Evolution Combinatoire, Algorithmique des Chromosomes

Eric Tannier

Université Lyon 1 Claude Bernard
Thèse en vue de l'obtention de
l'habilitation à diriger des recherches

Évolution combinatoire,

Algorithmique des chromosomes

Eric Tannier

Jury:
Dominique Mouchiroud (présidente)
Nadia El-Mabrouk (rapportrice)
Mathieu Blanchette (rapporteur)
Hugues Roest-Crollius (rappoteur)
Hélène Touzet (examinatrice)

Lyon, le 21 juillet 2011

Sommaire

Avant-propos

Ce mémoire est une présentation de certains de mes travaux de recherches effectués en bioinformatique au laboratoire de Biométrie et Biologie Évolutive de l'université de Lyon 1 depuis 2003, date de mon arrivée en post-doctorat à l'INRIA, suivie de la stabilisation de mon poste un an après. J'ai conçu le document comme une présentation peu formelle, préférant insister sur le contexte, les enjeux, les liens entre les méthodes et la manière dont je vois ces recherches s'intégrer dans une filiation aux sources multiples et entrecroisées.

Je renvoie aux articles pour la plupart des détails techniques, je ne les expose ici que s'ils ne sont pas encore publiés, ou s'ils sont utiles à la compréhension d'une histoire que j'ai trouvée intéressante à raconter. Ces histoires, mêmes si elles n'appartiennent pas toujours de plein droit à mes recherches, y sont reliées par une philosophie qu'on a rarement l'occasion d'exposer en détails. Je prends donc l'occasion de cet exercice universitaire pour raconter comment Alfred Sturtevant ou Seymour Benzer ont inventé le paradigme méthodologique dans lequel j'ai évolué. J'insiste sur les communications multidirectionnelles entre les disciplines, aussi bien que sur la nécessité des recherches prospectives dans une discipline.

Et j'expose certains de mes résultats qui illustrent particulièrement cette démarche.

Aucune des recherches présentées ici n'est issue d'un travail solitaire. Je remercie tous ceux qui y ont participé. En premier lieu les membres du laboratoire BBE, qui m'ont patiemment appris leur discipline, m'ont encouragé et soutenu quand j'en ai eu besoin, et m'ont permis de faire de la bio-algorithmique pas toujours intégralement bio dans un environnement exceptionnel, motivant et passionnant. Ensuite, tous les collègues français, canadiens, hongrois et bosniens qui ont participé à ces recherches. Merci aussi à Matthieu Blanchette, Nadia El-Mabrouk, Dominique Mouchiroud, Hugues Roest-Collius et Hélène Touzet pour avoir accepté de participer au

jury d'évaluation de cette thèse. Enfin, not least, mes plus inconditionnels soutiens, celles pour qui j'ai raison même quand j'ai tort, les filles de ma petite famille.

Toute mon activité professionnelle n'est pas représentée dans ce document. Entre autres activités qui ne seront qu'évoquées bien que certaines aient pris une place parfois importante, j'ai animé un groupe de travail national en génomique comparée pendant plusieurs années, organisé le programme d'une conférence satellite de Recomb consacrée à la génomique comparée, enseigné à l'INPG, à l'INSA, à l'université catholique, à l'université populaire de Lyon et à l'institut Gulbenkian de Lisbonne, écrit un article de vulgarisation et publié des ouvrages de mathématiques ou d'histoire.

J'ai aussi eu l'occasion d'encadrer des étudiants, à qui je dois donner une place particulière ici puisqu'il s'agit d'un mémoire pour l'obtention d'une habilitation à diriger des recherches : merci à Yoan Diekmann, Simon Carrignon, Marilia Braga, Celine Scornavacca, Renaud Lenne, Damir Hasic, Matthieu Perrinel, Adrien Rougny, Bamba Gaye, Louis Fippo, Coralie Gallien.

Le titre de ce mémoire peut être réarrangé selon diverses combinaisons. C'est l'ensemble des combinaisons possibles ainsi que les opérations qui servent à les obtenir qui rendent le mieux compte de son contenu.

Introduction

Seymour Benzer fut un des grands biologistes moléculaires du XXe siècle. Avançant sur les mêmes pistes qu'un Francis Crick, il contribua à la compréhension de la structure des gènes et des mutations qui les affectent. Dans la communauté des théoriciens et algorithmiciens des graphes, il est célèbre (ou au moins, il mériterait de l'être) pour avoir inventé les graphes d'intervalles, qui depuis son article en 1959 (Benzer, 1959), ont été le sujet de centaines de publications mathématiques. À vrai dire, si l'invention ne fait pas de doute, l'antériorité est discutable. On peut trouver à Seymour Benzer au moins deux prédécesseurs et de bonnes raisons pour qu'il les ait ignorés : le premier est un paragraphe en Allemand du mathématicien hongrois Hajós, explicite sur la structure des graphes d'intervalles et le problème de leur reconnaissance, paru en 1957 dans les actes d'une conférence (Hajos, 1957). Mais on peut surtout attribuer à Norbert Wiener la même invention dès 1914 (Fishburn & Monjardet, 1992), puisque l'auteur, dans le langage des *Principia Mathematica* de Russell et Whitehead, sûrement beaucoup plus illisible que l'Allemand de Hajós pour Benzer, ordonne les cliques maximales du graphe d'incompatibilité d'un ordre d'intervalles, ce qui se révélera la propriété cruciale des graphes d'intervalles pour résoudre le problème de leur reconnaissance. Comme toute invention, celle des graphes d'intervalles est le fruit d'une histoire plutôt que d'un individu, mais l'article de Benzer (1959) concentre une multitude de pistes méthodologiques dans lesquelles s'inscrit le travail présenté dans ce mémoire. Un petit aperçu de son contexte et de son contenu ne sera pas déplacé en introduction.

Morgan (1926) avait décrit au début du siècle les chromosomes comme des arrangements linéaires de gènes. Il avait inventé avec Sturtevant la *cartographie* génétique (voir le chapitre 2), en associant les paires de gènes dont les produits sont le plus souvent observés ensemble. Il restait à comprendre la structure des gènes eux-mêmes, et guidé par la récente découverte de la

structure de l'ADN (Watson & Crick, 1953), Seymour Benzer réalisa une expérience qui permettrait d'argumenter en faveur d'un arrangement également linéaire de sous-unités (ce qui est confirmé par toutes les connaissances actuelles). Il existait dans les cultures du virus T4 une forme sauvage qui accomplissait une certaine fonction, et des formes mutantes qui en étaient incapables. En recombinant deux virus mutants, la forme sauvage pouvait être régénérée à condition que les sous-unités mutées ne s'intersectent pas. On pouvait donc savoir par l'expérience si deux mutations indépendantes s'intersectaient ou non.

Comment savoir si les sous-unités du gène s'organisent selon un structure linéaire ? Seymour Benzer utilise la comparaison avec un enregistrement musical : si on copie cet enregistrement, des défauts plus ou moins longs peuvent apparaître, et une copie sans défaut peut être obtenue par comparaison de plusieurs copies à condition que les défauts ne s'intersectent pas (on n'est pas loin de la correction d'erreurs dans la transmission de codes (Dumas et al., 2007) !). Benzer imagine alors que si les sous-unités d'un gène s'arrangent linéairement comme les notes d'un morceau de musique, alors les mutations sont des intervalles et ce sera visible sur la structure de l'ensemble des intersections. En effet, certains jeux d'intersections sont interprétables comme des intersections entre des intervalles sur une droite, d'autres non : si A intersecte B et C mais pas D, B intersecte A et D mais pas C, C intersecte A et D mais pas B, et D intersecte B et C mais pas A, alors il n'existe aucun ensemble de quatre intervalles A B, C, et D qui vérifie ces relations. C'est, sans le nommer, imaginer le graphe d'intervalles. En effet, on peut représenter le jeu d'intersections par un graphe, dans lequel les sommets sont les mutations, et les arêtes relient les paires de mutations qui s'intersectent. C'est le principe du "graphe d'intersection". Si les mutations peuvent être représentées par des intervalles sur une droite, le graphe est un graphe d'intervalles. Tous les graphes ne sont pas dans ce cas : l'exemple précédent produit un cycle à quatre sommets, incompatible avec une telle représentation.

Seymour Benzer pose, sans utiliser le vocabulaire des graphes, le problème de la reconnaissance des graphes d'intervalles. Dans un langage d'algorithmicien, il s'énonce de la façon suivante : étant donné un graphe, existe-t-il un ensemble d'intervalles sur une droite, dont chaque intervalle correspond à un sommet, et dont deux intervalles s'intersectent quand il y a une arête entre les deux sommets correspondants ? Benzer imagine une solution fondée sur la construction d'une matrice binaire, la "matrice de recombinaison", qui, en termes modernes et en échangeant les 1 et les 0, est la matrice d'adjacence du

graphe. Il énonce une condition nécessaire et suffisante pour que sa structure puisse être représentée par des intervalles sur une droite : il existe une permutation des colonnes telle que les 0 après la position diagonale sont consécutifs. L'intuition est bonne, et le résultat, donné par Benzer sans démonstration, est maintenant bien connu des mathématiciens [1]. Il avoue que ce n'est pas encore une bonne solution à cause du grand nombre de permutations possibles, mais qu'elle est bien suffisante pour son propre exemple, puisqu'il arrive à trouver une permutation qui prouve que le graphe issu de ses expériences est bien un graphe d'intervalles, tendant à confirmer l'hypothèse d'une structure linéaire pour les gènes (ou plutôt, à ne pas l'infirmer).

Cette mention d'une solution qui, du point de vue théorique, ne semble pas satisfaisante, mais qu'il arrive à résoudre sur son exemple, n'est pas le moindre intérêt de l'article de Benzer. Elle illustre la nécessité théorique de définir ce qu'est une "bonne" méthode, une préoccupation algorithmique que l'on rencontre souvent dans les articles mathématiques avant les années 1960. En particulier elle est présente dans deux autre articles importants pour ce mémoire, ceux de Sturtevant & Novitski (1941) et de Cayley (1849), détaillés pages 10 et 52. Le mathématicien Jack Edmonds théorisera ce que veut dire une "bonne" solution dans les années 1960, à savoir un algorithme de complexité bornée par un polynôme de la taille des données, définition qui sera à la base de la théorie de la complexité, cadre dans lequel se résoudront par la suite les problèmes de Benzer, Sturtevant ou Cayley.

Fulkerson & Gross (1965) reprennent le problème de Benzer quelques années plus tard et inventent pour le résoudre la propriété des uns consécutifs [2], qui réapparaîtra souvent dans ce mémoire : un graphe est d'intervalles si et seulement si les colonnes de la matrice d'incidence de ses cliques maximales peuvent être ordonnées de façon à ce que sur chaque ligne, les 1 soient consécutifs.

Mais revenons à Benzer, qui ne s'est pas contenté d'un seul coup de pouce à l'algorithmique, puisque son article contient aussi la résolution, pour son exemple, d'une généralisation du problème de la reconnaissance des graphes d'intervalles. On dit souvent que la modélisation biologique par des problèmes

1. Mais la preuve la plus ancienne dont j'aie connaissance n'est que des années 70 et apparaît dans dans l'ouvrage "Graphs and Genes", de Mirkin & Rodin (1984).

2. Là encore, on peut discuter sur l'identité des premiers auteurs de cette invention, le problème apparaît en filigrane dans des études de datation en archéologie depuis la fin du XIXe siècle (Kendall, 1969). Sans reparler de la propriété de Nobert Wiener de 1914, dont ce résultat est une reformulation en termes matriciels.

de mathématiques discrètes s'accorde mal avec le caractère toujours bruité, partiellement erroné, souvent incomplet, des données issues du monde vivant. Benzer prouve le contraire. S'il a d'abord testé son hypothèse de linéarité sur 19 mutants, en effectuant tous les croisements deux à deux, il disposait en fait de 145 mutants, et le nombre de croisements à effectuer était trop important pour disposer d'une matrice complète. Certaines entrées sont inconnues, et pourraient être indifféremment 1 ou 0. Il résout alors, sans le poser formellement, ce qu'on appelle aujourd'hui le "problème du sandwich", et qui consiste à décider s'il existe une façon de compléter la matrice pour obtenir un arrangement linéaire des sous-unités. Ce problème sera réinventé tel quel dans les années 1990 à des fins, entre autres, de cartographie des gènes (Voir le chapitre 2, où je mentionne certaines généralisations du problème des uns consécutifs, qui permettent de modéliser des pertes de gènes, des données manquantes ou une part d'erreur dans les données).

Le fil qui relie l'article de Benzer au travail présenté ici n'est pas précisément biologique, la nature des questions et des données est différente, mais il est plutôt méthodologique. Le problème de la structure des gènes avait engendré des objets de mathématiques discrètes qui devaient, par la suite et après d'intensifs développements qui les éloignaient de leur discipline de naissance (la biologie), resservir à la génomique : les graphes d'intervalles et matrices aux uns consécutifs ont servi entre autres à cartographier ou assembler les chromosomes, et sont utilisés dans ce mémoire pour prédire leur passé.

Ce n'était pas la première fois que la biologie moléculaire inventait sans l'avoir prémédité, un domaine mathématique important. Plus tôt, Sturtevant & Novitski (1941) avaient, plus de quarante ans avant que des algorithmiciens ne s'en emparent, formalisé le problème de la transformation d'un arrangement de gènes en un autre au moyen d'inversions (problème qui sera développé amplement au chapitre 3). Il n'est pas certain que les développements récents sur l'algorithmique dans les graphes d'intervalles, bien que leurs domaines d'applications soient nombreux, aient été aussi utiles à la biologie que la découverte de Seymour Benzer ne l'a été pour les mathématiques discrètes. La bioinformatique computationnelle, qui pourrait parfois se revendiquer d'une "biologie appliquée aux mathématiques" aussi bien que de l'opposé, révèle des correspondances plus riches que la conception et l'utilisation d'outils nouveaux pour l'observation de la vie. C'est une science nourrie de plusieurs traditions, dont la rencontre est due à l'essor simultané des ordinateurs et des techniques de séquençage.

Dans ce mémoire, il est question à la fois de complexité algorithmique et d'évolution des génomes. Les deux y sont évidemment liées, mais pas inféodées. Certains résultats manqueront de formalisme pour certains, d'autres d'utilité pour d'autres. J'espère malgré tout les faire coexister en harmonie.

Le mémoire est organisé en quatre chapitres. Le premier délimite les objets dont il sera question, à savoir les mutations de la structure des génomes, responsables de l'évolution et de la diversité des organisations chromosomiques. Les deux suivants détaillent deux techniques différentes, dont j'ai participé à l'élaboration, pour décrire, modéliser et prédire cette évolution structurale. En l'absence –actuelle– d'une bonne modélisation, la cartographie, détaillée au chapitre 2, est une technique utile pour prédire les configurations ancestrales et les grands évènements évolutifs qui ont façonné les génomes de bactéries, animaux, plantes ou levures sur une échelle de temps de quelques centaines de millions d'années. Le troisième chapitre présente les tentatives pour modéliser l'évolution structurale, qui permettent une description plus quantitative, et une compréhension plus poussée de ses mécanismes, bien qu'un modèle largement utilisable soit toujours à construire. Enfin, la dernière partie présente des perspectives méthodologiques sur l'intégration de plusieurs échelles d'évènements évolutifs, depuis les mutations d'un nucléotide jusqu'à celles qui affectent des génomes complets, en passant par les histoires individuelles des gènes, et l'utilisation de tels modèles en phylogénie.

Chapitre 1

Les mutations structurales

1.1 Découverte

En étudiant les déviations de la seconde loi de Mendel, Thomas Morgan avait découvert que les gènes, jusque-là entités abstraites portant les mutations qui expliquaient les caractères mendéliens, étaient portés par des morceaux de chromosomes, et s'arrangeaient le long de ceux-ci en ordre linéaire. C'est en étudiant une déviation de cette loi de Morgan qu'un de ses étudiants, Alfred Sturtevant, découvrit un type de mutation particulière (Sturtevant, 1921), qui affectait non les gènes eux-mêmes, mais leur organisation : certains segments de chromosomes de drosophile étaient parfois arrangés dans un ordre dans les chromosomes de certaines souches, et dans l'ordre inverse dans d'autres. Cette prédiction purement calculatoire, simplement fondée sur des cartes génétiques et l'observation des taux de recombinaison, sera confirmée par l'observation des chromosomes polytènes par Dobzhansky & Sturtevant (1938). L'hybridation de différentes souches de drosophiles montrait des boucles dans les concaténats de chromosomes, la signature d'une inversion. Dobzhansky & Sturtevant (1938) utilisèrent tout de suite ces mutations pour comprendre l'histoire des populations de mouches étudiées, et construisirent dans la même étude une phylogénie moléculaire de 17 souches de drosophiles, en prédisant les relations de parenté et les arrangements de tous les génomes ancestraux.

Ce mécanisme de remaniement des chromosomes sera reconnu comme universel, affectant toutes les branches du vivant, et fera dire à François Jacob dans une conférence en 2000 : “Les chromosomes, ces structures naguère

encore considérées comme pratiquement intangibles, sont en réalité l'objet de remaniements permanents, la molécule de l'hérédité est raboutée, modifiée, coupée, rallongée, raccourcie, retournée". Malgré cela, les remaniements, comparés aux mutations ponctuelles, sont considérés comme des évènements rares. Par exemple, plusieurs millions de mutations ponctuelles séparent les génomes humains et chimpanzés, tandis que les gros remaniements, visibles au microscope, ne dépassent pas la dizaine.

1.2 Motivations

1.2.1 Phylogénie

Leur caractère rare devrait faire des réarrangements des bons marqueurs phylogénétiques. Mais bien que la phylogénie moléculaire ait commencé avec ces types de mutations, il existe peu de problèmes phylogénétiques finalement résolus de cette façon. Les raisons sont multiples.

Par exemple, les possibilités d'homoplasie sont assez mal connues. On trouve dans la phylogénie de primates de Dutrillaux et al. (1986) deux inversions partagées par le chimpanzé et le gorille, et pas par l'humain. Dutrillaux et al. (1986) conclut à la présence de ces inversions dans certains individus de l'espèce ancestrale et à un tri de lignée différent. Mais à regarder de plus près (avec les séquences), les points de cassures ne semblent pas correspondre (Goidts et al., 2005; Kehrer-Sawatzki et al., 2005) et des inversions proches et indépendantes sont finalement plus probables.

Ensuite, la structure des génomes ne permet pas de résoudre des problèmes phylogénétiques anciens, à cause de la saturation du signal. Bien que l'espace des configurations possibles soit immense et les évènements relativement rares, deux espèces éloignées (comme un primate et une drosophile par exemple) n'ont quasiment plus de couples de gènes adjacents en commun.

L'absence d'un bon modèle, enfin, rend le signal, même s'il existe, illisible si les mutations se sont accumulées. On trouve la trace de cette difficulté chez Sturtevant lui-même (Sturtevant & Novitski, 1941), qui peine à s'y retrouver quand il examine des génomes un peu éloignés avec seulement neuf marqueurs : "... for each such sequence there was determined the minimum number of successive inversions required to reduce it to the ordinal sequence chosen as "standard". For numbers of loci above nine the determination of this minimum number proved too laborious, and too uncertain, to be carried

out...". Aujourd'hui, ce problème trouve des solutions pour un faible nombre de marqueurs (c'est le sujet du chapitre 3), mais leurs possibilités phylogénétiques sont encore restreintes. Par exemple, une méthode basée sur les réarrangements donne facilement les primates et carnivores dans un même clade, à l'exception des rongeurs (Alekseyev & Pevzner, 2009), ce qui est contraire à la plupart des études basées sur la morphologie ou les séquences. Et malgré la découverte plus récente des mutations ponctuelles sur les séquences, les modèles y sont beaucoup plus nombreux, mieux étudiés, leurs défauts mieux analysés. Il y a donc toutes les raisons de leur faire plus confiance.

La structure des génomes est tout de même porteuse d'une partie de leur histoire qu'il est possible d'exploiter dans un cadre phylogénétique. Par exemple, elle prédit l'orthologie entre deux gènes de façon plus sûre que leur séquence (voir la section 1.3). Prendre en compte l'organisation des gènes sur le génome en plus de la séquence des gènes est un des challenges de ce domaine. C'est le sujet du chapitre 4, qui présente des perspectives plus que des résultats.

1.2.2 Évolution

Les mutations structurales portent également des informations sur les modes de spéciations. Une inversion peut par exemple interrompre la recombinaison entre les chromosomes porteurs d'une configuration et les autres. Cet arrêt de recombinaison entraîne alors une divergence différentielle entre des parties différentes des génomes. Si on constate une corrélation entre divergence en séquence entre deux espèces et présence d'un réarrangement, il est probable que le réarrangement ait joué un rôle particulier dans la spéciation. Plusieurs cas ont été proposés dans des espèces de drosophiles (Rieseberg, 2001), mais la fréquence des observations répondant à ce mécanisme reste controversée (Coghlan et al., 2005).

L'arrêt de la recombinaison ne favorise pas que la divergence des espèces, mais aussi celle des sexes au sein d'une espèce. L'étude des scénarios de remaniements chromosomiques permet alors aussi de suivre la différentiation progressive des chromosomes sexuels (Lemaitre et al., 2009a), et de soutenir le rôle important des inversions dans l'arrêt de la recombinaison.

Les grands remaniements sont aussi très présents dans les génomes des cellules tumorales. La plupart des conséquences fonctionnelles décrites des réarrangements sont des maladies. Quelques tentatives d'application de principes issus de leur modélisation ont été reportées (Raphael et al., 2003; Ra-

phael & Pevzner, 2004; Ozery-Flato & Shamir, 2008), peu nombreuses pour l'instant probablement par manque de séquences. Ce manque devrait disparaître dans les années à venir, et probablement de nouveaux problèmes méthodologiques surgiront de ce type de données.

Mais il est possible qu'un grand nombre de remaniements soient neutres, bien que certaines associations entre réarrangement et phénotype aient été observées. Par exemple, un criquet d'Australie montre des changements de taille et de forme en fonction de la présence de deux inversions (Coghlan et al., 2005).

1.2.3 Le caryotype, moteur de l'évolution des génomes

La structure des génomes peut aussi expliquer en partie l'évolution des séquences : chaque méiose est accompagnée d'au moins un évènement de recombinaison par bras chromosomique, et le nombre d'évènements supplémentaires dépend de la taille des chromosomes. En conséquence, le taux de recombinaison est plus important dans les petits chromosomes que dans les grands. La recombinaison étant une force motrice de l'évolution des séquences en partie en raison de la conversion génique biaisée (la recombinaison suppose un appariement entre allèles, et en cas de différence, la conversion de l'un par l'autre se fait préférentiellement en faveur de bases G ou C, du moins chez les vertébrés), il est possible que les modifications du caryotype aient des conséquences sur la composition des séquences. On s'attend à ce qu'un petit chromosome s'enrichisse en bases G+C plus qu'un grand.

Par exemple, on voit sur la figure 1.1 (a) une bonne corrélation entre le taux de GC moyen des gènes sur un chromosome avec la taille dudit chromosome chez le poulet. Les oiseaux sont connus pour avoir un caryotype très stable, ce qui explique que cette corrélation soit plus visible que pour d'autres espèces. Et le caryotype est stable au point que le taux de GC moyen des gènes humains corrèle aussi avec la taille des chromosomes de poulet qui portent leurs orthologues [1] (figure 1.1 (b)). Ce qui veut dire que le taux de GC humain porte la trace de la configuration des chromosomes de l'ancêtre commun de l'humain et du poulet (proto-amniote).

On peut donc exploiter cette corrélation dans le cadre d'une reconstruction de chromosomes ancestraux. Par exemple, on arrive bien à recons-

1. Cette remarque est due à Laurent Duret, envers qui je m'excuse de ne pas avoir pu pousser ce travail aussi loin qu'il l'aurait mérité.

