Université Bordeaux I
Laboratoire Bordelais de Recherche en Informatique

Habilitation à diriger des Recherches
Specialité Informatique

par

Cédric Chauve

À la recherche des génomes perdus : Modèles, Méthodologies et Algorithmes pour la Comparaison de Génomes

Habilitation à Diriger des Recherches défendue le 20 juin 2011.

Après avis de

M. Alain Denise	Professeur	Université Paris XI
M. Éric Rivals	Directeur de Recherches	CNRS (UMR 5506)
M. David Sankoff	Professeur	Université d'Ottawa

Devant le jury composé de

M. Alain Denise	Professeur	Université Paris XI	Rapporteur
M. Serge Dulucq	Professeur	Université Bordeaux I	Examinateur
M. Guillaume Fertin	Professeur	Université de Nantes	Examinateur
M. Guy Mélançon	Professeur	Université Bordeaux I	Président
M. Éric Rivals	Directeur de Recherches	CNRS (UMR 5506)	Rapporteur

À Marni, Madeleine et Éléanor.

Remerciements

Je ne peux commencer ces remerciements autrement qu'en exprimant toute ma reconnaissance et mon admiration à Anne Bergeron. Je ne trouve pas de mots appropriés pour décrire ce que je lui dois, mais je sais que son influence sur les travaux décrits dans ce mémoire, et sur ma pratique scientifique en général, est incommensurable.

Je suis très honoré que David Sankoff ait accepté de rapporter ce mémoire. Ses travaux fondateurs sur l'analyse algorithmique des réarrangements génomiques sont à l'origine de la plupart des questions abordées dans ce mémoire.

Serge Dulucq m'a fait découvrir le monde de la recherche. Je lui suis reconnaissant de m'avoir donné cette chance et d'avoir guidé mes premiers pas, qui se sont poursuivis jusqu'à cette habilitation. Je suis reconnaissant à Éric Rivals et Alain Denise d'avoir accepté d'écrire, dans des délais assez courts, un rapport sur mon mémoire et de participer à la soutenance. Le parcours d'Alain, de la combinatoire bordelaise vers la bioinformatique, a été un modèle encourageant quand j'ai décidé de suivre le même cheminement, et je suis sensible au regard bienveillant qu'il porte sur mes travaux. Au moment de choisir mon sujet de recherche en DEA, j'ai longuement hésité à travailler sous la direction de Guy Mélançon et je suis honoré qu'il préside aujourd'hui ce jury d'habilitation. Avec Guillaume Fertin, nous avons débuté dans la recherche en bioinformatique ensemble. Je suis heureux que nos chemins se croisent de nouveau durant cette soutenance.

Les travaux présentés dans ce mémoire résultent de collaboration avec de nombreux collègues et étudiants. Je pense avoir été extrèmement chanceux d'avoir pu travailler avec autant de collaborateurs et je les remercie tous chaleureusement.

J'ai passé cinq années merveilleuses au LaCIM. L'ambiance y fut toujours amicale et stimulante. J'en remercie chaleureusement tous les membres et je ne peux m'empêcher d'avoir une pensée émue pour Pierre Leroux.

Une des raisons qui m'ont poussées à entreprendre la rédaction d'une habilitation à diriger des recherches était la perspective de la soutenir au LaBRI, notamment devant les membres de l'équipe de combinatoire énumérative. Avec un peu de recul, je m'estime privilégié de l'éducation scientifique que j'ai reçue durant mes années de doctorat passées dans cette équipe et je suis heureux de pouvoir exprimer ma reconnaissance à ses membres.

Je profite maintenant de cette conclusion de mon parcours universitaire pour rappeler à mes parents, à mon frère et à sa petite famille combien leur présence est importante et combien j'apprécie leur soutien sans faille.

Finalement, je n'aurais rien fait de tout cela sans la présence à mes côtés de Marni, sans ses encouragements constants, sa disponibilité permanente et sa patience. Je mesure aussi les sacrifices qu'elle a consentis pour me permettre de mener à bien ces travaux, tout en fondant une famille formidable. Je ne peux conclure ces remerciements qu'en lui exprimant avec amour toute ma gratitude.

Table des matières

Introduction

1

1.1 Introduction générale

Ce mémoire est avant tout destiné à convaincre un jury de mon aptitude à diriger des recherches en informatique. Pour ce faire, j'ai choisi de présenter, de manière synthétique, une partie des résultats que j'ai obtenus, durant les dix dernières années, dans le domaine de la *biologie computationnelle*[1]. En écrivant ce mémoire, mon intention n'est donc pas de présenter dans le détail l'ensemble de ces résultats, mais d'en donner un aperçu relativement non-technique, et surtout de replacer ces recherches dans le contexte plus large de la biologie computationnelle. Je m'autoriserai donc des digressions non directement reliées à mes résultats, souvent plus longues que la description de ces derniers, mais qui me semblent pertinentes pour les replacer dans leur contexte. J'espère ainsi convaincre un lecteur non-spécialiste, disposant cependant d'un bagage minimal en algorithmique et mathématiques discrètes, de l'intérêt des problèmes abordés et de la pertinence des éléments de réponse apportés.

Pour commencer très généralement, je rappellerai le but de la biologie computationnelle : répondre, partiellement la plupart du temps, à des questions biologiques en utilisant des méthodes informatiques. La motivation des mes travaux est donc biologique, bien que je ne sois pas biologiste de formation. La question de biologie qui m'a intéressé est la suivante :

> Étant donné un ensemble de génomes, descendant tous d'un génome ancestral commun, quels événements évolutifs peuvent expliquer les différences structurelles entre ces génomes ?

Pour justifier l'intérêt de cette question, je me contenterai de citer Theodosius Dobzhansky, dont le livre fondateur *Genetics and the Origin of Species* (1937) a initié plus de soixante dix ans de recherches intenses en génétique et génomique évolutives : "Rien en biologie n'a de sens, si ce n'est à la lumière de l'évolution".

[1] Traduction littérale de l'anglais *computational biology*, en cours au Québec, et caractérisant mieux la nature de ma recherche que le terme *bio-informatique*, d'usage plus courant en France.

D'un point de vue scientifique, cette question est extraordinairement ambitieuse. Par définition, l'évolution est passée et ne peut plus être observée. En conséquence, aucun élément de réponse ne peut être validé avec certitude. Même avec les immenses progrès de la génomique expérimentale, on peut au mieux espérer démontrer et reproduire en laboratoire des mécanismes d'évolution de certains génomes. Mais toute hypothèse sur la nature des événements évolutifs qui ont façonné les génomes actuels ne peut résulter que d'un processus d'*inférence* basé sur l'analyse de données génomiques.

Longtemps le domaine réservé de chercheurs biologistes, le domaine de la génomique évolutive a été investi depuis une cinquantaine d'années par un nombre croissant de mathématiciens[2], participant ainsi au développement de la biologie computationnelle. On peut expliquer ce phénomène relativement récent en partie par des développements technologiques majeurs, à la fois en génomique et en informatique. Dans le domaine de la génomique, les progrès des techniques de séquençage ont permis une croissance exponentielle de la masse de données disponibles, à l'échelle de génomes complets notamment, mises à la disposition de la communauté scientifique dans des bases de données à accès non restreint. De par leur masse et leur nature numérique, il est naturel de traiter ces données en utilisant des moyens informatiques, et donc des méthodes et modèles mathématiques et algorithmiques. Les progrès des capacités de calcul ont permis à la fois de traiter ces masses de données importantes et de développer des modèles mathématiques complexes. De plus, l'intérêt porté par les mathématiciens aux problèmes de biologie computationnelle, a permis le développement rapide d'un corpus d'algorithmes efficaces dédiés spécifiquement au traitement des données génomiques.

Dans une telle approche, le rôle des mathématiciens est fondamental. En effet, pour produire des éléments de réponse pertinents pour une question biologique donnée, dans le cadre d'un processus d'inférence informatique, il est nécessaire que le coeur de la "machine à inférer", à savoir les modèles mathématiques et les algorithmes, soit adapté à la nature des données fournies en entrée et à la question biologique initiale. Il faut donc sans cesse chercher à trouver un point d'équilibre entre les trois aspects de la modélisation, du développement théorique et de l'implémentation et application aux données. Cela n'interdit pas de se laisser entraîner sur des chemins de traverse, comme l'étude d'un jeu de données particulièrement intéressant qui incite à laisser quelque peu de côté les aspects plus formels, ou encore de questions mathématiques non directement reliées à l'analyse de données biologiques. Au contraire, ces exercices sont souvent nécessaires pour une compréhension plus fine des aspects appliqués ou des propriétés mathématiques sous-jacentes aux algorithmes et de leur prises en compte dans l'analyse de données réelles. Il s'agit là d'un rôle fondamental dévolu aux mathématiciens, qui peut prendre plusieurs formes, comme la validation (théorique, statistique, par simulations) de méthodes et modèles, l'analyse de la complexité et de l'exactitude d'algorithmes, la description explicite d'un cadre théorique général. Ce travail peut aussi participer à l'interprétation biologique de résultats informatiques, comme l'illustre la controverse en cours sur la réutilisation des régions de cassures durant l'évolution et la notion de région génomique fra-

[2]Le terme *mathématicien* est ici pris au sens large, et inclut les algorithmiciens et informaticiens théoriciens.

gile (artefact des méthodes d'analyse des réarrangements génomiques ou propriété génomique fondamentale de l'évolution des génomes ?).

J'ai pleinement conscience que les résultats que je présente dans ce mémoire sont loin d'avoir atteint le point d'équilibre décrit plus haut et sont fortement orientés vers des aspects théoriques. Mes premiers résultats notamment sont purement théoriques. J'ai cependant investi beaucoup de temps dans l'apprentissage de la biologie et de la génomique, et ma recherche s'est sensiblement orientée ces dernières années vers des aspects plus méthodologiques et appliqués. J'espère convaincre le lecteur, et le jury, d'avoir progressé vers une recherche plus équilibrée, et d'être apte à diriger des recherches de qualité en biologie computationnelle dans les années à venir.

1.2 Organisation du mémoire.

Ce mémoire se compose de deux parties. La partie principale comporte cinq chapitres consacrés à l'étude des *réarrangements génomiques*. La seconde partie se compose d'un appendice contenant un chapitre portant sur un sujet quelque peu différent de l'analyse de génomes complets, mais connexe : l'évolution des familles de gènes. Pour des raisons de cohérence du document, j'ai occulté mes quelques résultats sur la comparaison de structures secondaires d'ARN.

Le chapitre 2 contient un rapide survol des différents mécanismes d'évolution des génomes et de la notion de réarrangement génomique. On introduit ensuite dans le chapitre 3 différentes notions de structures combinatoires conservées entre génomes, avec une emphase sur la notion d'intervalles communs qui est centrale dans les problèmes et résultats présentés dans les deux chapitres suivants. En effet, le fil directeur de la plupart des résultats présentés dans les chapitres 4 et 5 est d'utiliser les structures génomiques potentiellement conservées durant l'évolution comme données additionnelles dans certains problèmes de génomique évolutive, comme le calcul des distances génomiques ou la reconstruction de génomes ancestraux.

Dans le chapitre 4, on s'intéresse à des algorithmes de calcul de *distances génomiques* et de *scénario d'évolution par réarrangements génomiques* entre une paire de génomes. Historiquement, il s'agit de mes premiers travaux de recherche en biologie computationnelle, dans un domaine très actif à Montréal, sous l'impulsion notamment de David Sankoff. Traditionnellement, le calcul de scénario d'évolution se traduit algorithmiquement par des questions d'optimisation combinatoire centrées sur un critère de *parcimonie*. La première partie de ce chapitre survole cette approche, laquelle a fait l'objet de très nombreux travaux théoriques et appliqués, et fait maintenant partie de la boîte à outils classique de la génomique comparée. Dans une seconde partie, on s'attarde plus en détail sur les algorithmes de calcul de *scénarios parfaits*. La notion de scénario parfait repose sur le principe qu'un intervalle commun à deux génomes doit être conservé durant l'évolution depuis leur plus proche ancêtre commun. On recherche alors à calculer des scénarios/distances qui respectent le critère de parcimonie parmi les scénarios qui ne cassent aucun intervalle commun. Les résultats présentés,

obtenus en collaboration avec Anne Bergeron, Séverine Bérard, Annie Chateau et Christophe Paul notamment, sont de nature essentiellement algorithmique, mais mènent naturellement aux travaux décrits dans le chapitre suivant.

L'idée que la comparaison de paires de génomes permet de détecter des structures potentiellement présentes dans leur plus proche ancêtre commun soulève naturellement la question suivante : étant donné un ensemble de génomes, peut-on inférer, à partir de leurs structures conservées, l'organisation du génome de leur plus proche ancêtre commun ? Cette question peut être vue comme une tentative d'inférer l'organisation de génomes ancestraux en se basant sur un modèle d'évolution minimal, non lié à un modèle de réarrangements génomiques. Depuis 2007, j'ai consacré la majeure partie de mon temps à l'étude de cette question de *paléogénomique*, dans le cadre d'une collaboration fructueuse avec Éric Tannier et Aïda Ouangraoua notamment. Je présente nos travaux sur ce sujet dans le chapitre 5. Notre approche repose sur le principe que la reconstruction de génomes ancestraux est, du point de vue méthodologique, un problème similaire à l'*assemblage* (au sens large, incluant la *cartographie*) de génomes.

Cette partie principale se termine par un court chapitre de conclusion.

En appendice, on trouvera de plus un chapitre décrivant mes travaux sur l'évolution des familles de gènes.

Génomes et réarrangements génomiques

2

Ce chapitre porte principalement sur la représentation combinatoire des génomes et de l'évolution de génomes par réarrangements génomiques. Je renvoie le lecteur à [Graur 2000] entre autres pour une description plus détaillée des aspects moléculaires de l'évolution des génomes.

2.1 Introduction

Le génome est un ensemble de molécules d'ADN qui contient l'information génétique présente dans la cellule. Il est en général composé de plusieurs paires de *chromosomes*, chaque chromosome étant constitué de de deux *brins* d'ADN, adoptant une structure de double hélice. Chaque brin d'ADN est une chaîne de *nucléotides* (A pour adénine, C pour cytosine, G pour guanine et T pour thymine). Ces chromosomes sont porteurs des *gènes*, qui encodent pour la plupart l'information nécessaire à la synthèse de *protéines*, mais contiennent aussi de nombreuses séquences d'ADN qui ont une fonction biologique sans pour autant encoder des protéines, que l'on appelle les *séquences fonctionnelles non codantes*.

Les génomes présents dans les cellules d'un organisme subissent constamment des modifications, du fait de processus extérieurs, comme l'irradiation par exemple, ou de processus internes à la cellule. Ces changements sont en général réparés par divers mécanismes biochimiques, mais certains échappent à leur vigilance et, dans de très rares cas, se transmettent à la descendance. L'accumulation de tels changements peut alors mener à un phénomène de *spéciation*, c'est-à-dire d'apparition d'une nouvelle espèce. Il s'agit là d'une vision simpliste de l'évolution, qui ignore par exemple les variations génomiques intra-espèces, mais qui est adaptée à la nature des questions biologiques abordées dans ce mémoire. Les modifications d'un génome sont variées. À très petite échelle, on trouve les *mutations ponctuelles*, qui modifient très légèrement la séquence d'ADN qui forme un chromosome en modifiant, supprimant ou insérant un ou quelques nucléotides (*indels*) ou les répétitions de petites séquences comme les micro-satellites. A l'extrême opposé du spectre, certains changements sont dramatiques, comme la *duplication de génome complet*. Ce dernier mécanisme appartient à la classe des *réarrangements génomiques* qui modifient le nombre de chromosomes d'un génome (son

caryotype) ou l'organisation d'un ou plusieurs de ses chromosomes. Contrairement aux mutations ponctuelles et indels, est aussi un les réarrangements génomiques sont des événements rares et qui modifient profondément la structure des génomes. L'évolution par réarrangements génomiques est au coeur des problèmes et résultats décrits dans la suite de ce document.

2.2 Les génomes comme objets combinatoires

Le principe général de représentation combinatoire d'un génome qu'on suivra dans le reste du document est le suivant (voir Figure 2.1 pour une illustration) :

> Un génome est un *multi-ensemble* de chromosomes ; un chromosome est une *séquence signée* de *marqueurs génomiques*.

Il est nécessaire d'utiliser la notion de multi-ensemble pour prendre en compte un génome qui vient de subir une duplication (de gène, chromosome ou de génome complet). Dans la suite de ce document, nous parlerons cependant d'ensemble pour alléger quelque peu l'écriture. Un chromosome étant égal à son inverse (au sens des inversions définies dans la suite de ce chapitre), plusieurs représentations équivalentes d'un même génome existent. Cela n'a cependant pas d'impact en général sur les problèmes et résultats présentés.

La notion de marqueur génomique sera décrite plus précisément par la suite, et peut être appréhendée premièrement d'un point de vue combinatoire : un marqueur est un nombre entier (positif ou négatif) dans une séquence d'entiers signés. Si l'on s'intéresse par exemple aux gènes et à l'évolution de leur position le long des chromosomes, chaque marqueur représente un gène : l'étiquette du marqueur encode la *famille de gènes* auquel ce gène appartient (le génome humain comporte quelques milliers de familles de gènes) et le signe encode le brin d'ADN qui porte ce gène. Les travaux décrits dans ce mémoire prennent plutôt le point de vue génique, et analysent des génomes représentés par des séquences dont la longueur est d'au plus quelques milliers d'éléments, définies sur un alphabet de quelques milliers de marqueurs.

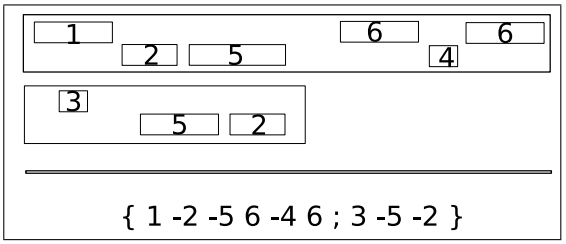

FIG. 2.1 – Exemple de représentation d'un génome par des séquences signées. Les boîtes étiquetées par des nombres représentent des marqueurs, leur position, haute ou basse, indiquant sur quel brin d'ADN le marqueur se trouve. Le génome comporte deux chromosomes et neuf marqueurs, appartenant à six familles.

Finalement, il convient d'étendre cette notion de représentation combinatoire d'un génome au cas d'un *ensemble de génomes*. En effet, les travaux décrits dans les chapitres suivants protent essentiellement sur la *comparaison de deux ou plusieurs génomes*.

Il existe d'autres modèles de représentation de génomes, adaptés au cas où le résultat du processus de *séquençage* est ambigu ou incomplet. Une approche naturelle consiste à traduire l'incertitude regardant l'organisation d'un génome en l'encodant par des structures qui représentent un *ensemble de génomes*. On peut par exemple utiliser des séquences (non signées) si la position des marqueurs sur les brins d'ADN n'est pas connue, ou des *ensembles de marqueurs* si la position des marqueurs le long des chromosomes n'est pas connue [Ferretti 1996]. Dans le cas de génomes obtenus par cartographie physique, le *PQ-arbre*, qui sera présenté dans le chapitre suivant, est une structure combinatoire adaptée à inclure une certains dose d'ambiguïté sur la position des marqueurs le long des chromosomes [Alizadeh 1995]. Il est aussi possible que plusieurs cartes physiques d'un même génome ne soient pas compatible avec un ordre linéaire des gènes ou marqueurs. On peut alors utiliser une structure d'*ordre partiel* pour représenter ces cartes [Zheng 2005a, Zheng 2006, Blin 2007a] et évaluer le degré d'incertitude par le nombre de permutations signées encodées par de telles structures.

