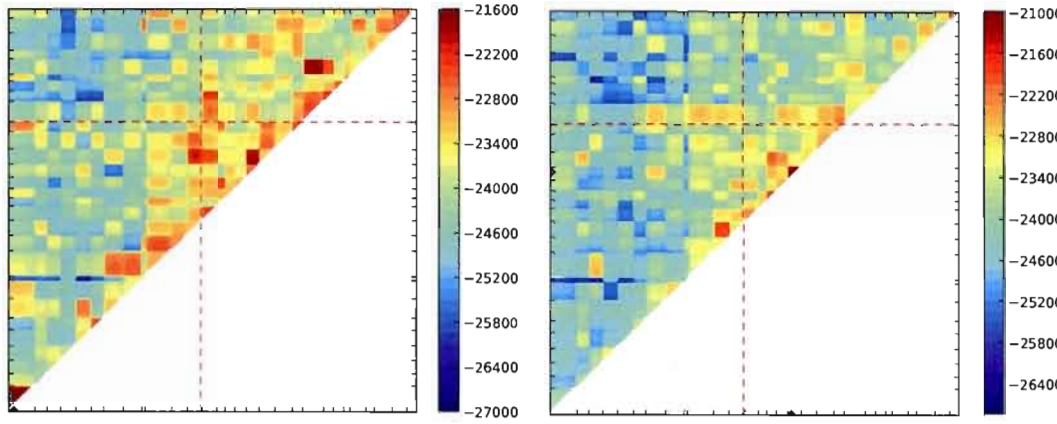
(a) Donnée H, simulation 4

(b) Donnée H, simulation 5

(c) Donnée H, simulation 6

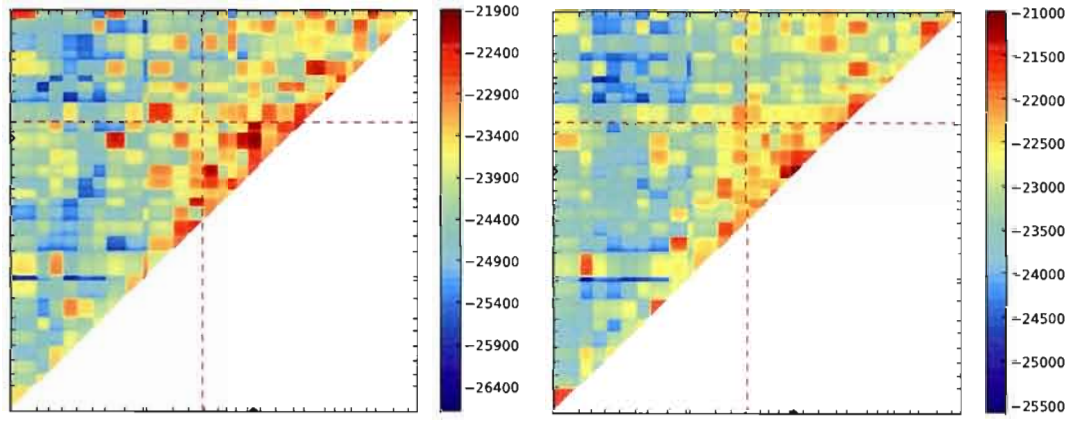
(d) Donnée H, simulation 7

Figure 5.13 Résultats avec PyArg pour les données H. Les marqueurs utilisés sont les 30 plus équidistants parmi les 60 plus polymorphiques. On présente ici les simulations 4 à 7 de 1 000 graphes par intervalle chacune.