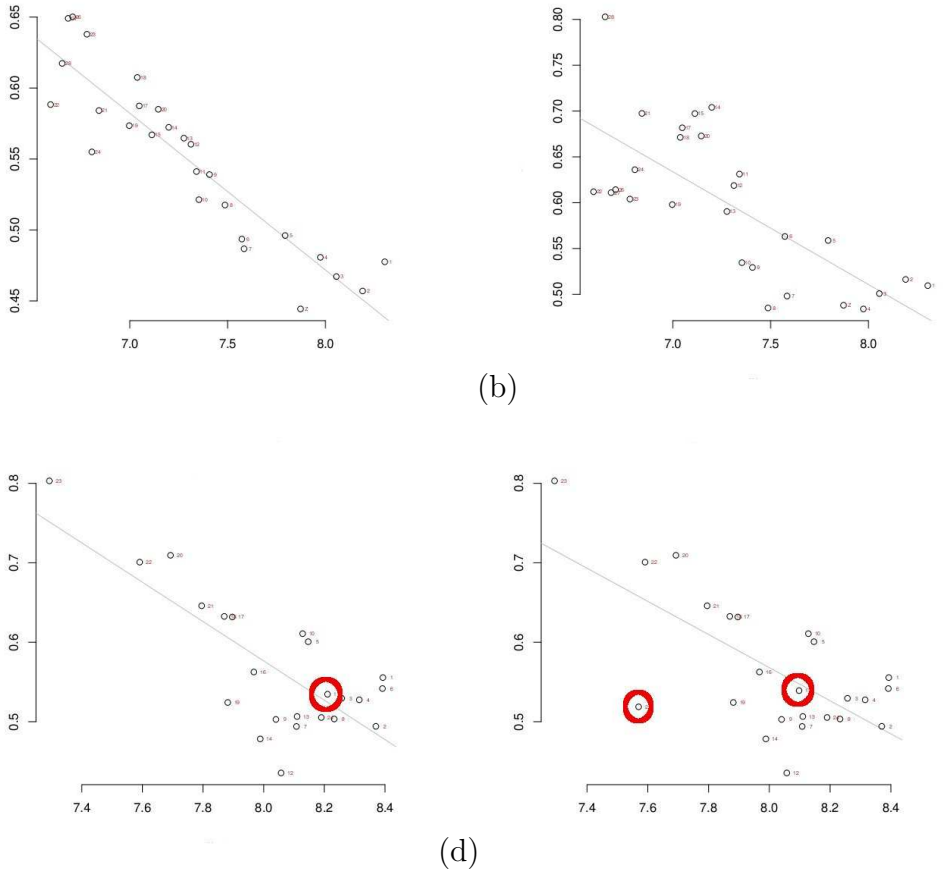


FIGURE 1.1 – Chaque point est un chromosome : (a) et (b) de poulet (c) et (d) de l'ancêtre boréoeuthérien selon la configuration de Wienberg (2004) (c) ou Richard et al. (2003) (d). En abscisse (échelle logarithmique), la taille des chromosomes (estimée à partir des chromosomes humains pour l'ancêtre), et en ordonnée, le taux moyen de bases G+C en troisième position des codons des gènes de poulet (a), des orthologues humains (b), (c), (d). Entourés en rouge, les points différents entre les deux hypothèses (c) et (d).

truire le caryotype ancestral des mammifères boréoeuthériens (qui incluent les primates, rongeurs, carnivores, férongulés, et qui excluent afrothériens, xénarthres, marsupiaux, pour les méthodes, voir le chapitre 2). Seules certaines petites incertitudes demeurent, comme la présence d'un ou de deux chromosomes dans l'ancêtre contenant les parties homologues à des chromosomes 12, 22 et 10 humains. Par exemple, Richard et al. (2003) les suppose en deux morceaux, tandis que Wienberg (2004) n'en voit qu'un (figure 1.2). Il

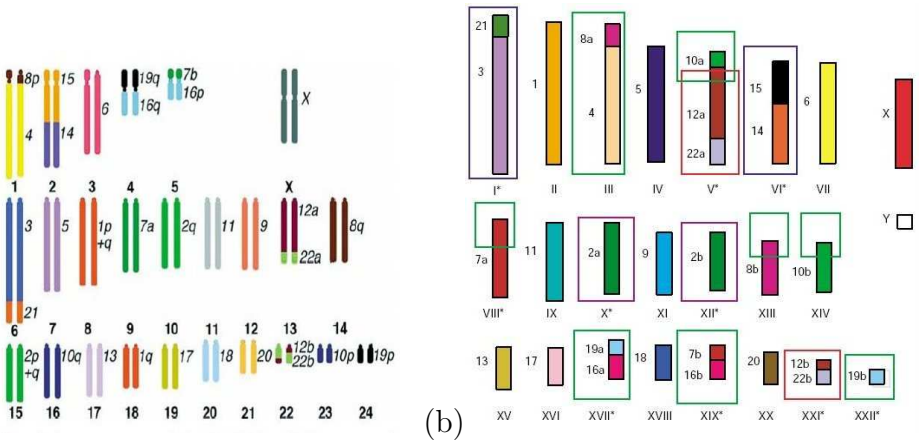


FIGURE 1.2 – Deux configurations alternatives pour le caryotype du proto-boreoeuthérien. (a) Celle de Richard et al. (2003), et (b) celle de Wienberg (2004) (la figure est une reproduction de celles des articles cités). L'une infère 24 chromosomes, l'autre 22, et une différence est la fusion, ou non, entre des morceaux 12-22 et 10, les numéros faisant référence aux chromosomes humains actuels, et le problème est toujours ouvert. L'autre différence, le chromosome 1 en une ou deux pièces, a été tranchée par la suite en faveur de l'unité ancestrale.

est difficile de discriminer ces deux hypothèses par des critères de parcimonie, les deux nécessitant le recours à de l'évolution convergente. Mais la corrélation avec le GC donne une préférence à la configuration de Wienberg (2004), comme le montre la figure 1.1 (c) et (d). Les graphes (c) et (d) ne diffèrent que par le regroupement de deux points (c) en un seul (d) (les points entourés en rouge), conséquence de la fusion de deux morceaux de chromosomes en un seul, selon les hypothèses. Dans l'hypothèse (d), un point est éloigné de la corrélation, c'est à dire qu'il porte un GC trop faible pour sa taille. On peut donc préférer son intégration à un chromosome plus grand (c), ce qui donne

un argument en faveur de l'hypothèse Wienberg (2004).

L'étude de l'évolution structurale peux donc donner des informations importantes sur les mutations ponctuelles, et vice-versa.

1.3 Définitions

Il n'est pas toujours facile de discriminer les types de mutations. Les remaniements ont d'abord été implicitement définis par la technique qui permettait de les repérer : hybridation ou bandes chromosomiques, et observation au microscope. Aujourd'hui, alors que nous sommes en présence des séquences génomiques, les limites sont devenues plus floues. Les inversions sont considérées comme des mutations structurales, mais il peut exister des inversions de toutes tailles, depuis quelques nucléotides jusqu'à la quasi intégralité des chromosomes.

Une autre technique permet de redéfinir les mutations structurales, par contraste cette fois : ce sont celles qui ne sont pas modélisées par les algorithmes d'alignement de séquence. Les séquences "qui s'alignent" sont supposées dépourvues de réarrangements. Définition toujours dépendante des algorithmes, de leurs paramètres et de la puissance des ordinateurs, comme la précédente était sans doute dépendante de la puissance des microscopes. Mais définition opérationnelle. Certains algorithmes d'alignement commencent à inclure les inversions (Schoeniger & Waterman, 1992; Vellozo et al., 2006; Ledergerber & Dessimoz, 2008), mais leurs possibilités ne sont pas encore assez étendues pour prétendre servir de critère discriminant entre types de mutations.

La plupart des études actuelles sur l'organisation des génomes règlent cette question de définition en choisissant des *unités structurales*, marqueurs génomiques supposés dépourvus des mutations à étudier, et tels que leur organisation sur le génome s'explique exclusivement par ces mutations. Ce choix n'est pas anodin, par exemple lorsqu'il s'agit ensuite de présenter des statistiques sur certaines mutations sans avoir entièrement démontré ces deux présupposés (voir par exemple la controverse sur l'évolution des chromosomes de mammifères dans Pevzner & Tesler (2003); Sankoff & Trinh (2005); Sankoff (2006), dont les arguments sont développés dans la section 3.3.).

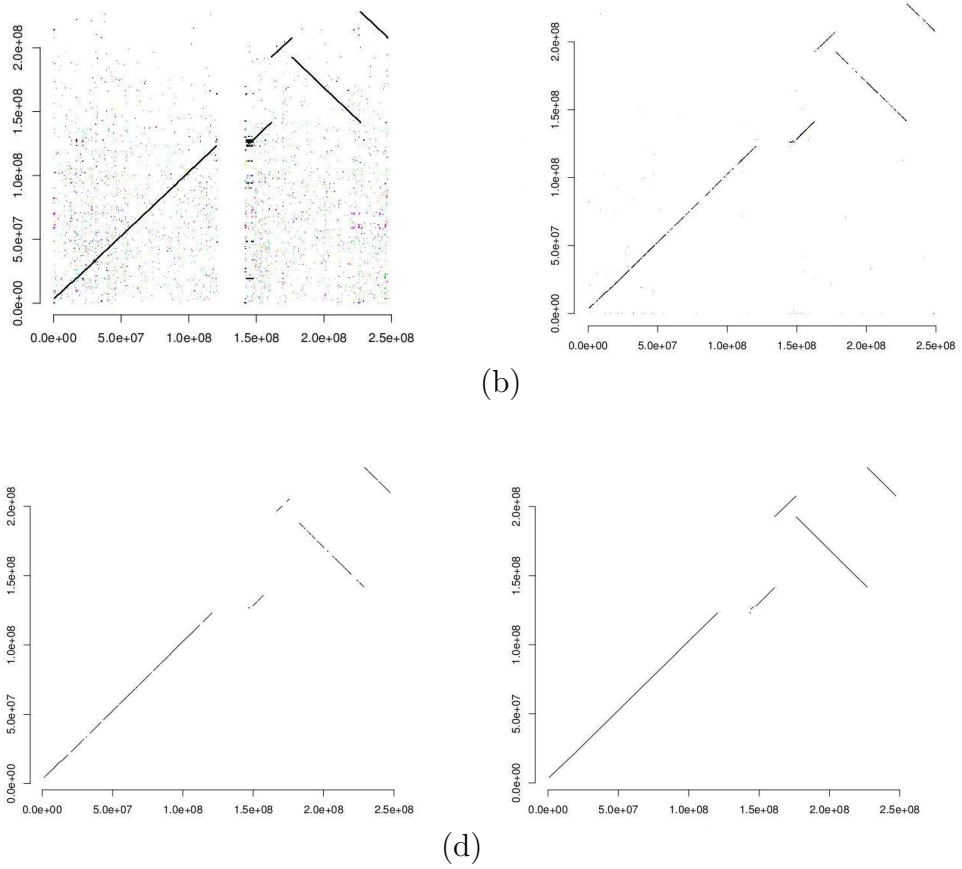
Le choix du gène ou de l'exon comme unité structurale présente l'avantage de pouvoir détecter des homologies de manière plus fine qu'avec une simple séquence d'ADN, en utilisant la similarité des séquences protéiques.

Des études sont possibles à une grande échelle évolutive. L'inconvénient est que les gènes ou exons ne couvrent qu'une petite proportion de certains génomes, et qu'on ne peut espérer retracer l'histoire complète de tous les génomes en n'utilisant que les parties codantes. D'autre part, utiliser les gènes implique de dépendre des méthodes de détection des homologies, de regroupement des gènes en familles d'homologues, et d'assignation des relations d'orthologies et de paralogies entre gènes au sein d'une famille. La marge d'erreur cumulée pour la succession de ces tâches n'est pas négligeable. Par exemple Muffato et al. (2010), pour étudier la structure des génomes ancestraux, éliminent deux tiers des familles de la base Ensembl-Compara sur des critères simples de confiance accordée dans les familles ou les phylogénies. Le chapitre 4 contient des réflexions prospectives sur l'amélioration possible des familles et des phylogénies, notamment avec des données d'organisation des génomes.

Plutôt que des unités fonctionnelles, les gènes, on peut construire des unités évolutives en recherchant les éléments conservés dans les génomes qu'on examine. Les gènes vont souvent être inclus parce qu'une unité fonctionnelle est souvent conservée, mais d'autre éléments peuvent être intégrés. Des alignements génomiques couvrant la quasi-totalité des génomes et un nombre important d'espèces existent pour certains clades (Rhead et al., 2010). On peut donc les utiliser comme unités structurales. L'inconvénient de ces alignements est qu'ils sont fondés sur la similarité de séquences parfois très petites, et que cette similarité peut être due à des paralogies autant qu'à des orthologies, ce qui brouille l'étude des mutations responsables de leur organisation.

Par exemple, la figure 1.3 (a) montre les alignements issus de l'UCSC (Rhead et al., 2010) qui ont une occurrence sur le chromosome 1 humain, et une occurrence dans le génome du macaque. L'orthologie avec le chromosome 1 du macaque est bien visible, mais le nuage de points qui ne correspond pas au chromosome 1 pointe probablement des paralogies ou des similarités dues au hasard, et témoigne des nombreuses répétitions ou régions de faibles complexité présentes dans les génomes de primates. La figure 1.3 (b) montre la même information mais issue de paires de gènes orthologues selon la base Ensembl-Compara. Les grandes lignes sont les mêmes et le bruit est nettement diminué, mais les erreurs restent nombreuses et la couverture des génomes un peu morcelée.

Pour détecter les parties orthologues d'un jeu d'alignements ou de gènes, on construit habituellement des "blocs de synténie". Cette appellation re-



(a) (b)

(c) (d)

FIGURE 1.3 – Pour les quatre graphes, l'ordonnée est le chromosome 1 humain, et l'abscisse tous les chromosomes du macaque superposés. Un point représente une similarité, détectée selon quatre techniques différentes. Les couleurs représentent des chromosomes différents du macaque (le noir est le chromosome 1). Sont représentés en (a) les alignements issus de l'UCSC, en (b) les gènes orthologues selon Ensembl-Compara, en (c) les blocs orthologues selon Mercator, en (d) les blocs orthologues construit par une méthode de détection de composantes connexes à partir des alignements de l'UCSC.

couvre une multitude de définitions et de méthodes de constructions. Les blocs de synténie sont souvent des segments orthologues aux extrémités mal définies. L'élimination des paralogues et erreurs supposés impose l'élimination des blocs trop petits. Par exemple, ne garder que les blocs de plus de 100kb permet de se passer de l'étude de la plupart des duplications chez les mammifères (les duplications segmentaires récentes détectées chez les primates ou les rongeurs dépassent très rarement 100kb (Bailey et al., 2002)). Mais on entre là dans le domaine de l'arbitraire, de l'empirique et du compromis. En faisant augmenter ce chiffre, on se prive d'une bonne partie des génomes, et en le faisant baisser, on intègre une plus grande diversité de mutations. Sur la figure 1.3 (c), une méthode de détection de l'orthologie part des similarités entre régions codantes comme dans la partie (b) et intègre une information de synténie (Paten et al., 2008). En (d), j'ai pris les alignements de l'UCSC de la partie (a) en entrée, et un algorithme relie toutes les paires d'alignements distants de moins de 300kb dans les deux espèces, et élimine les composantes connexes du graphe obtenu qui concernent moins de 300kb dans l'une des deux espèces. Ainsi, les duplications segmentaires et régions de faible complexité sont gommées, toutes concernant des résolutions inférieures à 300kb. On voit apparaître, débarrassées de tout le bruit, les trois inversions qui séparent les chromosomes 1 de l'humain et du macaque. On ne voit plus la zone fortement impliquée dans des duplications qui suit le centromère humain. En faisant varier les seuils, on peut donc se concentrer sur différents types et tailles de mutations.

Gènes, alignements, blocs, quel que soit le type d'unité employé, un élément sera génériquement désigné sous le terme de *marqueur*. L'histoire des génomes est alors divisée en trois : l'histoire interne du marqueur, faite de substitutions, et de petites mutations responsables d'indels internes ; l'histoire individuelle du marqueur dans le génome ensuite, marquée par des évènements d'insertions, duplications, pertes ou transferts de marqueurs entiers. Et enfin les évènements qui modifient l'organisation des marqueurs dans un génome, inversions, translocations, fusions, fissions, et autres remaniements. Les frontières entre ces trois types d'histoire ne sont pas étanches, et il peut par exemple arriver que plusieurs marqueurs soient perdus, dupliqués ou transférés ensemble, mais en l'état actuel de la recherche, elles sont le plus souvent étudiées séparément (voir pourtant le chapitre 4).

1.4 La paléogénomique

Une fois les termes plus ou moins définis et les unités choisies et détectées, il s'agit de prédire le passé des génomes actuels. J'emprunte le titre de cette section à Muffato & Roest-Crollius (2008), qui est employé dans une acception différente de celle qui est peut-être la plus courante, à savoir le séquençage de génomes anciens. Je l'utilise ici dans un cadre un peu plus général que Muffato & Roest-Crollius (2008) : celui de la prédiction computationnelle de la structure des génomes ancestraux, quelle que soit la méthode employée. Comme aucune molécule d'ADN n'est conservée au-delà de quelques centaines de milliers d'années, tous les génomes d'un passé plus lointain ne sont pour l'instant accessibles que de cette façon.

J'inclus donc la possibilité d'écrire un modèle d'évolution structurale par remaniements et de construire une séquence ancestrale qui soit la plus vraisemblable selon ce modèle. La description de ce mode de calcul est décrit dans le chapitre 3. C'est l'idée la plus ancienne (Sankoff, 1989), et sans doute aussi le mode de prédiction qui a le plus d'avenir, mais pourtant pas forcément la méthode la plus efficace à l'heure actuelle. En effet, plus les mutations impliquent une quantité importante d'ADN (et par là plus elles modifient la structure des génomes), plus rarement elles sont observées, et moins les lois de leurs apparitions sont connues. Si pour les substitutions, on sait estimer le taux de mutation, le degré de polymorphisme et un bon nombre de modèles existe pour décrire leur rôle dans l'évolution des génomes, rien de tout cela n'est acquis pour les autres mutations, insertions, délétions, duplications, et encore moins pour les grands remaniements. Même à une grande échelle évolutive, où les remaniements sont constatés en nombre important, il reste difficile de modéliser leurs effets sur la structure des génomes. L'espace des structures possibles est quasiment infini, comparé à l'espace des bases possibles pour un site donné, seulement de taille 4. C'est pourquoi la modélisation est encore balbutiante, et les résultats qui font état de taux de mutations pour des variants structuraux commencent à peine à voir le jour. Les résultats qui ont une vraie base théorique sont souvent limités à des comparaisons de deux génomes, ou d'un faible nombre de marqueurs.

Ce manque actuel de bons modèles et de bon algorithmes pour l'optimisation sur ces modèles ont ouvert la voie à une alternative moins ambitieuse mais pour l'instant plus efficace : la cartographie, décrite au chapitre 2. Le principe est d'oublier les processus de modification de l'organisation des génomes –les réarrangements–, et de se contenter de décrire l'évolution de

caractères, comme la proximité de deux marqueurs, ou plus généralement le regroupement de plusieurs marqueurs sur une même région génomique. Cette approche permet de reconstruire efficacement des génomes ancestraux à partir de tous les génomes disponibles.

Ces deux pistes –cartographie, modélisation– sont *a priori* disjointes et ont des objectifs différents : la première produit des résultats avec les données disponibles, tandis que la seconde est plus prospective. Mais la combinaison des deux est intéressante : on peut ainsi, sur un arbre phylogénétique, reconstruire un génome ancestral à chaque noeud par des méthodes de cartographie et étudier l'évolution sur chaque branche en ne comparant que deux génomes. Un exemple de résultat ainsi obtenu sera mentionné à la fin du chapitre 3.

Chapitre 2

La cartographie des génomes ancestraux

2.1 Principes de la cartographie des génomes

Les techniques de cartographie génétique ou physique, employées jadis par Alfred Sturtevant ou Simon Benzer, ont depuis été perfectionnées et largement utilisées pour connaître l'organisation des génomes actuels. J'essaie d'en retirer quelques principes pour ensuite les recycler dans la construction de génomes ancestraux.

La cartographie consiste à collecter des observations donnant une indication de proximité entre marqueurs qu'on sait présents dans un génome, puis à proposer une organisation de ce génome qui, le mieux possible, colle à toutes les observations. Les deux étapes, qui font appel à des principes méthodologiques disjoints, peuvent être implémentées de multiples façons, adaptées au type de données, au type d'information recueillie dans la première étape, et aux types de problèmes calculatoires soulevés.

Les premières cartes génétiques ont été réalisées par l'équipe de Thomas Morgan au début du siècle (Sturtevant, 1913). Les marqueurs étaient alors des caractères phénotypiques mendéliens, dont l'association par paires violaient la loi d'indépendance énoncée par Mendel. Une quantité de recombinaison peut être estimée à partir de la fréquence d'association de deux caractères, cette quantité étant supposée être proportionnelle –avec des variations dues aux taux de recombinaison– à la distance qui sépare, sur le chromosome, les gènes responsables de la variation de ces caractères. Un espace métrique peut

donc être construit sur l'ensemble des marqueurs, et le caractère approximativement additif de la distance permet d'organiser les marqueurs en "groupes de liaisons" linéaires, compatible avec la structure des chromosomes. La première étape est donc une estimation de distances à partir de l'expérience, et la seconde le plongement de la métrique obtenue dans un espace linéaire [1].

Il existe aujourd'hui divers moyens de collecter des informations sur l'organisation de marqueurs dans un génome. Par exemple, des techniques d'hybridation permettent, étant donnés plusieurs découpages du génome en segments, de savoir si deux marqueurs sont souvent détectés sur le même segment, ou non. On peut estimer une distance –physique, cette fois– entre ces deux marqueurs à partir de la fréquence de leur apparition commune. Le plongement de la métrique dans un espace linéaire s'effectue souvent en construisant un graphe dans lequel les sommets sont les marqueurs, et les arêtes pondérées reflètent la distance calculée. Puis en résolvant un problème proche de celui du voyageur de commerce, en trouvant un chemin ou un ensemble de chemins qui minimise la somme des distances empruntées (Servin et al., 2010) [2].

Une autre expérience consiste à recueillir par hybridation, pour un ensemble de segments et un ensemble de marqueurs, une information de présence ou absence de chaque marqueur sur chaque segment (Karp, 1993). Pour linéariser ce type d'information, on construit un autre objet combinatoire : une matrice binaire dont les colonnes sont les marqueurs et les lignes les segments. Une entrée vaut 1 si le marqueur a été détecté sur le segment, et 0 sinon. Le problème consiste cette fois à trouver un arrangement des colonnes tel que sur chaque ligne, les entrées 1 soient consécutives, pour retrouver les segments. Il est connu sous le nom de "problème des uns consécutifs" : dans une matrice, un *mauvais zéro* est une entrée i, j qui vaut 0, telle que sur la même ligne i, il existe deux entrées i, j_1 et i, j_2 valant 1,telles que $j_1 < j < j_2$. Deux matrices qui ne diffèrent que par une permutation de leurs colonnes sont

1. Avec la contrainte supplémentaire à cette époque qui était le débat scientifique encore ouvert sur la structure qui organisait les gènes. Il fallait donc en plus prouver que la métrique était bien compatible avec une structure linéaire, ce qui n'allait pas de soi (Sturtevant et al., 1919). Aujourd'hui on s'affranchit de cette contrainte, en supposant connue la structure des chromosomes, mais si les tests étaient réalisés, ils révéleraient peut-être des surprises sur la façon dont on collecte des informations d'organisation.

2. En écho à la note précédente, une méthode de voyageur de commerce ne garantit pas le meilleur plongement d'une métrique dans une droite. Il se peut que les extrémités d'un chemin appartenant à la solution soit très peu distantes dans le graphe.

dites *équivalentes*. Une matrice a la *propriété des uns consécutifs* si elle est équivalente à une matrice sans mauvais zéro. Malgré l'organisation linéaire des chromosomes, la plupart des matrices issues de données biologiques n'ont pas cette propriété, à cause des erreurs ou artefacts de manipulations. Dans ce cas, on utilise souvent des problèmes d'optimisation comme celui de chercher la plus grande sous-matrice qui a la propriété des uns consécutifs. C'est une forme de généralisation du voyageur de commerce.

La cartographie est donc un double problème : le premier, biologique, est d'imaginer une expérience qui puisse collecter efficacement de l'information sur l'organisation d'un génome ou d'une partie d'un génome. C'est le problème de la *collecte*. Le second est mathématique, il consiste à assembler l'information pour proposer une organisation qui prenne en compte toute l'information. C'est le problème de l'*arrangement* des marqueurs. On retrouve ce schéma dans le travail de Simon Benzer décrit en introduction : une expérience permettait de savoir si deux mutations d'un gène de virus s'intersectaient ; une analyse mathématique permettait d'assembler cette information en une structure linéaire, par le moyen d'un graphe d'intervalles cette fois. Certains problèmes mathématiques posés par l'arrangement de marqueurs, plus nombreux et riches que ne le laisse entrevoir ce trop rapide résumé, ont été formalisés par Karp (1993), qui lançait à l'époque un programme de recherche en combinatoire pour la cartographie physique. Pour les génomes ancestraux, il est toujours d'actualité.

2.2 Cartographie ancestrale

Les cartes génomiques ont permis, à l'ère du séquençage massif, de guider l'assemblage des génomes, parfois séquencés avec une faible couverture. Les principes de la cartographie ont aussi été à la base de certaines techniques d'assemblage.

Mais la recherche sur la cartographie a également connu une nouvelle jeunesse dans l'étude de l'organisation des génomes ancestraux. Comme les similarités entre organismes peuvent nous renseigner sur l'homologie de certains caractères, c'est-à-dire leur présence chez un ancêtre commun, on peut recueillir des informations sur la séquence d'un ancêtre en observant les similarités entre descendants et/ou apparentés. Par exemple, un caractère peut être la proximité de deux ou plusieurs gènes. En constatant la présence ou absence d'un tel caractère dans les génomes actuels, on peut modéliser son

évolution sans se prononcer sur les évènements évolutifs qui expliquent cette présence ou absence. Et en inférant un tel caractère dans un génome ancestral, on peut collecter des informations de synténie et s'en servir pour faire de la cartographie.