Dans les chapitres suivants, nous supposerons que les génomes étudiés sont sans ambiguïtés, mais nous utiliserons le modèle des PQ-arbres pour représenter des génomes ancestraux ambigus.

2.3 Réarrangements génomiques

Les réarrangements génomiques sont des événements évolutifs qui modifient la structure des chromosomes d'un génome. On distingue deux classes de réarrangements. Les réarrangements *équilibrés* ne modifient pas le contenu en marqueurs d'un génome mais modifient l'organisation des marqueurs le long des chromosomes, et parfois le caryotype (nombre de chromosomes). Les réarrangements *déséquilibrés* résultent en l'ajout ou la suppression de marqueurs. On s'intéresse ici aux aspects combinatoires (voir [Lemaitre 2008b, Lemaitre 2008a] par exemple pour un survol accessible des principaux mécanismes moléculaires des réarrangements génomiques).

Les réarrangements équilibrés. Les réarrangements équilibrés qui modifient le nombre de chromosomes sont la *fusion* et la *fission*. La fusion réunit deux chromosomes en joignant deux de leurs *télomères*[1] respectifs. La fission est l'opération inverse, qui coupe un chromosome en deux.

Les principaux réarrangements équilibrés qui préservent le caryotype sont l'*inversion*, la *translocation* (souvent appelée *translocation réciproque*) et la *transposition*. L'inversion

[1] Le terme télomère désigne une extrémité de chromosome.

est un réarrangement *intra-chromosomique*, qui n'affecte donc qu'un seul chromosome. Elle consiste à inverser l'ordre des marqueurs d'un segment de chromosome (un *intervalle*) tout en changeant le signe de chaque marqueur de ce segment. La translocation est un réarrangement *inter-chromosomique* dans lequel deux chromosomes échangent du matériel génomique et plus précisément des segments télomériques (un par chromosome). Finalement la transposition se caractérise par le déplacement d'un segment de génome (là encore un intervalle d'une séquence encodant un chromosome) vers une autre position, qui peut être sur son chromosome d'origine, ou sur un autre chromosome.

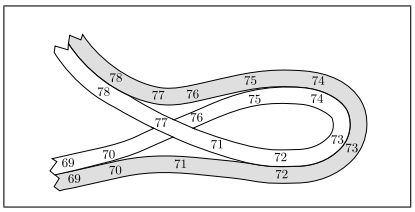

FIG. 2.2 – Inversion observée par en 1938 par Dobzhansky et Sturtevant en comparant les génomes de deux espèces de drosophiles. Le segment chromosomique gris est encodé par la séquence 69 70 71 72 73 74 75 76 77 78 et le segment blanc par 69 70 -76 -75 -74 -73 -72 -71 77 78. Adapté de [Dobzhansky 1938].

Les réarrangements déséquilibrés. Les réarrangements déséquilibrés modifient le *contenu en marqueurs* d'un génome. Les deux réarrangements déséquilibrés que nous allons prendre en compte sont la *suppression* de marqueurs et la *duplication* de marqueurs. Une suppression concerne en général un segment chromosomique et résulte en la disparition des marqueurs localisés sur ce segment. Il existe plusieurs types de duplications, dont les trois principales sont les *duplications en tandem*, les *duplications segmentales*, et les *duplications de génome complet*. Une duplication en tandem recopie un segment de génome à sa suite, résultant initialement en deux copies similaires contiguës. Une duplication segmentale recopie un segment de génome de manière non-contiguë. Finalement une duplication de génome complet résulte en une duplication de tous les chromosomes. Il s'agit d'un événement évolutif majeur, qui a sans doute joué un rôle moteur dans l'apparition des animaux vertébrés [Ohno 1970].

Points de cassure. La plupart des ces réarrangements modifient les chromosomes d'un génome en cassant ce génome aux extrémités d'un ou deux segments génomiques et en réparant ces cassures. Les zones où ces cassures se produisent sont appelées *points de cassures* (*breakpoints* en anglais) et sont l'objet de nombreuses études qui tentent de comprendre si ces zones ont des propriétés qui favorisent les cassures (voir [Sankoff 2009] pour une revue récente).

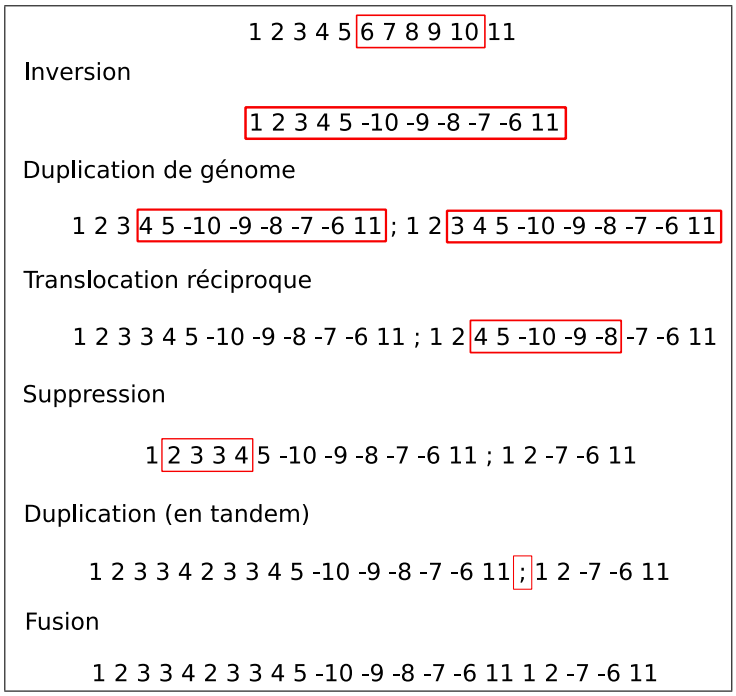

FIG. 2.3 – Exemple d'évolution par réarrangements génomiques. Le point de départ est un génome à un chromosome et onze gènes appartenant à onze familles distinctes. Les segments concernés par les réarrangements sont encadrés.

2.4 Marqueurs génomiques

Dans cette section, nous présentons rapidement les différents types de marqueurs génomiques (*gènes homologues* et *blocs homologues*[2]) et les propriétés, combinatoires et évolutives de ces marqueurs.

On note $\{G_1, \ldots, G_k\}$ un ensemble de k génomes et $\mathcal{A}$ leur plus proche ancêtre commun. Étant donné un alphabet $\mathcal{M}$ de marqueurs, on représente chaque génome par un ensemble de séquences signées sur $\mathcal{M}$.

Propriétés générales. Un jeu de marqueurs génomiques doit premièrement satisfaire deux propriétés naturelles : un marqueur représente un segment de chromosome défini par ses *coordonnées génomiques* de début et de fin et le brin portant ce segment (indiqué par son signe, $+$ ou $-$), et deux marqueurs situés sur un même chromosome doivent être disjoints (un acide nucléique donné ne peut appartenir à deux marqueurs). De plus, on suppose que l'ordre des marqueurs dans la séquence signée représentant un chromosome respecte l'ordre des segments génomiques qu'ils représentent le long de ce chromosome.

[2]En général appelés *blocs de synténie*, ce qui repose sur une interprétation incorrecte du terme "synténie" [Passarge 1999].

Si les propriétés purement combinatoires énoncées ci-dessus sont aisées à vérifier, il n'en est est pas de même des propriétés *évolutives* qu'un jeu de marqueurs idéal doit satisfaire. Premièrement, les marqueurs de même étiquette $x \in \mathcal{M}$ dans $\{G_1, \ldots, G_k\}$ (i.e. toutes les occurrences de x ou $-x$ dans ces génomes) représentent des segments génomiques qui *descendent* d'un unique[3] segment $\mathcal{A}_x$ du génome ancestral $\mathcal{A}$. Deuxièmement, pour assurer la cohérence avec les propriétés des marqueurs dans les génomes étudiés, on suppose que pour tout $x, y \in \mathcal{M}$, $x \neq y$, les segments $\mathcal{A}_x$ et $\mathcal{A}_y$ sont disjoints. Ces conditions assurent par exemple que dans un scénario d'évolution tous les génomes intermédiaires sont bien définis.

Une famille de marqueurs qui est présente dans chacun des k génomes $\{G_1, \ldots, G_k\}$ est qualifiée d'*universelle*. Une famille de marqueurs est *unique* si chaque génome contient au plus une occurrence de cette famille. Un génome représenté avec des marqueurs uniques et universels peut être vu comme une sorte de *permutation signée*, si ce n'est que cette permutation peut être séparée en plusieurs morceaux (les chromosomes). Ces notions sont importantes car certains problèmes algorithmiques de réarrangements génomiques sont difficiles (i.e. NP-complets ou NP-difficiles) si les données ne sont pas basées sur des marqueurs uniques et universels, mais peuvent être résolus en temps polynomial dans le cas contraire [Fertin 2009].

Caractères homologues et orthologues. Un ensemble de caractères génomiques (gènes par exemple) qui ont évolué à partir d'un caractère ancestral commun par des événements de spéciation et duplication forment une *famille homologue*. Deux membres de cette famille sont *orthologues* si ils vérifient une propriété évolutives simple : ils descendent de leur plus proche caractère ancestral commun via un événement de *spéciation* (voir le chapitre A en appendice pour une présentation plus détaillée de l'évolution des familles de gènes). La notion de caractères homologues ou orthologues ne se limite pas aux gènes et s'applique à toute famille de segments génomiques, ou plus généralement à toute famille d'objets biologiques dont l'évolution peut se décrire en termes de spéciation et duplication (comme par exemple les interactions entre protéines [Pinney 2007]).

Les familles de gènes homologues. Utiliser comme marqueurs génomiques les familles de gènes homologues est une approche naturelle, car ils satisfont toutes les propriété désirables des marqueurs génomiques énoncées plus haut. Il existe de plus des bases de données de familles de gènes couvrant de nombreux génomes [Vilella 2009].

Cependant, déterminer si des gènes sont homologues ou orthologues est une question difficile. Le principal problème est que la définition d'homologie/orthologie est basée sur une histoire évolutive non disponible et qu'il faut donc tenter de reconstruire. L'approche suivie en général comporte deux étapes. Premièrement, les séquences d'ADN des gènes sont comparées pour les regrouper en ensembles de séquences suffisamment similaires pour faire l'hypothèse d'un gène ancestral commun. Ces groupes forment les familles homologues. Ensuite, pour chaque famille homologue, un *arbre phylogénétique* (appelé un *arbre de gènes*) est reconstruit pour identifier les événements de spéciations et de duplications.

[3] Cette condition d'unicité peut être discutée, mais a été utilisée dans tous mes travaux.

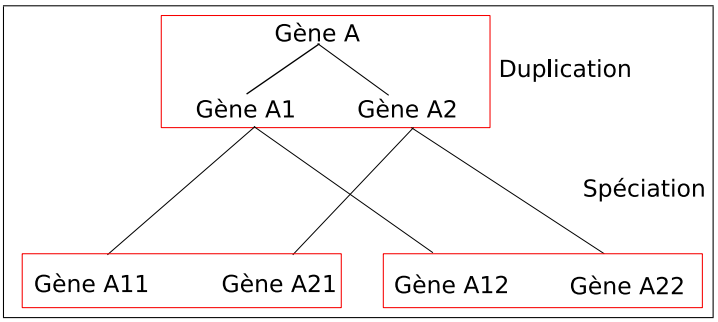

FIG. 2.4 – Évolution d'un gène A. Le caractère génomique considéré dans cet exemple est le gène A et ses descendants. Un premier événement de duplication se produit dans le génome ancestral, résultant en deux copies $A1$ et $A2$ de ce gène. Un événement de spéciation résulte en deux génomes descendants de cet ancêtre, représentés par les boîtes en bas de la figure. Le premier génome contient les copies $A11$ et $A21$ du gène, et le second génome les copies $A12$ et $A22$. $A11$ et $A12$ sont des gènes orthologues : ils divergent tous les deux du gène ancestral $A1$ via une spéciation. Pour les mêmes raisons, $A21$ et $A22$ sont orthologues via le gène $A2$. Par contre $A11$ et $A21$ ne sont pas orthologues : ils divergent du gène ancestral A via la duplication qui a mené à $A1$ et $A2$. Descendant d'un même gène ancestral, les quatre gènes $A11$, $A12$, $A21$ et $A22$ font partie de la même *famille de gènes*.

Il est important de noter que cette approche repose sur le principe de la détection de gènes ayant évolué à partir d'un gène ancestral commun via la similarité (directe ou transitive) de leurs séquences d'ADN. On ne peut donc pas détecter, comme membres d'une famille homologue, les gènes descendants du gène ancestral de la famille mais qui ont été modifiés durant l'évolution par des mutations ou des réarrangements internes qui empêchent de détecter une similarité de séquence significative. Ainsi, si la notion de famille homologue est évolutive, son implémentation repose essentiellement sur une notion de conservation de la structure combinatoire (ici la séquence) des membres de cette famille.

Pour conclure sur l'utilisation des gènes comme marqueurs, on peut relever un problème important, quand les génomes que l'on veut étudier sont peu denses en gènes. Par exemple, chez les mammifères, les gènes ne représentent qu'une très petite fraction des génomes (moins de 2% du génome humain). Il est alors naturel de remettre en question l'inférence d'hypothèses évolutives basée sur des marqueurs offrant une faible couverture des génomes étudiés, ou non distribués uniformément le long de ces génomes. L'utilisation de *blocs homologues* offre une solution alternative.

Les blocs homologues. Une première approche de la notion de blocs homologues consiste à appliquer la définition de familles de gènes homologues sans se limiter aux segments des génomes qui portent des gènes. Ainsi, une famille de blocs homologues est un ensemble de segments de génomes qui descendent d'un segment ancestral commun unique par des événements de spéciation et de duplication, sans réarrangements chevauchant et avec des mutations et réarrangement internes limités. Tout comme pour les gènes, on infère donc les familles de

blocs homologues en termes de similarité de séquence. Cependant, contrairement aux gènes dont les coordonnées génomiques sont connues, tout au moins pour les génomes bien annotés, les coordonnées de familles de blocs homologues ne sont pas connues a priori.

La plupart des méthodes de calcul de familles blocs orthologues reposent sur des techniques de *chaînage d'ancres*. Cette approche consiste à détecter des segments de génome homologues ou orthologues (les ancres), puis à définir des blocs en agglomérant les groupes d'ancres qui présentent une structure combinatoire *colinéaire* (c'est-à-dire similaire en termes d'ordre relatif le long des chromosomes, d'espacement et d'orientation) [Ma 2006] ou de bonnes propriétés statistiques [Hampson 2005]. La détection d'ancres génomiques est un problème difficile, qui nécessite d'aligner des génomes complets [Blanchette 2007, Paten 2008a, Dubchak 2009] et suppose une bonne conservation de la similarité des séquences génomiques durant l'évolution. Quand les génomes étudiés sont trop divergents, les séquences génomiques non-codantes voient leur similarité décroître, et on utilise alors des gènes homologues comme ancres [Simillion 2008, Murat 2010, Fostier 2011].

Discussion. Le problème du calcul des marqueurs est sans doute plus central qu'il ne le paraît. Ces marqueurs sont en effet les données de base de toute analyse d'un ensemble de génomes. Pour reprendre un adage essentiel en biologie computationnelle, *"garbage in, garbage out"* : toute erreur dans leur construction a un impact sur l'ensemble des analyses ultérieures. La construction de ces marqueurs est une étape préliminaire, souvent occultée dans les publications car relativement technique et très en amont des résultats "intéressants".

Pour certains ensembles de génomes il est bien connu que la construction de marqueurs est difficile, comme les génomes de plantes par exemple dont l'évolution comporte plusieurs duplications de génomes. Ce problème se pose aussi pour les vertébrés par exemple, du fait de l'accroissement du nombre de génomes disponibles [Consortium 2009]. En effet, l'augmentation du nombre de génomes séquencés accroît la possibilité qu'une occurrence d'un bloc soit supprimée par un réarrangement (interne ou chevauchant) spécifique à un génome. Ainsi, si il est possible de détecter des familles de blocs uniques et universels qui couvrent une grande partie (plus de 90%) d'un petit nombre de génomes de mammifères [Ma 2006, Zhao 2009], les mêmes méthodes appliquées à une dizaine de génomes de mammifères résultent en une forte chute de la couverture par des blocs uniques et universels, du fait de la disparition de chaînes d'ancres colinéaires et de longueur significative conservées dans tous les génomes.

Le calcul, la validation ou la comparaison de familles de marqueurs sont des problèmes très importants qui requièrent sans doute l'exploration de nouvelles approches [Hachiya 2009, Belcaid 2007].

2.5 Conclusion

Pour conclure ce chapitre, on peut rappeler le principe fondamental suivant : les réarrangements génomiques agissent comme des opérations d'édition sur des objets combina-

toires simples, des ensembles de permutations signées. La construction d'un jeu de marqueurs génomiques représente une première étape essentielle, qui dans les faits repose sur la détection de structures combinatoires conservées (chaînes d'ancres par exemple) dans un ensemble de génomes, soumise à des réarrangements génomiques limités. Le chapitre suivant élargit cette approche à la conservation de *groupes de marqueurs*.

Contributions. La plupart des travaux présentés dans les trois chapitres suivants ont inclus une réflexion initiale sur la modélisation des génomes, des marqueurs et des réarrangements. Concernant la notion de marqueurs génomiques, la discussion présentée dans ce mémoire constitue l'ébauche d'un texte de synthèse sur ce sujet. Ma seule contribution publiée sur ce sujet est l'article suivant.

[Belcaid 2007] M. Belcaid, A. Bergeron, A. Chateau, C. Chauve, Y. Gingras, G. Poisson et M. Vendette. *Exploring Genome Rearrangements using Virtual Hybridization.* Proceedings of 5th Asia-Pacific Bioinformatics Conference, APBC 2007, pages 205–214. Imperial College Press, 2007.

Structures génomiques conservées

3

L'évolution par réarrangements génomiques, d'un point de vue combinatoire, résulte en des modifications des ensembles de séquences signées que l'on utilise pour représenter les génomes. Cela amène naturellement à se poser la question de la détection des segments de génomes qui ne sont pas détruits par ces événements, et donc potentiellement conservés durant l'évolution. Ces structures peuvent ensuite être utilisées comme données supplémentaires pour proposer des scénarios d'évolution (chapitre 4) ou des génomes ancestraux (chapitre 5). Les principaux objets introduits dans ce chapitre[1] sont les *intervalles communs* et les *PQ-arbres*.

3.1 Intervalles communs : un point de vue combinatoire

Le modèle le plus simple de groupe de marqueurs conservé dans un ensemble de génomes repose sur une conservation dans tous les génomes du contenu en marqueurs, de l'ordre relatif des marqueurs et de l'orientation des marqueurs. Il s'agit de la notion de *segment conservé*, qui fut historiquement la première notion de structure conservée utilisée dans l'analyse de génomes (voir [Sankoff 1997, Sankoff 2005] par exemple). Un segment conservé composé de deux marqueurs est une *adjacence conservée*.