(a) Donnée H, simulation 8

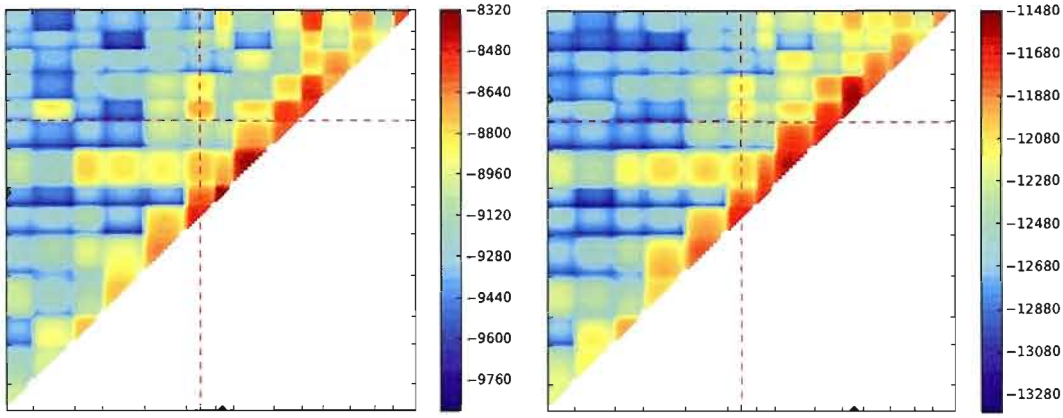
(b) Donnée H, simulation 9



(c) Donnée H, simulation 10

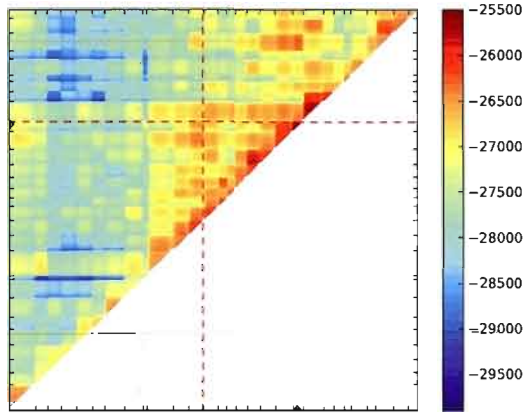
(d) Donnée H, combinaison des 10 simulations

Figure 5.14 Résultats avec PyArg pour les données J. Les marqueurs utilisés sont les 30 plus équidistants parmi les 60 plus polymorphiques. On présente ici les simulations 8 à 10 ((a) à (c)) de 1 000 graphes par intervalle chacune. La figure (d) représente la combinaison des dix simulations et représente une simulation de 10 000 graphes par intervalle.



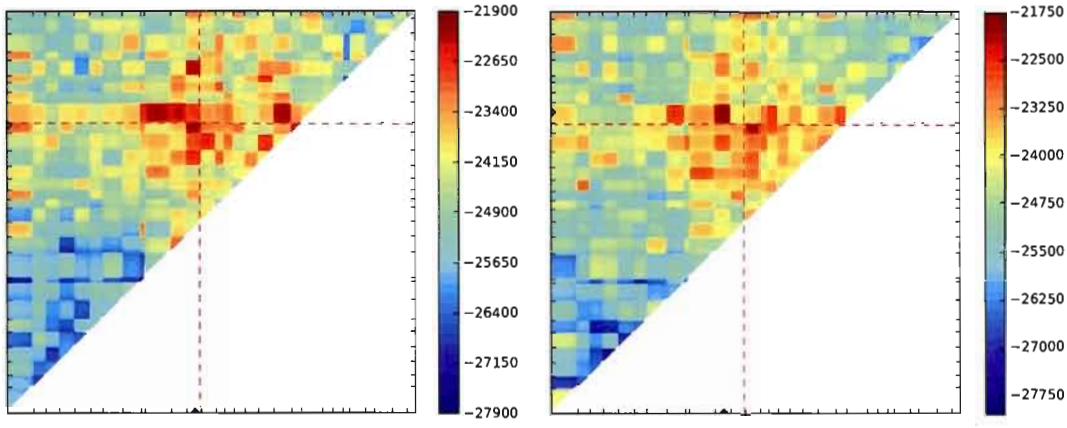
(a) Donnée H, combinaison des 4 simulations