La technique en deux étapes est la même que pour la cartographie des génomes actuels : on recueille tout d'abord des informations partielles sur le contenu en marqueurs et leur organisation dans un génome ancestral en comparant les organisations actuelles de ses descendants ou apparentés, puis on rassemble ces informations pour obtenir un arrangement ancestral de marqueurs. La collecte, faite principalement d'expérimentation biologique dans le cas de la cartographie des génomes actuels, s'est muée pour les génomes ancestraux en un problème de bioinformatique, et les spécificités des données qui sont produites de cette façon imposent de nouvelles façons de les arranger.

Je présente ici plusieurs façons d'implémenter ce principe de la cartographie ancestrale, en montrant plus ou moins formellement comment on peut choisir des marqueurs ancestraux, collecter des relations de synténie entre les marqueurs, et enfin arranger les marqueurs selon les relations collectées. C'est une sorte de revue des techniques utilisées, ou d'idées pour les pousser plus loin. L'intérêt de cette énumération est de faire le lien entre plusieurs études indépendantes. Dans les contributions auxquelles j'ai participé (Chauve & Tannier, 2008; Tannier, 2009; Ouangraoua et al., 2009; Chauve et al., 2010; Murat et al., 2010; Gavranović et al., 2011), principalement issues d'une collaboration avec Cedric Chauve, et dont les résultats sont décrits à la section 2.3, nous avons essayé d'utiliser dans un même cadre méthodologique tous les principes présentés, afin de recueillir le plus de signal possible tout en conservant une bonne spécificité, et de l'assembler avec le plus de précaution possible.

2.2.1 Marqueurs ancestraux

Donner une structure à des marqueurs ancestraux suppose d'abord de les définir et de les trouver. La définition d'une mutation structurale, en partie arbitraire ou dépendante des techniques, biologiques ou bioinformatique, comme on l'a vu à la section 1.3, définit un découpage des génomes actuels en segments (marqueurs, gènes) qui, sur l'échelle de temps étudiée, n'ont pas subi de telles mutations.

Ces marqueurs sont reliés par des relations d'homologie. Un ensemble de

marqueurs homologues est une *famille*.

Pour inférer un contenu en marqueurs dans un organisme ancestral, on peut se baser soit sur un comptage des marqueurs par famille dans chaque espèce Csurös & Miklós (2009), soit sur des phylogénies qui décrivent les relations de parenté au sein d'une famille. Dans tous les cas, on suppose connue la phylogénie des organismes ou des espèces (elle est nécessaire ne serait-ce que pour bien identifier les espèces ancestrales qui sont supposées porter les génomes qui seront reconstruits). Dans tout le document, j'emploie indistinctement les termes phylogénie des organismes, des espèces, arbre des espèces.

L'utilisation des phylogénies des marqueurs passe par la *réconciliation* des arbres phylogénétiques, c'est-à-dire l'explication des noeuds d'un arbre d'une famille de marqueurs par des évènements datés dans la phylogénie des organismes. Ces évènements sont la spéciation, la duplication, l'apparition, la perte, ou le transfert d'un marqueur. La figure 2.1 montre un exemple de réconciliation d'un arbre d'une famille de gènes.

La réconciliation qui minimise le nombre d'évènements est facile à calculer en l'absence de transferts (Zmasek & Eddy, 2001), un peu plus complexe avec des transferts (Hallett et al., 2004; Doyon et al., 2010). Des modèles d'évolution de gènes avec ces évènements ont été décrits par Arvestad et al. (2003); Sennblad et al. (2007); Akerborg et al. (2009); Tofigh et al. (2011), et permettent d'établir un contenu ancestral en gène, ou une distribution de contenus possibles, assortis de leur probabilités.

Dans la plupart des reconstructions actuelles de génomes ancestraux, c'est la version parcimonieuse qui est choisie, permettant de disposer d'une information binaire sur la présence ou l'absence d'un marqueur dans un génome ancestral (Ma et al., 2008; Bertrand et al., 2010; Muffato et al., 2010; Gavranović et al., 2011). Ou même un cas trivial de la réconciliation, où tous les marqueurs sont présents exactement une fois dans tous les génomes étudiés (Ma et al., 2006; Chauve & Tannier, 2008; Chauve et al., 2010), ce qui met duplications, pertes ou transferts à l'écart, mais permet sur des clades restreints d'étudier des jeux de données avec une couverture correcte des génomes.

2.2.2 La synténie ancestrale

La *synténie*, dans un sens étymologique, historique et rigoureux, est la relation qui unit deux segments génomiques portés par le même chromo-

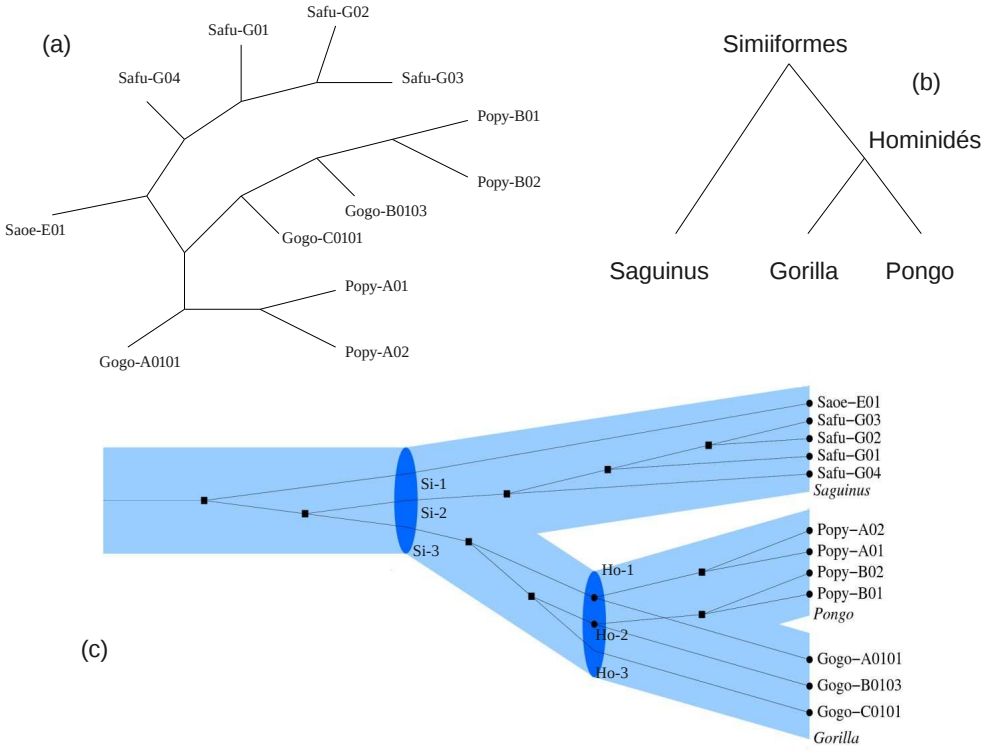


FIGURE 2.1 – (a) Phylogénie d'une famille de 12 gènes, appartenant à trois espèces, dont la phylogénie est représentée en (b). (c) La réconciliation de l'arbre des gènes : chaque noeud de l'arbre trouve une position dans l'arbre des espèces, et des évènements sont assignés à tous les noeuds : duplications (carrés) ou spéciations (ronds). Le contenu d'un génome ancestral est représenté par les ellipses en bleu foncé : les deux ancêtres de la phylogénie des espèces (le proto-simiiforme et le proto-hominidé) contiennent trois gènes selon la méthode de parcimonie (les Si-x et Ho-x). Cette partie de la figure est tirée de Sennblad et al. (2007).

some. Le titre de cette section déroge déjà à cette définition, puisqu'il sera question d'autres types de relations entre marqueurs génomiques (adjacence ou proximité sur un chromosome), qui impliquent la synténie mais sont plus restrictives, et qu'on regroupe souvent sous le terme de synténie.

Différentes formes de synténie

Ces définitions sont valables pour des marqueurs actuels ou ancestraux, et sont illustrées par la figure 2.2. La synténie, dans son sens générique, est une relation qui unit un ensemble de marqueurs génomiques selon des informations de positions. Elle peut se décliner en plusieurs définitions plus précises :

- La *synténie*, dans son sens originel, c'est à dire l'appartenance à un même chromosome.
- L'*adjacence*, c'est à dire l'immédiate proximité de deux marqueurs (l'absence d'autres marqueurs entre les deux).
- L'*appartenance à un intervalle* pour un groupe de marqueurs, c'est à dire l'absence (à considérer avec plus ou moins de rigueur) de marqueurs qui ne font pas partie du groupe, dans l'intervalle entre le premier et le dernier marqueur sur le génome. Pour deux marqueurs, l'appartenance à un intervalle est une adjacence.
- La *distance* entre deux marqueurs peut aussi, par extension, être appelée un signal synténique, et donc entrer dans cette collection de synténies. Une distance finie correspond à une appartenance à un même chromosome.

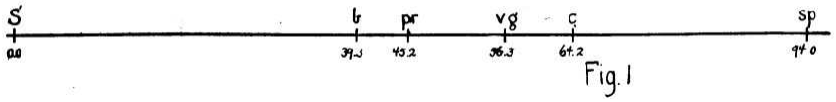


FIGURE 2.2 – Adjacence, intervalle, synténie, distances illustrées par une des premières cartes génétiques, ordonnant des marqueurs sur le chromosome 2 d'une drosophile, figure recopiée à partir de l'article de Sturtevant (1917). Ici, les marqueurs (qui déterminent la couleur et la forme des ailes des drosophiles) b et pr sont adjacents, tandis que pr, vg, c forment un intervalle, et S' et c sont en synténie, et sont distants de 64.2, dans une unité qui mesure le nombre de crossing-over observés entre les marqueurs par rapport au nombre d'observations.

Détection d'une synténie ancestrale

Étant donnés les différents types de synténie définis, et un jeu de marqueurs ancestraux, il existe plusieurs principes de détection d'une synténie dans un génome ancestral. C'est ici que la cartographie d'un génome actuel ou d'un génome ancestral se séparent. Les différents types de synténie sont illustrés par la figure 2.3.

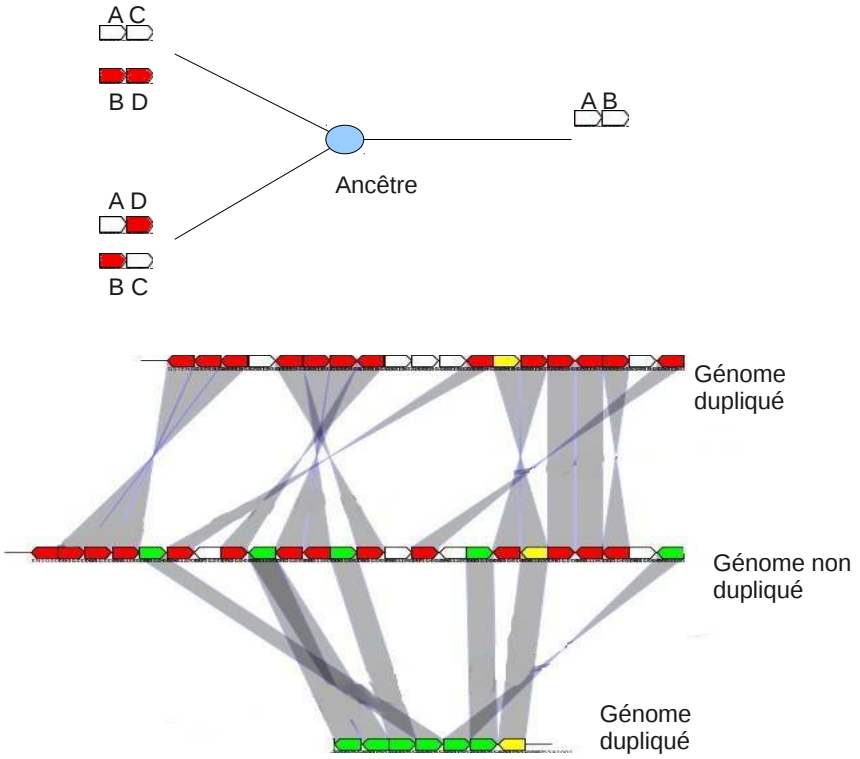


FIGURE 2.3 – Détection de synténies ancestrales : en haut, AB est une adjacence soutenue et BD est une adjacence fiable pour l'ancêtre figurant sur le noeud interne de la phylogénie. En bas, le segment du centre, appartenant à un génome non dupliqué, et les deux segments l'encadrant, appartenant tous les deux à un génome dupliqué (depuis la spéciation entre les deux), forment une double synténie.

Synténie soutenue. Une synténie est dite *soutenue* pour un génome ancestral GA si elle est présente dans au moins deux génomes dont le chemin évolutif dans un arbre phylogénétique contient l'espèce portant GA. Les deux espèces qui soutiennent la synténie peuvent être deux descendants ou un descendant et une espèce d'un groupe frère.

Dans la plupart des reconstructions de génomes ancestraux, la collection de synténies ancestrales ne contient que des synténies soutenues.

Synténie fiable. Mais on peut aussi détecter plus finement les adjacences ancestrales, en introduisant une petite quantité de mécanismes évolutifs, poussant vers les modèles présentés dans le chapitre suivant.

Par exemple, selon un principe décrit par Zhao & Bourque (2009), si on détecte deux adjacences dans un génome $G1$, notées AB et CD, puis les deux adjacences AC et BD dans un autre génomes $G2$, alors on pensera à une inversion, ou une translocation si les marqueurs A, B, C et D ne sont pas synténiques, sur le chemin évolutif qui relie ces deux génomes. Si un troisième génome $G3$ présente, lui l'adjacence AB, on pensera que l'inversion s'est produite sur le chemin entre $G2$ et le point P où la branche de $G3$ rejoint le chemin évolutif entre $G1$ et $G2$. Dans ce cas, le raisonnement implique aussi la présence de l'adjacence CD au point ancestral P, alors même qu'elle n'est pas soutenue par sa présence dans deux génomes. On appelle ce type de synténie *fiable*.

Double synténie. Ce type de synténie est un hybride des synténies soutenues et fiables : avec la connaissance d'un évènement évolutif, la duplication de grands segments génomiques (jusqu'à la duplication globale de génome), on détecte des traces conservées dans des génomes actuels. On appelle une *double synténie* l'appartenance d'un groupe de marqueurs à un même intervalle A, dans un génome qui n'a pas subi de duplication, et l'appartenance de leurs orthologues à deux intervalles disjoints B et C, dans un génome dupliqué.

Les intervalles B et C sont alors paralogues, tandis que A est orthologue à B et C, mais B et C peuvent ne présenter aucune similarité suite à des pertes de matériel génomique suivant une duplication. Ils portent pourtant la trace de cette duplication.

Les méthodes permettant de retrouver des doubles synténies sont nombreuses, et, détectant un signal synténique pour des marqueurs ancestraux,

peuvent être utilisée pour la cartographie ancestrale, comme dans Jaillon et al. (2004); Kohn et al. (2006); Nakatani et al. (2007); Tannier (2009); Chauve et al. (2010).

Support phylogénétique et statistique d'une synténie ancestrale

Si la détection des synténies ancestrales ne se heurte pas à des problèmes algorithmiques insurmontables, on n'a pas forcément la même confiance en toutes les synténies ancestrales. Cette confiance est proportionnelle par exemple au nombre d'espèces qui la soutiennent, pondéré par le signal phylogénétique : une synténie présente chez l'humain et le chien soutiendra mieux la synténie chez l'ancêtre des mammifères qu'une synténie présente chez l'humain et le macaque.

Il existe diverses façons de donner un score aux synténies ancestrales selon leur présence/absence ou valeur aux feuilles d'une phylogénie, s'inspirant de modèles d'évolution d'adjacences (Ma et al., 2006) ou de caractères continus (McArdle & Rodrigo, 1994).

La confiance accordée à une synténie ancestrale dépend aussi de la probabilité de la retrouver par hasard, et qu'elle ne soit pas due à de l'homologie. Par exemple, deux gènes peuvent se retrouver par hasard côte à côte dans deux génomes, et vont soutenir à tort une synténie ancestrale. C'est moins probable si quinze gènes se retrouvent contigus. Plusieurs batteries de tests statistiques sont ainsi capables de discriminer le signal du bruit, réalisés par Durand & Sankoff (2003); Raghupathy et al. (2008); Raghupathy & Durand (2009); Grusea (2010).

2.2.3 L'arrangement des marqueurs ancestraux

Après la phase de collecte, on dispose, pour une espèce ancestrale, d'un jeu de marqueurs et de relations entre eux, de type adjacence, appartenance à un intervalle ou à un chromosome, distance. Il faut construire un génome ancestral, c'est-à-dire arranger les marqueurs ancestraux le long de chromosomes, de façon à respecter, autant que possible, les relations collectées entre les marqueurs, pondérées par un signal phylogénétique ou un degré de significativité statistique.

À ce stade, à nouveau, que le génome soit actuel ou ancestral est indifférent. Sturtevant (1913) par exemple a construit ses premières cartes génétiques en calculant des distances entre marqueurs, puis en plaçant ses mar-

queurs sur une droite en respectant au maximum les distances. C'est un problème qu'on peut voir aujourd'hui comme le plongement d'un espace métrique dans une droite avec distorsion minimale (Badoiu et al., 2005), même si les distances génétiques ont des distorsions propres qui ont fait l'objet d'une littérature importante depuis la fonction de Haldane (1919). Pour les cartes physiques, la formulation comme un plongement d'espace métrique est rarement employée, le voyageur de commerce lui est substitué (Servin et al., 2010; Muffato et al., 2010), probablement pour des raisons de disponibilité de bons logiciels. Linéariser une collection d'adjacences (sans les distances entre paires de marqueurs) peut aussi se résoudre par plongement ou voyageur de commerce, c'est un cas particulier de métrique (Ma et al., 2006; Bertrand et al., 2010; Muffato et al., 2010). Mais le caractère discret de cette métrique pourrait ouvrir de nouvelles pistes, comme l'optimisation de la largeur de bande des graphes, qui est un analogue du plongement d'un graphe dans une ligne avec une distorsion minimum. Ces pistes ne sont pas explorées pour l'instant à ma connaissance.

Une collection d'intervalles, comprenant potentiellement des adjacences, appelle une résolution de type "uns consécutifs" ou de variantes (Chauve & Tannier, 2008; Chauve et al., 2010; Murat et al., 2010; Gavranović et al., 2011). L'introduction d'une souplesse dans les intervalles, où certains marqueurs sont manquants (absence de descendant d'un marqueur ancestral dans un génome actuel, défauts d'annotation, souplesse dans la définition d'intervalle) se modélise par un problème de "sandwich" (utilisé par Benzer (1959), voir en introduction) sur une matrice sur un ensemble à trois éléments $\{0, 1, X\}$. La propriété *SC1P* (*uns consécutifs sandwich*) est définie de la même manière que la propriété des uns consécutifs, c'est celle des matrices équivalentes à une matrice sans mauvais zéro. Ou, en d'autres termes, le principe est de remplacer toutes les occurrences des X par 1 ou 0 de façon à obtenir la propriété des uns consécutifs pour la matrice binaire qui en résulte (figure 2.4).

On peut imaginer plusieurs autres généralisations de la propriété des uns consécutifs. Par exemple, la propriété des uns circulaires (Hsu & McConnell, 2003) permet d'arranger des marqueurs sur un chromosome circulaire. La multiplicité des colonnes permet d'ajouter des duplications ou la proximité d'un télomère (Wittler & Stoye, 2010; Chauve et al., 2011). Les problèmes d'optimisation associés sont nombreux (plus grand sous-matrice qui vérifie la

	a	b	c	d
	7	10	11	12
13	1	1	0	1
17	1	1	0	0
42	X	1	1	0

	a	b	c	d	e	f
	1	6	7	8	12	13
4	1	1	1	0	X	0
12	0	1	1	1	X	1
13	X	X	1	1	0	1
16	0	1	1	1	X	0
43	X	X	0	1	1	0

FIGURE 2.4 – Deux matrices ternaires. Chaque colonne est un marqueur ancestral. Chaque ligne est un intervalle, où 1 veut dire que le marqueur est dans l'intervalle, 0 veut dire qu'il n'y est pas, et X veut dire qu'on ne sait pas. À gauche, la matrice a la propriété SC1P, et à droite, non.

propriété, plus petit nombre de mauvais zéros,...), sans qu'il existe encore de bons programme qui soit capable de les résoudre tous efficacement comme pour le voyageur de commerce.

Si les synténies ancestrales sont des relations d'appartenance à un même chromosome (la synténie dans son sens premier), alors la phase d'arrangement des marqueurs est un peu différente : on peut grouper les marqueurs dans des chromosomes sans les ordonner. Dans un graphe, dans lequel les sommets sont les marqueurs et les arêtes relient deux marqueurs synténiques, les composantes connexes, ou les composantes d'un clustering plus fin tenant compte des supports statistiques et phylogénétiques des synténies, sont des "groupes de liaisons" (Muffato et al., 2010) contenant des marqueurs, mais que les informations recueillies ne permettent pas d'ordonner.

2.3 Bactéries, Plantes, Animaux, Levures

Une méthode de cartographie ancestrale a été pour la première fois décrite en bioinformatique par Bergeron et al. (2004), et indépendemment appliquée aux poissons téléostes par Jaillon et al. (2004), puis appliquée aux streptophytes (Adam et al., 2007), aux mammifères (Ma et al., 2006; Chauve & Tannier, 2008), aux vertébrés (Kohn et al., 2006; Nakatani et al., 2007), chordés (Putnam et al., 2008), eumétazoaires (Putnam et al., 2007), aux céréales (Murat et al., 2010), aux levures (Tannier, 2009; Chauve et al., 2010; Bertrand et al., 2010), aux paramécies (Ma et al., 2008), ou à l'ensemble des eucaryotes (Muffato et al., 2010). Dans des travaux à venir, nous l'appliquons

encore aux amniotes, à d'autres plantes angiospermes, et à des bactéries. Différentes techniques ont été inventées et affinées à chacune de ces études, souvent indépendantes les unes des autres, et quelques développements théoriques ont permis une meilleure compréhension de leur efficacité et de leurs limites (Stoye & Wittler, 2009; Wittler & Stoye, 2010).

Ces études sont plus ou moins prospectives, plus ou moins fiables et produisent des génomes ancestraux plus ou moins assemblés. Par exemple, les travaux de Putnam et al. (2007) et Putnam et al. (2008) proposent des ancêtres bien assemblés sur des critères pas encore éprouvés, et n'utilisent la comparaison que de deux génomes. A l'opposé, Muffato et al. (2010) proposent une comparaison de tous les génomes disponibles, utilisent des critères fiables, et obtiennent des génomes peu assemblés (par exemple quelques centaines de blocs ancestraux pour le génome proto-amniote, qu'il reste à assembler en un nombre plus réduit de chromosomes).

Chaque domaine, ou royaume du vivant a ses spécificités évolutives (Coghlan et al., 2005). Mais les méthodes de cartographie, présentées ainsi de façon très générique, peuvent prédire des génomes ancestraux dans une bonne diversité de clades, pourvu qu'un signal synténique soit conservé. Les seuls ajustements nécessaires sont le signalement des spécificités évolutives comme les duplications globales de génomes, ou la forme des résultats attendus, par exemple l'arrangement en chromosomes linéaires ou circulaires.

J'ai ainsi participé à la construction de plusieurs caryotypes ancestraux, dont voici quelques représentations et une description de leurs spécificités.

Le caryotype du proto-boreoeuthérien est le plus étudié. C'est l'ancêtre des mammifères euthériens à l'exception des xénarthres et afrothériens. La figure 2.5 représente le résultat d'un arrangement de marqueurs universaux (chaque marqueur a exactement un homologue dans toutes les espèces) avec une approche basée sur les intervalles et adjacences soutenus et les uns consécutifs (Chauve & Tannier, 2008). L'assemblage est sans doute proche des chromosomes ancestraux, puisqu'il présente 27 morceaux, tandis que la cytogénétique prévoyait entre 23 et 25 chromosomes. L'intégration de marqueurs non dupliqués mais non nécessairement universaux est réalisée avec un problème de sandwich (SC1P, voir section 2.2.3) dans Gavranović et al. (2011), et l'assemblage propose un nombre similaire de morceaux de génomes, avec l'apparition d'un chromosome 12-22-10 (non représenté sur la figure) comme dans certaines études cytogénétiques (voir section 1.2.3, figure 1.2).