Les intervalles communs sont obtenus en supprimant les contraintes de conservation de l'ordre et des signes des segments conservés. Comme nous allons le voir, les intervalles communs de permutations ont une combinatoire très riche, que l'on peut aussi aborder via les approches plus générales des *familles faiblement partitives* et de la *décomposition modulaire de graphes* (ici de permutations) [de Montgolfier 2003, Bui-Xuan. 2008, Bui-Xuan 2005, Paul 2006, Bergeron 2008a].

Soit $\{G_1, \ldots, G_k\}$ un ensemble de génomes, représentés sur un alphabet $\mathcal{M} = \{1, 2 \ldots, n\}$ de marqueurs uniques et universels.

Définition 1 Soient $\mathcal{S} \subseteq \mathcal{M}$ un ensemble de marqueurs et C un chromosome d'un génome G.

[1] Basé en grande partie sur l'article de synthèse [Bergeron 2008b].

1. Un intervalle $C[i,j]$ de C est une *occurrence* de $\mathcal{S}$ si tous les éléments de $\mathcal{S}$ apparaissent dans $C[i,j]$. L'*empreinte* $c(C[i,j])$ de $C[i,j]$ est l'ensemble des marqueurs présents sur cet intervalle.
2. Une occurrence $C[i,j]$ de $\mathcal{S}$ sur C est sans *trou* si $c(C[i,j]) = \mathcal{S}$. Sinon, un trou est un segment maximal de $C[i,j]$ ne contenant aucun élément de $\mathcal{S}$.
3. $\mathcal{S}$ est un *intervalle commun* pour $\{G_1, \ldots, G_k\}$ si $\mathcal{S}$ a une occurrence sans trou dans chacun des G_i.
4. Les singletons $\{g\}$ sont appelés intervalles communs *triviaux*. Un intervalle commun qui n'est inclus dans aucun autre intervalle commun est *maximal*.

G1 = 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
G2 = -6 -5 4 13 14 -15 16 1 -3 9 -10 11 12 -7 8 -2
G3 = -13 -4 5 -6 -12 -8 -7 2 1 -3 9 10 11 14 -15 16

FIG. 3.1 – Intervalles communs de trois génomes (représentant les chromosomes X des génomes de l'humain (G_1), de la souris (G_2) et du rat (G_3) étudiés dans [Bourque 2004a]) composés d'un chromosome chacun. Chaque intervalle souligné dans G_3 est une occurrence d'un intervalle commun non trivial. Les intervalles communs non triviaux et non-maximaux sont donc $\{4,5\}$, $\{5,6\}$, $\{4,5,6\}$, $\{14,15\}$, $\{15,16\}$, $\{14,15,16\}$, $\{9,10\}$, $\{10,11\}$, $\{9,10,11\}$ et $\{7,8\}$. Le seul intervalle commun maximal est $\{1,2,\ldots,16\}$.

Il existe $O(n^2)$ intervalles communs pour $\{G_1, \ldots, G_k\}$, ce qui peut représenter un nombre important (n peut être de l'ordre de plusieurs milliers pour des génomes de mammifères par exemple [Zhao 2009]). Il existe cependant une structure de données permettant d'encoder l'ensemble des intervalles communs à un ensemble de génomes, le *PQ-arbre*, introduite par Booth et Lueker [Booth 1976]. On suppose dans un premier temps que $\{G_1, \ldots, G_k\}$ sont des permutations (des génomes *unichromosomaux*), dont l'une est la permutation identité $1\ 2\ \ldots\ n$. De plus, comme la notion d'intervalle commun n'utilise pas les signes, on peut se contenter de travailler avec des permutations non signées.

Définition 2
1. Deux ensembles $\mathcal{S}_1$ et $\mathcal{S}_2$ se *chevauchent* si leur intersection est non-vide et aucun n'est inclus dans l'autre.
2. Un intervalle commun de $\{G_1, \ldots, G_k\}$ est *fort* si il ne chevauche aucun autre intervalle commun.

Dans l'exemple donné en Figure 3.1, les intervalles communs forts sont les singletons, $\mathcal{M}$, $\{4,5,6\}$, $\{14,15,16\}$, $\{9,10,11\}$ et $\{7,8\}$, et ils contiennent tous les intervalles communs. Les intervalles communs forts ont deux propriétés importantes : il existe $O(n)$ intervalles communs forts, et, de par leur définition, ils peuvent être organisés en une structure d'*arbre d'inclusion*, dont on suppose les nœuds ordonnés de sorte que lire les feuilles de gauche à droite résulte en l'identité. On appelle cet arbre *l'arbre des intervalles forts*. Pour passer de l'arbre des

intervalles forts au PQ-arbre, on en partitionne les nœuds internes en *nœuds P* et *nœuds Q* de manière à respecter la propriété suivante.

Propriété 1
1. L'union d'enfants consécutifs d'un nœud Q est un intervalle commun pour $\{G_1, \ldots, G_k\}$.
2. L'union d'un sous-ensemble propre des enfants d'un nœud P n'est pas un intervalle commun pour $\{G_1, \ldots, G_k\}$.

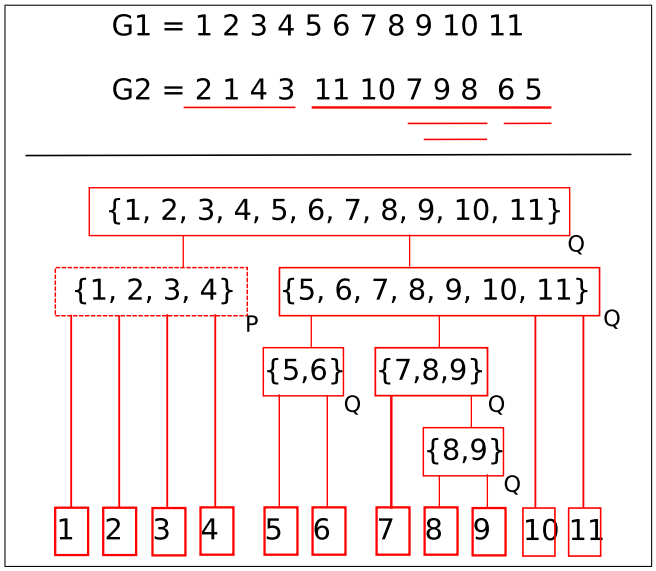

FIG. 3.2 – Le PQ-arbre des intervalles communs de deux permutations. Les intervalles forts sont soulignés dans G_2. Le seul nœud P est représenté en pointillés. Tous les autres nœuds internes sont des nœuds Q. Chaque nœud est étiqueté par l'intervalle fort correspondant. L'intervalle commun non fort $\{5, 6, 7, 8, 9\}$ est encodé par les deux premiers enfants du nœud Q $\{5, 6, 7, 8, 9, 10, 11\}$.

Il n'est pas difficile de prouver que cette partition en nœuds P et Q des nœuds internes de l'arbre des intervalles forts est bien définie et unique (elle ne dépend pas de l'ordre initial imposé aux nœuds de l'arbre des intervalles forts). De plus, comme le nombre d'intervalles forts (i.e. de nœuds internes) de cet arbre est en $O(n)$, le PQ-arbre occupe un espace en $O(n)$.

Finalement, le point 1 de la Propriété 1, conjugué à la définition des intervalles forts, permet de prouver que le PQ-arbre est une structure qui encode tous les intervalles communs pour $\{G_1, \ldots, G_k\}$. En d'autres termes, les intervalles forts forment une *base* de l'ensemble des intervalles communs.

Proposition 2 Tout intervalle commun de $\{G_1, \ldots, G_k\}$ est soit un intervalle fort, soit l'union des intervalles forts associés à un ensemble d'enfants consécutifs d'un nœud Q du PQ-arbre de $\{G_1, \ldots, G_k\}$.

Il est important de remarquer qu'un PQ-arbre T est une structure plane (i.e. l'ordre des enfants d'un nœud est important), qui représente une permutation : on note $p(T)$ la permutation obtenue en lisant ses feuilles de gauche à droite. Une propriété fondamentale du PQ-arbre associé à $\{G_1, \ldots, G_k\}$ est de représenter en fait une *classe d'équivalence de permutations*, à savoir l'ensemble des permutations ayant les mêmes intervalles communs que $\{G_1, \ldots, G_k\}$.

Définition 3
1. Deux PQ-arbres sont *équivalents* si on peut transformer le premier en le second par une série d'opérations consistant en (1) inverser un nœud Q (i.e. inverser l'intervalle fort correspondant[2]) ou (2) changer, arbitrairement, l'ordre des enfants d'un nœud P.
2. Deux permutations p_1 et p_2 sont équivalentes pour un PQ-arbre T si il existe deux arbres T_1 et T_2 équivalents à T tels que $p(T_1) = p_1$ et $p(T_2) = p_2$.

Propriété 3 Soit $\mathcal{T}$ un PQ-arbre des intervalles communs de $\{G_1, \ldots, G_k\}$. Une permutation p appartient à la classe d'équivalence des permutations associées à $\mathcal{T}$ si et seulement si tous les intervalles communs de $\{G_1, \ldots, G_k\}$ sont aussi des intervalles de p.

Finalement, pour définir le PQ-arbre de génomes multichromosomaux, il suffit de prendre en compte leurs intervalles communs maximaux : chacun définit un ensemble de k permutations, dont on peut calculer le PQ-arbre, et il ne reste plus qu'à regrouper tous ces PQ-arbres sous une racine qui est un nœud P (ou Q si il n'y a que deux intervalles communs maximaux).

3.2 Algorithmes, extensions et applications

Dans cette seconde section nous nous intéressons aux aspects algorithmiques des intervalles communs et PQ-arbres, à quelques extensions de cette notion, notamment au cas des familles de marqueurs non uniques et universels, et à des applications de la détection de structures conservées en génomique.

3.2.1 Calculer les intervalles communs et le PQ-arbre

Le premier algorithme optimal de calcul des intervalles communs d'un ensemble de permutations est dû à Uno et Yagiura [Uno 2000] : il permet de calculer les N intervalles communs de deux permutations de taille n en temps $O(n + N)$. Cependant, cet algorithme est assez difficile à implémenter car il repose sur des structures de données complexes. Heber et Stoye [Heber 2001] ont étendu cet algorithme pour calculer les intervalles communs de k permutations. L'idée principale de leur algorithme est d'utiliser une autre base de l'ensemble des

[2]Sans changer de signe car on ne prend en compte ici que des permutations non signées.

intervalles communs, les intervalles communs *irréductibles*, définis comme les intervalles communs qui ne sont *pas* l'union de deux intervalles communs chevauchants. Calculer le PQ-arbre associé à un ensemble de permutations peut se faire en adaptant les techniques initiales de *réduction* introduites par Booth et Lueker dans [Booth 1976]. Cette approche a été explorée dans [Landau 2005]. Cependant, là encore ces approches sont difficiles à implémenter, notamment car l'article initial de Booth et Lueker contient quelques erreurs dans sa description de l'opération de réduction.

Dans [Bergeron 2008a], nous avons proposé une approche radicalement différente du calcul des intervalles communs d'un ensemble de permutations. L'idée centrale consiste à les définir en termes d'une autre familles d'intervalles qui ne sont, en toute généralité, *pas* des intervalles communs, appelés les *générateurs*.

Définition 4 Un générateur des intervalles communs de $\{G_1, \ldots, G_k\}$ est une paire (R, L) de tableaux de n éléments telle que

1. $R[i] \geq i$ et $L[i] \leq i$ pour $i \in \{1, 2, \ldots, n\}$,
2. $\{i, \ldots, j\}$ est un intervalle commun pour $\{G_1, \ldots, G_k\}$ si et seulement si $\{i, \ldots, j\} = \{i, \ldots, R[i]\} \cap \{L[j], \ldots, j\}$.

Un générateur est *canonique* si tous les $\{i, \ldots, R[i]\}$ et $\{L[i], \ldots, i\}$, pour $i \in \{1, 2, \ldots, n\}$, sont des intervalles communs.

Nous avons montré que ces générateurs existent toujours, et que l'ensemble des intervalles forts définit naturellement l'unique générateur canonique. Nous avons aussi proposé des algorithmes simples (n'utilisant que des structures de données et de contrôle élémentaires, telles que boucles et tableaux) et optimaux pour calculer un générateur, les intervalles forts ou énumérer les intervalles communs à partir d'un générateur. Nous avons aussi montré comment obtenir le PQ-arbre, en temps optimal, à partir d'un générateur.

3.2.2 Autre modèles de structures conservées

L'utilisation de familles de marqueurs uniques et universels est plus l'exception que la norme en génomique comparée. Par exemple, si on veut utiliser les gènes comme marqueurs, imposer les contraintes d'unicité et d'universalité nécessite de déterminer les gènes orthologues, et donc de reconstruire des arbres phylogénétiques, un problème difficile [Felsenstein 2004], alors que n'imposer aucune de ces contraintes permet de se contenter du calcul de familles homologues. Il est alors naturel de rechercher des structures génomiques conservées dans des ensembles de séquences signées qui peuvent comporter des *marqueurs répétés* dans un même génome.

Intervalles communs de séquences. Le point (3) de la Définition 1 peut s'appliquer tel quel pour définir les intervalles communs de séquences. La principale différence conceptuelle

est qu'un intervalle commun $\mathcal{S}$ peut avoir plusieurs occurrences dans un même génome. On qualifie de *maximale* une occurrence de $\mathcal{S}$ qui n'est pas incluse dans une autre occurrence de $\mathcal{S}$. Deux occurrences maximales d'un intervalle commun ne peuvent donc pas se chevaucher.

Par exemple, si $G_1 = 6\ 5\ 3\ 2\ 5\ 2\ 4$ et $G_2 = 1\ 2\ 2\ 5\ 3\ 2\ 3\ 4\ 5\ 2\ 3$, alors $\{2, 5\}$ est un intervalle commun pour $\{G_1, G_2\}$ avec une occurrence maximale dans G_1 (le segment 2 5 2) et deux dans G_2 (les segments 2 2 5 et 5 2).

Chaque segment des génomes $\{G_1, \ldots, G_k\}$ peut être occurrence maximale d'au plus un intervalle commun, et le nombre maximal d'intervalles communs (et d'occurrences maximales) est donc en $O((n_1 + \cdots + n_k)^2)$, où n_i est la taille (nombre de marqueurs) de G_i. Cependant, il n'existe, à ma connaissance, aucun équivalent de bases de taille linéaire ou de structure de données permettant d'encoder les intervalles communs de séquences avec répétition en espace linéaire. Cela a pour conséquence que les algorithmes de calcul de ces intervalles communs sont radicalement différents du cas des permutations, et sont inspirés d'algorithmes mis au point pour détecter l'ensemble des empreintes (*fingerprints* en anglais) d'une séquence [Didier 2007, Schmidt 2004]. Ces techniques permettent de calculer les intervalles communs de plusieurs génomes en temps $O(N^2)$.

Équipe de gènes et occurrences inexactes. Pour conclure ce survol des modèles de structures génomiques conservées, on peut remarquer que les intervalles communs souffrent d'un défaut majeur, surtout dans le cas de familles de marqueurs répétés : ils ne permettent pas de détecter des occurrences non-exactes pouvant comprendre des trous. Une approche naturelle consiste à détecter des ensembles de marqueurs ayant des occurrences dans tous les génomes étudiés comportant des trous de longueur bornée par un paramètre δ (des δ-occurrences). De tels ensembles de marqueurs sont appelés des δ-équipes (en anglais, la dénomination courante est *max-gap clusters*). Leur étude a été initiée, dans le cas des permutations, dans [Luc 2003, Béal 2004, He 2005]. Le cadre combinatoire naturel pour ces objets est assez proche des intervalles communs de permutations et repose sur des concepts de décomposition modulaire que l'on peut appliquer en fait à diverses familles d'objets combinatoires [Habib 2004, Boyer 2005, Paul 2006]. Cependant, du fait des nombreux chevauchements possibles entre occurrences d'équipes différentes, le nombre d'équipes de gènes ou d'occurrences maximales peut croître exponentiellement avec le nombre de génomes considérés ou si les familles de marqueurs homologues contiennent des marqueurs répétés, et les algorithmes de détection de telles structures conservées ont une complexité exponentielle en temps et en espace [Pasek 2005, Ling 2009]

Le groupe de Jens Stoye a exploré une approche alternative, dans la quelle la taille des trous dans les occurrences inexactes n'est pas bornée. Pour réduire le nombre d'occurrences et permettre ainsi des calculs en temps acceptable, ils ont proposé plusieurs approches reposant sur la détection d'une occurrence médiane ou de référence [Böcker 2009, Jahn 2010, Chauve 2006], utilisant là encore des techniques algorithmiques adaptées du calcul d'empreintes de séquences.

3.2.3 Quelques applications

Comme nous l'avons vu dans le chapitre 2, les réarrangements génomiques sont susceptibles de modifier profondément l'organisation des gènes le long des chromosomes. Cependant, il est bien connu que certains groupes de gènes sont contraints d'être regroupés pour pouvoir fonctionner [Overbeek 1999]. L'exemple classique de ce phénomène est l'opéron des génomes bactériens, un groupe de gènes co-transcrits et co-régulés [Brouwer 2008], et il existe de plus en plus d'indices que ce principe de lien entre localisation et fonction biologique s'applique aussi aux génomes eucaryotes (voir [Batada 2007] par exemple). Les modèles de structures génomiques conservées décrits ci-dessus, notamment les intervalles communs avec marqueurs répétés et les équipes de gènes, offrent une base théorique solide pour la détection de telles régions, par exemple pour découvrir de nouveaux opérons par comparaison avec des opérons déjà annotés dans des génomes de référence [Pasek 2005, Schmidt 2007]. À ma connaissance, la seule utilisation des intervalles communs et PQ-arbres dans le cas de génomes eucaryotes (une comparaison des génomes de l'humain et du rat), en dehors des problèmes de scénarios parfaits décrits dans le chapitre suivant, est due à Landau, Parida et Weimann [Landau 2005].

Le principe d'un lien entre conservation de la fonction biologique et conservation de l'organisation des gènes est à la base de la notion d'*homologues positionnels* introduite notamment dans [Burgetz 2007] (voir l'article de synthèse [Dewey 2011]). La question est d'inférer des paires de gènes potentiellement orthologues en comparant deux génomes dont seules les familles de gènes homologues sont connues. L'idée est que des homologues qui partagent un contexte génomique similaire sont plus susceptibles d'avoir une fonction similaire et d'être orthologues ; il s'agit là d'une sorte de renversement de l'hypothèse que les gènes orthologues sont plus susceptibles de conserver une même fonction biologique, qui mène à un problème de calcul d'un *couplage* entre deux génomes [Swidan 2006a]. Dans [Burgetz 2007], les auteurs attaquent cette question en utilisant uniquement la notion d'adjacence conservée. Dans [Blin 2006] nous avons montré comment la notion d'intervalles communs avec marqueurs répétés pouvait s'appliquer naturellement à ce problème. Nous avons introduit le problème d'optimisation suivant : calculer la couverture maximale (i.e. non-extensible) d'une paire de génomes utilisant le nombre minimum d'intervalles communs. Ce problème est NP-difficile, mais nous avons proposé une heuristique efficace et qui a produit des résultats probants sur des génomes de γ-protéobactéries et sur une comparaison entres les génomes de l'humain et de la souris.