(b) Donnée H, vraisemblance composite



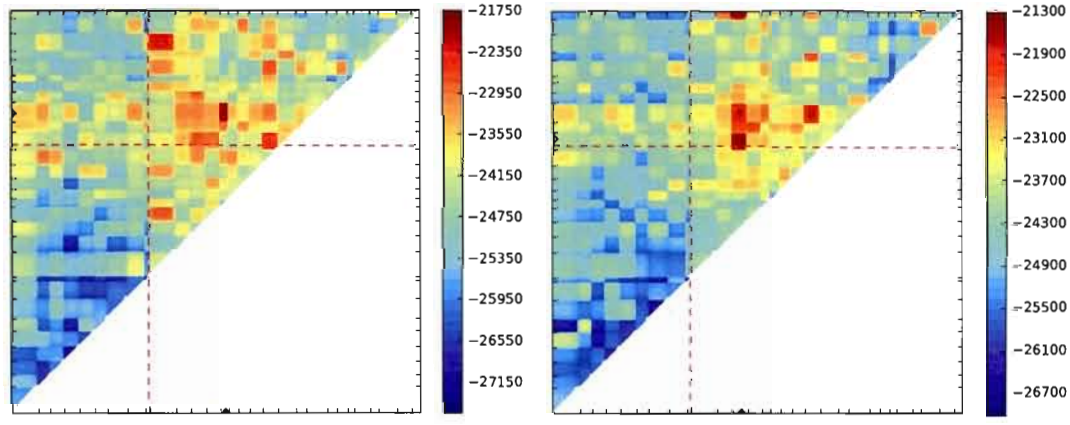
(c) Donnée H, vraisemblance composite

Figure 5.15 La figure (a) représente la combinaison des quatre simulations des données H de 5 000 graphes chacune (présentées sur la figure 5.12) en utilisant 15 marqueurs. La figure (b) représente la vraisemblance composite de ces quatre simulations. Et la figure (c) représente la vraisemblance composite des dix simulations de 1 000 graphes chacune des données H, utilisant 30 marqueurs.



(a) Donnée H, simulation 1 avec les solutions

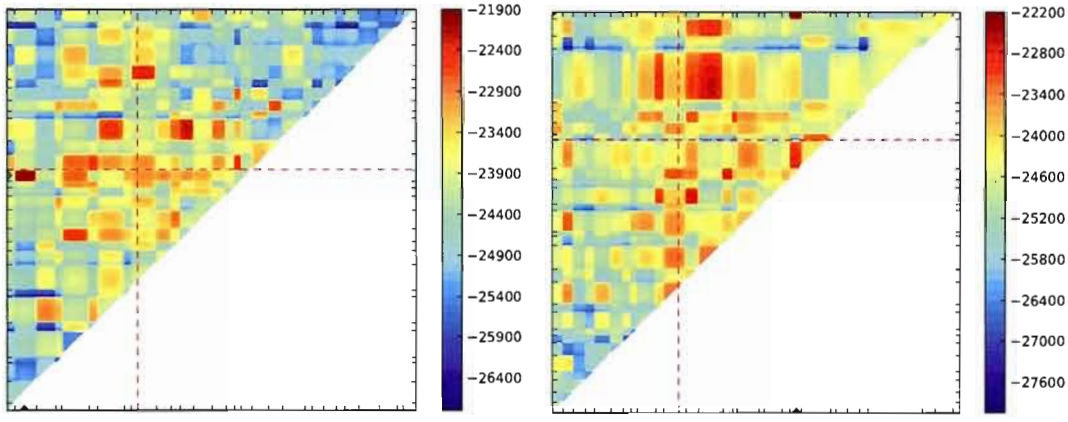
(b) Donnée H, simulation 2 avec les solutions



(c) Donnée G, simulation 1 avec les solutions

(d) Donnée G, simulation 2 avec les solutions

Figure 5.16 Résultats obtenus avec PyArg en utilisant les vrais génotypes aux marqueurs cherchés pour les données H et G. Les marqueurs utilisés sont les 30 plus équidistants parmi les 60 plus polymorphiques. On présente ici les deux simulations de 500 graphes par intervalle chacune pour les données H et G.