Le génome ancestral des amniotes (oiseaux et mammifères) est plus difficile à obtenir à cause de la plus grande profondeur évolutive, et du fait que

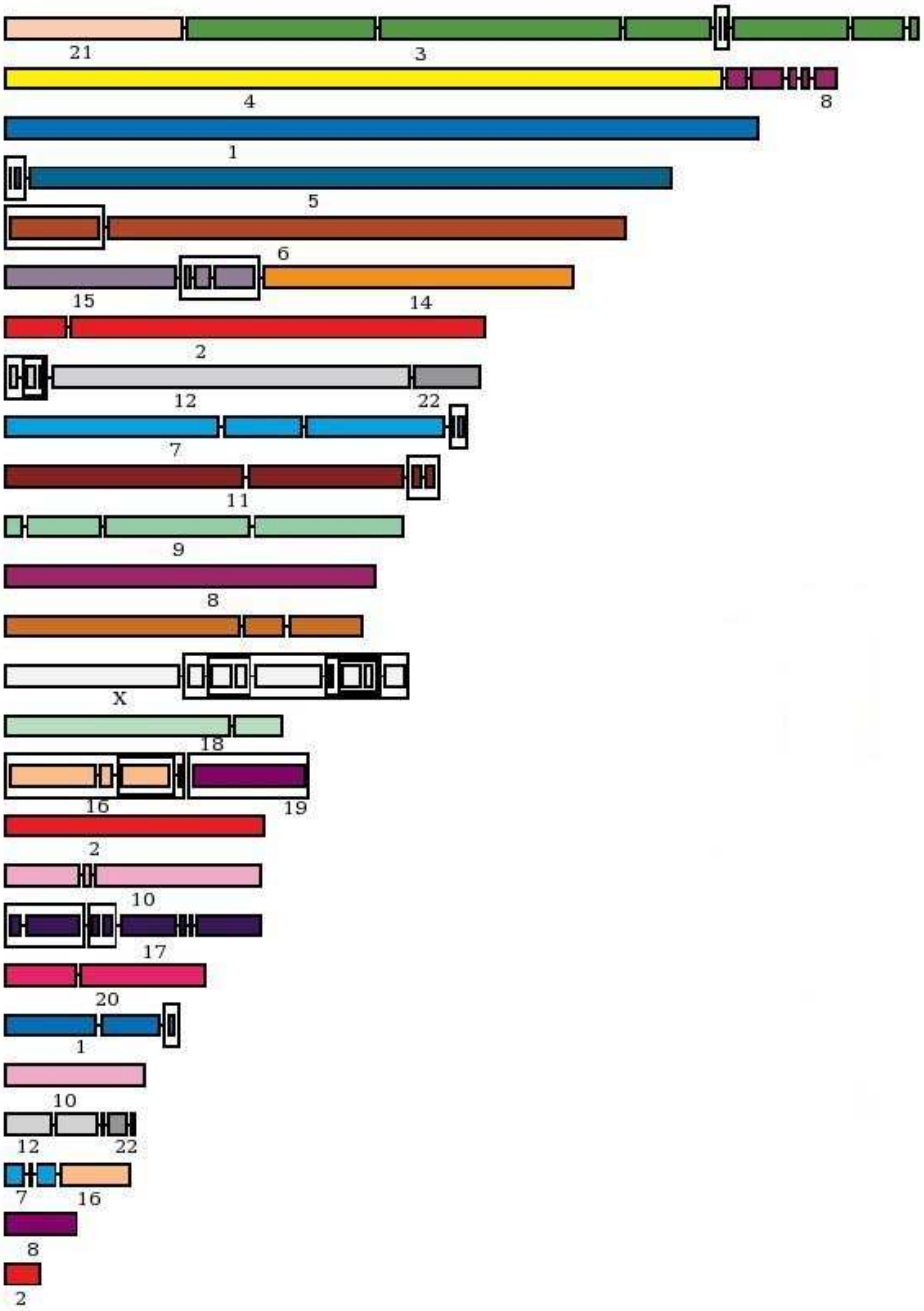


FIGURE 2.5 – Caryotype du proto-boreoeuthérien, colorié selon l'homologie avec le génome humain. Figure tirée de Chauve & Tannier (2008). La représentation emboîtée permet de représenter des incertitudes sur l'ordre de certains marqueurs, si l'information manque dans les génomes actuels pour établir une certitude sur certaines adjacences.

les plus proches parents dont le génome est séquencé sont des poissons qui ont subi une duplication globale de génome, des pertes massives de gènes, et des réarrangements intra-chromosomiques en nombre important. Certaines caractéristiques importantes de l'ancêtre ne sont accessibles que par des calculs de doubles synténies. Et en conséquence, la représentation finale est un graphe, dessiné sur la figure 2.6. Chaque sommet du graphe est lui-même un arrangement de marqueurs réalisé avec des marqueurs universaux sur les amniotes et la résolution d'un problème de uns consécutifs.

Les chromosomes de plantes angiospermes et leur évolution sont représentés sur les figures 2.7 (monocotylédones) et 2.8 (eudicotylédones). L'ordre des gènes a été obtenu avec la variante "sandwich" de la méthode des uns consécutifs, autorisant la perte de gènes (Gavranović et al., 2011). Cette méthode permet donc de reconstruire à la fois des génomes ancestraux de mammifères avec des marqueurs éventuellement absents et des génomes ancestraux de plantes ayant subi une duplication globale suivie de pertes massives de gènes. C'est pour cette raison que j'insiste sur le caractère générique du cadre méthodologique exposé ici, qui permet de construire des outils utilisables dans plusieurs royaumes du vivant aux spécificités évolutives diverses. La triplication originelle des eudicotylédones est retrouvée avec des méthodes de recherche de doubles synténies (section 2.2.2), en prenant un même génome comme génome dupliqué et non dupliqué (sur la figure 2.3 en bas, les trois segments font donc partie du même génome). Ceci illustre encore la généralité des méthodes et la disponibilité d'une boîte à outils conséquente, mais aussi l'adaptation nécessaire de cette boîte à certains cas particuliers, pour l'instant inaccessibles à une complète automatisation. L'ordre des gènes dans les chromosomes ancestraux de monocotylédones et de dicotylédones ont permis de retrouver des régions communes à l'ancêtre des angiospermes et à imaginer une histoire évolutive des deux grands groupes de plantes à fleurs. D'autre part, les mécanismes évolutifs semblent très différents dans les deux clades : si les caryotypes de céréales évoluent exclusivement par des fusions consistant en l'intégration d'un chromosome dans un autre (et les deux sont souvent des paralogues issus d'une duplication globale de génomes), ce réarrangement découvert récemment est complètement absent chez les dicotylédones.

Un caryotype ancestral de levure est représenté sur la figure 2.9. Ses descendants sont *Kluyveromyces thermotolerans*, *Saccharomyces klyuveri*, *Ashbya gossypii*, et *Kluyveromyces lactis*, tandis que *Zygosaccharomyces rouxii* est utilisée comme espèce externe. Cette proposition, publiée dans Chauve et al. (2010) corrige l'hypothèse d'une étude de Jean et al. (2009) fondée

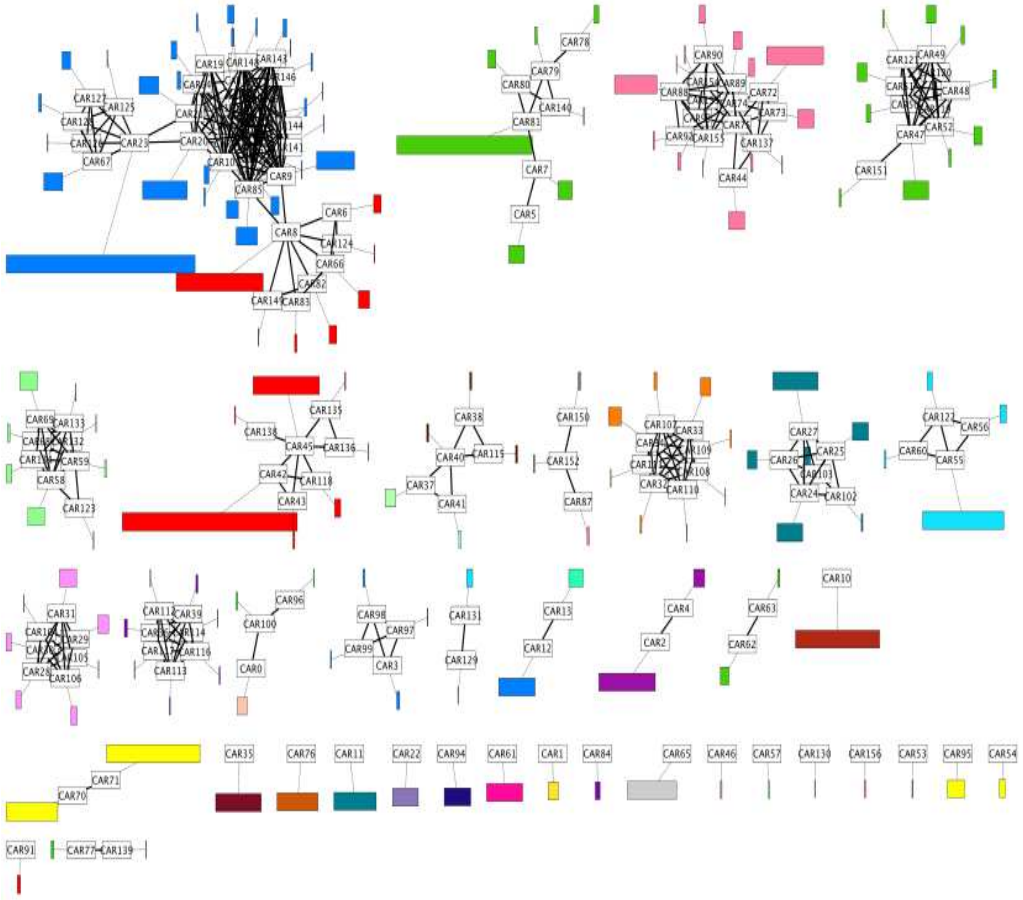


FIGURE 2.6 – Caryotype du proto-amniote, colorié selon l'homologie avec les chromosomes du poulet. Figure tirée d'un article en préparation. Certaines parties sont sur un même chromosome, mais pas ordonnées, d'où une représentation sous forme de graphe. Les sommets sont notés "CAR", pour "Contiguous ancestral region". Chaque CAR est un arrangement linéaire de marqueurs ancestraux.

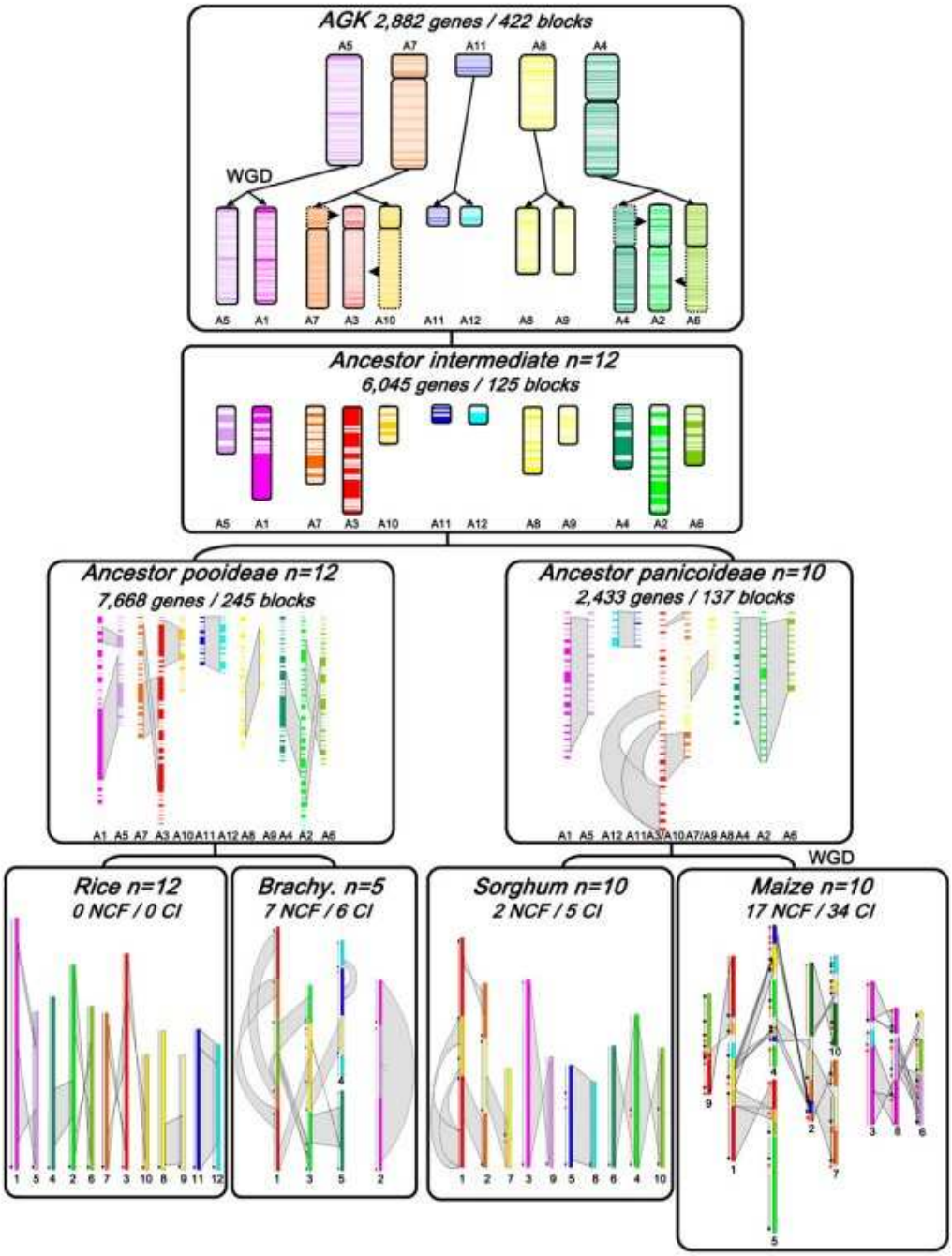


FIGURE 2.7 – Caryotypes ancestraux et actuels de céréales, avec leurs grands évènements évolutifs. CI annotes des inversions chromosomiques, NCF des fusions de chromosomes par un mécanisme qui intègre un chromosome entier dans le centromère d'un autre. Figure tirée de Murat et al. (2010).

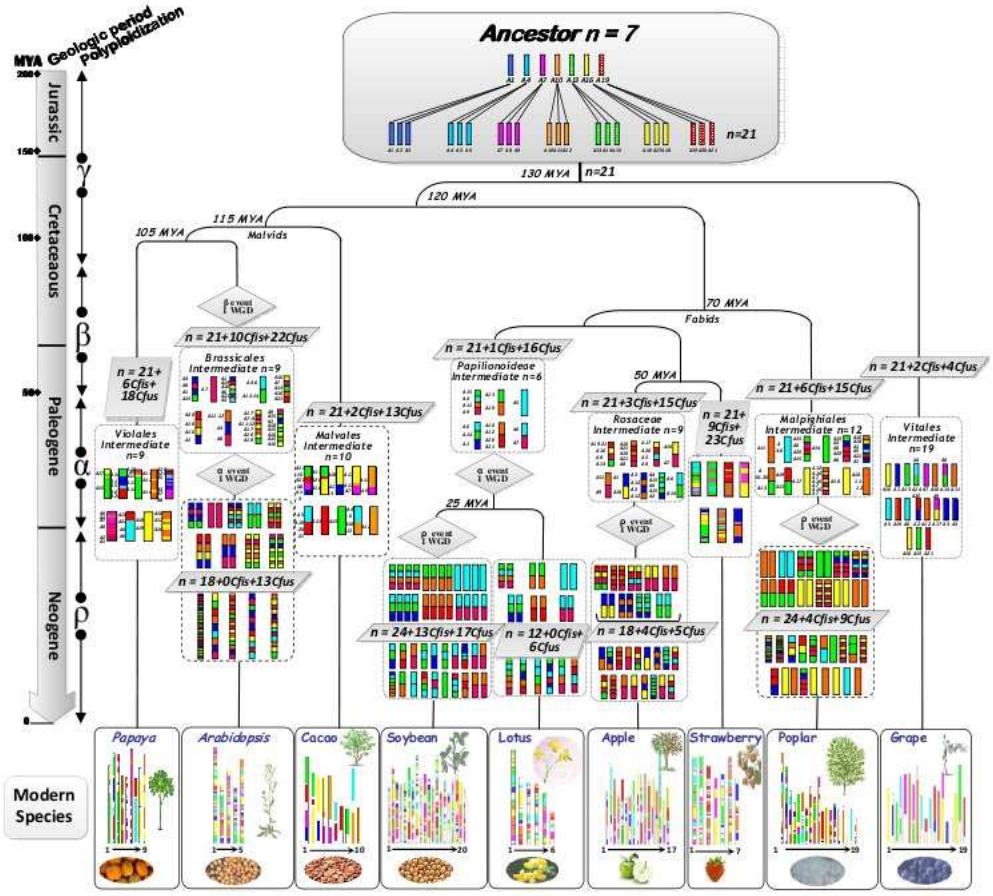


FIGURE 2.8 – Caryotypes ancestraux et actuels de dicotylédones, avec leurs grands évènements évolutifs. WGD annote une duplication globale de génomes, Cfis des fissions, Cfus des fusions de chromosomes. Figure tirée d'un article en préparation.

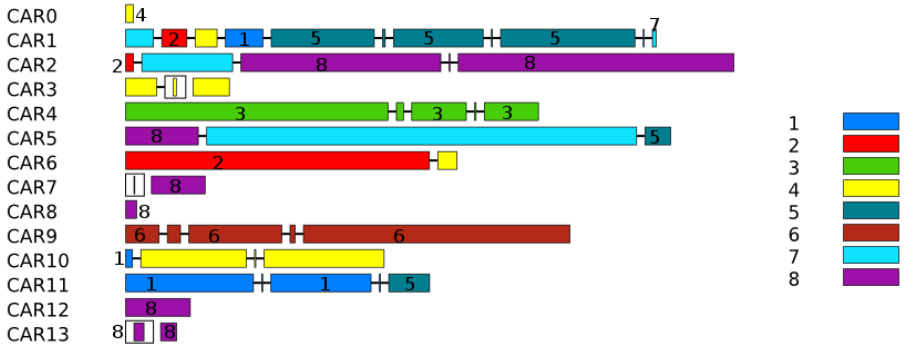


FIGURE 2.9 – Chromosome ancestral d'une famille de levures. La figure est publiée dans l'article Chauve et al. (2010). Les couleurs renvoient aux chromosomes de *Saccharomyces kluyveri*.

sur l'optimisation combinatoire, plus proche des modèles décrits dans le chapitre 3. Mais mon opinion sur la plus grande fiabilité de notre configuration est plus fondée sur ma confiance dans la méthode que dans une étude approfondie de sa validité (voir à ce sujet la section 2.4). L'obtention de ce caryotype repose sur l'utilisation de la quasi-totalité des développements méthodologiques décrits ici : adjacences et intervalles soutenus, adjacences fiables, double synténie (utilisation d'une espèce au génome dupliqué comme groupe externe, *Saccharomyces cerevisiae*, comme les poissons sont utilisés pour résoudre certaines parties du génome ancestral des amniotes).

Une variante des techniques d'arrangement permet de reconstituer des chromosomes circulaires, et donc de s'intéresser à l'évolution bactérienne. La figure 2.10 représente un chromosome ancestral de deux espèces d'aquificales (*Aquifex aeolicus* et *Hydrogenobaculum*), reconstruit en utilisant un jeu de données de 7 espèces. C'est un ensemble de segments plus ou moins longs, dont l'ordre interne est connu, mais l'arrangement des segments eux-mêmes le long du génome est inconnu, par manque d'informations dans les espèces actuelles. La méthode utilise le problème des "uns circulaires", une généralisation des uns consécutifs. Le résultat est un arbre, appelé *arbre PC* (Hsu & McConnell, 2003), qui est ici représenté par les branches internes au cercle, et permet de montrer l'incertitude concernant certaines parties du génome, comme la représentation emboîtée de la figure 2.5 représentait l'homologue de cet arbre PC dans les structures linéaires, l'*arbre PQ*.

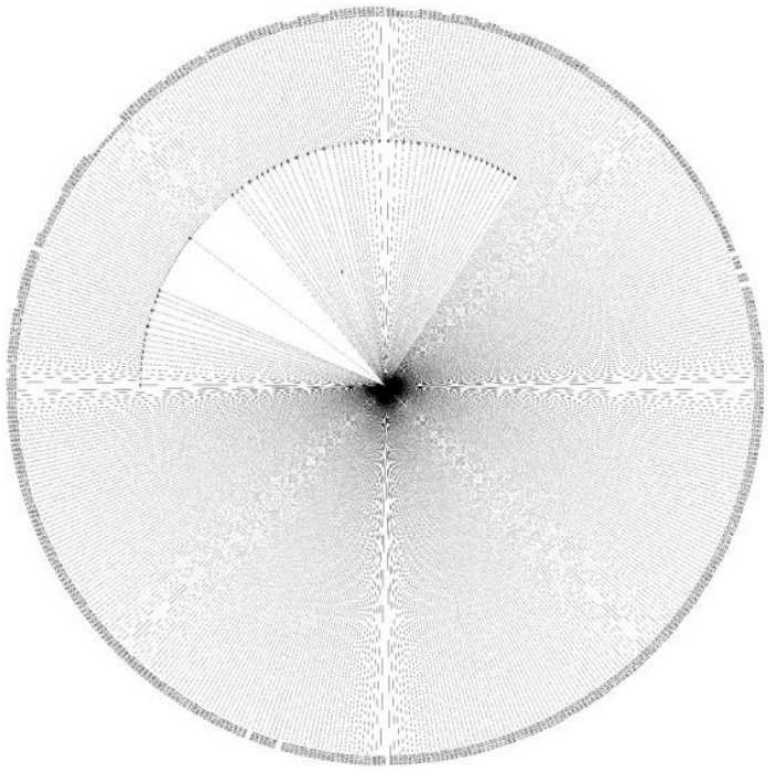


FIGURE 2.10 – Chromosome ancestral d'une bactérie aquifex, sous forme d'arbre PC. Les feuilles de l'arbre sont les gènes, ordonnés en cercle. Les noeuds, à l'intérieur du disque, correspondent à des ensembles de gènes qu'on prédit contigus dans le génome, sans se prononcer sur l'ordre exact de leur apparition. Figure tirée du rapport de stage de A. Rougny.

2.4 Les possibilités des méthodes de reconstruction de génomes ancestraux

"Unfortunately for users, an area can become quite cluttered with mediocre tools, because many people find it much easier and more fun to develop a new program than to adequately verify that it actually improves upon earlier work, and using another person's software is sometimes treated like using their toothbrush." écrit Webb Miller en parlant de la génomique comparée (Miller, 2001).

Ce qu'il décrit pourrait s'appliquer à la reconstruction de génomes ancestraux, où produire une hypothèse est relativement aisé, tandis que son évaluation est difficile. Le séquençage de l'ADN ancien concerne une partie trop récente de l'évolution pour espérer connaître la vérité sur les génomes représentés dans la section précédente.

On peut évaluer une méthode avec des données simulées, mais les modèles d'évolution structurales ne sont pas encore bien développés (voir le chapitre 3). La plupart des méthodes montrent un très bon comportement sur des données simulées avec des modèles très simplifiés (Ma et al., 2006; Alekseyev & Pevzner, 2009; Muffato et al., 2010), alors même qu'ils produisent des résultats contradictoires. C'est en partie un effet de mesure : si 99% des adjacences d'un génome sont correctement reconstruites, les 1% d'erreurs, sur 20000 gènes, font 200 différences, et deux génomes qui varient de 200 adjacences peuvent être aussi différents que le sont ceux de l'humain et de la souris, alors même qu'ils présentent tous deux un excellent score pour la reconstruction.

Une petite controverse scientifique a secoué la paléogénomique en 2006. Elle portait sur le génome d'un proto-mammifère, qui avait été reconstruit par des cytogénéticiens à l'aide de techniques d'hybridation in-situ. Avec l'arrivée de séquences de génomes complets, les premiers génomes ancestraux issus de la bioinformatique sont apparus (Bourque & Pevzner, 2002; Bourque et al., 2004). Mais ils ne ressemblaient pas à ceux, plus anciens, plus travaillés, soutenus par plus d'espèces, de la cytogénétique (Richard et al., 2003; Yang et al., 2003; Froenicke et al., 2003; Wienberg, 2004). Deux articles à peine polis sont parus dans le même numéro de Genome Research, l'un posant les résultats de la cytogénétique comme référence et réfutant ceux de la bioinformatique (Froenicke et al., 2006), l'autre défendant les résultats de la génomique en minimisant les différences observées (Bourque et al., 2006). Un

dernier article est venu conclure le débat en affirmant la nécessité d'une approche interdisciplinaire (Rocchi et al., 2006). Plusieurs approches purement bioinformatiques ont finalement retrouvé les résultats de la cytogénétique (Ma et al., 2006; Chauve & Tannier, 2008; Alekseyev & Pevzner, 2009; Muffato et al., 2010; Gavranović et al., 2011) (voir pourtant la petite incertitude mentionnée en section 1.2.3 et 2.3), et la controverse s'est éteinte.

A mon avis, ce débat n'est pas complètement sans intérêt, et montre d'une part que l'interdisciplinarité, si importante fût-elle en général, n'a pas apporté la solution dans ce cas. Il est maintenant admis que les résultats de la cytogénétique sont une référence et ils sont utilisés comme tel dans toutes les méthodes bioinformatiques qui s'attaquent à la reconstruction de cet ancêtre. Ils fournissent d'ailleurs aux différentes méthodes un moyen précieux d'évaluation. Ensuite, si le problème ne venait pas de différences de disciplines ou de nature des données (puisque maintenant il n'y a pratiquement plus de divergence), c'est qu'il venait des méthodes. Les premières reconstructions du génome de l'ancêtre des mammifères utilisant des séquences génomiques et la bioinformatique étaient fondées sur une solution à problème d'optimisation combinatoire qui tentait de minimiser le nombre de réarrangements expliquant les différences de structure entre génomes (Bourque et al., 2004). Dans ce type de méthode, l'optimalité des solutions n'est pas garantie, et le nombre de solutions optimales ou légèrement sous-optimales est gigantesque, deux solutions pouvant théoriquement être très éloignées. Le choix d'une solution au hasard, ou même d'un échantillon dont on ne maîtrise pas la loi, est donc forcément une source d'erreur.