La notion d'intervalles communs a aussi des applications dans le calcul de distances et de mesures de dissimilarités génomiques. En effet, les intervalles communs sont apparus, du moins dans le cadre d'applications en génomique comparée, comme un concept important pour le tri de permutations signées par inversion [Bergeron 2002b, Bergeron 2005]. De plus, le nombre d'intervalles communs entre deux permutations a été utilisé pour définir une mesure de leur (dis)similarité, un principe introduit dans [Bergeron 2006b] qui généralise la notion classique de *distance des points de cassures* (*breakpoint distance* en anglais). Finalement, il sont à la base de la notion de distance/scénario *parfait*, qui sera décrit en détail dans le chapitre suivant

Dans le cadre plus général des génomes contenant des marqueurs répétés, on peut aussi relever deux points importants, qui mènent à d'autres applications des intervalles communs. Premièrement, la plupart des problèmes de comparaison (calcul de distance génomique notamment) entre deux génomes sont de complexité polynomiale si les marqueurs utilisés sont uniques et universels, mais NP-complets si des marqueurs peuvent être répétés (voir [Fertin 2009] pour un traitement plus complet de ce sujet). Deuxièmement, si on connaît les paires de marqueurs orthologues un-à-un (i.e. sans duplication postérieure à la spéciation définissant la relation d'orthologie) entre deux génomes avec marqueurs répétés, on peut se ramener à un problème de comparaison de génomes *sans marqueurs répétés*. En effet, la notion d'orthologie reposant sur une origine ancestrale commune, toute paire de marqueurs orthologues définit en fait une famille de marqueurs uniques et universels. Il suffit alors de supprimer les marqueurs restants spécifiques à l'un ou l'autre des génomes pour obtenir deux permutations [Blin 2005, Bourque 2005a]. Ces deux propriétés mènent naturellement au problème suivant : étant donnés deux génomes avec marqueurs répétés et un modèle de distance ou (dis)similarité, raffiner ces familles pour obtenir des familles de marqueurs uniques et universels qui minimisent la distance/dissimilarité entre les deux permutations résultantes. Sans surprise, tous les problèmes de ce type, pour les modèles classiques, sont NP-difficiles [Blin 2004, Chen 2005, Blin 2007b, Fertin 2009, Angibaud 2009], mais des heuristiques et algorithmes d'approximation ont été développés par le groupe de Tao Jiang [Chen 2005, Fu 2007, Fu 2008], donnant de bons résultats en pratique, tout comme des approches à base de programmation booléennes [Angibaud 2008, Angibaud 2007].

3.3 Conclusion

Nous venons de survoler un ensemble de modèles de structures génomiques conservées, organisé en une hiérarchie dominée par les modèles autorisant les occurrences inexactes dans des génomes avec marqueurs répétés. Ces modèles sont riches en propriétés algorithmiques et combinatoires, qui n'ont été qu'esquissées ici, ignorant plusieurs modèles comme les intervalles *conservés* [Bergeron 2006b], les intervalles communs *emboîtés* [Hoberman 2005a], ou encore les modèles adaptés à la prise en compte d'une duplication de génome complet et basés sur le principe de *synténies doublement conservées*, que nous allons aborder dans le chapitre 5. Nous avons aussi laissé de côté le vaste domaine des propriétés statistiques de ces modèles [Durand 2003, Hoberman 2005b, Raghupathy 2008, Raghupathy 2009], qui nécessiterait un chapitre complet. Finalement, une caractéristique fondamentale des modèles décrits dans ce chapitre est qu'ils sont *combinatoirement bien définis*, basés sur une définition non-ambiguë de la notion de structure conservée. D'autres méthodes suivent une approche plus *constructive* qui ne repose pas sur une telle définition mais sont centrées sur une méthode de calcul de régions génomiques conservées [Zheng 2005b, St-Onge 2005].

Tous ces modèles, à l'exception des intervalles communs de permutations, ont été définis dans le but d'analyser des jeux de données de génomes bactériens et ont été appliqués, en général avec succès. Ils pourraient cependant sans doute servir de base à des applications à

d'autres problèmes et jeux de données. Par exemple, pour résoudre le problème de la disparition de longs segments colinéaires d'ancres dans le calcul de familles de blocs homologues due au nombre croissant de génomes vertébrés disponibles (chapitre 2), on pourrait penser à utiliser une approche basée sur la notion plus souple d'équipes de gènes.

Contributions. Détection d'intervalles communs et de structures conservées.

- [StOnge 2005] K. StOnge, A. Bergeron et C. Chauve. *Fast identification of gene clusters in prokaryotic genomes*. CompBioNets 2005 : Algorithms and Computational Methods for Biochemical and Evolutionary Networks. College Publications, 2005.
- [Bergeron 2008a] A. Bergeron, C. Chauve, F. de Montgolfier et M. Raffinot. *Computing Common Intervals of K Permutations, with Applications to Modular Decomposition of Graphs*. SIAM Journal on Discrete Mathematics, vol. 22, no. 3, pages 1022–1039, 2008. Version préliminaire publiée dans les actes de ESA 2005.
- [Chauve 2006] C. Chauve, Y. Diekmann, S. Heber, J. Mixtacki, S. Rahmann et J. Stoye. *On Common Intervals with Errors*. Rapport technique 2006-02, Technische Fakultät der Universität Bielefeld, 2006.

Utilisation des intervalles communs dans divers problèmes de génomique.

- [Blin 2005] G. Blin, C. Chauve et G. Fertin. *Genes Order and Phylogenetic Reconstruction : Application to γ-Proteobacteria*. Comparative Genomics, RECOMB 2005 International Workshop, RCG 2005, pages 11–20. Springer, 2005.
- [Blin 2006] G. Blin, A. Chateau, C. Chauve et Y. Gingras. *Inferring Positional Homologs with Common Intervals of Sequences*. Comparative Genomics, RECOMB 2006 International Workshop, RCG 2006, pages 24–38. Springer, 2006.

Complexité de la comparaison de génomes avec marqueurs répétés.

- [Blin 2004] G. Blin, C. Chauve et G. Fertin. *The breakpoint distance for signed sequences*. CompBioNets 2004 : Algorithms and Computational Methods for Biochemical and Evolutionary Networks, pages 3–16. College Publications, 2004.
- [Blin 2007b] G. Blin, C. Chauve, G. Fertin, R. Rizzi et S. Vialette. *Comparing Genomes with Duplications : A Computational Complexity Point of View*. IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 4, no. 4, pages 523–534, 2007. Version préliminaire publiée dans les actes de IWBRA 2006.

Scénarios d'évolution par réarrangements génomiques

4

Nous allons maintenant aborder le problème du calcul de scénarios d'évolution (et donc de distances) en termes de réarrangements génomiques. Il s'agit là d'une question naturelle, par exemple pour proposer une hypothèse constructive, ou plus modestement des aspects quantitatifs, de la divergence entre une paire de génomes, dans le cadre d'un modèle de réarrangements génomiques donné. Les questions posées dans [Sankoff 1992a, Sankoff 1992b] ont été très vite résolues par des algorithmes efficaces proposés par Hannenhalli et Pevzner dans deux articles fondateurs de ce qui est connu maintenant comme la *théorie HP* [Hannenhalli 1995a, Hannenhalli 1995b]. Cependant, loin de clore le problème, ces succès initiaux ont suscité un vif intérêt dans la communauté algorithmique, que ce soit pour la recherche d'algorithmes plus efficaces ou plus simples, ou dans le cadre de modèles de réarrangements génomiques alternatifs.

Dans la première section de ce chapitre, je vais essayer de retracer les grandes lignes de cette histoire, en m'efforçant de mettre en avant le lien entre les aspects algorithmiques et le modèle de réarrangements génomiques dans lequel on se place. Ma contribution dans ce domaine, qui se compose essentiellement d'algorithmes de calcul de scénarios *parfaits*, c'est-à-dire qui préservent les intervalles communs, sera abordée dans la section suivante.

Nous n'aborderons ici que le problème de calcul de scénarios entre *deux génomes*, représentés par des ensembles de séquences signées sur un alphabet de marqueurs uniques et universels. Les aspects techniques et bibliographiques seront limités au minimum nécessaire à la compréhension des grandes lignes des algorithmes évoqués, et le lecteur peut se référer au livre [Fertin 2009] pour une bibliographie plus étoffée pointant vers les articles contenant les détails techniques de ces algorithmes, ainsi que pour une revue des nombreux autres problèmes de calcul de scénarios de réarrangements génomiques.

4.1 L'approche classique : scénarios parcimonieux

La principale contrainte imposée aux scénarios que nous décrivons dans cette section est la *parcimonie* : parmi l'ensemble des scénarios entre deux génomes donnés, on ne s'intéresse qu'à ceux qui minimisent le nombre de réarrangements génomiques.

Génomes unichromosomaux et inversions. Il s'agit là du problème le plus simple : on considère deux génomes $G = g_1\ g_2 \dots g_n$ et $H = h_1\ h_2 \dots h_n$ ayant chacun un chromosome et n marqueurs signés, et le modèle d'évolution se limite à un seul type de réarrangement génomique, à savoir l'inversion. On rappelle qu'une inversion est un réarrangement qui inverse l'ordre des marqueurs d'un segment (intervalle) d'un chromosome tout en changeant le signe des marqueurs inversés. Un scénario de G à H est donc une suite d'inversions qui transforme G en H. Le but est donc de calculer un scénario *parcimonieux* de G à H, c'est-à-dire un scénario qui utilise un nombre minimum d'inversions. La *distance d'inversions*, $d_I(G, H)$ est le nombre minimum d'inversions nécessaires pour transformer G en H. Ce problème[1] a été formalisé comme un problème d'optimisation combinatoire dans [Sankoff 1992a] et son intérêt en phylogénétique illustré dans [Sankoff 1992b]. Pour des raisons de clarté d'exposition, et sans perte de généralité, on suppose aussi que $g_1 = h_1 = 1$ et $g_n = h_n = n$.

L'objet central dans le calcul de scénarios par inversions est le *graphe des points de cassure*, un graphe aux arêtes bicolorées dont le nombre de cycles permet, la plupart du temps, de déterminer la distance d'inversion. Il se définit en termes d'*adjacences* dans les génomes G et H.

Définition 5 Soit $G = g_1\ g_2 \dots g_n$ un génome unichromosomal ayant n marqueurs.

1. Le génome G', défini sur l'alphabet de marqueurs $\{1_h, 1_t, 2_h, 2_t, \dots, n_h, n_t\}$, est obtenu en remplaçant chaque marqueur g de G par $g_t\ g_h$ (resp. $g_h\ g_t$) si g est positif (resp. négatif).
2. Une paire $\{i_a, j_b\}$, où $i, j \in \{1, 2, \dots, n\}$ et $a, b \in \{h, t\}$, est une *adjacence* de G si et seulement si $i \neq j$ et i_a et j_b sont consécutifs dans G'. On note $A(G)$ l'ensemble des adjacences de G. Un marqueur qui n'appartient à aucune adjacence est un *télomère*.
3. Le graphe des point de cassures de deux génomes G et H, $B(G, H)$, a pour ensemble de sommets $V = \{1_h, 1_t, 2_h, 2_t, \dots, n_h, n_t\}$ et pour ensemble d'arêtes E l'union de $A(G)$ (arêtes de type G) et de $A(H)$ (arêtes de type H).

$B(G, H)$ est donc composé de deux sommets isolés (1_t et n_h, les télomères) et d'un ensemble de cycles *alternants* : le long de chaque cycle, les arêtes de type G et H alternent. On note $c(B(G, H))$ son nombre de cycles. G et H sont égaux si et seulement si $B(G, H)$ contient $n + 1$ composantes connexes : les deux sommets isolés 1_t et n_h, et $n - 1$ cycles de longueur

[1] Comme G et H sont en fait des *permutations signées* et que l'on peut supposer que $H = 1\ 2 \dots n$, ce problème est aussi appelé *tri de permutations signées*.

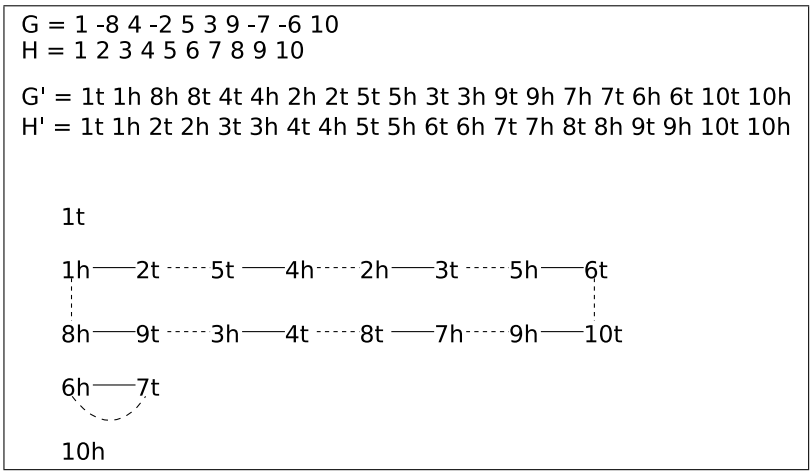

FIG. 4.1 – Un exemple de graphe des points de cassures, comportant deux cycles. Les arêtes de type H sont en pointillés et les arêtes de type G en lignes pleines.

2. On peut donc reformuler le problème qui nous intéresse en un problème de transformation de graphe : calculer une suite parcimonieuse d'inversions qui maximise le nombre de cycles dans le graphe des points de cassure résultant. Il n'est pas difficile non plus de prouver qu'une inversion sur G ajoute au plus un cycle à $B(G,H)$; par exemple, une inversion qui crée une adjacence de $A(H)$, c'est-à-dire qui rend consécutifs deux marqueurs qui le sont dans H, crée un cycle de longueur 2 dans le graphe des points de cassures. Cette propriété implique alors que

$$d_I(G,H) \leq n - 1 - c(B(G,H)). \tag{4.1}$$

Une inversion qui crée un cycle dans $B(G,H)$ est qualifiée de *gloutonne*. Il est naturel de se demander si $d_I(G,H) = n - 1 - c(B(G,H))$ et si l'approche consistant à effectuer des inversions gloutonnes arbitraires produit toujours un scénario parcimonieux. Cela n'est malheureusement pas le cas comme le montre le simple exemple où $G = 1\ 3\ 2$ et $H = 1\ 2\ 3$, car il n'existe alors aucune inversion gloutonne. Comprendre les obstructions combinatoires bloquant l'approche gloutonne est au cœur de la théorie HP. La solution proposée par Hannenhalli et Pevzner repose sur la définition de structures du graphe des points de cassure, appelées *obstacles* (*hurdles* en anglais) et *forteresses*, qui doivent être éliminées par des inversions non-gloutonnes.

La description initiale de cet aspect de la théorie HP est complexe (voir par exemple [Pevzner 2000]) et ce n'est que récemment qu'une description alternative simple a été proposée, qui repose sur l'utilisation du PQ-arbre d'un sous-ensemble des intervalles conservés de G et H [Bergeron 2005]. Il est aussi intéressant de remarquer que le nombre d'inversions requises pour éliminer obstacles et forteresses est en général faible et que la probabilité d'observer de telles structures dans une permutation signée aléatoire est très faible [Bergeron 2002a, Swenson 2008][2]. Cela explique pourquoi la borne supérieure (4.1) est

[2]Même si on en observe dans les données réelles.

en général une bonne approximation de la distance d'inversion. Cette "course d'obstacles" n'est sans doute pas terminée, mais a déjà mené à des algorithmes très efficaces, qui seront utilisés dans le cadre du calcul de scénarios parfaits.

Théorème 4
1. Calculer $d_I(G, H)$ peut se faire en temps et espace $O(n)$ [Bergeron 2005].
2. Calculer un scénario par inversions parcimonieux entre G et H peut se faire en temps $O(n^{3/2}\sqrt{\log(n)})$ et espace $O(n)$ [Tannier 2007].

Génomes multichromosomaux, inversions, translocations, fusions et fissions. Si la limitation aux génomes unichromosomaux est pertinente pour certains jeux de données, comme les génomes de mitochondries [Sankoff 1992b] ou de chloroplastes, ou encore les chromosomes X des mammifères, elle ne permet pas de comparer des génomes nucléaires d'eucaryotes. Les opérations naturelles pour comparer de tels génomes sont, outre les inversions, les translocations, fusions et fissions. On note $d_{HP}(G, H)$ la distance correspondante.

En 1995, Hannenhalli et Pevzner [Hannenhalli 1995b] annonçaient un algorithme polynomial pour calculer un scénario parcimonieux, en termes de ces quatre réarrangements, entre deux génomes multichromosomaux donnés G et H. L'idée principale, surprenante au premier abord, est la suivante : il est possible d'ordonner les chromosomes de G (resp. H) pour former deux génomes unichromosomaux G_u et H_u tels que $d_{HP}(G, H) = d_I(G_u, H_u)$. Cette opération d'agglomération de G et H (*capping* en anglais) est simple, mais la preuve de validité de ce principe initialement décrite dans [Hannenhalli 1995b] était complexe et incomplète, et a nécessité plusieurs corrections, celle de Jean et Nikolski [Jean 2007] étant la dernière en date. Plus récemment, reprenant l'approche suivie avec succès dans le cas des génomes unichromosomaux, Bergeron, Mixtacki et Stoye ont montré comment $d_{HP}(G, U)$ peut s'exprimer simplement en termes des cycles et chemins de longueur paire de $B(G, H)$ et du PQ-arbre d'un ensemble d'intervalles conservés de G et H, ce qui résulte en un algorithme de calcul de $d_{HP}(G, H)$ en temps $O(n)$ [Bergeron 2009].

Double-Cut-and-Join (DCJ). Les travaux de Bergeron, Mixtacki et Stoye pour simplifier la théorie HP [Bergeron 2005, Bergeron 2009] reposent sur le principe suivant : les distances d_I et d_{HP} peuvent en fait se décrire comme une distance simple à calculer en termes de cycles et chemins du graphe des points de cassure, la *distance DCJ*, corrigée par un paramètre lisible sur un PQ-arbre.

Pour présenter le modèle *Double-Cut-and-Join*, il est nécessaire d'augmenter la modélisation combinatoire utilisée pour représenter les génomes. Premièrement, un chromosome peut être linéaire (ce qui était le cas jusqu'à présent) ou *circulaire* (les marqueurs le long de ce chromosome sont organisés selon un ordre *cyclique*). Deuxièmement, un télomère p est représenté par une adjacence $\{p, T\}$, appelée une adjacence *télomérique*, où T est un symbole n'appartenant pas à l'alphabet des marqueurs. Finalement, il existe en permanence une adjacence $\{T, T\}$ disponible, représentant un chromosome virtuel.

Définition 6 Un DCJ opère sur deux adjacences $\{p, q\}$ et $\{r, s\}$ et remplace ces deux adjacences soit par $\{p, r\}$ et $\{q, s\}$, soit par $\{p, s\}$ et $\{q, r\}$. Un DCJ n'est pas défini quand trois ou quatre des éléments $p,\ q,\ r,\ s$ sont égaux à T.

On définit donc un DCJ en termes d'adjacences, et non plus de segments de génomes. L'opération DCJ permet de modéliser tous les réarrangements vus précédemment (inversion, translocation, fusion, fission sont en fait des DCJ) et d'en définir de nouveaux :

- L'*extraction* d'un segment de chromosome qui est transformé en chromosome circulaire en reliant ses deux extrémités, (si $p,\ q,\ r,\ s$ sont sur un même chromosome, dans cet ordre, et que $\{p, r\}$ et $\{q, s\}$ sont remplacées par $\{p, s\}$ et $\{q, r\}$) incluant la *circularisation* d'un chromosome linéaire complet (si de plus $p = s = T$).
- L'*insertion* dans un chromosome d'un chromosome circulaire, qui inclue la *linéarisation* d'un chromosome circulaire (le complément de l'extraction et de la circularisation).