(a) Donnée A, simulation avec les solutions (b) Donnée J, simulation avec les solutions

Figure 5.17 Résultats obtenus avec PyArg en utilisant les vrais génotypes aux marqueurs cherchés pour les données A et J. Les marqueurs utilisés sont les 30 plus équidistants parmi les 60 plus polymorphiques. On présente ici la simulation de 500 graphes par intervalle chacune pour les données A et J.

5.4 Discussion

Les résultats obtenus avec la méthode de cartographie MapArg et la mesure r^2 justifient nos efforts d'adaptation de MapArg à la cartographie de caractère polygénique. Bien souvent ces deux méthodes n'arrivent pas à détecter un des TIM, et ce, malgré le fait que les ensembles de données simulées ne sont pas complexes (il n'y a ni phénocopie, ni pénétrance incomplète).

Il est difficile de savoir ce qui améliorerait nos résultats, car le temps ne nous a pas permis de faire un grand nombre de simulations. Bien que le programme en parallèle nous ait permis de faire beaucoup de simulations en peu de temps, il demande la gestion de plusieurs ordinateurs ce qui complique la tâche.

Il reste encore quelques pistes à explorer ; par exemple, il faudrait faire encore plus de simulations avec les différentes bases de données ; nous avons exploré seulement avec les données H, mais peut-être aurions-nous de meilleurs résultats avec une autre. Le

nombre de marqueurs et le nombre de graphes construits par intervalle sont peut-être insuffisants.

De plus, nous avons vu que l'utilisation des vrais génotypes nous permet d'obtenir de meilleures estimations des positions des TIM1 et TIM2. Il faudrait, par conséquent, trouver un moyen d'améliorer l'inférence des TIM1 et TIM2, car, pour l'instant, on observe que le maximum de vraisemblance est souvent situé près de la diagonale, c'est-à-dire lorsque le TIM1 et le TIM2 sont situés dans le même intervalle. Ce qui pourrait être expliqué par notre façon d'inférer les TIM1 et TIM2 ; rappelons qu'on infère aux personnes affectées deux allèles mutants et deux allèles sains aux personnes non-affectées. En insérant les TIM1 et TIM2 qui ont des allèles identiques, dans le même intervalle, c'est comme si on insérait qu'un seul marqueur. Si, de plus, cet intervalle est proche d'un intervalle contenant un des marqueurs cherchés, le graphe construit peut alors avoir une plus grande vraisemblance ; il a été plus «facile» de construire des graphes en plaçant les TIM1 et TIM2 dans cet intervalle.

Terminons en mentionnant que la distribution proposée utilisée pour construire les graphes n'est peut-être pas la distribution optimale. Fearnhead et Donnelly (2001) ont démontré quelle était la distribution proposée optimale et ont proposé une estimation de cette distribution qui semble offrir de meilleurs résultats que celle de Griffiths et Marjoram (1996). Avec un peu plus de temps, il aurait été possible de l'utiliser.

CONCLUSION

Notre objectif était de proposer une adaptation de la méthode de cartographie génétique fine MapArg à la cartographie de caractère polygénique. Nous souhaitions aussi tester cette adaptation, établir ses limitations et la comparer aux méthodes de cartographie génétique existantes.

MapArg suppose que le caractère est influencé par un seul gène ; en utilisant un échantillon de cas et de contrôles, cette méthode estime la fonction de vraisemblance de la position de ce gène. Pour ce faire, elle doit inférer le génotype au gène cherché, à partir du phénotype, pour chacune des séquences de l'échantillon. Pour simplifier notre adaptation de MapArg à la cartographie de caractère polygénique, nous avons choisi de commencer par supposer que le caractère est influencé par deux gènes et d'utiliser un échantillon d'haplotypes ce qui, comparé aux diploïdes, facilite la modélisation. Nous devions modéliser l'interaction entre les gènes et déterminer comment ils affectent le caractère. De plus, il a fallu décider de la façon dont nous allions inférer les génotypes des TIM1 et TIM2 à partir des phénotypes.