Je pense par conséquent que la solution au débat et à l'évaluation des résultats se trouve pour l'instant dans l'examen des principes fondant les méthodes. En l'absence d'un modèle dans lequel on puisse placer sa confiance, on devrait éviter l'optimisation sur des problèmes *ad-hoc*, y compris voyageur de commerce ou sous-matrices aux uns consécutifs. De cette manière, on pourrait comprendre le résultat à partir des données, sans l'intermédiaire d'une boîte noire dont on ne maîtrise pas bien le déroulement. Pour l'instant, aucune méthode n'en est pleinement capable, et les miennes ne font pas exception, puisque les solutions aux problèmes de uns consécutifs ou leurs variantes nécessitent toujours de l'optimisation. Mais sa part devient de plus en plus réduite. Par exemple, pour reconstruire le génome ancestral des boréoeuthériens dans Gavranović et al. (2011), une matrice à 1724 colonnes et 89023 lignes est construite, et retirer seulement 0.3% des lignes suffit à produire une sous-matrice SC1P. C'est un gain substantiel par rapport à une

étude antérieure (Chauve & Tannier, 2008) qui ne prenait pas en compte les possibles absences de marqueurs dans certains génomes, et nécessitait le retrait de 3% des lignes de la matrice. Une manière d'éviter l'optimisation serait de ne pas utiliser un signal s'il est en conflit avec un autre. Et pour conserver du signal, les conflits peuvent être évités à l'aide de tests statistiques sur le signal détecté, ou du traitement explicite de certaines situations, comme la perte de gènes, plutôt que son inclusion dans une masse de donnée bruitée (ce qui est fait avec le problème du sandwich par exemple).

Le problème n'est donc pas clos. Et prendre l'évaluation au sérieux, poser les propriétés qu'on désire pour les méthodes et les résultats, expliciter les hypothèses utilisées dans les méthodes, permettra d'établir une confiance dans les résultats et de les utiliser pour déduire des propriétés sur les modes d'évolution structurale. Et par là, construire des modèles d'évolution réalistes. C'est le sujet du chapitre suivant.

Chapitre 3

Les modèles d'évolution structurale

La cartographie est pour l'instant le seul moyen fiable de construire des génomes ancestraux. Mais les techniques mises en oeuvre sont éloignées des processus évolutifs, et ne donnent pas, en plus des séquences ancestrales, leur explication en termes d'évènements évolutifs qui les séparent des génomes actuels. Pour ça, il faut construire des modèles d'évolution structurale des génomes, et être capable de faire des calculs dans ces modèles. C'est une tâche qui occupe des groupes de mathématiciens et informaticiens depuis maintenant plusieurs décennies, et dont les succès se mesurent bien sûr à quelques résultats encourageants sur la compréhension de l'évolution des chromosomes, mais aussi à la découverte d'un champ de recherche en mathématiques discrètes et en optimisation combinatoire, qui tisse des relations intéressantes avec l'algèbre des corps finis, l'algorithmique dans les graphes, la combinatoire des permutations, ou d'autres problèmes jusque-là isolés.

3.1 La combinatoire des réarrangements génomiques

3.1.1 Un nouveau champ d'investigation mathématique

La première apparition d'un énoncé de combinatoire des réarrangements est la phrase déjà citée de Sturtevant & Novitski (1941) : "... for each such sequence there was determined the minimum number of successive inversions

required to reduce it to the ordinal sequence chosen as "standard". For numbers of loci above nine the determination of this minimum number proved too laborious, and too uncertain, to be carried out..." Mais cette phrase, formulation d'un problème d'optimisation combinatoire, est restée lettre morte à cette époque où peu de mathématiciens pensaient en termes d'algorithmes, et presque aucun en termes de "bons" algorithmes. Sturtevant avait pourtant fait appel à un mathématicien de son université, Morgan Ward, qui a sans doute aidé à la formalisation du problème. Mais il s'est limité à des tentatives d'énumération ou de comptage des couples de permutations à une distance donnée, sans s'intéresser aux algorithmes qui calculeraient cette distance. Ce n'est que plus de 40 ans plus tard que des tentatives de résolution ont été publiées, dans un contexte différent et sans référence à Sturtevant, puisqu'il s'agissait d'expliquer les différences d'ordre de gènes dans des chromosomes circulaires (Watterson et al., 1982). La recherche a réellement démarré dans le domaine sous l'impulsion de David Sankoff (Sankoff, 1989), qui a exploré les différentes variantes et les possibilités d'applications.

Une forme générale du problème est : étant donnés deux génomes et un type de remaniement, transformer un génome en l'autre en utilisant un nombre minimum de remaniements du type donné.

Les variantes sont ensuite construites en instanciant les définitions de génome (permutation, chaîne de caractères, ensemble d'adjacences), le type de réarrangement (inversions, transpositions, translocations, duplications, pertes, fusions, fissions, ou combinaisons de ces opérations, éventuellement pondérées), ou en généralisant l'énoncé à plus de deux génomes : dans ce cas, on peut résoudre la "petite parcimonie", en ajoutant un arbre phylogénétique dans les données du problème et en minimisant le nombre de réarrangements le long de ses branches, ou la "grande parcimonie", qui consiste à construire l'arbre qui minimise le score de petite parcimonie. S'il y a trois génomes, les deux sont équivalents, et le problème est un cas particulier de celui de la "médiane" génomique. Puis, les solutions n'étant jamais uniques, on doit s'intéresser à leur énumération, et quand elles sont trop nombreuses pour être énumérées, à l'exploration de l'espace des solutions (structuration, comptage, échantillonnage).

J'ai participé à la recherche dans ce domaine par plusieurs travaux algorithmiques (Tannier & Sagot, 2004; Sagot & Tannier, 2005; Tannier et al., 2007; Braga et al., 2007; Diekmann et al., 2007; Braga et al., 2008; Tannier et al., 2008; Lenne et al., 2008; Bérard et al., 2008; Tannier et al., 2009; Bérard et al., 2009; Gavranović & Tannier, 2010), une revue pour l' *Encyclopedia of*

Algorithms (Tannier, 2008), un article de vulgarisation pour *La Recherche* (Tannier, 2010), ainsi que par l'écriture d'un ouvrage, qui se veut exhaustif sur les problèmes de ce type traités dans la littérature jusqu'en 2008 (Fertin et al., 2009). L'ouvrage n'est pas destiné à des biologistes, et n'est pas un recueil de méthodes qu'ils pourraient utiliser selon les besoins dictés par les données. Un tel ouvrage n'existe pas encore, et ce qu'on peut imaginer de son contenu n'est sans doute pas encore assez fourni pour l'envisager. Mais le livre de Fertin et al. (2009) illustre comment la biologie, comme je l'ai déjà écrit en introduction, peut être appliquées aux mathématiques. L'ouvrage est principalement destiné aux mathématiciens et informaticiens, et la biologie est l'inspiratrice des problèmes et des variantes, un aiguillon pour la recherche, qui a produit autant de résultats mathématiques intéressants que de retours fructueux à la biologie. Ces derniers existent, heureusement, et je détaille l'un d'entre eux, auquel j'ai participé, dans la section 3.1.2.

Mais tout d'abord, je voudrais insister ici sur cet aspect "biologie appliquée" en montrant comment la combinatoire des réarrangements génomiques a parfois permis la résolution de certains problèmes mathématiques qui s'étaient posés en d'autres termes auparavant. Je le ferai en relatant trois ensembles de résultats, moins directement liés à mes propres contributions sur le sujet [1], mon apport étant simplement de les réunir dans ce mémoire pour les raconter (la plupart des développements étant plus récents que le livre de Fertin et al. (2009), ces histoires n'y sont pas racontées, les trois sections qui suivent pourraient être une base de sa mise à jour).

La factorisation des permutations

Le premier problème remonte au début du XIXe siècle et aux études de Cauchy sur les permutations. La forme aboutie de sa théorie est publiée en 1844 (Cauchy, 1844) et stipule que toute permutation est

– décomposable de façon unique en "cycles", et

1. Ou alors, elles concernent mon travail de rapporteur pour des conférences ou des revues. C'est par définition un travail de l'ombre, effectué dans l'anonymat et qui n'est pas destiné à être valorisé autrement que par la satisfaction de voir publiés des articles de meilleure qualité. Mais poussé par les réflexions d'Aaron Darling sur le sujet (`http://secretmicrobe.org/who-owns-your-review`) qui milite pour la reconnaissance de ce travail, je m'autorise à raconter l'histoire de la résolution du problème de Cayley, non pas en tant qu'auteur des articles qui annoncent maintenant avoir résolu ce problème vieux de 150 ans (Amir & Levy, 2010), mais parce que j'ai fait pour les auteurs, en tant que rapporteur d'articles, le rapprochement entre le problème de Cayley et leur contribution.

– décomposable en un produit de “transpositions” (c’est-à-dire l’échange de positions de deux éléments [2]),

et que le nombre minimum de transpositions dans un produit est égal à la taille de la permutation moins le nombre de cycles dans la décomposition. D’autre part, si la décomposition en transpositions n’est pas unique, le nombre de transpositions a toujours la même parité pour une permutation donnée. C’est sa *signature*.

Ce résultat, devenu classique et enseigné en premier cycle en mathématiques, a eu une descendance abondante, et parmi celle-ci, une note de Cayley (1849) qui étend cette décomposition en un produit de transpositions à des “permutations d’éléments pas nécessairement différents”. Pour éviter la confusion je continue à appeler *permutation* une séquence d’éléments tous différents, et *chaîne* une séquence d’éléments pas nécessairement différents. Cayley essaie donc de définir la signature d’une chaîne de caractères. Cette note de Cayley est intéressante pour les algorithmiciens, parce que la décomposition d’une chaîne est définie de deux façons différentes et équivalentes :

– comme le nombre minimal de transpositions dont le produit est une permutation obtenue en différenciant les éléments égaux de la chaîne ;
– comme le nombre minimal d’échanges de deux éléments nécessaire pour transformer la chaîne donnée en un arrangement de référence contenant les mêmes éléments.

Cayley écrit alors “the former enunciation is based upon and indicates a direct method of determining the minimum number of inversions [3] requisite in order to obtain a given permutation [4] ; but the latter is, in simple cases, of the easiest application.” Il donne donc une solution algorithmique au problème de la décomposition, qui consiste à énumérer toutes les permutations qui différencient les éléments égaux de la chaîne. Mais il admet implicitement que

2. Cette définition de transposition ne correspond pas au réarrangement biologique. Transposition a d’ailleurs au moins trois acceptions possibles dans des domaines qui touchent ce mémoire, puisque la définition biologique se réfère plutôt aux éléments transposables, celle des informaticiens des réarrangements au déplacement d’un segment quelconque de génome, et celle, utilisée dans la section présente, des mathématiciens des permutations. Pour ne rien arranger, Cayley (1849) appelle les transpositions des inversions, qui ont aussi un sens particulier pour les réarrangements génomiques, et un autre dans la terminologie moderne des permutations. Amir et al. (2009), qui se placent dans le cadre de l’algorithmique des réarrangements, inventent une nouvelle terminologie pour ces opérations : les “échanges d’éléments”.

3. transpositions.

4. chaîne.

cette méthode, même si elle a l'avantage d'être définie formellement, n'est pas forcément la meilleure à utiliser en pratique (et en effet le nombre de cas à examiner peut être de l'ordre de la factorielle de la taille de la chaîne). On peut trouver plus facilement le nombre de transpositions à appliquer pour "trier" une chaîne dans des cas simples, bien qu'il n'ait pas d'algorithme pour calculer ce nombre dans le cas général. Cayley donne l'exemple de la chaîne 1212, pour laquelle énumérer toutes les permutations nécessite d'écrire les permutations *acbd*, *bcad*, *adbc*, *bdac*, et à les décomposer toutes en transpositions, tandis que trouver une transposition qui transforme 1212 en 1122 est immédiat.

Il y a donc un problème algorithmique posé par Cayley dans cette note, qui semble resté sans réponse pendant plus de 150 ans. L'article est parfois cité, mais apparemment c'est à chaque fois pour lui attribuer indûment le résultat de Cauchy, et pas pour s'intéresser au problème qui y est posé.

Et c'est dans le cadre de la combinatoire des réarrangements génomiques, mêlée d'algorithmique sur les chaînes de caractères (le problème est déjà bien éloigné de la génomique à cause d'un réarrangement assez irréaliste), que le problème a été redécouvert par Amir et al. (2009), avec un résultat de NP-complétude pour le problème de Cayley (sa deuxième formulation ne mène finalement à aucun bon algorithme), sauf si les transpositions ont un poids proportionnel à la distance des éléments échangés, auquel cas il existe un bon algorithme.

Mélange des jeux de cartes, retournement de crêpes

Quand les génomes sont des permutations, les problèmes de réarrangements s'inscrivent dans le cadre computationnel de la factorisation des permutations formalisé par Jerrum (1985). Certains problèmes relèvent aussi de l'étude des "graphes de Cayley" dont les sommets sont des permutations et les arêtes relient deux permutations si on peut transformer l'une en l'autre en une opération.

D'autres problèmes mathématiques s'inscrivent dans ce cadre, sans toujours sortir de l'isolement et former une théorie conséquente. Certains se sont découvert a posteriori des affinités avec les réarrangements, comme les piles de crêpes et le mélange des cartes à jouer.

Le renversement des piles de crêpes est un problème célèbre posé sous forme d'un jeu combinatoire (Dweighter, 1975) : étant donné un tas de crêpes de tailles toutes différentes, on cherche à les réarranger de façon à ce qu'au-

cune crêpe ne soit placée au-dessus d'une plus petite, en un nombre minimum d'opérations. Chaque opération doit consister à renverser la partie haute du tas, à partir d'une position choisie. C'est un cas contraint des réarrangements génomiques par inversions, qui aura aussi sa version dans laquelle les crêpes sont "orientées", comme les gènes (Cohen & Blum, 1995). Cette opération, qui oblige le segment renversé à contenir une extrémité de la permutation, est même intéressante du point de vue de la génomique, puisqu'il est possible que les inversions contenant un télomère aient un statut particulier. Les rapprochements sont pourtant restés à l'état d'idées inabouties et la convergence des deux problèmes n'est que méthodologique pour l'instant. Par contre, l'expansion de la combinatoire des réarrangements génomiques et la multiplication des variantes a permis de repartir de cette contrainte et de l'appliquer à d'autres types de réarrangements, comme les transpositions (Dias & Meidanis, 2002), dans une étude qui se situe dans le champ des réarrangements génomiques.

Le mélange des cartes à jouer est un exemple moins connu mais plus intéressant pour ses relations avec les réarrangements. Le geste utilisé par les joueurs pour mélanger les cartes consiste souvent à couper le jeu en deux, puis réunir les deux parties en les entrecroisant le plus possible (figure 3.1). Il est bien connu des joueurs de black-jack qu'effectuer cette opération une ou

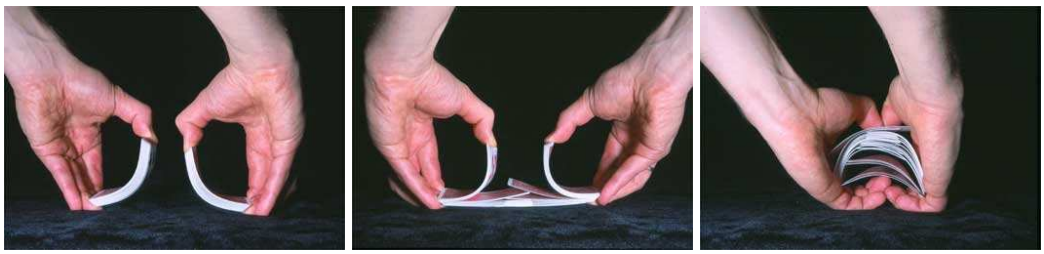


FIGURE 3.1 – Mélange de cartes à jouer.

deux fois n'élimine pas toute trace de l'ordre antérieur dans l'arrangement obtenu, et qu'on peut alors encore prédire en partie quelles cartes vont être tirées les premières. On peut calculer le nombre de fois où l'on doit mélanger de cette façon pour atteindre un ordre qui soit indépendant de l'ordre de départ. Le problème est plutôt difficile, et il semble que $\frac{3}{2}\log_2 n$ gestes soient suffisants pour mélanger n cartes, soit 7 coups pour un jeu de 52 cartes (Bayer & Diaconis, 1992). Ce qui, selon les auteurs de ce résultat, qui n'ont pas fréquenté que les laboratoires de mathématiques, est en dessous de ce que

les croupiers pensaient être un mélange raisonnable, et certains tenteraient d'en profiter pour gagner en prédisant l'ordre d'apparition des cartes.

Un résultat amusant de Diaconis et al. (1983) et pas sans conséquence pour ce qui nous occupe montre enfin que l'aléatoire et l'imprécision dans le geste sont indispensables à un bon mélange : si on imagine un croupier qui sache entrecroiser exactement les cartes des deux moitiés de jeu (une carte sur deux est d'une moitié, l'autre de l'autre), un ordre indépendant de l'ordre de départ n'est pas accessible et l'ordre initial est retrouvé au bout de quelques opérations. Si le processus est vraiment aléatoire, c'est-à-dire qu'à chaque position du jeu de carte, on a une chance sur deux de trouver une carte venant d'un tas et une chance sur deux de l'autre, on attend une distribution normale des tailles des suites de cartes appartenant à un tas et se retrouvant consécutives dans l'arrangement produit par un mélange (Wald & Wolfowitz, 1940).

La relation avec l'évolution des génomes n'est pas frappante, mais non moins réelle et intéressante. Elle a été découverte par Bernt et al. (2011). Si on déroule en imagination le geste inverse de celui du mélange des cartes, il s'agit de séparer un jeu de cartes en deux tas, en décidant pour chaque carte, avec une probabilité $\frac{1}{2}$, à quel tas elle doit être affectée. Puis les deux tas sont fusionnés en apposant l'un sur l'autre. Le processus de duplication totale du génome fonctionne quasiment de la même façon (figure 3.2) : deux copies d'un chromosome sont construites, puis pour chaque gène, une des deux copies est perdue (je considère ici qu'aucun gène n'est gardé en deux copies). Ensuite, les deux chromosomes peuvent être fusionnés comme on le constate parfois chez les plantes (Murat et al., 2010). Si le nombre de duplications nécessaire à bien mélanger les génomes n'est sans doute pas un problème biologique de grande importance, on peut, dans l'ordre des gènes obtenu après une duplication, mesurer comme dans Diaconis et al. (1983) le degré d'aléatoire dans la répartition des gènes dans une copie d'un chromosome ou une autre. La question de l'indépendance du choix de la copie perdue d'un gène par rapport à ses voisins est, elle, une question biologique débattue. C'est celle du croupier trop adroit qui entrecroiserait strictement une carte sur deux ou qui, un peu maladroit et plus proche des données biologiques, laisserait de longues suites de cartes consécutives provenant d'un seul tas.

Le débat de savoir si ce regroupement de gènes est du à la délétion concomitante de plusieurs gènes ou à une sélection pour garder proches des groupes de gènes consécutifs n'est pas tranché (Sankoff et al., 2010).

Dans les génomes mitochondriaux, les duplications suivies de pertes aléa-

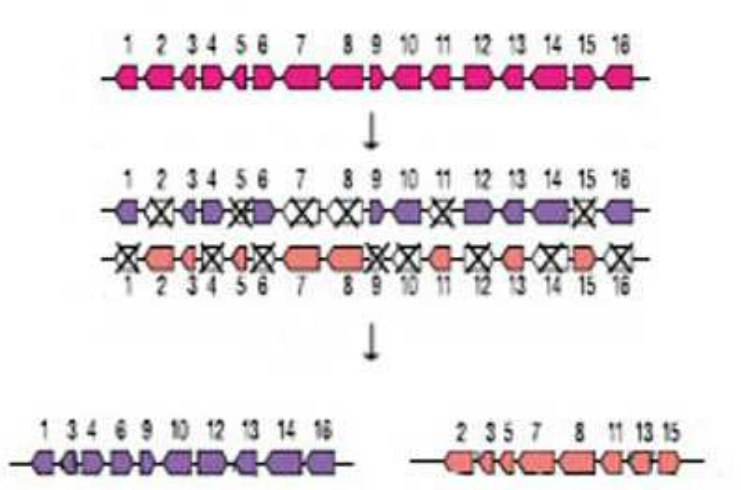


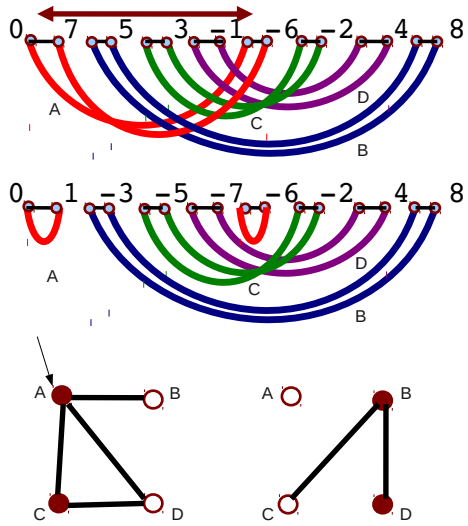
FIGURE 3.2 – Modèle pour la duplication d'un chromosome de levure, à partir de Kellis et al. (2004). Le processus consiste à dupliquer tous les gènes, puis supprimer une copie de chaque, au hasard. Si on remonte le temps, on a l'ensemble des gènes en deux parties, qu'on mélange avec le même procédé qu'un jeu de cartes.

toires sont peut-être des réarrangements courants. Dans ce cas, certains résultats sur les mélanges de cartes peuvent être directement appliqués, comme l'énumération des scénarios qui transforment un ordre en un autre par cette opération (Grinstead & Snell, 2006).

Matrices positive de GF[2]

Le modèle mathématique le plus simple pour comparer des ordres de gènes est la permutation. Les permutations signées, où chaque élément porte un signe positif ou négatif (voir figure 3.3, en haut à gauche), sont un modèle plus réaliste puisqu'elles permettent de modéliser le sens de lecture des gènes ou marqueurs, tout en rendant, une fois n'est pas coutume, les problèmes computationnels plus simples.

La comparaison de deux permutations signées π et σ permet de construire un *graphe de comparaison* qui contient deux sommets par éléments de la



	A	B	C	D
A	1	1	1	1
B	1	0	0	0
C	1	0	1	1
D	1	0	1	1

	A	B	C	D
A	0	0	0	0
B	0	1	1	1
C	0	1	0	0
D	0	1	0	1

FIGURE 3.3 – Effet d'une inversion sur une permutation (en haut à gauche), ainsi que sur le graphe de comparaison. Effet d'une complémentation locale sur le graphe des chevauchements correspondant (en bas à gauche). Le graphe des chevauchements après la complémentation locale est le graphe des chevauchements de la permutation après l'inversion. À droite, effet d'une "élimination gaussienne" sur la matrice d'adjacence du graphe. La matrice après l'opération d'élimination est la matrice d'adjacence du graphe après complémentation locale. L'inversion est donc aussi une opération algébrique, et une opération de minoration de graphe.

permutation, et une arête pour chaque adjacence, soit dans π, soit dans σ (le graphe, composé de cycles disjoints, est dessiné en dessous des permutations de la figure 3.3). Le *graphe des chevauchements* (en bas à gauche sur la figure 3.3) est alors construit à partir du graphe de comparaison : c'est le graphe des chevauchements (intersection, mais pas inclusion) des intervalles décrits par les cycles du graphe de comparaison (sur la figure, le cycle A décrit l'intervalle contenant $\{7, 5, 3, 1\}$ et le cycle B décrit l'intervalle contenant $\{5, 3, 1, 6, 2, 4\}$, donc ils se chevauchent. Par contre B ne chevauche pas C ou D. Les sommets sont coloriés en noir si les arêtes du cycle correspondant se croisent dans le graphe de comparaison, et en blanc sinon. La *matrice des chevauchements* (à droite sur la figure 3.3) est la matrice d'adjacence

du graphe des chevauchements (les entrées sur la diagonales sont 1 pour les sommets noirs, 0 pour les blancs).

Hannenhalli & Pevzner (1999) ont remarqué qu'une inversion dans la permutation (inversion de l'ordre des éléments dans un segment de la permutation, accompagnée de l'inversion de tous les signes dans le segment) avait l'effet d'une complémentation locale dans le graphe des chevauchements, c'est-à-dire que si une inversion est appliquée à la permutation, les couleurs et les relations de voisinage sont complémentées dans le voisinage d'un sommet du graphe des chevauchements. Sur la figure 3.3, l'inversion de l'intervalle décrit par le cycle A correspond à la complémentation du sommet A et de son voisinnage dans le graphe.