Ce modèle de réarrangements a été introduit dans [Yancopoulos 2005], puis étudié en détail dans [Bergeron 2006a]. La distance DCJ entre G et H, notée $d_{DCJ}(G, H)$, est le nombre minimal de DCJ nécessaire pour transformer G en H. Le résultat central de la théorie DCJ est que la distance se lit facilement sur le graphe des points de cassure.

Théorème 5 [Bergeron 2006a] Soit $p_e(B(G, H))$le nombre de chemins de longueur paire dans $B(G, H)$.

$$d_{DCJ}(G, H) = n - (c(B(G, H) + p_e(B(G, H))/2).$$

De plus, un scénario DCJ parcimonieux peut être calculé en temps $O(n)$.

Pour expliquer l'algorithme de calcul d'un scénario parcimonieux, on définit un *DCJ glouton* comme un DCJ qui crée une adjacence ou un télomère de H. Il existe toujours un DCJ glouton. En effet, il suffit de prendre dans le génome courant deux marqueurs p et r appartenant à des adjacences (possiblement télomériques) $\{p, q\}$ et $\{r, s\}$, et qui sont adjacents dans H, et de créer $\{p, r\}$ par un DCJ qui crée aussi $\{q, s\}$ (possiblement au prix de créer un chromosome circulaire). De plus, un DCJ accroît le nombre de cycles (resp. chemins pairs) du graphe des points de cassure d'au plus un (resp. deux). Il en résulte que l'approche gloutonne fonctionne pour calculer un scénario DCJ parcimonieux.

Comparaison entre le modèle DCJ et le modèle HP. D'un point de vue combinatoire, le modèle DCJ évacue les subtilités liées aux obstacles, forteresses et *capping*. Il est donc naturel de s'interroger sur la nature de cette complexité combinatoire : nécessité ou artefact d'un modèle trop strict ? Le modèle DCJ est plus simple du fait de la possibilité de créer des chromosomes temporaires circulaires. Si la création de chromosomes circulaires a été observée assez fréquemment dans des génomes de cellules cancéreuses par exemple [Gebhart 2008], il n'en est pas de même pour des réarrangements évolutifs, notamment chez les eucaryotes où tous les chromosomes (sauf les organelles) sont linéasires. Cela pose donc la question pertinente de l'équilibre à trouver entre un modèle à la combinatoire complexe et un modèle simple aux

propriétés évolutives discutables. On peut remarquer que la création de chromosomes circulaires permet d'effectuer des transpositions en deux opérations DCJ. Une limitation naturelle de ce modèle, qui en renforce la pertinence évolutive, consiste alors à imposer l'intégration immédiate d'un chromosome circulaire nouvellement créé [Yancopoulos 2005, Kovác 2010].

La comparaison HP/DCJ est aussi intéressante dans le cadre de la controverse sur la réutilisation des points de cassure. En effet, cette statistique découle directement de la distance entre deux génomes G et H et du nombre de points de cassures dans un scénario optimal. La théorie HP simule les translocations, fusions et fissions par des renversements, et tout réarrangement comporte donc deux points de cassure, alors que le modèle DCJ autorise explicitement des réarrangements utilisant moins de deux points de cassures. Une étude instructive présentée dans [Bergeron 2008c] a montré que tout en restant parcimonieux, en jouant sur les différents types de DCJ, la statistique "réutilisation des points de cassure" n'est pas bien définie et peut varier grandement.

L'ensemble de tous les scénarios parcimonieux. Pour conclure cette section, il est nécessaire d'évoquer la question de la multiplicité des solutions optimales, qui se pose dans de nombreux problèmes d'optimisation combinatoire. La question se pose comme suit : étant donné deux génomes G et H et un modèle donné (HP ou DCJ ici), calculer (ou compter) tous les scénarios parcimonieux transformant G en H.

Dans [Bergeron 2002a], nous avons posé ce problème pour le modèle limité aux génomes unichromosomaux et aux inversions. Nous avons montré que le nombre de scénarios optimaux peut croître comme $n!$, et notamment qu'il existe des scénarios (appelés *scénarios commutants*) pour lesquels les inversions peuvent être appliquées dans n'importe quel ordre sans changer le résultat final, à savoir la transformation de G en H. Nous avons de plus proposé de représenter l'ensemble des scénarios optimaux en un ensemble de *traces de monoïde partiellement commutatif*, utilisant une notion de *commutation* entre inversions consécutives non-chevauchantes, sans toutefois explorer les propriétés algorithmiques de cette représentation. Cette étude a été menée dans [Braga 2008a, Braga 2008b, Badr 2010]. Dans le cadre du modèle DCJ, les concepts combinatoires adaptés à la représentation de l'ensemble des scénarios optimaux sont différents, notamment car la notion de commutation entre inversions consécutives (qui repose sur la vision d'une inversion comme un *ensemble* de gènes inversés) n'est pas adaptée à la représentation des réarrangements en termes d'adjacences cassées puis réparées [Braga 2009b, Ouangraoua 2009a].

4.2 Scénarios parfaits

La notion de *scénario parfait* a été introduite par Figeac et Varré dans [Figeac 2004], dans le cadre des génomes unichromosomaux et des inversions. Le principe général est simple : si un groupe de marqueurs forme un intervalle commun de G et H, on peut supposer que ces marqueurs étaient contigus dans l'ancêtre commun de ces deux génomes. Il est donc

raisonnable de supposer que ce groupe de marqueurs a été préservé, en tant qu'intervalle, durant l'évolution. Cette contrainte combinatoire additionnelle est bien entendu discutable, tout comme l'est en fait le critère de parcimonie pure, qui s'avère non-respecté si on considère plus de deux génomes par exemple, et nous reviendrons sur ce point en conclusion de ce chapitre. Plus que sur l'exploration de l'impact de cette contrainte sur l'analyse de données, les résultats présentés dans cette section sont centrés sur la compréhension de son intégration dans le calcul de scénarios d'évolution.

4.2.1 Génomes unichromosomaux et inversions

Pour commencer, nous allons définir la notion de scénario préservant les intervalles communs de G et H. Une inversion peut être vue comme un intervalle d'une permutation signée, ou plus simplement un ensemble de marqueurs (les marqueurs inversés). On peut donc définir une notion de *chevauchement* entre une inversion et un intervalle : une inversion chevauche un intervalle si leur intersection est non-vide, mais qu'aucun n'est inclus dans l'autre. Par exemple, si $G = 2\ -3\ 4\ 7\ 1\ 6\ 5$, l'inversion du segment 7 1 6 chevauche l'intervalle -3 4 7. Un scénario transformant G en H est parfait si aucune inversion ne chevauche un intervalle commun de G et H. Un scénario parfait est *optimal* si il est de longueur (nombre d'inversions) minimum parmi tous les scénarios parfaits.

Scénarios parfaits optimaux. Figeac et Varré ont montré qu'il existe toujours un scénario parfait transformant G en H, mais que le calcul d'un scénario parfait optimal est un problème NP-complet [Figeac 2004] dans le cas où tous les intervalles communs de G et H sont forts (i.e. le PQ-arbre ne comporte alors que des nœuds P). Dans la suite de ce paragraphe, nous allons montrer que le problème du calcul d'un scénario parfait optimal, dans le cas général, est en fait de complexité paramétrée [Bérard 2007]. L'idée principale de l'algorithme de calcul d'un scénario parfait optimal que nous allons décrire est d'utiliser le PQ-arbre des intervalles communs de G et H comme un guide.

Théorème 6 [Bérard 2007] Un scénario transformant G en H est parfait si et seulement si chaque inversion est soit un intervalle commun fort de G et H, soit l'union d'intervalles forts enfants d'un nœud P du PQ-arbre des intervalles communs de G et H.

Il découle du Théorème 6 que l'on peut calculer un scénario parfait en prenant en compte les nœuds du PQ-arbre des intervalles communs de G et H (dénoté $\mathcal{T}$ à partir de maintenant) un par un et en décidant (1) pour un nœud P ou Q si il doit être inversé et (2) pour un nœud P, quels renversements de groupes d'enfants de ce nœud doivent être inclus dans le scénario.

Pour ce faire, il est nécessaire de raffiner les propriétés combinatoires de $\mathcal{T}$. Pour simplifier l'exposition, on suppose que H est la permutation identitée et que $\mathcal{T}$ est plongé dans le plan de sorte que $p(\mathcal{T}) = H$. On peut alors associer une *permutation quotient* à chaque nœud de $\mathcal{T}$ définie par l'ordre relatif des éléments minimaux des sous-arbres des enfants de ce nœud.

Il s'ensuit que la permutation (non-signée) quotient d'un nœud Q ayant k enfants est soit $1\ 2\ \ldots k$ (le nœud est *croissant*), soit $k\ k-1\ \ldots 1$ (le nœud est *décroissant*). Il est aisé de remarquer que deux nœuds Q qui forment une arête ne peuvent pas être de même type. La permutation quotient d'un nœud P ayant k enfants est une permutation de longueur k *simple*, c'est-à-dire n'ayant aucun intervalle commun non-trivial avec la permutation identitée $1\ 2\ \ldots k$.

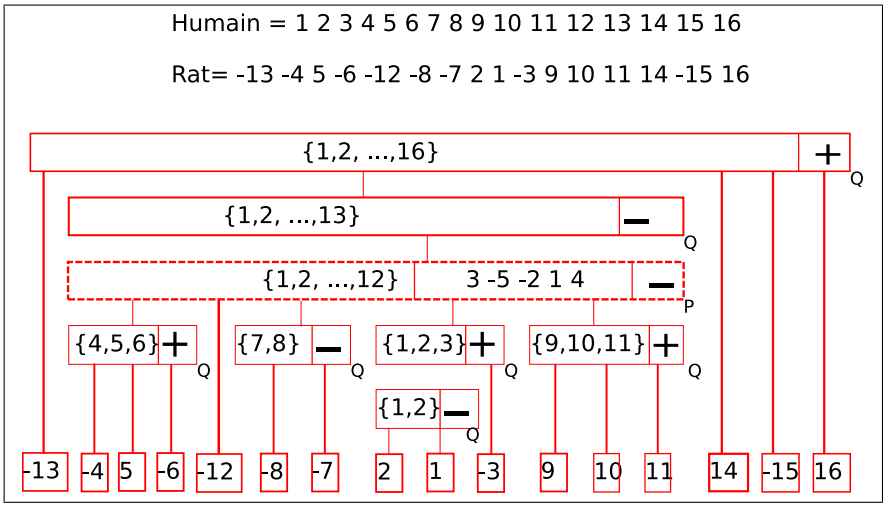

FIG. 4.2 – Un PQ-arbre augmenté avec la permutation quotient de son unique nœud P et le signe de tous ses nœuds internes. Les deux permutations signées représentent les chromosomes X des génomes humain et du rat. Adapté de [Bérard 2007].

On peut maintenant *signer* les nœuds de $\mathcal{T}$ comme suit : une feuille étiquetée x reçoit le signe de x dans G, un nœud Q croissant (resp. décroissant) est signé $+$ (resp. $-$), et un nœud P reçoit le signe de son parent si ce dernier est un nœud Q. Certains nœuds P ne reçoivent donc pas de signe, et comme nous allons le voir, les arêtes incidentes à deux nœuds P (les P-arêtes) sont les obstructions à un algorithme polynomial.

Lemme 7 [Bérard 2007] Soit $\mathcal{I}$ un intervalle commun fort de G et H, ayant un nœud Q $\mathcal{J}$ pour parent. Si $\mathcal{I}$ et $\mathcal{J}$ sont de signes différents, alors $\mathcal{I}$ appartient à tout scénario parfait optimal transformant G en H.

Si on suppose que tous les nœuds de $\mathcal{T}$ ont reçu un signe, ce qui suppose qu'il n'y a pas de P-arêtes, il ne reste alors qu'à déterminer les réarrangements parfaits définis comme union d'intervalles forts enfants d'un nœud P de $\mathcal{T}$ (Théorème 6). Dans ce cas, pour un nœud P donné $\mathcal{I}$ ayant k enfants, on fait *remonter* les signes de ses enfants sur les éléments correspondants de sa permutation quotient et on calcule un scénario par inversion transformant cette permutation signée en $1\ 2\ \ldots k$ (si $\mathcal{I}$ est positif) ou $k\ k-1\ \ldots 1$ si $\mathcal{I}$ est négatif (Théorème 4). Finalement, si certains nœuds P ne sont pas signés, on ne sait actuellement pas faire mieux qu'essayer toutes les façons de les signer, ce qui donne un algorithme paramétré par le nombre de P-arêtes, qui sert de base au résultat suivant.

Théorème 8 [Bérard 2007] Calculer un scénario parfait entre G et H peut se faire en temps $O(2^p n^{3/2}\sqrt{\log(n)})$, où p est le nombre de P-arêtes et a valeur au plus $n-1$.

L'algorithme décrit dans [Bérard 2007] peut être modifié pour obtenir une complexité en termes de *chemins de P-arêtes*, sans que cela représente toutefois une amélioration théorique importante [Bérard 2008]. Il peut aussi être utilisé pour calculer un scénario qui ne préserve qu'un sous-ensemble des intervalles communs de G et H [Bérard 2007].

Plus intéressant, dans [Bouvel 2009], nous avons montré que la *complexité en moyenne* de cet algorithme est polynomiale (sous-quadratique en fait). Ce résultat découle du fait que le PQ-arbre des intervalles communs de deux permutations aléatoires a une forte probabilité d'avoir une structure très simple, avec un unique nœud P, la racine.

Théorème 9 [Bouvel 2009] Pour n grand, le PQ-arbre des intervalles communs de deux permutations G et H a la structure suivante, avec probabilité 1 : la racine est un nœud P et tout sous-arbre de la racine est soit une feuille soit un nœud interne connecté à deux feuilles (une *cerise*). De plus, la distribution du nombre c de cerises est donnée par $P(c) = \frac{2^c}{e^2 c!}$, ce qui résulte en un nombre moyen de 2 cerises par arbre.

Ce résultat précis a été obtenu par une analyse combinatoire fine des propriétés énumératives des permutations simples. Il n'est cependant pas surprenant. En effet, si l'on rappelle que $\mathcal{T}$ décrit la décomposition modulaire du graphe de permutation défini par G et H [Paul 2006], le Théorème 9 ne fait que formaliser une propriété bien connue et aisée à prouver des graphes en général : un graphe aléatoire ne contient que des modules triviaux.

Un corollaire immédiat du Lemme 7 est que si $\mathcal{T}$ ne contient pas de nœud P, alors il n'existe essentiellement qu'un seul scénario parfait, qui est donc optimal, défini par les intervalles communs forts de G et H. De plus, les inversions de ce scénario peuvent être réarrangées dans n'importe quel ordre (i.e. commuter [Bergeron 2002a]) sans changer le fait qu'elles transforment G en H. Nous avons introduit cette classe de permutations signées dans [Bérard 2004], après avoir remarqué que malgré leur structure combinatoire assez exceptionnelle, elles apparaissaient dans des jeux de données de génomes de mammifères [Bérard 2007, Bourque 2004b].

Le PQ-arbre de deux permutations sans nœud P est un objet classique en combinatoire (un arbre de Schröder) et dans [Bouvel 2009], nous avons utilisé cette propriété pour donner des formules asymptotiques précises pour la longueur d'un scénario parfait entre deux permutations commutantes ($1.2n$) et la longueur moyenne d'une inversion dans un tel scénario ($1.02\sqrt{n}$).

La question naturelle des scénarios optimaux parfaits qui sont *aussi* parcimonieux a été posée pour la première fois dans [Bérard 2004], où nous avons caractérisé cette propriété dans le cas des permutations commutantes. Ces résultats ont été étendus au cas général dans [Diekmann 2007].

Finalement, on peut noter que le nombre de scenarios parfaits optimaux est difficile à calculer. Les renversements provenant de nœuds Q sont aisés à prendre en compte, mais compter le nombre de réarrangements de nœuds P est un problème ouvert. On peut néammoins relever quelques propriétés évidentes. Premièrement le nombre de nœuds internes joue un rôle important, car les inversions spécifiques à deux nœuds différents peuvent être permutées sans problème, et le nombre de scénarios parfaits optimaux comporte donc un facteur factoriel en le nombre de nœuds internes. Deuxièmement, la présence de nœuds P de grand degré est un autre facteur important, là encore pouvant induire un facteur factoriel (en le degré). Par contre, si on a peu de nœuds internes et des nœdus P de petit degré – des caractéristiques attendues pour des génomes proches – alors le nombre de scénarios est faibles. D'un point de vue algorithmique, l'intégration des méthodes de calcul de scénarios parfaits et des techniques de calcul de tous les scénarios par inversions [Braga 2008a, Braga 2008b, Badr 2010] ne semble pas poser de difficulté technique majeure.

4.2.2 Génomes multichromosomaux et DCJ

La notion de scénario parfait ne s'étend pas immédiatement au modèle DCJ. En effet, la particularité de ce modèle est d'autoriser la création temporaire de chromosomes circulaires. D'un côté, si on interdit de telles opérations, on retombe sur le modèle HP. D'un autre côté, toute extraction d'un chromosome circulaire détruit un intervalle commun car ses marqueurs ne forment plus un segment chromosomique contigu.

Dans [Bérard 2009], nous avons introduit une définition de scénario DCJ parfait basée sur la notion suivante : un intervalle commun à deux génomes est conservé si ses marqueurs sont répartis en au plus un segment chromosomique linéaire et un nombre non limité de chromosomes circulaires. On peut extraire des marqueurs d'un intervalle commun pour former un ou plusieurs chromosomes circulaires tant que l'on n'intègre pas ces marqueurs pour former un second segment de chromosome. En termes d'adjacences, le point de vue naturel pour le modèle DCJ, cela se traduit comme suit : à tout moment et pour tout intervalle commun, il existe au plus deux adjacences $\{p, q\}$ et $\{r, s\}$ telles que deux de ces marqueurs appartiennent à l'intervalle commun et les deux autres n'en font pas partie. Nous avons donc pris le parti, discutable, de considérer que les chromosomes circulaires sont un artefact du modèle DCJ utile pour modéliser des réarrangements plus complexes, et non des structures créées durant l'évolution.

D'un point de vue algorithmique, le calcul des scénarios DCJ parfaits est NP-difficile, ce qui n'est pas surprenant. Ce qui est surprenant est le fait que la difficulté du problème est due aux nœuds Q du PQ-arbre[3] des intervalles communs de G et H, alors que d'un autre côté, le cas d'un PQ-arbre ne comportant que des nœuds P peut être traité en temps polynomial. Avant de présenter notre résultat, on introduit les familles *faiblement séparables* d'intervalles communs : une famille d'intervalles communs est faiblement séparable si tout intervalle fort

[3]En fait il faut utiliser une variante du PQ-arbre, le PC-arbre, mais nous passons sur les détails techniques.

est l'union de deux intervalles communs qui se chevauchent (et correspond donc à un nœud Q dans le PQ-arbre).