Des ensembles de données ont été simulés pour nous permettre de tester notre adaptation. Une séquence était affectée par le caractère si elle possédait deux allèles mutants, sinon cette séquence n'était pas affectée par le caractère. Pour l'inférence des génotypes à partir des phénotypes, nous avons choisi qu'une séquence affectée possèderait les deux allèles mutants, tandis qu'une séquence non atteinte du caractère possèderait deux allèles sains. Cela revient à supposer que la maladie est doublement récessive et les allèles mutants rares. Nous étions conscients de l'erreur d'inférence, mais nous ne pensions pas qu'elle aurait un impact sur l'estimation de la position des TIM1 et TIM2.

Les résultats obtenus avec la mesure d'association r^2 et MapArg pour les ensembles de données simulées, ont montré la nécessité d'adapter les méthodes de cartographie à la réalité des caractères polygéniques, car souvent ces deux méthodes ne trouvaient pas un des deux marqueurs cherchés. Le programme PyArg nous a permis de tester notre adaptation avec des temps de calcul raisonnables. Les premiers résultats obtenus étaient tout de même encourageants, nous arrivions à trouver un des gènes causaux. Nous avons tenté de faire plus de graphes par intervalle, d'utiliser la vraisemblance composite pour tenter de diminuer la variabilité observée en faisant plusieurs répétitions de l'estimation de la vraisemblance.

C'est en utilisant les génotypes aux marqueurs cherchés pour l'inférence que nous avons obtenu les meilleurs résultats. L'erreur faite lors de l'inférence était donc significative. Bien sûr, d'autres facteurs pourraient améliorer notre estimation. L'utilisation d'une meilleure distribution proposée pour la construction des graphes pourrait aider. Mais pour éventuellement utiliser l'adaptation proposée, il faut trouver un moyen d'améliorer l'inférence des génotypes aux gènes causaux. Il faudrait aussi tester l'adaptation de MapArg avec différentes modélisations de l'interaction des gènes causaux.

GLOSSAIRE

ADN. Formé de deux brins en forme d'hélice reliés par des paires de bases azotées, il est la matière première des chromosomes.

Allèle. Les différentes versions possibles d'un même gène.

Base azotée. Molécule formant l'ADN, il en existe quatre notées G, C, T et A. La base G est en paire avec la base C, et la base T est en paire avec la base A.

Caractère. On nomme caractère une particularité biologique d'origine génétique. La couleur des yeux, le groupe sanguin et la taille d'un individu sont des caractères.

Cartographie fine. La cartographie fine fait référence aux méthodes de cartographie génétique permettant de trouver de très courte région d'un chromosome en association avec un certain caractère.

Centimorgan (cM). Unité de mesure de positionnement sur le génome. Une distance de 1 cM entre deux gènes équivaut à un taux de recombinaison de 1% entre ces gènes.

Centromètre. Partie du chromosome où débute le phénomène de réplication. Le chromosome reste alors attaché à sa copie par le centromètre jusqu'à la deuxième division cellulaire lors de la méiose.

Chromosome. Formé d'une longue molécule d'ADN, le chromosome ressemble à un bâtonnet. Le génome humain est composé de 46 chromosomes.

Coalescence. Un événement de coalescence a lieu dans une généalogie lorsque deux lignées possèdent le même ancêtre commun à un moment dans le passé.

Codominance. Il y a codominance si les allèles à un gène s'expriment de la même façon et influencent tous deux le phénotype. Par exemple, lorsque les allèles A et B sont en présence l'un de l'autre au gène du groupe sanguin, il y a codominance.

Coségrégation. La coségrégation de deux allèles situés sur un même chromosome signifie qu'ils ont été transmis ensemble à un gamète lors de la méiose.

Déséquilibre de liaison. Deux gènes sont en déséquilibre de liaison, lorsqu'ils sont situés assez près l'un de l'autre pour permettre d'observer une association entre ces gènes dans la population. En opposition à l'équilibre de liaison.

Diploïde. Une cellule composée de paires de chromosomes homologues est dite : cellule diploïde.