Plus tard, Hartman & Verbin (2006) ont remarqué qu'une inversion dans une permutation avait un effet similaire à l'élimination gaussienne effectuée sur le corps à deux éléments dans la matrice des chevauchements. Ainsi, le nombre minimum d'inversions pour transformer une permutation en une autre peut s'exprimer en fonction du rang dans $GF[2]$ d'une matrice binaire. Le gain pour le problème initial n'est pas immédiatement flagrant, puisque celui-ci avait été résolu plusieurs années auparavant par Hannenhalli & Pevzner (1999) par un algorithme de meilleure complexité que celle du calcul d'un rang de matrice. Hartman & Verbin (2006) espéraient que cette formulation permettrait de résoudre des problèmes plus difficiles, comme le tri par transpositions [5], mais leur trouvaille vaut peut-être plus par l'apport possible de la combinatoire des réarrangement génomiques à des problèmes d'algèbre sur les corps finis. Un cas particulier du théorème de Hannenhalli & Pevzner (1999) généralise des résultats de Lempel (1975) ou Seroussi & Lempel (1980) sur la factorisation des matrices dans un corps fini. Il donne une caractérisation des matrices binaires carrées B telles que $B = PLO$, où P est une matrice de permutation, L est une matrice triangulaire supérieure, et O est une matrice orthogonale.

C'est une proposition pour étendre la notion de matrice définie positive (habituellement définie sur le corps des réels) au corps à deux éléments. Apparemment, c'est une piste qui n'est pas explorée par les algébristes et qui pourrait déboucher sur des nouvelles méthodes pour factoriser des matrices binaires.

5. Espoir sans doute maintenant déçu par une démonstration de NP-complétude à paraître (Bulteau et al., 2011).

3.1.2 Structurer l'espace des solutions

Conscient d'avoir ennuyé une partie de mon lectorat en m'éloignant de la bioinformatique, j'y reviens pas à pas, en montrant comment la combinatoire des réarrangements génomiques peut finalement être utile pour l'étude de l'évolution des génomes. Car toujours guidée par l'application qui lui a donné naissance, la combinatoire des réarrangements génomiques semble avoir, ces dernières années, acquis suffisamment de maturité pour pouvoir faire ce pour quoi elle a été imaginée.

On a remarqué assez tôt que le problème initial, trouver un scénario qui transforme un génome en un autre avec un nombre minimum de réarrangements, donnait des solutions quasiment inopérantes pour la biologie car dans la plupart des cas, le nombre de solutions optimales était si grand qu'en donner une seule apparaissait vide de sens, et les donner toutes était à la fois impossible et inutile. Les résultats ne valaient donc que pour des calculs de distance génomiques, et n'aidaient pas à comprendre les remaniements.

Plusieurs pistes ont alors été explorées, soit pour rajouter des contraintes biologiques et réduire l'espace des solutions (Bérard et al., 2005; Diekmann et al., 2007), structurer l'espace des solutions pour en donner un aperçu sans l'énumérer (Bergeron et al., 2002; Braga et al., 2008), ou échantillonner des solutions, si possible selon une loi uniforme ou au moins maîtrisée (Ajana et al., 2002; Larget et al., 2002; Durrett et al., 2004; Miklós & Tannier, 2010).

Échantillonnage presque uniforme

Je vais développer sans trop formaliser un résultat récent sur l'échantillonnage dans un espace de scénarios de réarrangements, parce qu'il me paraît être d'une part un exemple rare d'un algorithme polynomial d'échantillonnage approché dans le cadre d'une application, du moins en bioinformatique, et d'autre part une façon opérante d'étudier les types divers de mutations structurales.

La technique "Markov chain Monte Carlo" (MCMC) permet de résoudre de façon approximative des problèmes d'échantillonnage. Elle consiste à construire une chaîne de Markov dans l'espace qu'on cherche à échantillonner, et à lui donner une structure qui garantisse un état stationnaire qui corresponde à une distribution prescrite sur l'espace.

Par exemple, s'il existe un grand nombre de scénarios de réarrangements qui transforment un génome en un autre, comme c'est souvent le cas (figure

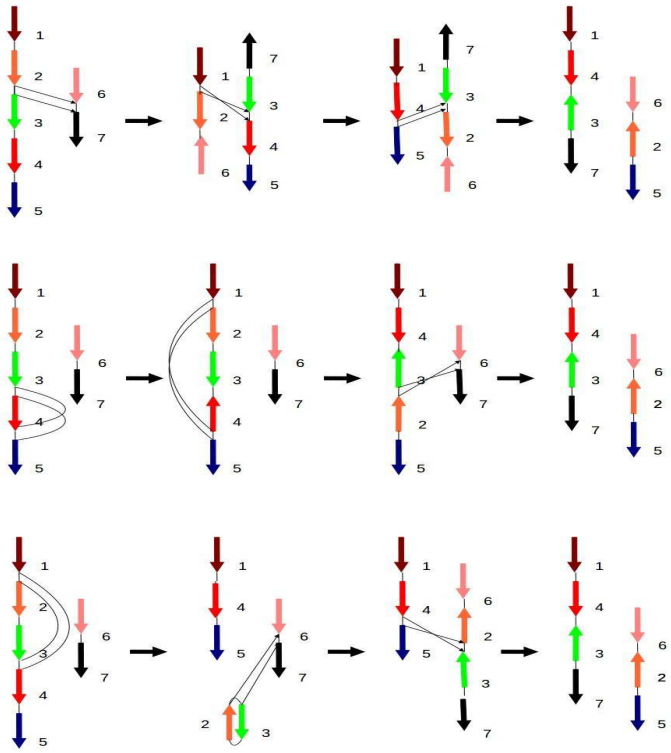


FIGURE 3.4 – Trois scénarios de réarrangements transformant le génome de gauche en le génome de droite. Les trois sont parcimonieux, mais ont des propriétés différentes. Le premier contient trois translocations réciproques, le deuxième deux inversions et une translocation réciproque, le troisième consiste en l'excision d'un segment de chromosome, et sa réinsertion dans un chromosome différent, suivis d'une translocation réciproque. L'examen d'un seul scénario ne permet donc pas de déduire un mode d'évolution.

3.4), les états de la chaîne de Markov sont les scénarios, et les transitions sont des passages d'un scénario à un autre. Si un scénario est une séquence de réarrangements successifs, une transition peut par exemple consister à modifier un petit nombre d'éléments de cette séquence (figure 3.5).

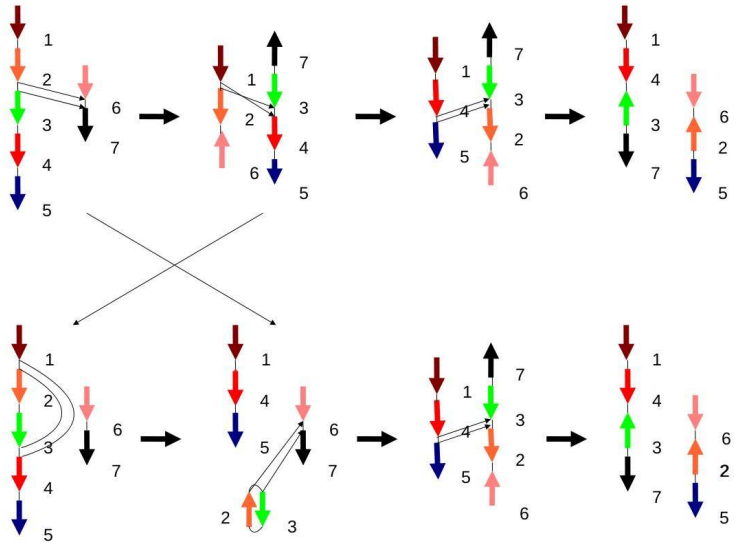


FIGURE 3.5 – Deux scénarios différents transformant le génome de gauche en le génome de droite. Le troisième réarrangement est identique dans les deux scénarios. En modifiant une partie du scénario, on peut obtenir l'autre. Cette propriété est générale : on peut transformer n'importe quel scénario en un autre avec des modifications d'un petit nombre de réarrangements.

Un théorème de Hastings (1970) garantit qu'une chaîne de Markov (ergodique) dont on note les probabilités de transition $T(x, y)$, converge vers son état stationnaire Π si on donne à l'acceptation d'un nouvel état y, venant d'un état x, la probabilité

$$P(x, y) = \min\left(1, \frac{\Pi(y)T(x, y)}{\Pi(x)T(y, x)}\right).$$

Mais la vitesse de convergence est *a priori* inconnue. Un théorème de Sinclair (1992) borne le nombre de pas à effectuer avant d'échantillonner dans un espace avec une distribution proche de Π. Il consiste à construire le

graphe de Markov $G = (V, E)$, où les sommets sont les états de la chaîne de Markov (les scénarios), et les arcs sont les transitions (transformation d'un scénario en un autre avec une petite modification). Ensuite, il faut construire un système de chemins $\Gamma(x, y)$ entre tout couple de sommets x et y. Si on peut borner

$$\kappa = max_{ab \in E} \frac{1}{\Pi(b)T(a,b)} \sum_{x,y|ab \in \Gamma(x,y)} \Pi(x)\Pi(y)|\Gamma(x,y)|$$

par une fonction polynomiale des données, alors on peut garantir que la chaîne de Markov échantillonne dans l'espace avec une distribution D telle que la distance de variation

$$dv(D, \Pi) = \sum_x \frac{1}{2}|D(x) - \Pi(x)| \leq \epsilon,$$

après avoir exploré un nombre d'état borné par un polynôme des données et de $1/\epsilon$.

Les résultats de convergence polynomiale d'une chaîne de Markov ne sont pas nombreux. Ainsi, on peut échantillonner approximativement en temps polynomial dans l'ensemble des couplages d'un graphe, dans l'ensemble des extensions linéaires d'un ordre partiel, ou estimer le permanent d'une matrice, alors que sous réserve que $P \neq NP$, il n'existe pas de solution exacte en temps polynomial aux problèmes de comptages associés à tous ces objets. Quelques autres exemples viennent compléter ce petit ensemble de classiques des mathématiques discrètes, mais il n'est pas courant de pouvoir structurer un espace plus complexe, qui viendrait de la modélisation d'un problème appliqué.

C'est pourtant le cas des réarrangements. Ou en tous cas, de ceux qu'on appelle les *doubles couper-coller*[6]. Pour décrire ce réarrangement, on considère un gène comme un arc orienté reliant deux sommets (les extrémités du gène), et un génome comme un ensemble d'adjacences (arêtes sans orientation) entre extrémités de gènes. Le graphe qui réunit les arcs et les adjacences est constitué de chemins et de cycles, chaque composante représentant un chromosome (figure 3.4). Un chromosome peut donc être circulaire ou linéaire. Un génome est dit linéaire s'il na que des chromosomes linéaires.

6. DCJ pour double cut-and-join en anglais. Je ne connais aucune traduction française, j'attends la parution d'une thèse sur le sujet pour voir si j'ai bien accordé mon mot composé.

Un *double couper-coller* sur un génome consiste en l'application d'une des opérations suivantes :

- ajouter une adjacence entre deux sommets qui n'en ont aucune ;
- effacer une adjacence ;
- effacer deux adjacences et rajouter deux autres adjacences reliant les quatre mêmes sommets ;
- effacer une adjacence et relier par une nouvelle adjacence l'un des deux sommets concernés à un sommet distinct qui n'a aucune adjacence.

Ce réarrangement, comme illustré sur la figure 3.4, peut être une inversion, une translocation réciproque, mais aussi une excision d'un chromosome, une fusion, une fission. Il existe toujours un moyen de transformer un génome en un autre qui porte les mêmes gènes avec ce type d'opération (un moyen trivial est de n'employer que les deux premières variantes, en effaçant toutes les adjacences du premier génomes puis en ajoutant toutes celles du second). La longueur de la plus petite séquence de DCJ qui transforme un génome G_1 en un autre G_2 est appelée la distance DCJ, notée $d_{DCJ}(G1, G2)$. On cherche ici à échantillonner de façon uniforme dans l'ensemble de toutes les séquences de DCJ de taille $d_{DCJ}(G_1, G_2)$.

Le réalisme de ce modèle est discutable et discuté. Dans les génomes eucaryotes, les excisions de chromosomes circulaires semblent être, au mieux, rares et suivies immédiatement de leur réinsertion dans un chromosome linéaire (un tel scénario est proposé pour l'évolution du chromosome X des mammifères par Ross et al. (2005)). En effet, comme la quasi-totalité des chromosomes sont linéaires, il ne serait pas bien vu de proposer des scénarios d'évolution où pullulent les chromosomes circulaires dans les génomes ancestraux. Dans les génomes procaryotes, si l'excision d'un plasmide n'est pas exclue, et si une minorité importante de génomes semble porter des chromosomes secondaires avec lesquels les réarrangements sont possible, ce n'est sans doute pas le cas le plus fréquent.

On peut restreindre le modèle pour le rendre plus réaliste, en définissant un *double couper-coller linéaire* sur un génome linéaire qui consisterait à ne donner en résultat que des génomes linéaires. C'est d'ailleurs une définition dont l'invention est bien antérieure à celle du cas général (Hannenhalli & Pevzner (1995) pour le cas linéaire, et Bergeron et al. (2006) pour le cas général). On pourrait aussi demander à ce que dans un scénario, un double couper-coller non linéaire soit immédiatement suivi par la réinsertion du chromosome circulaire (c'est la définition de Yancopoulos et al. (2005)). Mais c'est un peu prématuré (ce sera fait dans la section 3.2), parce que le modèle géné-

ral a des propriétés mathématiques intéressantes. Alors que certains calculs sont un peu compliqués dans le modèle linéaire, ils deviennent faisables dans le modèle général.

Le mathématicien est comme le fou qui cherche sous le lampadaire non pas parce que ce qu'il y a à trouver s'y trouve, mais parce que c'est éclairé. Mais la différence avec le fou, ce n'est pas qu'il n'est pas fou, mais que sa recherche peut être fructueuse. Il peut paraître déraisonnable de s'intéresser à un modèle moins réaliste mais plus facile d'accès. C'est pourtant cette recherche prospective qui engendrera parfois des résultats. C'est le cas ici. Si le but est d'échantillonner dans l'espace des scénarios vraisemblables, c'est à dire linéaires, un premier pas peut consister à échantillonner dans l'espace des scénarios invraisemblables. L'utilisation de ce premier pas pour faire les suivants est décrit dans la section 3.2. Ici, je donne le résultat sur le cas général des doubles couper-coller.

Théorème 1 (Miklós & Tannier (2011)) *Pour deux génomes G_1 et G_2 définis sur les mêmes gènes, et un petit réel ϵ, il existe une chaîne de Markov qui, en un nombre de pas borné par un polynôme du nombre des gènes et de $1/\epsilon$, échantillonne dans l'espace des scénarios de doubles couper-coller transformant G_1 en G_2 selon une distribution dont la distance de variation à la distribution uniforme est plus petite que ϵ.*

Autrement dit, on peut échantillonner (et par conséquent compter) dans l'espace des scénarios de réarrangements avec une bonne approximation et un temps d'exécution raisonnable. La chaîne de Markov qui permet de prouver ce résultat est relativement complexe (Miklós & Tannier, 2011). Avec István Miklós, nous pensons qu'une plus simple marche aussi bien, mais sans pouvoir le prouver. Soit la chaîne de Markov M définie par la transition qui consiste à effacer deux réarrangements consécutifs dans un scénario, et à les remplacer par deux autres.

Conjecture 1 (Miklós & Tannier (2011)) *Le théorème précédent est vrai pour la chaîne M.*

Quoiqu'il en soit, nous disposons maintenant d'une méthode pour échantillonner dans l'espace des scénarios de doubles couper-coller. C'est un résultat théorique intéressant, et il reste à faire ce pas vers la biologie plusieurs fois promis. C'est l'objet de la section 3.2.

3.2 Modèles probabilistes et échantillonnage statistique

Il est frappant que bien que leur découverte soit plus ancienne que celle des mutations locales, les mutations structurales n'ont que peu de cadre mathématique propre à les étudier quantitativement. Les premiers modèles qui ne soient pas seulement fondés sur la parcimonie et un comptage des opérations sont très récents, et considèrent l'inversion comme seul mécanisme d'évolution (Larget et al., 2002, 2005a,b) et tous les génomes comme circulaires. Les transpositions sont intégrées dans des modèles plus récents (Miklós, 2003), ainsi que les génomes linéaires et les translocations réciproques (Durrett et al., 2004).

Les possibilités computationnelles de toutes ces études sont limitées. Celles-ci sont appliquées à des génomes mitochondriaux ou chloroplastiques (Larget et al., 2002, 2005a,b), des virus (Larget et al., 2005a), ou à des clades restreints de bactéries (Darling et al., 2008). Néanmoins, les premiers résultats quantitatifs sur la pertinence des solutions parcimonieuses sont ainsi établis (Durrett et al., 2004). On peut étudier la fréquence des inversions symétriques autour de l'origine de réplication (Darling et al., 2008), ce qui était impossible par l'obtention d'un seul scénario parcimonieux. Des distributions maîtrisées d'organisations de génomes ancestraux peuvent être calculées.

Le programme de recherche pour méthodologistes proposé par Coghlan et al. (2005) commence ici à enregistrer certains résultats. Il s'agit de donner une base quantitative à l'observation de modes différents d'évolution selon les domaines ou royaumes du vivant : les génomes de plantes subissent beaucoup de duplications, ceux des mouches peu de translocations,...

J'ai participé à l'élaboration d'un échantillonneur bayésien dans l'espace des scénarios de réarrangements comprenant des inversions, translocations, fusions et fissions, capable de comparer deux génomes de plusieurs centaines de marqueurs (Miklós & Tannier, 2010). Si ce type d'algorithme est possible aujourd'hui, c'est en partie grâce à la recherche sur la combinatoire des réarrangements génomiques. En effet, elle utilise l'échantillonneur efficace sur les DCJ décrit à la section 3.1.2. Puis une technique d'élimination des scénarios peu crédibles (trop de chromosomes circulaires) par "tempérance parallèle" (parallel tempering). C'est une illustration de l'utilité de la recherche sous les lampadaires, un mauvais modèle pouvant être tout de même bon à étudier s'il a de bonnes propriétés mathématiques. Quelques propriétés théoriques

de l'échantillonneur dans l'espace des scénarios réalistes ont aussi été obtenues grâce à des connaissances sur la combinatoire des réarrangements, et en particulier ses relations avec la théorie des graphes (Miklós & Tannier, 2010).

L'échantillonneur reste pour l'instant limité à la comparaison de deux génomes. On peut alors tirer profit des méthodes de cartographie exposées au chapitre 2. En reconstruisant un génome ancestral, on peut comparer les modes d'évolution dans plusieurs lignées.

Par exemple, en triant différents types de réarrangements dans les scénarios entre

- *Saccharomyces cerevisiae* et son ancêtre suivant immédiatement la duplication globale de son génome, il y a une centaine de millions d'années ;
- *Saccharomyces kluyveri* et l'ancêtre décrit par la figure 2.9 ;
- *Homo sapiens* et son ancêtre proto-euarchontoglire (l'ancêtre de l'humain et de la souris, qui a une structure identique à l'ancêtre boreoeuthérien de la figure 2.5, aucune différence d'organisation n'a été reportée entre ces deux ancêtres sans doute séparés par un petit intervalle de temps) ;
- *Mus musculus* et son ancêtre proto-euarchontoglire,

on obtient les résultats reportés dans la table 3.1. Les chiffres les plus marquants sont ceux qui différencient l'évolution des génomes dans les branches humaine et murine depuis leur ancêtre commun. Il semble que le ratio inversions/translocations réciproques soit dans un cas très en faveur des inversions, dans l'autre des translocations.

	Fus.	Fis.	Rec. Transl.	Transl.	Rev.
anc. levure → Scerevisiae	0	0	86.5	6.7	61.3
anc. levure → Skluyveri	0	0	66.9	5.2	17.6
anc. euarch → Human	3.3	0.3	7.8	10.9	61.5
anc. euarch → Mouse	6.6	0.6	112.9	53.4	34.5

TABLE 3.1 – Nombres moyens de réarrangements selon leurs types : fusions, fissions, translocations réciproques, translocations non réciproques (ou télomériques), inversions.

Ces résultats restent à valider, il faudra les comparer avec des expériences indépendantes, tester la validité de cette approche, mais il semble qu'elle puisse déjà détecter des modes d'évolution différents.

Ces caractéristiques pourront par la suite être intégrées aux modèles en donnant des probabilités différentes aux types de remaniements, et en autorisant des paramètres différents sur chaque branche.

3.3 Distribution des cassures

3.3.1 Réutilisation des points

Compter les réarrangements selon leur type n'est pas la seule connaissance qu'on peut tirer des échantillons de scénarios. Par exemple, on peut aussi observer la distribution de la taille des inversions, de leur positionnement autour de l'origine de réplication chez les bactéries, ou d'une origine chez les eucaryotes.

Un autre sujet très débattu sur lesquels les scénarios de remaniements peuvent apporter des informations est celui de la fragilité de certaines régions génomiques, d'une propension différentielle le long du génome à subir des remaniements.

Une petite controverse scientifique a par exemple porté sur le calcul du "taux de réutilisation des points de cassure" dans les génomes de mammifères, les parties prenant ou non ce chiffre pour preuve d'une existence de régions fragiles. Ce qu'on appelle *points de cassure* ce sont les parties du génome qui ne sont pas dans les marqueurs utilisés pour les comparer. Ce sont donc les régions où on a identifié des réarrangements, un marqueur étant défini comme une région sans réarrangement. Comme le remarquent Lemaitre & Sagot (2008), ces points de cassure ne sont ni des points ni de cassure. En effet, les marqueurs ne couvrent jamais la totalité d'un génome, et on devrait plutôt parler de "région de cassure". D'autre part, cette définition étant purement comparative, on sait qu'un réarrangement est responsable de la présence de cette région dans un génome, mais sans *a priori* sur la branche sur laquelle il est arrivé. Mais le vocable est bien établi, alors continuons de l'utiliser. Le *taux de réutilisation des points de cassure* est le rapport entre le nombre de cassures minimum nécessaire pour transformer un génome en un autre au moyen de réarrangements et le nombre de points de cassures dans un génome. Les inversions et translocations réciproques cassent les génomes en deux régions, donc le nombre de cassures minimum nécessaire est supérieur au nombre de réarrangements nécessaires. Selon Pevzner & Tesler (2003), un fort taux de réutilisation des points de cassure (proche de 2) est

un argument pour l'existence de régions fragiles, dont on ne peut pourtant pas connaître la localisation. C'est le modèle *FBM*, pour "Fragile breakpoint model". Un faible taux (proche de 1) plaiderait pour une solidité uniforme des régions génomiques, et un modèle d'apparition des remaniements selon une loi uniforme, le *RBM*, pour "Random breakage model".

Il y a plusieurs objections à ce raisonnement, dont certaines ont été soulevées dans des commentaires à l'article de Pevzner & Tesler (2003).

- L'interprétation du taux dépend beaucoup de la taille des points, ou plutôt des régions de cassure. En effet, plusieurs cassures dans une même région ne contredirait pas le RBM si la région est très grande. Cette objection est prise en compte par Pevzner & Tesler (2003), qui effectuent un test dépendant de la taille moyenne des régions de cassures, en supposant qu'il est très improbable que les cassures se concentrent dans les grandes régions. Cette possibilité n'est pas testée à ce jour.
- L'interprétation du taux dépend aussi de l'hypothèse de l'inexistence de cassures à l'intérieur des marqueurs. C'est-à-dire que les petits réarrangements sont ignorés dans le calcul du taux. La fragilité ne concernerait alors que les grands réarrangements. Mais dans ce cas, l'arbitraire de la définition d'un grand réarrangement, qui n'est pas fondée sur un mécanisme biologique mais plutôt sur des limites techniques (voir la section 1.3) rend un peu incertain le résultat.
- Enfin on peut s'interroger sur le calcul du taux lui-même. Il a d'abord été calculé en supposant que tous les réarrangements cassaient deux fois le génome (Pevzner & Tesler, 2003), ce qui est incorrect pour les fusions, fissions, ou translocations et inversions télomériques. Bergeron et al. (2008) montrent que le choix du scénario est crucial : le taux peut varier du simple au double selon le scénario choisi. Ici, les méthodes d'échantillonnage décrites à la section 3.2 peuvent donner une réponse utile. Le taux de réutilisation se calcule sur un scénario. On peut faire une estimation de la moyenne sur tous les scénarios en échantillonnant (ils sont la plupart du temps beaucoup trop nombreux pour pouvoir les énumérer). Le calcul donne une estimation qui se trouve entre l'estimation de Pevzner & Tesler (2003) et celle de Bergeron et al. (2008).

La controverse n'est donc pas tout à fait éteinte, et si les première et troisième objections peuvent trouver une réponse dans un traitement statistique, la seconde renvoie à la définition d'un réarrangement, qui trouvera peut-être une réponse dans l'étude des mécanismes. S'il se confirme qu'il existe des "micro-inversions" dont le mécanisme serait différent de celui qui provoque

les grands réarrangements chromosomiques, comme semblent le penser par exemple Chaisson et al. (2006) ou Ma et al. (2008), et si une statistique correcte tenant compte de la distribution des tailles des régions de cassures est décrite, alors l'argument de Pevzner & Tesler (2003) pourra être réévalué.