Théorème 10 [Bérard 2009]

1. Calculer un scénario DCJ parfait optimal est NP-difficile si les intervalles communs à préserver forment une famille faiblement séparable.
2. On peut calculer un scénario DCJ parfait optimal entre G et H en temps $O(2^q n^2)$ où q est le nombre de nœuds Q du PQ-arbre des intervalles communs de G et H.

On a donc un algorithme paramétré par le nombre de nœuds Q du PQ-arbre, un résultat dont la preuve est beaucoup plus technique que pour le cas des inversions et des génomes unichromosomaux. On peut cependant avancer une explication intuitive de ces propriétés. Dans le cas des inversions, traiter un nœud P non signé est difficile car on ne sait pas si il faut trier la permutation quotient vers l'identité ou l'identité inversée, et il n'est pas possible de prendre cette décision indépendamment des décisions prises dans le reste de l'arbre. Utiliser des DCJ offre une troisième voie : trier la permutation quotient vers l'identité *circulaire* et il est possible de prendre une décision optimale parmi les trois choix en temps polynomial. Dans le cas d'un nœud Q, le modèle n'utilisant que les inversions laisse peu de choix car on ne peut inverser que l'intervalle complet. Un DCJ peut cependant extraire un segment de cet intervalle pour le circulariser, tout en préservant cet intervalle. Cette liberté accrue est la cause de la difficulté dans le traitement des nœuds Q.

Pour résumer, le calcul de scénarios parfaits est beaucoup plus difficile dans le cas du modèle DCJ, même si en termes de complexité algorithmique il n'existe pas de différence fondamentale. Il est par contre intéressant d'observer le renversement qui voit des problèmes simples à résoudre avec les inversions devenir difficiles pour le modèle DCJ et vice-versa. Il s'agit d'un contre-exemple à la propriété, qui était généralement acceptée, que les modèles DCJ et HP étaient essentiellement équivalents. La raison principale de cette différence dans le cas des scénarios parfaits est que l'objet combinatoire central n'est plus le seul graphe des points de cassures mais le PQ-arbre des intervalles communs, qui "résiste" moins bien à l'impact de circulariser des ensembles de marqueurs.

4.3 Conclusion

Bien que très incomplet, ce chapitre illustre les progrès importants réalisés dans le calcul de distances et de scénarios d'évolution par réarrangements génomiques depuis 1992, quand les premiers problèmes furent formellement introduits. En particulier, la réflexion sur l'équilibre à trouver entres la pertinence des modèles et leurs propriétés combinatoires et algorithmiques, suscitée par des aspects combinatoires (obstacles et forteresses, scénarios parfaits) ou appliqués (la réutilisation des points de cassure) a joué un rôle moteur. L'intégration dans ces modèles combinatoires de propriétés moléculaires des réarrangements est une question difficile mais qui demande à être explorée (voir [Swidan 2006b] par exemple).

Pour revenir sur le calcul de scénarios parfaits, on peut tout d'abord s'interroger sur la pertinence de ce concept. En effet, il est possible qu'un phénomène d'évolution convergente par exemple induise la création d'intervalles communs (adjacences notamment) dans des lignées indépendantes. Une première réponse suit une approche mathématique : si l'on fait une hypothèse nulle d'ordres de gènes aléatoires, alors il est peu probable d'observer des intervalles communs autres que des adjacences [Xu 2008c]. Mais plus généralement, les techniques développées pour le calcul de scénarios parfaits sont adaptées au cas où on ne veut conserver qu'un sous-ensemble donné des intervalles communs [Bérard 2007]. D'un point de vue combinatoire, onq peut remarquer l'importance du PQ-arbre des intervalles communs. La question de réconcilier le graphe des points de cassures et ce PQ-arbre en un formalisme plus intégré est encore largement ouverte. L'analyse en moyenne du calcul d'un scénario parfait décrite dans [Bouvel 2009], qui repose sur les techniques de combinatoire énumérative et analytique formalisées notamment par Flajolet et Sedgewick [Flajolet 2009], accompagne une série de travaux, d'Andrew Wei Xu notamment, sur les propriétés en moyenne du graphe des points de cassure [Xu 2008c, Xu 2008b]. L'utilisation de ces techniques pour la génération de jeux de données simulées réalistes est sans aucun doute un sujet de recherche fécond (voir par exemple [Ponty 2006]). Finalement, on peut noter des propositions de notions alternatives de scénario parfaits [Braga 2009a, Ouangraoua 2010a], introduites dans le but de rendre ce concept plus réaliste du point de vue évolutif.

Nous avons laissé de côté certains aspects importants de ces problèmes, comme par exemple les méthodes de correction de distances (voir [Lin 2010] par exemple) ou les approches probabilistes [Larget 2005]. Ces dernières, bien que naturelles, souffrent cependant du fait que les propriétés stochastiques de l'évolution par réarrangements génomiques (probabilité des différents types de réarrangements par exemple, longueur des inversions, localisation des points de cassure, ...) ne sont pas connues. Une approche basée sur l'utilisation de l'ensemble des scénarios parcimonieux n'est pas réaliste du fait de leur nombre super-exponentiel. Les techniques d'échantillonnage qui sont apparues récemment (voir [Miklós 2010] notamment) ouvrent une piste prometteuse vers des progrès sur ces questions. Bien qu'essentiellement théorique pour le moment, l'utilisation des techniques de la combinatoire analytique pour la génération de données simulées réalistes peut aussi participer à des avancées sur cette question.

Pour conclure, il faut remettre la question de la comparaison de *deux génomes* dans le cadre plus général de l'analyse d'un *ensemble de génomes*. En effet, pour raffiner un scénario d'évolution (supposé correct) entre deux génomes G et H, il est intéressant de déterminer quel génome intermédiaire est leur plus proche ancêtre commun. La seule façon de faire est de prendre en compte un troisième génome (un *groupe extérieur*) qui résulte d'une spéciation antérieure. Si l'on maintient la contrainte de parcimonie, on vient de définir le problème du *médian*, dont une définition générale est la suivante : étant donné trois génomes G, H, O, déterminer un quatrième génome (l'ancêtre ou médian) dont la somme des trois distances à G, H et O est minimisée. L'heuristique MGR [Bourque 2002] pour ce problème, dans le modèle HP, est au cœur des résultats obtenus par le groupe de Pavel Pevzner sur l'évolution des génomes de mammifères [Bourque 2004b, Bourque 2005b, Murphy 2005]. Le problème du

médian est difficile [Tannier 2009], mais des progrès récents, basés sur les concepts développés pour la comparaison de deux génomes décrits dans ce chapitre, permettent d'obtenir des solutions exactes ou presque exactes en temps raisonnable [Xu 2008a, Xu 2009a, Xu 2009b, Zhang 2009, Zheng 2011, Mahmoody-Ghaidary 2011], certaines incluant même la notion de scénario parfait [Bernt 2006, Bernt 2008]. Cependant, là encore, le problème de la multiplicité des solutions optimales se pose, en théorie [Eriksen 2007] comme en pratique [Murphy 2005]. Les résultats décrits dans le chapitre suivant, sur la reconstruction de génomes ancestraux, offrent des possibilités de limiter l'espace des solutions.

Contributions. Calcul de de scénarios par inversions parcimonieux.

- [Bergeron 2002a] A. Bergeron, C. Chauve, T. Hartman et K. St-Onge. *On the properties of sequences of reversals that sort a signed permutation.* Journées Ouvertes en Biologie, Informatique et Mathématiques, JOBIM 2002, pages 99–108, 2002.

Calcul de scénarios parfaits.

- [Bérard 2004] S. Bérard, A. Bergeron et C. Chauve. *Conservation of Combinatorial Structures in Evolution Scenarios.* RECOMB 2004 International Workshop, RCG 2004, pages 1–14. Springer, 2004.
- [Bérard 2007] S. Bérard, A. Bergeron, C. Chauve et C. Paul. *Perfect Sorting by Reversals Is Not Always Difficult.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 4, no. 1, pages 4–16, 2007. Version préliminaire publiée dans les actes de WABI 2005.
- [Bérard 2008] S. Bérard, C. Chauve et C. Paul. *A more efficient algorithm for perfect sorting by reversals.* Information Processing Letters, vol. 106, no. 3, pages 90–95, 2008.
- [Bouvel 2009] M. Bouvel, C. Chauve, M. Mishna et D. Rossin. *Average-Case Analysis of Perfect Sorting by Reversals.* Combinatorial Pattern Matching, 20th Annual Symposium, CPM 2009, pages 314–325. Springer, 2009. Version étendue soumise à *Discrete Mathematics Algorithms and Applications*, en cours de révision.

Autres : médian et autres modèles de distance.

- [Chauve 2004] C. Chauve et G. Fertin. *On maximal instances for the original syntenic distance.* Theoretical Computer Science, vol. 326, no. 1-3, pages 29–43, 2004.
- [Jiang 2010] H. Jiang, C. Chauve et B. Zhu. *Breakpoint Distance and PQ-Trees.* Combinatorial Pattern Matching, 21st Annual Symposium, CPM 2010, pages 112–124. Springer, 2010.
- [Mahmoody-Ghaidary 2011] A. Mahmoody-Ghaidary, C. Chauve et L. Stacho. Tractability results for the double-cut-and-join multichromosomal median problem. Présenté comme poster à *IWOCA*, 2011.

Reconstruction de génomes ancestraux

5

Si on combine le principe que les intervalles communs de deux génomes étaient présents dans leur ancêtre commun avec l'idée d'utiliser un groupe extérieur pour fixer cet ancêtre le long du chemin évolutif entre ces deux génomes, on arrive alors naturellement à la question que nous traitons dans ce chapitre : étant donné un ensemble de génomes actuels, dont on connaît, au moins partiellement, la phylogénie, comment inférer l'organisation d'un génome ancestral représentant un noeud interne de leur phylogénie ?

Ce problème suscite actuellement de nombreux travaux, utilisant essentiellement deux approches *in silico* : le calcul de scénarios d'évolution en termes de réarrangement génomiques (le problème du médian) et des approches inspirées de méthodes cytogénétiques et de cartographie. Après une introduction, incluant un rapide survol de ces approches, nous allons examiner plus en détail le résultat des mes recherches sur ce sujet, et plus précisément pour l'inférence de segments de génomes ancestraux et de groupes synténiques ancestraux. Dans ce chapitre, nous allons nous concentrer sur des aspects méthodologiques plutôt qu'algorithmiques. Le fil directeur des résultats présentés ici est que l'inférence de l'organisation de génomes ancestraux est essentiellement un problème d'assemblage/cartographie de génomes, et qu'il existe donc un ensemble de méthodes et idées qui offrent une solide base de départ.

5.1 Introduction

Les génomes d'organismes éteints depuis des dizaines de millions d'années ne peuvent pas être séquençés, même si on découvre des restes fossiles, du fait de la dégradation des molécules d'ADN [Marota 2002]. Leur organisation ne peut être déterminée que par la comparaison des génomes de leurs descendants que l'on a pu séquencer et assembler. Leur représentation ne peut donc être qu'abstraite et décrite à différents niveaux de précision : caryotype, associations synténiques, contenu en gènes, groupes de gènes synténique, ordre de ces gènes le long des chromosomes ancestraux. Les applications de tels résultats vont bien au-delà de la seule génomique évolutive.

Ce problème, qui a été longtemps dominé par les approches essentiellement expérimentales

utilisées par les cytogénéticiens [Richard 2003, Wienberg 2004, Robinson 2008], a récemment été l'objet de nombreux travaux proposant des approches *in silico* (voir [Rascol 2007, Ferguson-Smith 2007, Faraut 2008, Muffato 2008] pour des articles de synthèse récents). Sans prétendre être exhaustif, on peut citer des reconstruction de génomes ancestraux de levures [Zheng 2008, Jean 2009, Gordon 2009], de plantes [Adam 2007, Salsé 2009a, Salsé 2009b, Murat 2010], de drosophiles [Bhutkar 2007], d'invertébrés [Putnam 2007, Putnam 2008], de vertébrés [Woods 2005, Catchen 2008, Muffato 2010], d'amniotes [Kohn 2006, Nakatani 2007] et de mammifères [Bourque 2002, Bourque 2004b, Bourque 2005b, Murphy 2005, Kemkemer 2006, Ma 2006, Kemkemer 2009, Zhao 2009, Alekseyev 2009]. Ce "déluge" de méthodes pose clairement le problème de leur comparaison, comme l'illustre par exemple le débat animé qui a eu lieu en 2006 autour des divergences des approches cytogénétiques et *in silico* pour la reconstruction du génome de l'ancêtre des mammifères placentaires [Froenicke 2006, Rocchi 2006, Bourque 2006].

Les méthodes cytogénétiques. Précédant la complétion du séquençage des premiers génomes d'animaux, les cytogénéticiens ont développé, dans les années 90, des techniques de reconstructions de génomes ancestraux basées sur la combinaison de méthodes expérimentales et de méthodes mathématiques [Murphy 2001, Richard 2003, Yang 2003, Wienberg 2004, Froenicke 2006][1].

L'apparition de la technique d'*hybridation fluorescente in-situ* (*FISH* en anglais) a été un progrès décisif. Sans rentrer dans les détails, cette technique permet, étant donné un génome de référence (molécule d'ADN ici, et non pas objet combinatoire), de détecter de larges segments de ce génome conservés dans d'autres génomes, via un processus d'hybridation. Chaque chromosome du génome de référence est coloré (avec une couleur spécifique), puis ce génome est mis en contact avec un autre génome (la cible) et les segments chromosomiques suffisamment similaires entre ces deux génomes s'hybrident, colorant ainsi des segments chromosomiques du génome cible. Cela permet de déterminer, pour un chromosome cible donné, avec quels chromosomes de référence il présente une similarité de séquence significative.

Comment utiliser cette technique pour inférer l'organisation d'un génome ancestral ? Supposons que deux segments de deux chromosomes différents du génome de référence s'hybrident sur un unique chromosome dans deux espèces cibles différentes. Le principe de *parcimonie de Dollo*, classique en inférence phylogénétique [Felsenstein 2004] et qui stipule qu'un caractère complexe peut être perdu mais non gagné, permet alors de faire l'hypothèse que ces deux segments appartenaient à un même chromosome (i.e. étaient *synténiques*) dans le plus proche ancêtre commun de ces deux génomes cibles, formant ainsi une *association synténique ancestrale*, décrite en termes de chromosomes de référence. De manière générale, on peut utiliser les résultats des expériences d'hybridation comme caractères discrets pour les méthodes classiques d'inférence phylogénétique[2].

[1] En particulier, l'article de synthèse [Murphy 2001] donne un aperçu intéressant des résultats obtenus juste avant l'apparition des premières méthodes *in silico*.

[2] En fait ces techniques utilisent la similarité de séquence (hybridation) entre segments chromosomaux

Les méthodes *in silico*. Ces méthodes utilisent comme données de base un ensemble de génomes séquencés et assemblés. En 2002, Bourque et Pevzner ont introduit MGR (Multiple Genome Rearrangement) qui permet de calculer des scénarios d'évolution par réarrangements (en utilisant le problème du médian dans le modèle HP) pour plusieurs génomes encodés sur un alphabet de marqueurs à une résolution de quelques centaines de milliers de bases [Bourque 2002]. MGR a permis d'analyser successivement les génomes de l'humain, de la souris et du rat [Bourque 2004b], puis des mêmes génomes et du poulet, pour proposer un premier ancêtre des mammifères placentaires (pour le sous-groupe des Euarchontoglires) à une telle résolution [Bourque 2005b]. Depuis, les progrès de méthodes *in silico* basées sur le calcul de scénarios de réarrangements parcimonieux et des problèmes de type médian ont été immenses, et je me contenterai de citer quelques articles récents qui les illustrent [Alekseyev 2009, Gavranovic 2010, Zheng 2008, Zheng 2011].

Une autre famille de méthodes *in silico* suit une approche différente, que l'on peut qualifier de "sans modèle" car elle ne repose pas sur un modèle explicite de réarrangements génomiques. Le principe général de ces méthodes, introduit dans [Bergeron 2004], consiste à inférer des *caractères génomiques* potentiellement ancestraux (par exemple des adjacences entre marqueurs [Ma 2006, Bertrand 2010], ou des intervalles communs ou conservés [Adam 2007, Stoye 2009]), puis à *assembler* ces caractères en un génome ancestral. Les principales implémentations de cette approche infèrent les caractères ancestraux en utilisant un algorithme de Fitch-Hartigan [Bergeron 2004, Adam 2007, Stoye 2009, Bertrand 2010] qui minimise le nombre de pertes/gains de caractères le long des branches de la phylogénie des génomes étudiés (on suppose que cette phylogénie est connue).

Comparaison des trois approches. L'approche cytogénétique permet de comparer de nombreux génomes, car il n'est nul besoin de la séquence des génomes, mais seulement des molécules d'ADN. Elle a été utilisée pour reconstruire des génomes ancestraux de mammifères à partir de jeux de données contenant plusieurs dizaines de génomes. Par contre, elle souffre de deux défauts importants. Pour que l'hybridation puisse se produire, les génomes référence et cible doivent être suffisamment proches pour avoir conservé assez de similarité de séquence. De plus, les segments qui peuvent s'hybrider sont en général longs (de l'ordre de quelques mégabases), et on ne peut donc pas détecter des associations de petits segments de chromosomes (à une résolution de quelques gènes par exemple). Les approches *in silico* ne peuvent se baser que sur un nombre limité de génomes séquencés et assemblés (et de qualité), mais peuvent cependant analyser ces génomes à une résolution bien supérieure.

Un débat animé et intéressant a eu lieu en 2006, dans les colonnes de Genome Research, entre un groupe de cytogénéticiens et le groupe de Pavel Pevzner [Froenicke 2006, Bourque 2006]. Il trouve sa source dans les différences entre l'ancêtre des mammifères placentaires obtenus par les méthodes cytogénétiques et MGR. Sans rentrer dans les détails, ce débat a tourné autour de questions d'acquisition de données, d'échantillonnage phylogénétique, de

(intervalles) pour poser l'hypothèse que ces segments sont orthologues : les segments hybridés peuvent donc être vus comme des marqueurs, exactement au sens défini dans le chapitre 2.

résolution et de la multiplicité de solutions obtenues avec MGR. Il a cependant été remarqué dans [Rocchi 2006] que la méthode (*in silico*) décrite dans [Ma 2006] obtenait des résultats en accord avec ceux des cytogénéticiens. Ce débat a servi de point départ aux travaux décrits dans la suite de ce chapitre.