Dominant. Un allèle est dit dominant si, peu importe l'autre allèle en sa présence, il s'exprime à travers le phénotype. Par exemple, les allèles A et B du gène du groupe sanguin sont chacun dominant.

Enjambement. Phénomène se produisant lors de l'appariement des chromosomes homologues durant la méiose et consistant en l'échange de certaines parties des chromosomes homologues entre eux. Les deux chromosomes homologues résultants sont donc formés du mélange des chromosomes parentaux.

Équilibre de liaison. Il y a équilibre de liaison entre deux gènes lorsque la fréquence des différents haplotypes possibles est seulement le produit des fréquences des allèles les composants.

Faux positif. Dans le contexte des études d'association, ce dit d'un marqueur déclaré à tort en association avec le caractère étudié.

Fonction de pénétrance. La fonction de pénétrance d'un gène influençant un caractère, est l'ensemble des probabilités d'être affecté par le caractère en fonction du génotype au gène.

Fondateur. En analyse de liaison, on appelle fondateurs les individus n'ayant pas de parents dans la généalogie.

Gamète. Cellule sexuelle haploïde contenant seulement un seul exemplaire de chacun des chromosomes. Chez l'être humain, les gamètes contiennent 23 chromosomes.

Gène. Un gène est composé d'un ensemble de paires de bases consécutives qui ensemble transmettent de l'information sur une tâche bien précise à la cellule.

Génome. Ensemble des informations nécessaires pour faire fonctionner un organisme et contenues dans chacune des cellules le composant. Le génome humain est composé de 46 chromosomes.

Génotype. On nomme la paire d'allèles qu'un individu possède pour un gène en particulier son génotype à ce gène. Par extension, on parle aussi du génotype d'un individu comme étant la paire d'allèles pour un ensemble de gènes.

Graphe de recombinaison ancestral (ARG). Généralisation du processus de coalescence permettant d'inclure des événements de recombinaison dans le modèle de généalogie.

Haploïde. Un haploïde est une cellule possédant qu'une seule copie de chacun des chromosomes. Les gamètes sont des cellules haploïdes.

Haplotype. Lorsqu'un seul chromosome est considéré, ou seulement une séquence de marqueurs extraite de ce chromosome est considérée, on parle alors d'haplotype.

Hétérozygote. Un individu est hétérozygote à un gène, si les deux allèles qu'il possède pour ce gène sont différents.

Homologues. Les chromosomes d'une même paire, ayant la même taille et les mêmes gènes, mais pas nécessairement les mêmes allèles, sont dits homologues.

Homozygote. Un individu est homozygote à un gène, si les deux allèles qu'il possède pour ce gène sont identiques.

Identique par descendance (IBD). On dit que deux individus possèdent un allèle identique par descendance si celui-ci provient exactement du même ancêtre.

Intervalle ancestral. Dans la méthode MapArg et dans l'adaptation proposée de celle-ci, un intervalle entre marqueurs est dit ancestral, si l'on retrouve à sa gauche et à sa droite des marqueurs ancestraux. Ces marqueurs ne seront pas nécessairement situés près de l'intervalle.

Locus. Un emplacement exact de paires de bases sur un chromosome se nomme locus (loci au pluriel).

MapArg. Méthode de cartographie génétique fine utilisant le graphe de recombinaison ancestral (Larribe, Lessard et Schork, 2002 ; Larribe et Lessard, 2008).

Marqueur génétique. Gène ou morceau de séquence d'ADN ayant une position connue sur le génome et possédant des variations (différents allèles) dans la population.

Méiose. Division cellulaire menant à la production des cellules sexuelles.

MRCA. De l'anglais « most recent common ancester », signifiant le plus récent ancêtre commun et faisant référence à l'ancêtre le plus récent de l'ensemble des individus (ou séquences) considérés.

Mutation. Nous appelons les erreurs pouvant se produire lors de la méiose — paire de bases oubliée, reproduite deux fois ou mal reproduite — des mutations. Un événement de mutation aura lieu, du présent vers le passé, dans un graphe de recombinaison ancestral, si une seule séquence, à cette étape, possède un allèle mutant à un marqueur et que l'étape suivante cet allèle est devenu sain (modèle de mutation non récurrente).