3.3.2 Hétérogénéité de la distribution

En attendant, il est possible, indépendemment de tout scénario de réarrangements, d'observer la distribution des régions d'un génome cassées par des réarrangements. Ce que prédisent Pevzner & Tesler (2003), c'est que même si cette distribution est conforme à une loi uniforme le long d'un génome, la nécessité d'avoir plusieurs cassures dans des mêmes régions contredit le RBM. Mais il se trouve que chez les mammifères ladite distribution n'est pas uniforme, ou du moins est liée à l'organisation du génome en termes de contenu en gènes et de sa composition en nucléotides.

Cette corrélation est mentionnée dans plusieurs études, parmi lesquelles Mongin et al. (2009) ou Lemaitre et al. (2009b), les deux premières analyses des points de cassure basées sur des génomes complets. En effet, les régions de cassures sont sur-représentées dans les régions denses en gènes. C'est un peu inattendu, parce qu'une première intuition voudrait qu'il soit moins dangereux pour le fonctionnement d'un génome d'accumuler des mutations là où il y a moins de gènes.

Si les articles de Mongin et al. (2009) et Lemaitre et al. (2009b) sont en accord sur les observations, elles diffèrent radicalement quant à leur interprétation. Mongin et al. (2009) expliquent l'absence de cassures dans les régions peu denses en gènes par la présence de grandes régions régulatrices, une hypothèse émise par Peng et al. (2006). Les mêmes auteurs (Mongin et al., 2010) utilisent ensuite cette même absence de cassures dans certaines régions génomiques pour détecter les régions régulatrices.

Dans Lemaitre et al. (2009b), nous avons donné une explication alternative et moins sélectionniste : s'il y a plus de cassures proches des gènes, c'est que ces régions sont plus souvent actives pour le fonctionnement du génome, plus souvent transcrites, plus tôt répliquées, et cela induirait une fragilité.

Actuellement, les hypothèses sont toujours en concurrence et aucun test décisif ne les a départagées. Des données indépendantes sur la distribution des régions régulatrices pourraient faire pencher la balance d'un côté ou de l'autre, ou au moins déterminer la part des deux explications, ou d'autres, dans la distribution surprenante des régions de cassures.

Chapitre 4

Descendance et dépendances

Instiller des relations évolutives (la descendance) dans des interactions, régulations, voisinage, co-expression, co-fonctionnement (la dépendance) des gènes, ou vice-versa, utiliser des informations sur la structure et le fonctionnement d'un génome pour étudier les relations évolutives entre les organismes, sont parmi mes projets de recherche à moyen terme. Ils consistent à intégrer dans un modèle évolutif le contenu en gènes (ou autres marqueurs), et par là les possibilités métaboliques d'un organisme, la séquence protéique ou nucléotidique des gènes, et des relations synténiques ou fonctionnelles.

Le projet a été rédigé sous le nom d'*Ancestrome* (reconstruction de génomes et autres "-omes" ancestraux) par Vincent Daubin, avec plusieurs équipes de recherche françaises. Je voudrais en explorer les aspects méthodologiques.

Je résume dans ce chapitre quelques idées qui pourront faire avancer ce projet, consistant à étudier l'évolution des génomes en enrichissant progressivement la définition d'un génome.

4.1 Le génome, ensemble de gènes

Étymologiquement, un génome est un ensemble de gènes. Pris comme tel, son évolution est marquée par des mutations à l'intérieur des gènes, et des mutations qui affectent le contenu en gènes : apparitions, duplications, transferts, pertes (je condense toutes ces opérations sous le sigle "ADTP").

Dans la section 2.2.1, j'ai évoqué le calcul du contenu en gènes d'une espèce ancestrale, étant donnés un arbre phylogénétique décrivant les rela-

tions entre les organismes (ou les espèces), et un arbre pour chaque famille de gènes. Ce calcul nécessite des réconciliations de toutes les familles de gènes avec l'hypothèse phylogénétique sur les organismes. Il utilise un modèle d'évolution des gènes par ADTP (Arvestad et al., 2003; Sennblad et al., 2007; Akerborg et al., 2009; Tofigh et al., 2011) ou des cas particuliers, négligeant les transferts ou ne retenant que les solutions parcimonieuses en termes d'évènements.

Dans cette démarche, les mutations à l'intérieur de la séquence des gènes, qui permettent de faire la phylogénie des familles, sont séparées des évènements de type ADTP. C'est une méthode en deux temps : on calcule la phylogénie d'une famille de gènes, puis on réconcilie la famille étant donnée cette phylogénie.

Mais il est possible que la réconciliation elle-même soit porteuse d'informations phylogénétiques. Il serait utile de se donner la possibilité de réviser la phylogénie en fonction des réconciliations, voire de construire la phylogénie selon un modèle qui intègre les petites et les grandes mutations. Cette intégration est une entreprise de recherche actuelle. En supposant l'indépendance entre petites et grandes mutations, on peut additionner les deux modèles et rechercher les phylogénies des familles de gènes qui maximisent une vraisemblance calculée en fonction des deux. C'est ce qui est réalisé dans des études en préparation de Boussau et al. (2011), pour des modèles ADP (sans transfert de gènes) ou de Szöllősi et al. (2011), pour des modèles ADTP. Puisque les réconciliations nécessitent la connaissance d'une phylogénie des organismes, la méthode peut aussi servir à explorer différentes hypothèses phylogénétiques de ce côté-là. C'est ce qui a été réalisé pour les mammifères (Boussau et al., 2011) et les cyanobactéries (Szöllősi et al., 2011).

Ma contribution consiste principalement, conformément au thème de ce mémoire, à ajouter à ce modèle des informations de structure.

4.2 Le génome, une structure

Un génome n'est plus seulement un ensemble de gène. C'est aussi un ensemble structuré en chromosomes.

Il est théoriquement possible d'intégrer aux modèles décrits dans la section précédente (mutations locales et ADTP) un modèle d'évolution par réarrangements tels que ceux présentés dans le chapitre 3. Mais ces modèles ne sont pas encore utilisable à grande échelle. Ils sont incapables de prendre

en compte des histoires complexes de gènes, incluant des évènements de type ADTP. Les modes d'évolution ne sont pas bien connus (taux et types de réarrangements, régions fragiles,...). Et les calculs ne sont efficaces que pour deux génomes ou un très petit nombre de gènes. De leur côté, les méthodes de cartographie du chapitre 2 ne sont pas fondées sur des modèles, rendant difficile leur intégration dans un modèle évolutif plus large.

Cette intégration de données de structure dans un modèle d'évolution des génomes est un saut méthodologique qui reste à accomplir. Les études prenant en compte des informations phylogénétiques et des informations de structures ne sont pas très nombreuses. Depuis le programme a défini il y a plus de dix ans par Sankoff & El-Mabrouk (2000), les succès sont toujours rares. Lajoie et al. (2007), Bertrand et al. (2008) et Lajoie et al. (2010) ont construit des modèles d'évolution pour des groupes de gènes se dupliquant en tandem. Ma et al. (2008) utilisent une méthode de parcimonie avec des phylogénies de gènes très contraintes, interdisant les pertes par exemple, et Muffato et al. (2010) ou Bertrand et al. (2010) utilisent des phylogénie dans un processus d'inférence d'adjacences ancestrales, sans modéliser la complexité des histoires de gènes.

L'intégration de la structure à l'étude de l'évolution des génomes peut pourtant être importante pour la phylogénie. Car la synténie porte des traces des relations de parentés entre gènes. Par exemple, Jun et al. (2009) notent que la synténie est une des meilleures méthodes pour prédire l'orthologie. Wapinski et al. (2007) construisent des familles de gènes et leurs arbres phylogénétiques à partir d'informations de divergence en séquence et de synténie. Un modèle d'évolution avec des informations de synténie permettra de construire des meilleures phylogénies de gènes, et par suite des meilleures phylogénies des espèces.

Je donne dans la fin de cette section un aperçu de la possibilité de l'intégration d'informations de séquences, de contenu et de synténie. Sans avoir pour l'instant de méthode phylogénétique qui utilise toutes ces informations, on peut montrer comment la synténie peut guider des hypothèses phylogénétiques. C'est un premier pas pour son utilisation dans des modèles.

On utilise un ensemble de génomes assemblés de mammifères (figure 4.1) dont on suppose la phylogénie connue (on retire les espèces dont les relations sont controversées).

Dans tous ces génomes, les gènes sont regroupés en familles dans la base

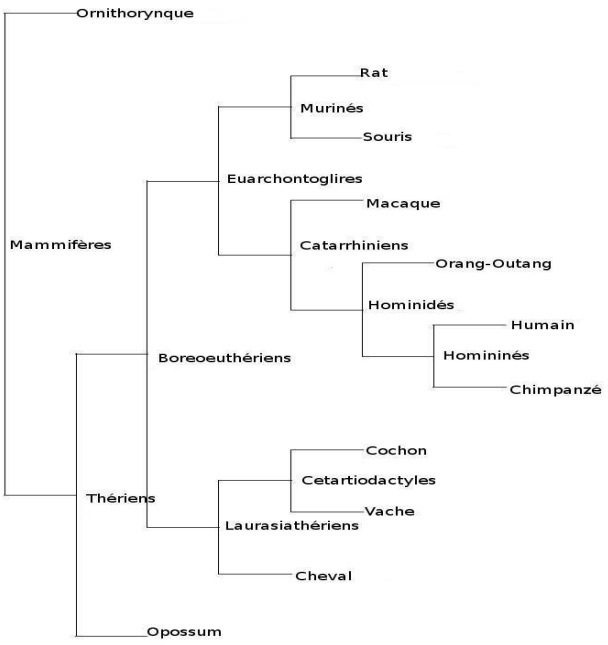


FIGURE 4.1 – L'arbre des espèces utilisé avec les noms des ancêtres annotés.

de donnée d'Ensembl (Vilella et al., 2009). Nous gardons ce découpage pour toutes les analyses, en ne conservant qu'un sous-ensemble de familles, sur des critères de longueur des alignements. Puis pour chaque famille conservée, trois arbres phylogénétiques réconciliés sont construits : le premier est celui d'Ensembl construit selon la méthode TreeBest, le deuxième est construit par PhyML (Guindon et al., 2010) et réconcilié en minimisant le nombre d'évènements de type ADP, et le troisième par PhylDOG, la méthode de Boussau et al. (2011) décrite à la section précédente, dont le principe consiste à optimiser les arbres selon un modèle incluant mutations ponctuelles, duplications et pertes de gènes.

Pour chaque ancêtre S dans la phylogénie des espèces, et chaque ensemble d'arbres phylogénétiques sur les mêmes familles de gènes, on reconstruit d'abord les gènes ancestraux selon une méthode de parcimonie. Les arbres d'Ensembl n'étant qu'approximativement réconciliés, on a besoin pour cela d'utiliser la méthode suivante sur chacun des trois jeux d'arbres : soit

un graphe G qui a pour sommets les gènes actuels et pour arêtes
- les relations d'orthologie quand elles concernent deux gènes dont l'ancêtre commun le plus récent appartient à S ou un descendant de S ;
- les relations de paralogies quand elles concernent deux gènes dont l'ancêtre commun appartient à un descendant de S (mais pas S).

Chaque composante de G est un gène dans S [1].

On dispose donc pour chaque ancêtre de la liste de ses gènes. On reconstruit ensuite certaines adjacences entre gènes d'une même espèce ancestrale par une méthode proche de celles de Bertrand et al. (2010) ou Muffato et al. (2010), plus restrictive dans sa définition pour tenter de limiter la quantité de signal conflictuel. La méthode est illustrée par le figure 4.2, qui reprend l'exemple de la figure 2.1 page 30 en ajoutant des informations de synténie.

Pour deux espèces actuelles $S1$ et $S2$, on recherche les ensembles de 4 gènes A, B, C, D tels que :
- A et B sont adjacents dans $S1$;
- C et D sont adjacents dans $S2$;
- A et C sont dans la même famille et leur dernier ancêtre commun $Anc1$ est un noeud de spéciation associé à une espèce ancestrale (un noeud de l'arbre des espèces) SA ;
- B et D sont dans la même famille et leur dernier ancêtre commun $Anc2$, distinct de $Anc1$, est un noeud de spéciation associé à SA.

Pour chaque quadruplet de gènes ainsi trouvé, on définit une adjacence ancestrale entre deux gènes ancestraux X et Y si
- $X = Anc1$ et $Y = Anc2$, ou
- X et Y appartiennent à la même espèce descendante de SA, X est un ancêtre de A (resp. C), et Y est un ancêtre de B (resp. D).

Par exemple, dans la figure 4.2, une adjacence est proposée entre les gènes Ho-1 et Ho-2, à la suite de l'examen des quatre gènes Gogo-A0101, Gogo-B0103, Popy-B02, Popy-A01.

Si les arbres et les familles sont bien construits, les gènes et adjacences ancestraux auront tendance à reconstruire les chromosomes ancestraux, ou au moins des parties, c'est-à-dire un signal linéaire. Les erreurs dans les arbres rendent certaines adjacences ancestrales impossibles à retrouver ou multiplient les adjacences et produisent des conflits entre elles : par exemple, un

1. On peut vérifier que dans le cas des arbres entièrement réconciliés, cette méthode correspond à une méthode de parcimonie qui compte le nombre de gènes que les réconciliations placent dans un noeud de l'arbre des espèces.

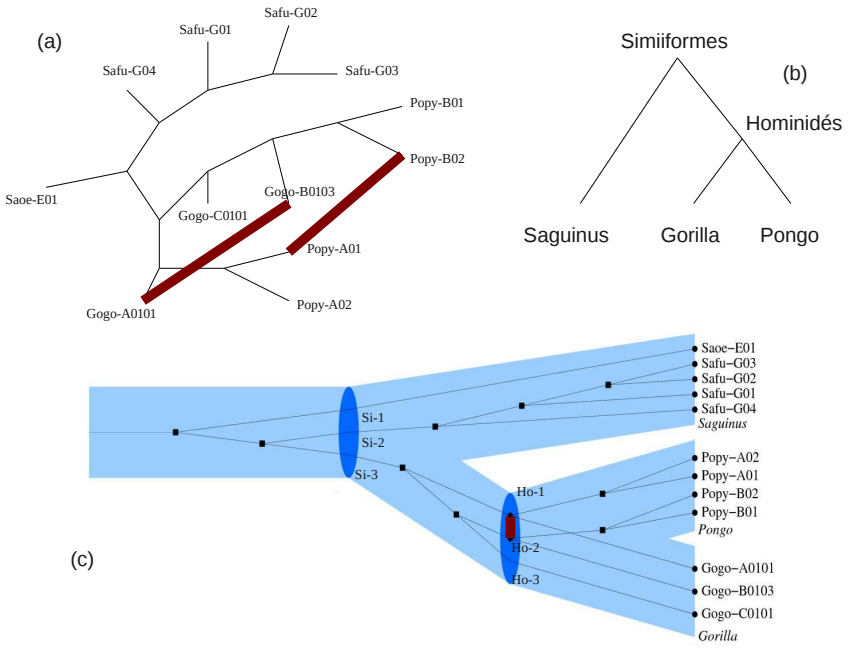


FIGURE 4.2 – Ajout de relations de dépendance au schéma de descendance de la figure 2.1 page 30. (a) Un arbre de gènes, et des adjacences entre ses feuilles. (b) Un arbre d'espèces. (c) La réconciliation de l'arbre de gènes et l'inférence d'une adjacence entre deux gènes ancestraux Ho-1 et Ho-2.

gène ancestral adjacent à plus de deux autres.

Pour évaluer la qualité d'un jeu d'arbres de gènes, on ne prend donc en compte que le signal compatible avec la linéarité attendue, et on supprime toutes les adjacences en conflit avec d'autres (c'est-à-dire toutes celles qui concernent un gène ancestral qui aurait au moins trois voisins). On compte donc le nombre d'adjacences non conflictuelles rapporté au nombre de gènes ancestraux trouvés.

Pour neuf ancêtres de la phylogénie de la figure 4.1 (il en manque un pour des raisons de place sur la page, il montre la même tendance que les autres), nous dessinons le nombre de gènes (les carrés en noir de la figure 4.3), le nombre d'adjacences non conflictuelles inférées (les carrés en rouge), et le rapport entre les deux, normalisé par le contenu maximal en gènes pour que les points apparaissent sur le même graphe (les ronds verts). Le calcul est fait avec un échantillon des familles de gènes basé sur une taille minimale des alignements de séquences, le nombre de gènes ne correspond donc pas à la totalité des gènes prédits pour une espèces ancestrale.

Si la deuxième place est disputée entre les arbres issus de PhyML et Ensembl, les arbres construits par PhylDOG, optimisés selon un modèle qui prend en compte les duplications et pertes de gènes, en plus de la séquence, donnent toujours un meilleur résultat sur ce jeu de données.

Ceci illustre la possibilité d'un modèle sur différents types de mutations des génomes. Il reste à intégrer la synténie, non plus comme un test, mais comme un moteur de la construction des arbres.

4.3 Le génome, un fonctionnement

Finalement, le génome n'est pas qu'un ensemble structuré de gènes mais c'est aussi un fonctionnement, des relations complexes entre les gènes qui peuvent aller de l'expression simultanée à la régulation, en passant par la participation à une même fonction, une participation dans un même complexe protéique, ou une relation physique élargie : synténie ou proximité dans la cellule.

La co-évolution des gènes peut refléter tous ces types de relations, et se traduire par des évolutions similaires de séquences, des duplications ou transferts simultanés (Tuller et al., 2010; Burleigh et al., 2010; Bansal et al., 2011).

On peut peut-être étendre l'étude de l'évolution des gènes à celle des

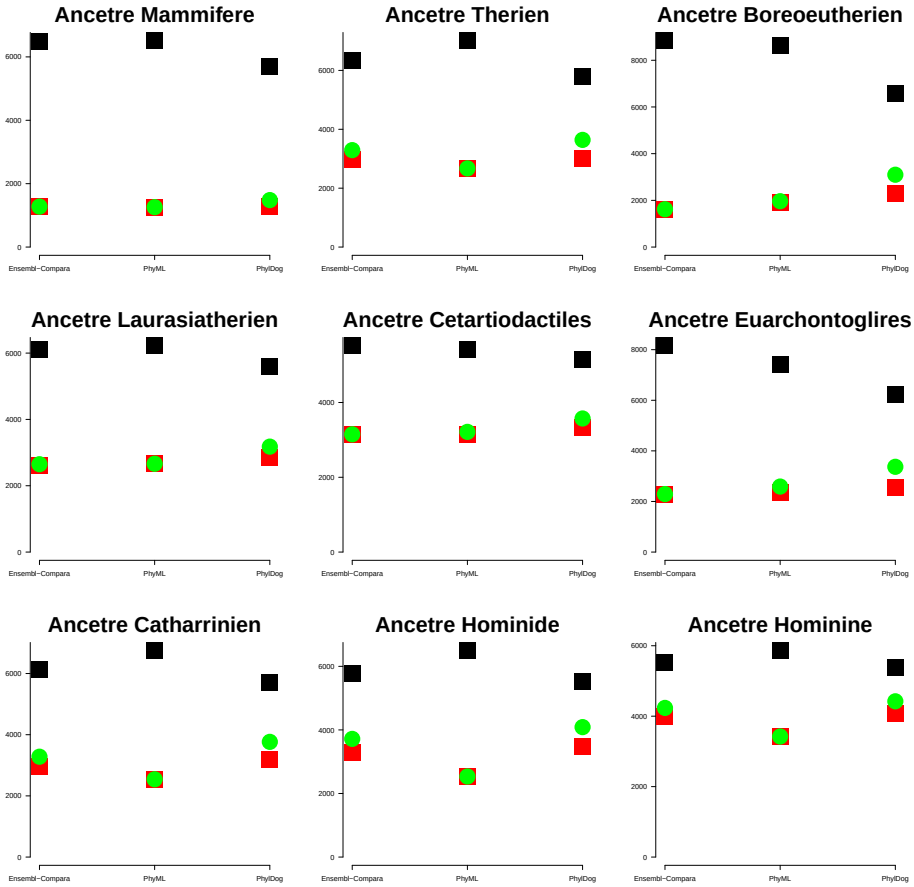


FIGURE 4.3 – En abscisse, les trois méthodes : de gauche à droite, Ensembl, PhyML, PhylDOG. En ordonnée, le nombre de gènes ancestraux (carrés noirs), le nombre d'adjacences non conflictuelles (carrés rouges), et le rapport entre les deux (ronds verts), normalisé par le nombre maximum de gènes. Chaque graphique donne les résultats pour un ancêtre dans l'arbre de la figure 4.1.

relations, voir si celles-ci évoluent, divergent, se dupliquent, quels sont les évènements qui leur sont propres (Clark et al., 2011). Pour les relations de synténie, il faudra modéliser les apparitions et disparitions d'adjacences, ainsi que leur transmission verticale au moment de la duplication d'un gène. Pour les autres relations, c'est un problème largement ouvert.

Bibliographie

Z. Adam, M. Turmel, C. Lemieux & D. Sankoff (2007). Common intervals and symmetric difference in a model-free phylogenomics, with an application to streptophyte evolution. *J Comput Biol* **14**, 436–445.

Y. Ajana, J. Lefebvre & E. Tillier (2002). Exploring the set of all minimal sequences of reversals –an application to test the replication-directed reversal hypothesis. In *Proceedings of Wabi'02*, vol. 2452 of *Lecture Notes in Computer Science*, pp. 300–315.

O. Akerborg, B. Sennblad, L. Arvestad & J. Lagergren (2009). Simultaneous bayesian gene tree reconstruction and reconciliation analysis. *Proc Natl Acad Sci U S A* **106**, 5714–5719.

M. Alekseyev & P. Pevzner (2009). Breakpoint graphs and ancestral genome reconstruction. *Genome Res* **19**, 943–957.

A. Amir & A. Levy (2010). String rearrangement metrics : A survey. In *Algorithms and Applications, essays dedicated to Esko Ukkonen on the occasion of his 60th birthday*, vol. 6060 of *Lecture Notes in Computer Science*, pp. 1–33.

A. Amir, T. Hartman, O. Kapah, A. Levy & E. Porat (2009). On the cost of interchange rearrangement in strings. *SIAM J. Comput.* **39**, 1444–1461.

L. Arvestad, A.-C. Berglund, J. Lagergren & B. Sennblad (2003). Bayesian gene/species tree reconciliation and orthology analysis using mcmc. *Bioinformatics* **19**, i7–15.

M. Badoiu, J. Chuzhoy, P. Indyk & A. Sidiropoulos (2005). Low-distortion embeddings of general metrics into the line. In *Proceeding of STOC '05*.

J. A. Bailey, Z. Gu, R. A. Clark, K. Reinert, R. V. Samonte, S. Schwartz, M. D. Adams, E. W. Myers et al. (2002). Recent segmental duplications in the human genome. *Science* **297**, 1003–1007.

M. S. Bansal, G. Banay, J. P. Gogarten & R. Shamir. (2011). Detecting highways of horizontal gene transfer. *Journal of Computational Biology* **sous presse**.

D. Bayer & P. Diaconis (1992). Trailing the dovetail shuffle to its lair. *Ann. Appl. Probability* **2**, 294–313.

S. Benzer (1959). On the topology of the genetic fine structure. *Proc Natl Acad Sci U S A* **45**, 1607–1620.

S. Bérard, A. Bergeron & C. Chauve (2005). Common structures in evolution scenarios. In *Proceedings of RECOMB-CG'04*, vol. 3388 of *Lecture Notes in BioInformatics*, pp. 1–15.

S. Bérard, A. Chateau, C. Chauve, C. Paul & E. Tannier (2008). Perfect dcj rearrangement. In *Proceedings of RECOMB-CG'08*, vol. 5267 of *Lecture Notes in Bioinformatics*, pp. 159–168.

S. Bérard, A. Chateau, C. Chauve, C. Paul & E. Tannier (2009). Computation of perfect dcj rearrangement scenarios with linear and circular chromosomes. *J Comput Biol* **16**, 1287–1309.

A. Bergeron, C. Chauve, T. Hartman & K. St-Onge (2002). On the properties of sequences of reversals thath sort a signed permutation. In *Proceedings of JOBIM'02*.

A. Bergeron, M. Blanchette, A. Chateau & C. Chauve (2004). Reconstructing ancestral gene order using conserved intervals. In *Algorithms in Bioinformatics, Proceedings of WABI'04*, vol. 3240 of *Lecture Notes in Bioinformatics*, pp. 14 – 25.

A. Bergeron, J. Mixtacki & J. Stoye (2006). A unifying view of genome rearrangements. In *Proceedings of WABI '06*, vol. 4175 of *Lecture Notes in Bioinformatics*, pp. 163–173.

A. Bergeron, J. Mixtacki & J. Stoye (2008). On computing the breakpoint reuse rate in rearrangement scenarios. In *Proceedings of RECOMB-CG 2008*, vol. 5267 of *Lecture Notes in Bioinformatics*, pp. 226–240.

M. Bernt, K.-Y. Chen, M.-C. Chen, A.-C. Chu, D. Merkle, H.-L. Wang, K.-M. Chao & M. Middendorf (2011). Finding all sorting tandem duplication random loss operations. *Journal of Discrete Algorithms* **9**, 32–48.