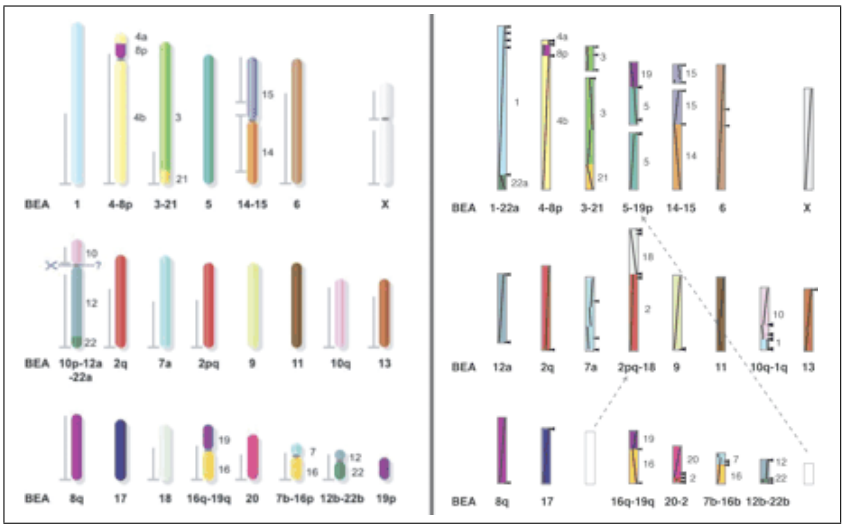

FIG. 5.1 – Les deux ancêtres de mammifères placentaires obtenus par une approche cytogénétique (à gauche) et par MGR (à droite). Les différentes couleurs représentent les chromosomes humains. Les lignes pointillées représentent les différences majeures entre les deux propositions d'ancêtres. [Froenicke 2006]

Survol de notre approche. L'approche que nous décrivons appartient au second groupe de méthodes *in silico*. Elle emprunte cependant une spécificité des approches cytogénétiques : tout caractère ancestral potentiel doit être conservé dans au moins une paire de génomes actuels dont le chemin évolutif passe par l'ancêtre recherché. Notre approche diffère en cela des approches basées sur la parcimonie de Fitch-Hartigan et nous discuterons en conclusion des limitations de cette approche. De plus, l'inférence d'un génome ancestral est décomposée en trois étapes : marqueurs génomiques, régions ancestrales contiguës (segments de chromosomes) et groupes synténiques. Finalement, elle repose sur des concepts développés pour la *cartographie* de génomes non-séquencés, et notamment les PQ-arbres.

Notre but en développant cette approche n'est pas tant de proposer une nouvelle méthode de reconstruction de génomes ancestraux que de formaliser des concepts et principes permettant ainsi de mieux comparer les nombreuses méthodes publiées récemment.

5.2 Régions Ancestrales Contiguës et la Propriété des Uns Consécutifs

On suppose ici qu'on veut reconstruire l'architecture d'un génome ancestral $\mathcal{A}$ à partir de deux ensembles de génomes : $\{\mathcal{D}_1, \ldots, \mathcal{D}_k\}$ sont des descendants de $\mathcal{A}$ et $\{\mathcal{O}_1, \ldots, \mathcal{O}_p\}$ descendent d'un ancêtre de $\mathcal{A}$ mais pas de $\mathcal{A}$ (*groupe extérieurs*). On suppose de plus qu'on a un ensemble de marqueurs $\mathcal{M} = \{m_1, \ldots, m_n\}$ qui sont uniques et universels pour $\{D_1, \ldots, D_k\}$ et aussi pour $\{O_1, \ldots, O_p\}$. Cette dernière condition n'est pas nécessaire mais facilite l'exposition. Deux génomes forment une *paire informative* si $\mathcal{A}$ appartient au chemin évolutif qui les joint.

On fait donc l'hypothèse que $\mathcal{A}$ contient une et une seule copie de chaque marqueur de $\mathcal{M}$, qui sert d'alphabet pour sa description. Idéalement, on voudrait donc partitionner $\mathcal{M}$ en ensembles totalement ordonnés, chaque ensemble étant un chromosome ancestral, ou plutôt un segment de chromosome ancestral. La notion de *région ancestrale contiguë* a été introduite par Ma [Ma 2006] pour désigner des segments de chromosomes ancestraux. Une Région Ancestrale Contiguë est un ensemble de marqueurs de $\mathcal{M}$ qui forment un intervalle dans $\mathcal{A}$. Dans la définition initiale de Ma, les marqueurs d'une Région Ancestrale Contiguë sont ordonnés. Dans [Chauve 2008b], nous avons relâché cette contrainte.

5.2.1 La Propriété des Uns Consécutifs et les PQ-arbres.

En adaptant un des principes des méthodes cytogénétiques, on définit un ensemble de marqueurs S comme *potentiellement contigu dans* $\mathcal{A}$ si il forme un intervalle dans une paire informative. Soient $\{S_1, \ldots, S_m\}$ m ensembles de marqueurs potentiellement contigus dans $\mathcal{A}$ obtenus en comparant toutes les paires informatives de génomes. On agglomère les S_i en une *matrice binaire* $\mathcal{R}$ comme suit : les colonnes de $\mathcal{R}$ sont indexées par les n marqueurs de $\mathcal{M}$ et chaque S_i définit une ligne de $\mathcal{R}$, les éléments de S_i donnant les entrées 1 de cette ligne.

Dans un premier temps, on suppose que $\{S_1, \ldots, S_m\}$ ne contient pas d'erreur : les marqueurs correspondent bien à des segments orthologues issus d'un segment ancestral et chaque S_i formait un intervalle dans $\mathcal{A}$. Si on admet que $\mathcal{A}$ n'avait que des chromosomes linéaires, alors $\mathcal{R}$ satisfait la *Propriété des Uns Consécutifs*.

Définition 7 Une matrice binaire $\mathcal{R}$ satisfait la Propriété des Uns Consécutifs si il existe une permutation de ses colonnes telle que les 1 de chaque ligne sont consécutifs.

Théorème 11 [Booth 1976, Habib 2000, McConnell 2004] Si $\mathcal{R}$ a la Propriété des Uns Consécutifs, alors l'ensemble des permutations qui gardent les 1 de chaque ligne consécutifs est la classe d'équivalence de permutations d'un PQ-arbre que l'on peut calculer en temps $O(n + m + e)$ où e est le nombre de 1 dans $\mathcal{R}$.

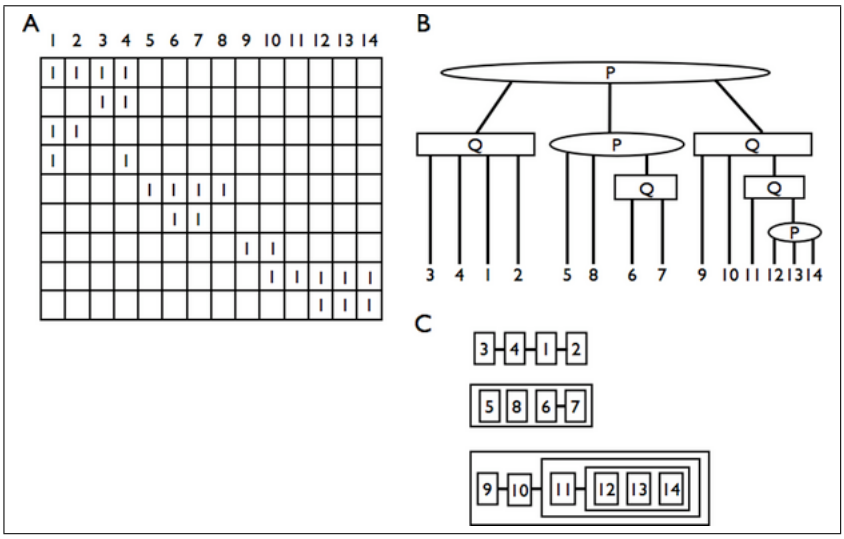

FIG. 5.2 – (A) Un exemple de matrice ayant la Propriété des Uns Consécutifs. (B) Le PQ-arbre correspondant. (C) Une représentation alternative où chaque ligne correspond à un sous-arbre de ce PQ-arbre et représente une Région Ancestrale Contiguë. [Chauve 2008b].

Il n'est pas surprenant de retrouver les PQ-arbres ici : si tous les S_i sont sans erreur, alors ils forment des intervalles communs de certains génomes D_i avec $\mathcal{A}$. Le PQ-arbre de la matrice $\mathcal{R}$ est alors utilisé comme une représentation, possiblement ambiguë, du génome ancestral $\mathcal{A}$. La notion de PQ-arbre en relation avec la Propriété des Uns Consécutifs a été très utilisée dans le cadre de la cartographie physique de génomes [Alizadeh 1995]. Le problème d'ordonner les marqueurs de $\mathcal{M}$ le long des chromosomes de $\mathcal{A}$ est en effet un problème de cartographie physique *in silico* (i.e. ne reposant pas sur des expériences d'hybridation) sur un génome disparu.

Si $\mathcal{R}$ n'a pas la Propriété des Uns Consécutifs, parceque soit certains marqueurs ne forment pas une famille orthologue, soit certains S_i ne correspondent pas à des ensembles qui étaient contiguës dans $\mathcal{A}$ (en cas d'évolution convergente par exemple), tout n'est pas perdu. McConnell a en effet introduit dans [McConnell 2004] le *PQR-arbre*, où les noeuds de type R encodent les parties de $\mathcal{R}$ qui font obstacle à la Propriété des Uns Consécutifs. Cet objet, lui aussi calculable en temps optimal, permet donc de définir un ensemble de Régions Ancestrales Contiguës sans conflit (les sous-arbres de la racine qui ne contiennent pas de noeud R). Pour éliminer les noeuds R, on peut alors utiliser des techniques d'optimisation combinatoire, comme conserver une sous-matrice maximale qui a la Propriété des Uns Consécutifs [Chauve 2008b]. De tels problèmes de transformation d'une matrice qui n'a pas la Propriété des Uns Consécutifs sont en général difficiles (voir [Dom 2008] pour un texte de synthèse sur ce sujet). Par exemple, si les S_i comportent tous deux marqueurs (adjacences), $\mathcal{R}$ est la matrice d'incidence d'un graphe, dont les sommets sont les marqueurs, que l'on cherche à transformer en une collection de chemins (les Régions Ancestrales Contiguës) en supprimant un nombre minimum d'arêtes. Ce problème de "Partition en un nombre Minimum de Chemins" (*Minimum Path*

Partition/Cover) un problème NP-complet classique, qui généralise le problème du Chemin Hamiltonien.

5.2.2 Applications aux mammifères placentaires.

Dans [Chauve 2008b] nous avons appliqué cette approche à la reconstruction du génome ancestral des mammifères placentaires. Nous avons obtenu des blocs orthologues à partir des alignements avec le génome humain de cinq génomes de mammifères placentaires (macaque, souris, rat, chien, vache), de l'opossum et du poulet, et disponibles sur le site du UCSC Genome Browser. Nos données ont permis d'identifier $n = 824$ blocs à une résolution de 200kb. Nous avons défini chaque S_i comme étant soit une adjacence conservée soit un intervalle commun maximal dans une paire informative. La matrice binaire $\mathcal{R}$ ainsi obtenue comporte $m = 1431$ lignes et ne satisfait pas la Propriété des Uns Consécutifs. Cependant il a suffi de supprimer 14 lignes de cette matrice pour satisfaire cette propriété (résultat optimal obtenu avec un algorithme de type "Branch-and-Bound"). Le PQ-arbre correspondant décrit 26 Régions Ancestrales Contiguës. Les associations synténiques obtenues sont toutes en accord avec les résultats des cytogénéticiens et indiquent une conservation générale de l'organisation de ce génome le long de la lignée menant au génome humain (Figure 5.3).

On peut souligner plusieurs points méthodologiques importants de ce travail. Premièrement, la matrice $\mathcal{R}$ satisfait presque la Propriété des Uns Consécutifs. Cela indique qu'elle contient sans doute peu de bruit et que l'approche de type optimisation combinatoire utilisée pour supprimer les lignes n'est sans doute pas sujette au problème d'un grand nombre de solutions optimales. Nous reviendrons sur ce point dans le paragraphe suivant. Le PQ-arbre est de plus peu ambigu : 778 des 824 marqueurs peuvent être ordonnés précisément le long du segment ancestral qui les contient. Si on s'était limité à utiliser des adjacences, le PQ-arbre n'aurait contenu que des noeuds Q (hors la racine) et aucune ambiguïté : le prix à payer pour introduire les intervalles communs comme possibles caractères ancestraux n'est donc pas trop élevé en termes de perte de résolution de l'ancêtre. Par contre le signal détecté par les intervalles communs est important : en appliquant une méthode similaire mais n'utilisant que des adjacences [Ma 2006], on obtient 34 Régions Ancestrales Contiguës. De plus, ces régions contiennent 5 adjacences, chacune spécifique à un génome de mammifère, et donc non conservée et considérées comme ancestrale du fait de l'utilisation de l'algorithme de Fitch-Hartigan pour inférer les adjacences potentiellement ancestrales. Supprimer ces adjacences résulte en 39 Régions Ancestrales Contiguës, un nombre bien supérieur au nombre de chromosomes attendu pour ce génome ancestral (24).

Ce travail montre que l'on peut combiner des éléments des méthodes cytogénétiques, de cartographie physique, de génomique comparée et d'optimisation combinatoire, et les appliquer sur un petit nombre de génomes séquencés pour obtenir un ancêtre bien résolu et supporté et en accord avec la cytogénétique (à deux fusions de segments près). Étant donné le statut de *standard* de l'ancêtre des mammifères placentaires dans la communauté cytogénétique, il s'agit là d'un résultat encourageant, qui suggère que les divergences discutées dans [Froenicke 2006, Bourque 2006, Rocchi 2006] étaient plus dues à des questions

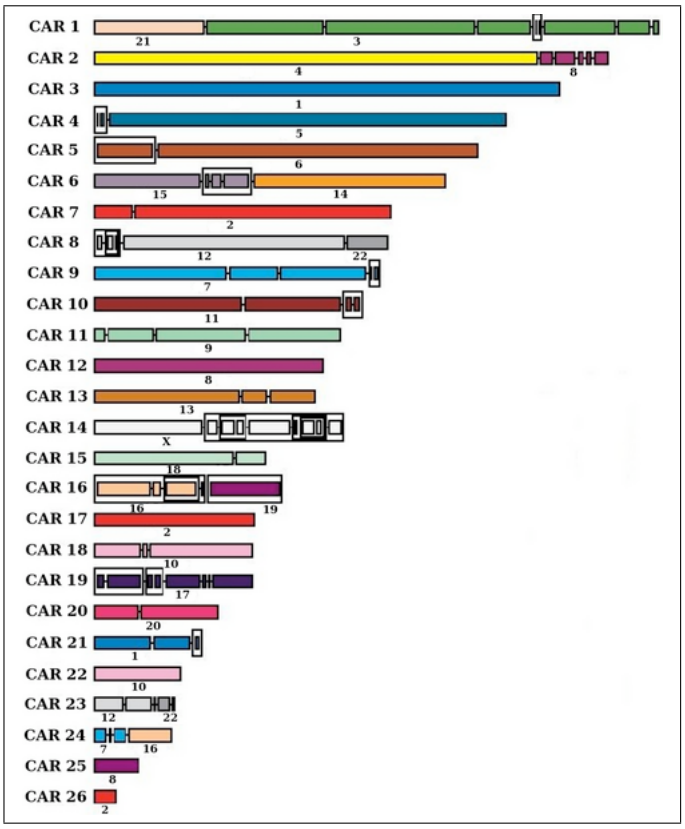

FIG. 5.3 – Les 26 Régions Ancestrales Contiguës du génome de l'ancêtre des mammifères placentaires obtenu dans [Chauve 2008b]. Les différentes couleurs représentent les chromosomes humains.

de méthodologie que de données ou de résolution. Cela illustre la nécessité de discussions méthodologiques en amont de la comparaison des résultats de méthodes différentes.

5.3 Extensions et autres applications.

Nous présentons maintenant des développements récents basés pour la plupart sur une exploration de la Propriété des Uns Consécutifs et de certaines de ces variantes.

Ensembles conflictuels. L'approche que nous venons de décrire comporte une phase d'optimisation combinatoire pour laquelle se pose le problème de la multiplicité des solutions optimales. La notion d'ensemble conflictuel, introduite dans [Bergeron 2004], s'avère utile pour étudier cet aspect. Un *ensemble conflictuel* est un ensemble $R = \{r_1, \ldots, r_k\}$ de lignes de $\mathcal{R}$ qui ne satisfait pas la Propriété des Uns Consécutifs ; cet ensemble est un *ensemble conflictuel*

minimal si tout sous-ensemble propre de R satisfait la Propriété des Uns Consécutifs. On peut donc les voir comme des obstructions minimales à la Propriété des Uns Consécutifs.

On peut utiliser cette notion de plusieurs manières. Dans [Bergeron 2004], toute ligne de $\mathcal{R}$ appartenant à un ensemble conflictuel minimal est supprimée, une approche radicale. Dans [Stoye 2009] le nombre d'ensembles conflictuels minimaux auxquels une ligne de $\mathcal{R}$ appartient est utilisé pour ordonner les lignes en vue d'un algorithme de Branch-and-Bound pour éliminer les conflits de $\mathcal{R}$. Cette approche suit le principe qu'une ligne apparaissant dans de nombreux conflits est susceptible à la fois d'être incorrecte (i.e. de ne pas représenter un intervalle du génome ancestral) et de devoir être supprimée dans toute solution optimale d'élimination des conflits.

Calculer (ou même compter) les ensembles conflictuels minimaux est un problème difficile. Par exemple, si $\mathcal{R}$ est de *degré* deux (tout ligne contient exactement deux 1), alors tout cycle du graphe dont $\mathcal{R}$ est la matrice d'incidence est un ensemble conflictuel minimal, et compter le nombre de cycles d'un graphe est un problème #P-complet. Nous avons décrit dans [Chauve 2010] une méthode de calcul de ces ensembles basée sur l'utilisation de *fonctions booléennes monotones*. Nous avons utilisée cette méthode pour analyser des données simulées et réelles et illustrer le potentiel de ce concept pour la compréhension des obstructions combinatoires à la Propriété des Uns Consécutifs que l'on rencontre en pratique.

Une approche alternative peut être de prendre en compte une notion plus fine d'obstruction à la Propriété des Uns Consécutifs, à savoir les motifs interdits de Tucker [Tucker 1972]. Ces motifs sont plus faciles à énumérer ques les ensembles conflictuels minimaux [Chauve 2011b].

La Propriété des Uns Consécutifs en Sandwich. Un problème du travail sur les mammifères placentaires est de se baser sur des marqueurs uniques et universels, et surtout d'imposer cette contrainte pour les génomes du poulet et de l'opossum. On peut argumenter que la définition de l'alphabet de description du génome ancestral doit être rigoureuse, mais le prix à payer est élevé en termes de couverture des génomes par les marqueurs car les génomes de l'humain et du poulet s'alignent en général mal en dehors des régions codantes. Dans un travail récent sur la reconstruction de génomes ancestraux de plantes [Murat 2010], Éric Tannier a introduit une variante de la Propriété des Uns Consécutifs qui permet de résoudre ce problème, comme nous l'avons illustré dans [Gavranovic 2011].

On considère maintenant que $\mathcal{R}$ est une matrice avec trois types d'éléments : 0, 1 et X, qui représente un caractère ambigu qui pourrait être soit 0 soit 1. $\mathcal{R}$ satisfait la Propriété des Uns Consécutifs en Sandwich si on peut remplacer un ensemble de X par des 1 (et les autres par des 0) pour obtenir une matrice binaire qui satisfait la Propriété des Uns Consécutifs. Décider si $\mathcal{R}$ satisfait cette propriété est, sans surprise, un problème NP-complet. Nous avons mis au point divers heuristiques et algorithmes, basés sur des techniques d'affinage de partition, de recherche locale et de Voyageur de Commerce, notamment pour ordonner les colonnes d'une telle matrice en minimisant le nombre de lignes qui comportent un *trou* (i.e. un segment de 0 encadré par deux 1) et la taille des trous.

Que représentent ces X ? Dans le problème qui nous intéresse, ils représentent des marqueurs absents d'un (ou deux) génomes d'une paire informative et qui auraient donc pu faire partie (ou non) d'un ensemble de marqueurs contigus dans le génome ancestral.