Phase. Lorsque les génotypes à deux marqueurs sont connus pour un individu, nous disons que la phase est elle aussi connue si nous savons quels allèles, aux deux marqueurs, sont situés sur le même chromosome.

Phénocopie. Facteur d'origine autre que génétique influençant la probabilité d'être affecté par un caractère génétique.

Phénotype. Le phénotype est le résultat observé de l'action d'un gène.

Polygénique. Une maladie d'origine génétique est dite polygénique lorsqu'elle est en fait influencée simultanément par plusieurs gènes.

Polymorphisme nucléotide simple (SNP). Courte séquence d'ADN où une seule paire de bases varie dans la population. Cette paire de bases est utilisée comme marqueur génétique et se nomme SNP.

Processus de coalescence. Processus stochastique utilisé pour la simulation de généalogie. Sa flexibilité permet la simulation de généalogie selon différents modèles génétiques. Une de ses généralisations, le graphe de recombinaison ancestral, permet d'inclure des événements de recombinaison dans la généalogie.

Python. Langage de programmation orienté objet.

PyArg. Nom du programme Python créé pour implanter la méthode MapArg et son adaptation proposée, utilisant la programmation en parallèle.

Récessif. Un allèle est dit récessif s'il s'exprime, à travers le phénotype, seulement en présence d'un allèle identique à lui.

Recombinaison. Un événement de recombinaison est observé entre deux gènes s'ils ne proviennent pas du même chromosome parental, c'est-à-dire qu'un nombre impair d'enjambements ont eu lieu entre ces gènes. Un événement de recombinaison aura lieu, du présent vers le passé, dans un graphe de recombinaison ancestral, si le matériel génétique d'une séquence provient de deux séquences parentales différentes.

Réplication. Phénomène de dédoublement des chromosomes observé lors des divisions cellulaires.

Taux de recombinaison. Le taux de recombinaison entre deux gènes est la proportion de recombinaison observée en général entre ces deux gènes.

TIM. Terme utilisé pour indiquer le gène cherché influençant le caractère étudié. De l'anglais, « trait influencing mutation ».

RÉFÉRENCES

Almgren, P., P.-O. Bendahl, H. Bengtsson, O. Hössjer, et R. Perfekt. 2003. *Statistics in genetics (lecture notes)*. Centre for Mathematical Sciences, Lund University.

Barhdadi, A. et M.-P. Dubé. 2010. « Testing for gene-gene interaction with ammi models », *Statistical applications in genetics and molecular biology*, vol. 9, no. 1, p. Article 2.

Biernacka, J. M., L. Sun, et S. B. Bull. 2005. « Simultaneous localization of two linked disease susceptibility genes », *Genetic Epidemioly*, vol. 28, no. 1, p. 33–47.

Boucher, G. 2009. « Intégration de la réalité diploïde et des modèles de pénétrance à une méthode de cartographie génétique fine ». Mémoire de maîtrise, Université du Québec à Montréal.

Campbell, N. A. et R. Mathieu. 1995. *Biologie*. Éditions du Renouveau Pédagogique.

Collins, F. S. et M. K. Mansoura. 2001. « The human genome project. revealing the shared inheritance of all humankind », *Cancer*, vol. 91, no. 1 Suppl, p. 221–5.

Cordell, H. J. 2002. « Epistasis : what it means, what it doesn't mean, and statistical methods to detect it in humans », *Human Molecular Genetics*, vol. 11, no. 20, p. 2463–2468.

Elston, R. C. 2000. « Introduction and overview, statistical methods in genetic epidemiology », *Statistical methods in medical research*, vol. 9, no. 6, p. 527–541.

Fearnhead, P. et P. Donnelly. 2001. « Estimating recombination rates from population genetic data », *Genetics*, vol. 159, no. 3, p. 1299–1318.

Griffiths, R. C. et P. Marjoram. 1996. « Ancestral inference from samples of dna sequences with recombination », *Journal of Computational Biology*, vol. 3, no. 4, p. 479–502.