D. Bertrand, M. Lajoie & N. El-Mabrouk (2008). Inferring ancestral gene orders for a family of tandemly arrayed genes. *J Comput Biol* **15**, 1063–1077.

D. Bertrand, Y. Gagnon, M. Blanchette & N. El-Mabrouk (2010). Reconstruction of ancestral genome subject to whole genome duplication, speciation, rearrangement and loss. In *Algorithms in Bioinformatics, proceedings of WABI'10*, Lecture Notes in Bioinformatics. Springer.

G. Bourque & P. Pevzner (2002). Genome-scale evolution : reconstructing gene orders in the ancestral species. *Genome Res.* **12**, 26–36.

G. Bourque, P. A. Pevzner & G. Tesler (2004). Reconstructing the genomic architecture of ancestral mammals : lessons from human, mouse, and rat genomes. *Genome Res* **14**, 507–516.

G. Bourque, G. Tesler & P. A. Pevzner (2006). The convergence of cytogenetics and rearrangement-based models for ancestral genome reconstruction. *Genome Res* **16**, 311–313.

B. Boussau, G. Szollosi, L. Duret, M. Gouy, E. Tannier & V. Daubin (2011). Simultaneous inference of gene trees and organism tree in the presence of duplications and losses. En préparation.

M. D. V. Braga, M.-F. Sagot, C. Scornavacca & E. Tannier (2007). The solution space of sorting by reversals. In *Proceedings of ISBRA'07*, vol. 4463 of *Lecture Notes in Bioinformatics*, pp. 293–304.

M. D. V. Braga, M.-F. Sagot, C. Scornavacca & E. Tannier (2008). Exploring the solution space of sorting by reversals, with experiments and an application to evolution. *IEEE/ACM Trans Comput Biol Bioinform* **5**, 348–356.

L. Bulteau, G. Fertin & I. Rusu (2011). Sorting by transpositions is difficult. ArXiv :1011.1157.

J. G. Burleigh, M. S. Bansal, O. Eulenstein & T. J. Vision. (2010). Inferring species trees from gene duplication episodes. In *Proceedings of ACM-BCB*, pp. 198–203.

A.-L. Cauchy (1844). *Exercices d'analyse et de physique mathématique, tome 3*, chap. Mémoire sur les arrangements que l'on peut former avec des lettres données, et sur les permutations ou substitutions à l'aide desquelles on passe d'un arrangement à un autre, pp. 151–252. Kessinger Publishing, LLC.

A. Cayley (1849). A note on the theory of permutations. *Philosophical Magazine* **34**, 527–529.

M. J. Chaisson, B. J. Raphael & P. A. Pevzner (2006). Microinversions in mammalian evolution. *Proc Natl Acad Sci U S A* **103**, 19824–19829.

C. Chauve & E. Tannier (2008). A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes. *PLoS Comput Biol* **4**, e1000234.

C. Chauve, H. Gavranovic, A. Ouangraoua & E. Tannier (2010). Yeast ancestral genome reconstructions : the possibilities of computational methods ii. *J Comput Biol* **17**, 1097–1112.

C. Chauve, M. Jan, M. Patterson & R. Wittler (2011). Tractability results for the consecutive-ones property with multiplicity. In *Proceedings of CPM'11*.

G. W. Clark, V. Dar, A. Bezginov, J. Yang & T. E. R. M (2011). Using coevolution to predict protein-protein interactions. *Methods in Molecular Biology* **sous presse**.

A. Coghlan, E. E. Eichler, S. G. Oliver, A. H. Paterson & L. Stein (2005). Chromosome evolution in eukaryotes : a multi-kingdom perspective. *Trends Genet* **21**, 673–682.

D. Cohen & M. Blum (1995). On the problem of sorting burnt pancakes. *Discrete Applied Mathematics* **61**, 105–120.

M. Csurös & I. Miklós (2009). Streamlining and large ancestral genomes in archaea inferred with a phylogenetic birth-and-death model. *Mol Biol Evol* **26**, 2087–2095.

A. E. Darling, I. Miklós & M. A. Ragan (2008). Dynamics of genome rearrangement in bacterial populations. *PLoS Genet.* **4**, e1000128.

P. Diaconis, R. L. Graham & W. M. Kantor (1983). The mathematics of perfect shuffles. *Adv. Appl. Math.* **4**, 175–196.

Z. Dias & J. Meidanis (2002). Sorting by prefix transpositions. In *Proceedings of SPIRE'02*, vol. 2476 of *Lecture Notes in Computer Science*, pp. 65–76.

Y. Diekmann, M.-F. Sagot & E. Tannier (2007). Evolution under reversals : parsimony and conservation of common intervals. *IEEE/ACM Trans Comput Biol Bioinform* **4**, 301–309.

T. Dobzhansky & A. H. Sturtevant (1938). Inversions in the chromosomes of drosophila pseudoobscura. *Genetics* **23**, 28–64.

J.-P. Doyon, C. Scornavacca, K. Y. Gorbunov, G. Szöllősi, V. Ranwez & V. Berry (2010). An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. In *Proceedings of Recomb-CG'10*.

J. Dumas, J. Roch, E. Tannier & S. Varrette (2007). *Théorie des codes*. Dunod.

D. Durand & D. Sankoff (2003). Tests for gene clustering. *J Comput Biol* **10**, 453–482.

R. Durrett, R. Nielsen & T. L. York (2004). Bayesian estimation of genomic distance. *Genetics* **166**, 621–629.

B. Dutrillaux, J. Couturier, L. Sabatier, M. Muleris & M. Prieur (1986). Inversions in evolution of man and closely related species. *Ann Genet* **29**, 195–202.

H. Dweighter (1975). Problem e2569. *American Mathematical Monthly* **82**, 1010.

G. Fertin, A. Labarre, I. Rusu, E. Tannier & S. Vialette (2009). *Combinatorics of genome rearrangements*. MIT press.

P. Fishburn & B. Monjardet (1992). Norbert wiener on the theory of measurement(1914, 1915, 1921). *Journal of Mathematical Psychology* **36**, 165–184.

L. Froenicke, J. Wienberg, G. Stone, L. Adams & R. Stanyon (2003). Towards the delineation of the ancestral eutherian genome organization : comparative genome maps of human and the African elephant (loxodonta africana) generated by chromsome painting. *Proc Royal Soc London* **270**, 1331–1340.

L. Froenicke, M. G. Caldés, A. Graphodatsky, S. Mueller, L. Lyons, T. Robinson, M. Volleth, F. Yang et al. (2006). Are molecular cytogenetics and bioinformatics suggesting diverging models of ancestral mammalian genomes ? *Genome Res.* **16**, 306–310.

D. R. Fulkerson & O. A. Gross (1965). Incidence matrices and interval graphs. *Pacific J. Math* **15**, 835–855.

H. Gavranović & E. Tannier (2010). Guided genome halving : provably optimal solutions provide good insights into the preduplication ancestral genome of saccharomyces cerevisiae. *Pac Symp Biocomput* **15**, 21–30.

H. Gavranović, C. Chauve, J. Salse & E. Tannier (2011). Mapping ancestral genomes with massive gene loss : a matrix sandwich problem. Bioinformatics, in press.

V. Goidts, J. M. Szamalek, P. J. de Jong, D. N. Cooper, N. Chuzhanova, H. Hameister & H. Kehrer-Sawatzki (2005). Independent intrachromosomal recombination events underlie the pericentric inversions of chimpanzee and gorilla chromosomes homologous to human chromosome 16. *Genome Res* **15**, 1232–1242.

C. Grinstead & J. Snell (2006). *Introduction to Probability.* American Mathematical Society.

S. Grusea (2010). Measures for the exceptionality of gene order in conserved genomic regions. *Advances in Applied Mathematics* **45(3)**, 359–372.

S. Guindon, J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk & O. Gascuel (2010). New algorithms and methods to estimate maximum-likelihood phylogenies : assessing the performance of phyml 3.0. *Syst Biol* **59**, 307–321.

G. Hajos (1957). Uber ein art von graphen. *Internat. Math. Nachr.* **2**, 65.

J. Haldane (1919). The combination of linkage values, and the calculation of distances between the loci of linked factors. *J. Genet.* **8**, 299–309.

M. Hallett, J. Lagergren & A. Tofigh (2004). Simultaneous identification of duplications and lateral transfers. In *Proceedings of RECOMB'04*.

S. Hannenhalli & P. Pevzner (1995). Transforming men into mice (polynomial algorithm for genomic distance problem). In *36th Annual Symposium on Foundations of Computer Science, IEEE Comput. Soc. Press, Los Alamitos, CA*, pp. 581–592.

S. Hannenhalli & P. Pevzner (1999). Transforming cabbage into turnip : polynomial algorithm for sorting signed permutations by reversals. *J. ACM* **46**, 1–27.

T. Hartman & E. Verbin (2006). Matrix tightness : A linear-algebraic framework for sorting by transpositions. In *Proceedings of SPIRE'06*, vol. 4209 of *Lecture Notes in Computer Science*, pp. 279–290.

W. Hastings (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika* **57**, 97–109.

W.-L. Hsu & R. M. McConnell (2003). PC trees and circular-ones arrangements. *Theoretical computer science* **296**, 99–116.

O. Jaillon, J.-M. Aury, F. Brunet, J.-L. Petit, N. S.-T. n, E. Mauceli, L. Bouneau, C. Fischer et al. (2004). Genome duplication in the teleost fish tetraodon nigroviridis reveals the early vertebrate proto-karyoty pe. *Nature* **431**, 946–957.

G. Jean, D. Sherman & M. Nikolski (2009). Mining the semantics of genome superblocks to infer ancestral architechtures. *J. Comput. Biol.* **16**, 1267–1284.

M. Jerrum (1985). The complexity of finding minimum-length generator sequences. *Theoretical Computer Science* **36**, 265–289.

J. Jun, I. I. Mandoiu & C. E. Nelson (2009). Identification of mammalian orthologs using local synteny. *BMC Genomics* **10**, 630.

R. Karp (1993). Mapping the genome : Some combinatorial problems arising in molecular biology. In *Proceedings of STOC'93*, pp. 278–285.

H. Kehrer-Sawatzki, C. A. Sandig, V. Goidts & H. Hameister (2005). Breakpoint analysis of the pericentric inversion between chimpanzee chromosome 10 and the homologous chromosome 12 in humans. *Cytogenet Genome Res* **108**, 91–97.

M. Kellis, B. Birren & E. S. Lander (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast *saccharomyces cerevisiae*. *Nature* **428**, 617–624.

D. G. Kendall (1969). Some problems and methods in statistical archaeology. *World Archaeology* **1**, 68–76.

M. Kohn, J. Högel, W. Vogel, P. Minich, H. Kehrer-Sawatzki, J. Graves & H. Hameister (2006). Reconstruction of a 450-my-old ancestral vertebrate protokaryotype. *Trends Genet* **22**, 203–210.

M. Lajoie, D. Bertrand, N. El-Mabrouk & O. Gascuel (2007). Duplication and inversion history of a tandemly repeated genes family. *J Comput Biol* **14**, 462–478.

M. Lajoie, D. Bertrand & N. El-Mabrouk (2010). Inferring the evolutionary history of gene clusters from phylogenetic and gene order data. *Mol Biol Evol* **27**, 761–772.

B. Larget, D. Simon & J. Kadane (2002). On a bayesian approach to phylogenetic inference from animal mitochondrial genome arrangements. *Journal of the Royal Statistical Society B* **64**, 681–693.

B. Larget, J. B. Kadane & D. L. Simon (2005a). A bayesian approach to the estimation of ancestral genome arrangements. *Mol Phylogenet Evol* **36**, 214–223.

B. Larget, D. L. Simon, J. B. Kadane & D. Sweet (2005b). A bayesian analysis of metazoan mitochondrial genome arrangements. *Mol Biol Evol* **22**, 486–495.

C. Ledergerber & C. Dessimoz (2008). Alignments with non-overlapping moves, inversions and tandem duplications in o(n4) time. *Journal of Combinatorial Optimization* **16 :3**, 263–278.

C. Lemaitre & M. Sagot (2008). A small trip in the untranquil world of genomes a survey on the detection and analysis of genome rearrangement breakpoints,. *Theoretical Computer Science* **395**, 171–192.

C. Lemaitre, M. D. V. Braga, C. Gautier, M.-F. Sagot, E. Tannier & G. A. B. Marais (2009a). Footprints of inversions at present and past pseudoautosomal boundaries in human sex chromosomes. *Genome Biol Evol* **1**, 56–66.

C. Lemaitre, L. Zaghloul, M.-F. Sagot, C. Gautier, A. Arneodo, E. Tannier & B. Audit (2009b). Analysis of fine-scale mammalian evolutionary breakpoints provides new insight into their relation to genome organisation. *BMC Genomics* **10**, 335.

A. Lempel (1975). Matrix factorization of symmetric matrices and trace-orthogonal bases of $GF(2^n)$. *SIAM J. Comput.* **4**, 175–186.

R. Lenne, C. Solnon, T. Stützle, E. Tannier & M. Birattari (2008). Reactive stochastic local search algorithms for the genomic median problem. In *Proceedings of Evocop'08*, vol. 4972 of *Lecture Notes in Computer Science*, pp. 266–276.

J. Ma, L. Zhang, B. Suh, B. Rany, R. Burhans, W. Kent, M. Blanchette, D. Haussler et al. (2006). Reconstructing contiguous regions of an ancestral genome. *Genome Res.* **16**, 1557–1565.

J. Ma, A. Ratan, B. J. Raney, B. B. Suh, L. Zhang, W. Miller & D. Haussler (2008). Dupcar : reconstructing contiguous ancestral regions with duplications. *J Comput Biol* **15**, 1007–1027.

B. McArdle & A. G. Rodrigo (1994). Estimating the ancestral states of a continuous-valued character using squared-change parsimony : An analytical solution. *Syst. Biol.* **43(4)**, 573–578.

I. Miklós (2003). MCMC genome rearrangement. *Bioinformatics* **19**, ii130–ii137.

I. Miklós & E. Tannier (2010). Bayesian sampling of genomic rearrangement scenarios via double cut and join. *Bioinformatics* **26**, 3012–3019.

I. Miklós & E. Tannier (2011). Approximating the number of double cut-an-join scenarios. Soumis.

W. Miller (2001). Comparison of genomic dna sequences : solved and unsolved problems. *Bioinformatics* **17**, 391–397.

B. G. Mirkin & S. N. Rodin (1984). *Graphs and Genes*, vol. 11 of *Biomathematics*. Springer.

E. Mongin, K. Dewar & M. Blanchette (2009). Long-range regulation is a major driving force in maintaining genome integrity. *BMC Evol Biol* **9**, 203.

E. Mongin, K. Dewar & M. Blanchette (2010). Mapping association between long-range cis-regulatory regions and their target genes using comparative genomics. In *Proceedings of RECOMB-CG'10*, vol. 6398 of *Lecture Notes in Computer Science*, pp. 216–227.

T. H. Morgan (1926). *The theory of the gene*. Yale University Press.

M. Muffato & H. Roest-Crollius (2008). Paleogenomics, or the recovery of lost genomes from the mist of times. *BioEssays* **30**, 122–134.

M. Muffato, A. Louis, C.-E. Poisnel & H. R. Crollius (2010). Genomicus : a database and a browser to study gene synteny in modern and ancestral genomes. *Bioinformatics* **26**, 1119–1121.

F. Murat, J.-H. Xu, E. Tannier, M. Abrouk, N. Guilhot, C. Pont, J. Messing & J. Salse (2010). Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. *Genome Res* **20**, 1545–1557.

Y. Nakatani, H. Takeda & S. Morishita (2007). Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res* **17**, 1254–1265.

A. Ouangraoua, F. Boyer, E. Tannier & C. Chauve (2009). Prediction of contiguous ancestral regions in the amniote ancestral genome. In *Proceedings if ISBRA'09*, vol. 5542 of *Lecture Notes in Computer Science*, pp. 173–185.

M. Ozery-Flato & R. Shamir (2008). Sorting cancer karyotypes by elementary operations. In *Proceedings of the 6th RECOMB Satellite Workshop on Comparative Genomics 2008*, vol. 5267 of *LNCS*, pp. 211–225.

B. Paten, J. Herrero, K. Beal, S. Fitzgerald & E. Birney (2008). Enredo and Pecan : genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res* **18**, 1814–1828.

Q. Peng, P. A. Pevzner & G. Tesler (2006). The fragile breakage versus random breakage models of chromosome evolution. *PLoS Comput Biol* **2**, e14.

P. Pevzner & G. Tesler (2003). Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc Natl Acad Sci U S A* **100**, 7672–7677.

N. Putnam, T. Butts, D. Ferrier, R. Furlong, U. Hellsten, T. Kawashima, M. Robinson-Rechavi, E. Shoguchi et al. (2008). The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**, 1064–1071.

N. H. Putnam, M. Srivastava, U. Hellsten, B. Dirks, J. Chapman, A. Salamov, A. Terry, H. Shapiro et al. (2007). Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* **317**, 86–94.

N. Raghupathy & D. Durand (2009). Gene cluster statistics with gene families. *Molecular Biology and Evolution* **26**, 957–968.

N. Raghupathy, R. Hoberman & . D Durand (2008). Two plus two does not equal three : Statistical tests for multiple genome comparison. *Journal of Bioinformatics and Computational Biology* **6(1)**, 1–22.

B. J. Raphael & P. A. Pevzner (2004). Reconstructing tumor amplisomes. *Bioinformatics* **20**, i265–i273.

B. J. Raphael, S. Volik, C. Collins & P. A. Pevzner (2003). Reconstructing tumor genome architectures. *Bioinformatics* **19**, ii162–ii171.

B. Rhead, D. Karolchik, R. M. Kuhn, A. S. Hinrichs, A. S. Zweig, P. A. Fujita, M. Diekhans, K. E. Smith et al. (2010). The ucsc genome browser database : update 2010. *Nucleic Acids Res* **38**, D613–D619.

F. Richard, M. Lombard & B. Dutrillaux (2003). Reconstruction of the ancestral karyotype of eutherian mammals. *Chromosome Res* **11**, 605–618.

L. H. Rieseberg (2001). Chromosomal rearrangements and speciation. *Trends Ecol Evol* **16**, 351–358.

M. Rocchi, N. Archidiacono & R. Stanyon (2006). Ancestral genome reconstruction : An integrated, multi-disciplinary approach is needed. *Genome Res.* **16**, 1441 – 1444.

M. T. Ross, D. V. Grafham, A. J. Coffey, S. Scherer, K. McLay, D. Muzny, M. Platzer, G. R. Howell et al. (2005). The DNA sequence of the human X chromosome. *Nature* **434**, 325–337.

M. Sagot & E. Tannier (2005). Perfect sorting by reversals. In *Proceedings of COCOON'05*, vol. 3595 of *Lecture Notes in Computer Science*, pp. 42–51.

D. Sankoff (1989). Mechanisms of genome evolution : models and inference. *Bulletin of the International Statistical Institute* **47**, 461–475.

D. Sankoff (2006). The signal in the genomes. *PLoS Comput Biol* **2**, e35.

D. Sankoff & N. El-Mabrouk (2000). Duplication, rearrangement and reconciliation. In *Comparative Genomics : Empirical and Analytical Approaches to Gene Order Dynamics, Map alignment and the Evolution of Gene Families* (D. Sankoff & J. H. Nadeau, eds.), vol. 1 of *Computational Biology*. Kluwer Academic Press.

D. Sankoff & P. Trinh (2005). Chromosomal breakpoint reuse in genome sequence rearrangement. *J Comput Biol* **12**, 812–821.

D. Sankoff, C. Zheng & Q. Zhu (2010). The collapse of gene complement following whole genome duplication. *BMC Genomics* **11**, 313.

Schoeniger & M. Waterman (1992). A local algorithm for DNA sequence alignment with inversions. *Bull. Math. Biol.* **54**, 521–536.

B. Sennblad, E. Schreil, A.-C. B. Sonnhammer, J. Lagergren & L. Arvestad (2007). Primetv : a viewer for reconciled trees. *BMC Bioinformatics* **8**, 148.

G. Seroussi & A. Lempel (1980). Factorization of symmetric matrices and trace-orthogonal bases in finite fields. *SIAM J. Comput.* **9**, 758–767.

B. Servin, S. de Givry & T. Faraut (2010). Statistical confidence measures for genome maps : application to the validation of genome assemblies. *Bioinformatics* **26**, 3035–3042.

A. Sinclair (1992). Improved bounds for mixing rates of markov chains and multicommodity flow. *Combinatorics, Probability and Computing* **1**, 351–370.

J. Stoye & R. Wittler (2009). A unified approach for reconstructing ancient gene clusters. *IEEE/ACM Trans Comput Biol Bioinform* **6**, 387–400.

A. H. Sturtevant (1913). The linear arrangement of six sex-linked factors in drosophila as shown by their mode of association. *Journal of Experimental Zoology* **14**, 43–59.

A. H. Sturtevant (1917). Genetic factors affecting the strength of linkage in drosophila. *Proc Natl Acad Sci U S A* **3**, 555–558.

A. H. Sturtevant (1921). A case of rearrangement of genes in drosophila. *Proc Natl Acad Sci U S A* **7**, 235–237.

A. H. Sturtevant & E. Novitski (1941). The homologies of the chromosome elements in the genus drosophila. *Genetics* **26**, 517–541.

A. H. Sturtevant, C. B. Bridges & T. H. Morgan (1919). The spatial relations of genes. *Proc Natl Acad Sci U S A* **5**, 168–173.

G. Szöllősi, B. Boussau, E. Tannier & V. Daubin (2011). Using a probabilistic model of duplication, transfer and loss to infer the phylogenetic tree of cyanobacterial genomes. En préparation.

E. Tannier (2008). Sorting signed permutations by reversal (reversal sequence). In *Encyclopedia of Algorithms* (M.-Y. Kao, ed.). Springer.

E. Tannier (2009). Yeast ancestral genome reconstruction : the possibilities of computational methods. In *Proceedings of RECOMB-CG'09*, vol. 5817 of *Lecture Notes in Computer Science*, pp. 1–12.

E. Tannier (2010). Le génome aux ordres des mathématiciens. *La Recherche* **439**, 54–56.

E. Tannier & M.-F. Sagot (2004). Sorting by reversals in subquadratic times. In *Proceedings of CPM'04*, vol. 3109 of *Lecture Notes in Computer Science*, pp. 1–13.

E. Tannier, A. Bergeron & M. Sagot (2007). Advances on sorting by reversals. *Discrete Applied Mathematics* **155**, 881–888.

E. Tannier, C. Zheng & D. Sankoff (2008). Multichromosomal genome median and halving problems. In *Proceedings of WABI'08*, vol. 5251 of *Lecture notes in Bioinformatics*, pp. 1–13.

E. Tannier, C. Zheng & D. Sankoff (2009). Multichromosomal median and halving problems under different genomic distances. *BMC Bioinformatics* **10**, 120.

A. Tofigh, M. Hallett & J. Lagergren (2011). Simultaneous identification of duplications and lateral gene transfers. *IEEE/ACM Trans Comput Biol Bioinform* **8**, 517–535.

T. Tuller, H. Birin, U. Gophna, M. Kupiec & E. Ruppin (2010). Reconstructing ancestral gene content by coevolution. *Genome Res* **20**, 122–132.

A. F. Vellozo, C. E. Alves & A. P. do Lago (2006). Alignment with non-overlapping inversions in o(n3)-time. In *Algorithms in Bioinformatics, Proceedings of WABI'06*, vol. 4175 of *LNCS*, pp. 186–196.

A. J. Vilella, J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin & E. Birney (2009). Ensemblcompara genetrees : Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* **19**, 327–335.

A. Wald & J. Wolfowitz (1940). On a test whether two samples are from the same population. *Ann. Math Statist* **11**, 147–162.

I. Wapinski, A. Pfeffer, N. Friedman & A. Regev (2007). Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics* **23**, i549–i558.

J. D. Watson & F. H. Crick (1953). Molecular structure of nucleic acids ; a structure for deoxyribose nucleic acid. *Nature* **171**, 737–738.

G. A. Watterson, W. J. Ewens, T. E. Hall & A. Morgan (1982). The chromosome inversion problem. *Journal of Theoretical Biology* **99**, 1–7.

J. Wienberg (2004). The evolution of eutherian chromosomes. *Curr Opin Genet Dev* **14**, 657–666.

R. Wittler & J. Stoye (2010). Consistency of sequence-based gene clusters. In *Proceedings of RECOMB Comparative Genomics*, vol. 6398 of *LNBI*, pp. 252–263.

S. Yancopoulos, O. Attie & R. Friedberg (2005). Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics* **21**, 3340–3346.

F. Yang, E. Alkalaeva, P. Perelman, A. Pardini, W. Harrison, P. O'Brien, B. Fu, A. Graphodatsky et al. (2003). Reciprocal chromsome painting among human, aardvark, and elephant (superorder afrotheria) reveals the likely eutherian ancestral karyotype. *PNAS* **100**, 1062–1066.

H. Zhao & G. Bourque (2009). Recovering genome rearrangements in the mammalian phylogeny. *Genome Res.* **19**, 934–942.

C. M. Zmasek & S. R. Eddy (2001). A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics* **17**, 821–828.

Table des matières