Griffiths, R. C. et S. Tavaré. 1994a. « Ancestral inference in population genetics », *Statistical Science*, vol. 9, no. 3, p. 307–319.

———. 1994b. « Simulating probability distributions in the coalescent », *Theoretical population biology*, vol. 46, p. 131–159.

Hein, J., M. H. Schierup, et C. Wiuf. 2005. *Gene Genealogies, Variation and Evolution : A Primer in Coalescent Theory*. Oxford University Press.

Hudson, R. R. 1983. « Properties of a neutral allele model with intragenic recombination », *Theoretical population biology*, vol. 23, no. 2, p. 183–201.

———. 2002. « Generating samples under a wright-fisher neutral model of genetic variation », *Bioinformatics*, vol. 18, no. 2, p. 337–338.

Kingman, J. F. C. 1982. « The coalescent », *Stochastic Processes and their Applications*, vol. 13, p. 235–248.

Lander, E. S. et N. J. Schork. 1994. « Genetic dissection of complex traits », *Science*, vol. 265, no. 5181, p. 2037–2048.

Larribe, F. et P. Fearnhead. 2010. « On composite likelihoods in statistical genetics », *Soumis*.

Larribe, F. et S. Lessard. 2008. « A composite-conditional-likelihood approach for gene mapping based on linkage disequilibrium in windows of marker loci », *Statistical applications in genetics and molecular biology*, vol. 7, p. Article 27.

Larribe, F., S. Lessard, et N. J. Schork. 2002. « Gene mapping via the ancestral recombination graph », *Theoretical population biology*, vol. 62, no. 2, p. 215–229.

Liang, K. Y., Y. F. Chiu, et T. H. Beaty. 2001. « A robust identity-by-descent procedure using affected sib pairs : multipoint mapping for complex diseases », *Human Heredity*, vol. 51, no. 1-2, p. 64–78.

Lindsay, B. G. 1988. « Composite likelihood mehods », *Contempory Mathematics*, vol. 80, p. 221–239.

Marchini, J., P. Donnelly, et L. R. Cardon. 2005. « Genome-wide strategies for detecting multiple loci that influence complex diseases », *Nature Genetics*, vol. 37, no. 4, p. 413–417.

McPeek, M. S. et A. Strahs. 1999. « Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping », *American Journal of Human Genetics*, vol. 65, no. 3, p. 858–875.

Mendel, G. 1866. « Versuche über plflanzenhybriden (expériementation sur les végétaux) », *Verhandlungen des naturforschenden Vereins Brünn*, vol. IV, p. 3–47.

Morris, A. P., J. C. Whittaker, et D. J. Balding. 2000. « Bayesian fine-scale mapping of disease loci, by hidden markov models », *American Journal of Human Genetics*, vol. 67, no. 1, p. 155–169.

Nordborg, M. 2007. *Handbook of Statistical Genetics*. T. 2, chapitre Coalescent Theory, p. 843–877. John Wiley and Sons, Ltd., 3 édition.

Nordborg, M. et S. Tavaré. 2002. « Linkage disequilibrium : what history has to tell us », *Trends in Genetics*, vol. 18, no. 2, p. 83–90.

Olson, J. M., J. S. Witte, et R. C. Elston. 1999. « Genetic mapping of complex traits », *Statistics in medicine*, vol. 18, no. 21, p. 2961–2981.

Stephens, M. et P. Donnelly. 2000. « Inference in molecular population genetics », *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 62, no. 4, p. 605–655.

Tajima, F. 1983. « Evolutionary relationship of dna sequences in finite populations », *Genetics*, vol. 105, no. 2, p. 437–460.

Varin, C. et P. Vidoni. 2005. « A note on composite likelihood inference and model selection », *Biometrika*, vol. 92, no. 3, p. 519–528.

Zöllner, S. et J. K. Pritchard. 2005. « Coalescent-based association mapping and fine mapping of complex trait loci », *Genetics*, vol. 169, no. 2, p. 1071–1092.