Nous avons appliqué ce modèle à l'inférence du génome ancestral des mammifères placentaires encore (du fait de son statut de standard). Nous avons inclu quinze génomes dans notre jeu de données (douze mammifères, l'opossum, le poulet et le mandarin) et défini des marqueurs non universels en termes d'alignement (obtenus sur le site Ensembl [Paten 2008a]) conservés dans au moins une paire informative. Cela a permis d'augmenter la couverture du génome humain par les marqueurs de 11%, passant de 24% de sa longueur totale à 35%, ce qui n'est pas négligeable. À notre grande surprise, la matrice résultante, qui est très grande ($n = 1724$ et $m = 89023$) satisfait presque la Propriété des Uns Consécutifs en Sandwich. En appliquant nos méthodes d'ordonnancement de ses colonnes, nous avons obtenu 26 Régions Ancestrales Contiguës qui là encore sont en accord avec les résultats de cytogénétique.

Autres variantes de la Propriété des Uns Consécutifs. Suivant la constatation que la la matrice $\mathcal{R}$ obtenue dans [Chauve 2008b] satisfait presque la Propriété des Uns Consécutifs, il est naturel de se demander quelles sont propriétés des 14 lignes supprimées. Un examen de ces lignes permet de voir qu'elles comportent en fait peu de trous et qu'ils sont courts en général. C'est en se basant sur ces éléments que nous avons étudié dans [Chauve 2009c] la Propriété des Uns Consécutifs avec Trous Bornés. Sans surprise, nous avons montré que le problème de décider si les colonnes d'une matrice binaire peuvent être ordonnées de sorte que chaque ligne comporte au plus k trous et que chaque trou est de longueur bornée[3] par un paramètre δ est NP-complet [Chauve 2009c] (sauf si $k = 2$ et $\delta = 1$, qui reste un problème ouvert), même si $\mathcal{R}$ est de degré borné [Manuch 2010]. Les preuves de ces résultats de complexité reposent en partie sur un lien nouveau et intéressant avec des problèmes de couverture d'hypergraphes, qui généralisent le problème de Partition en un nombre Minimum de Chemins.

Le cas de marqueurs non-uniques nécessite aussi une adaptation de la Propriété des Uns Consécutifs. Par exemple, si on veut éviter d'utiliser des arbres de gènes pour inférer des relations d'orthologie pour une famille donnée (chapitre 2 et appendice A), on peut se contenter d'estimer le nombre de copies du gène ancestral correspondant, en utilisant des approches probabilistes [Bie 2006]. Il faut alors gérer le fait que les colonnes de $\mathcal{R}$ peuvent avoir une multiplicité (le marqueur correspondant à une colonne peut apparaître plusieurs fois dans le génome ancestral), bornée cependant. Roland Wittler a étudié dans [Wittler 2010] le problème de décider cette variante de la Propriété des Uns Consécutifs. Il a prouvé que seul le cas des matrices de degré deux pouvait être résolu en temps polynomial.

Récemment, nous avons aussi utilisé la Propriété des Uns Consécutifs avec Multiplicité dans un cadre quelque peu différent, pour intégrer la notion de télomère. Une des questions que l'on se pose face à un ensemble de Régions Ancestrales Contiguës est de savoir lesquelles

[3] Il s'agit d'une fusion des concepts de Propriété des Uns Consécutifs et d'équipes de gènes décrits dans le chapitre 3.

sont des chromosomes ancestraux complets. Une façon de répondre à cette question est de décider quels marqueurs sont télomériques dans l'ancêtre $\mathcal{A}$ et de définir un chromosome ancestral comme une Région Ancestrale Contiguë contenant deux marqueurs télomériques à ses deux extrémités. Pour cela on peut définir un marqueur (virtuel) télomère T qui est présent avec multiplicité (tout comme on le fait dans le modèle DCJ) et intégrer ce marqueur comme une colonne de $\mathcal{R}$. Nous avons montré dans [Chauve 2011a] que le problème de décider si une telle matrice satisfait la Propriété des Uns Consécutifs avec Multiplicité peut être résolu en temps polynomial en travaillant sur le PQR-arbre de la matrice sans la colonne T. Cela permettra d'intégrer une notion génomique importante (les télomères) dans le cadre général décrit dans [Chauve 2008b].

Autres applications : levures et amniotes. Les génomes de levure présentent un cas intéressant pour la reconstruction de génomes ancestraux. Leurs génomes sont assez denses en gènes, ce qui facilite la construction de marqueurs, et leur évolution comporte une duplication de génome complet [Wolfe 1997, Kellis 2004]. Dans [Chauve 2009b], nous avons reconstruits deux génomes ancestraux : un ancêtre de quatre levures non-dupliquées, étudié dans [Jean 2009], et l'ancêtre pré-duplication de *S. cerevisiae* et *C. glabrata*. Cette étude s'est avérée intéressante d'un point de vue méthodologique. En effet, nous avons pu constater que l'ancêtre proposé dans [Jean 2009] est discutable, car il comporte de nombreuses adjacences spécifiques à certains de ses descendants et probablement non-ancestrales, dues là encore à un artefact d'une approche de type parcimonie (en termes de fusion de Régions Ancestrales Contiguës cette fois). De plus, l'application du problème du médian sur les données de [Jean 2009] montre une saturation en nombres de réarrangements et donc une perte probable de signal évolutif. L'ensemble de ces problèmes s'estompe (sans disparaître complètement toutefois) si l'on augmente la résolution des données et que l'on utilise des marqueurs couvrant bien les quatre génomes descendants de cet ancêtre. Dans le cadre de l'ancêtre pre-duplication, nous montrons au contraire que les approches de type médian [Gavranovic 2010] et de cartographie sont en accord et ont essentiellement une unique solution optimale (voir aussi [Gordon 2009]).

Finalement, un travail en cours s'intéresse au génome ancestral des amniotes (le groupe qui regroupe mammifères, marsupiaux, reptiles et oiseaux) [Ouangraoua 2011]. L'application de la méthode décrite dans [Chauve 2008b] résulte en 164 Régions Ancestrales Contiguës. Ce nombre (trop) important s'explique par plusieurs raisons. Premièrement, il est plus difficile d'obtenir de bons marqueurs du fait de la grande divergence en termes de similarité de séquence hors des gènes. De plus, la divergence évolutive entre oiseaux et mammifères, naturellement plus importante que parmi le groupe des mammifères, résulte en une perte de signal due à l'augmentation du nombre de réarrangements. La reconstruction de cet ancêtre demande donc d'intégrer des structures combinatoires moins contraintes que les Régions Ancestrales Contiguës, que nous décrivons dans la section suivante.

5.4 Groupes Synténiques Ancestraux et application aux amniotes

On veut donc corriger le fait que la construction de Régions Ancestrales Contiguës à partir des génomes amniotes actuellement disponibles résulte en un trop grand nombre de segments. Une approche qui découle naturellement des principes que nous avons suivis dans [Chauve 2008b] consiste à (1) définir une nouvelle notion de caractère ancestral, (2) définir comment la détecter dans les génomes de ses descendants, et (3) définir comment assembler ces caractères ancestraux en larges groupes.

La notion qui généralise naturellement le concept de contiguité est celle de *synténie* : un groupe de marqueurs est synténique dans un génome si ils appartiennent au même chromosome. On veut donc calculer des groupes de marqueurs synténiques dans un génome ancestral, ici celui des amniotes. Il s'agit en effet du plus ancien ancêtre vertébré dont aucun descendant n'a subi de duplication de génome complet et que l'on peut inférer avec les données disponibles actuellement.

La question maintenant est de définir une représentation combinatoire de tels groupes. On va utiliser un *graphe* : les sommets sont les marqueurs, les arêtes les paires de marqueurs supposés synténiques dans l'ancêtre, et chaque composante connexe de ce graphe représente un *Groupe Synténique Ancestral*. Pour intégrer les Régions Ancestrales Contiguës dans ce concept, on peut noter que chaque marqueur appartient à une unique région ancestrale et ce graphe induit donc un graphe dont les sommets sont ces régions et dont les arêtes et composantes connexes s'interprètent de manière similaire.

Détecter ces caractères chez les descendants de l'ancêtre est un problème plus délicat. On ne peut pas se contenter de détecter les paires de marqueurs synténiques dans une paire informative. En effet, le risque d'évolution convergente est trop important. Par exemple l'évolution de l'ancêtre des thériens vers l'opossum se caractérise par plusieurs fusions de chromosomes qui sont susceptibles de créer de l'évolution convergente pour un tel caractère. De plus la structure non-linéaire d'un graphe ne permet pas de détecter des faux positifs en termes de conflits ou obstructions comme on peut le faire dans le cadre de la Propriété des Uns Consécutifs. La piste que nous avons suivie est la suivante : on va chercher à détecter des groupes de marqueurs qui, étant donnés deux génomes G et H d'une paire informative, sont *contiguës dans* G et synténiques dans H, un concept introduit sous le nom de *Segment Autosomal Conservé* (*CSAM*) dans [Kumar 2001]. Étant donné un tel segment, on peut alors utiliser un test statistique défini dans [Durand 2003] pour en mesurer la pertinence et ne pas le prendre en compte si il ne passe pas le test. Sinon, l'ensemble des marqueurs présents dans G est supposé synténique dans l'ancêtre et on établit une arête entre toutes les paires qu'ils forment.

Dans le cas des amniotes, si on veut comparer une paire informative comportant un descendant de l'ancêtre et un "groupe extérieur", un problème supplémentaire se pose : les "groupe extérieurs"séquencés les plus proches sont les poissons téléostéens, dont l'évolution a

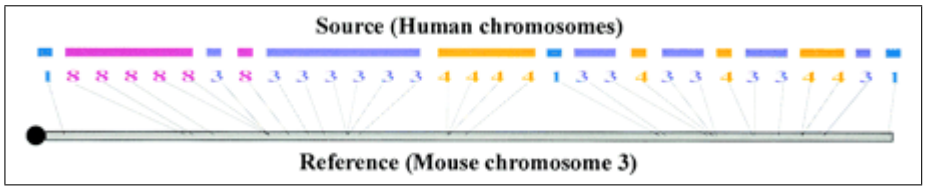

FIG. 5.4 – Exemple de CSAMs entre le chromosome 3 de la souris et le génome humain. [Kumar 2001]

été marquée par une duplication de génome complet peu après la divergence séparant poissons et amniotes [Jaillon 2004]. Cette duplication a de plus été suivie de nombreuses pertes de gènes et réarrangements [Sémon 2007b, Sémon 2007a, Hufton 2009, Sankoff 2010]. Finalement, les marqueurs amniotes n'ont pas d'occurrences dans les poissons, du fait de très faible similarité de séquence et nous devons donc utiliser des familles de gènes [Vilella 2009].

La notion de Segment Autosomal Conservé n'est donc pas adaptée à la détection de telles synténies et doit être modifiée. L'hypothèse d'une duplication de génome complet chez les téléostéens repose sur la mise en évidence d'une structure combinatoire particulière, la *Synténie Doublement Conservée*, que l'on peut définir intuitivement comme un segment de génome non-dupliqué dont les gènes ont des orthologues sur deux segments de chromosomes différents d'un génome potentiellement dupliqué. L'observation d'un tel phénomène à l'échelle d'un génome complet suggère alors fortement l'existence d'une duplication de génome complet. Pour définir des groupes synténiques en comparant un génome amniote et un génome de poisson, on va donc chercher des Synténies Doublement Conservées. Pour un segment amniote portant une Synténie Doublement Conservée, on peut donc faire l'hypothèse que les marqueurs qu'il porte étaient synténiques et définir un Groupe Synténique Ancestral.

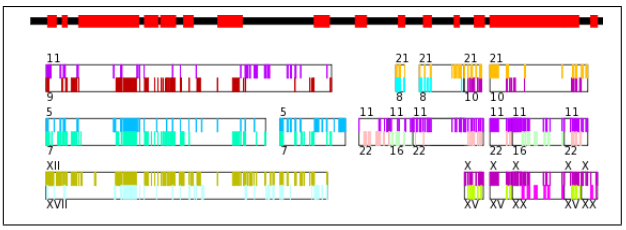

FIG. 5.5 – Exemple de Synténies Doublement Conservées (DCS) entre un segment de chromosome humain (les rectangles rouges représentent les marqueurs) et les génomes de trois téléostéens. Chaque boîte est une DCS, les barres verticales indiquant les gènes humains ayant des orthologues vers deux chromosomes de poisson. Par exemple, la première Synténie Doublement Conservée implique les chromosomes 11 et 9 du tetraodon. [Ouangraoua 2011]

Ce principe général est simple mais son implémentation n'est pas formalisée. Il a été utilisé pour analyser des génomes de plantes, de vertébrés et de levures, et dans chaque cas des méthodes ad-hoc (parfois manuelles) ont été utilisées. Dans [Ouangraoua 2011], nous proposons une méthode générique, plutôt stricte, reposant sur la détection d'équipes de gènes avec des paramètres δ différents pour les deux génomes (1 pour le génome humain, pour

conserver un caractère de contiguïté tout en prenant en compte des problèmes d'annotation de gènes par exemple, et ∞ pour le poisson, pour ne détecter qu'un caractère synténique). Malgré sa rigueur, cette méthode permet de détecter des Synténies Doublement Conservées qui couvrent une large partie du génome humain (Figure 5.6).

FIG. 5.6 – Couverture du génome humain par des DCS avec trois téléostéens. [Ouangraoua 2011]

Après avoir filtré, sur des critères de conservation phylogénétique, les arêtes entre mar-

queurs définies par ces Synténies Doublement Conservées on obtient un graphe ayant 39 composantes connexes, et donc 39 Groupes Synténiques Ancestraux, illustrés en Figure 5.7. L'analyse de ce génome ancestral est encore en cours, mais on peut déjà noter des différences importantes avec les deux seuls autres génomes ancestraux publiés [Kohn 2006, Nakatani 2007] en termes d'associations synténiques ancestrales (Table 5.1). L'approche que nous venons de survoler permet cependant de relier chaque association synténique à des segments bien définis des génomes actuels, ce qui nous a permis d'expliquer la plupart des ces différences et de relever plusieurs associations sans doute fausses dans les articles [Kohn 2006, Nakatani 2007], dues notamment à des erreurs de détection des Synténies Doublement Conservées.

[Kohn 2006]	[Nakatani 2007]	[Ouangraoua 2011]
18-19-27		18-27
1-24		1-24
21-23-26-32	26-32	21-26
17-Z	10-13-17-Z	
2-9-16	2-9	
1-3-14-18, 3-14	3-14	
4-22, 8-28		
	5-10, 1-7, 3-5	
		1-2, Z-22

Tab. 5.1 – Associations synténiques (en termes de chromosomes du poulet) dans trois propositions de génome ancestral des amniotes. [Ouangraoua 2011]

En conclusion, les travaux présentés dans cette section illustrent une approche possible pour régler le manque de signal de contiguïté pour cet ancêtre. Une solution alternative est de regrouper les Régions Ancestrales Contiguës en segments (linéaires donc) plus longs via une approche parcimonieuse par exemple [Gaul 2006, Munoz 2010]. Nous avons plutôt cherché à détecter un signal évolutif moins fort. La contrepartie est le danger de détecter en fait des caractères résultant d'un phénomène d'évolution convergente, que nous avons essayé d'éviter en combinant une méthode rigoureuse, des tests statistiques et un critère de conservation phylogénétique. Il s'agit cependant d'un chantier encore loin d'être terminé.

5.5 Conclusion

Comme le suggère la difficulté de reconstruire un génome amniote ancestral fiable, le problème de la reconstruction de génomes ancestraux est donc loin d'être résolu, notamment pour les vertébrés antérieurs aux mammifères placentaires. Dans ce chapitre nous avons ébauché un cadre méthodologique général qui semble adapté pour le génome ancestral des amniotes, pour lequel nous sommes encore loin d'un consensus. Les génomes ancestraux de plantes, et leur nombreuses duplications de génome complet sont un autre champ d'investigation qui commence juste à être exploré avec succès [Murat 2010].

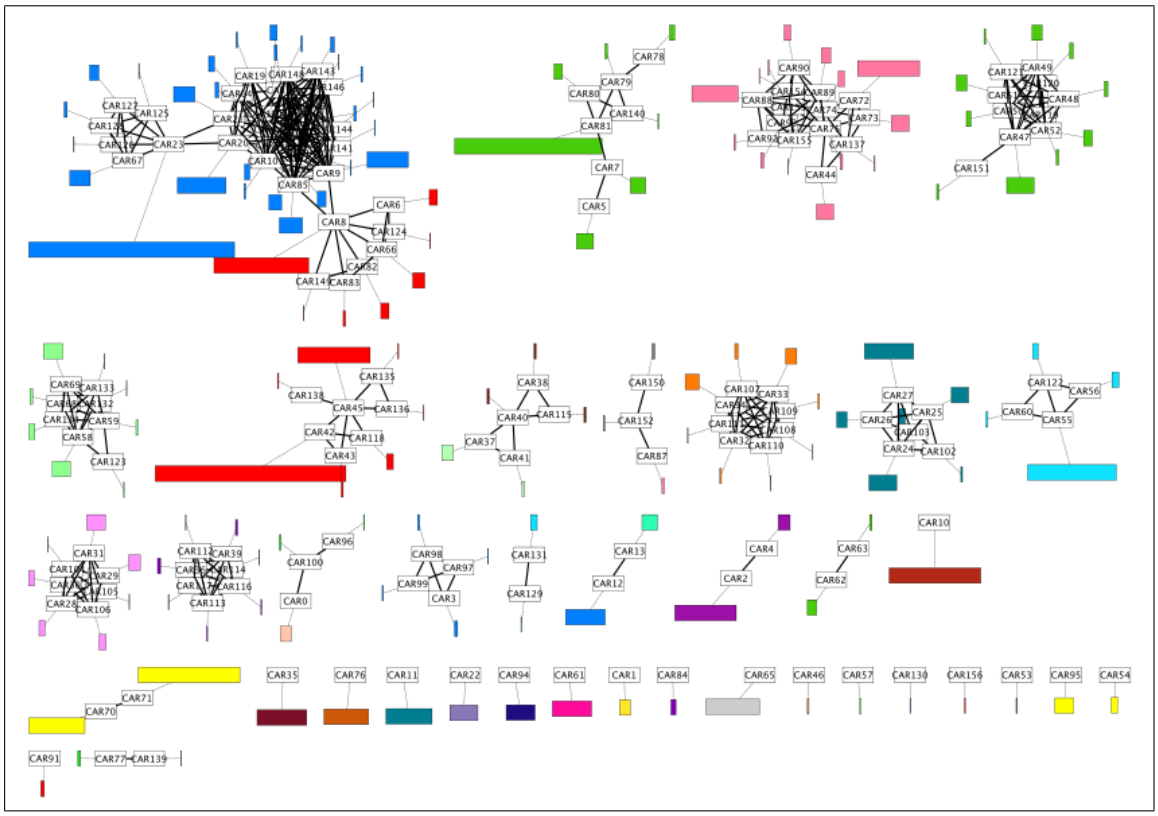

FIG. 5.7 – Les 39 Groupes Synténiques Ancestraux du génome ancestral des amniotes. Chaque rectangle représente une Région Ancestrale Contiguë. Les couleurs correspondent aux chromosomes du poulet. [Ouangraoua 2011]

Pour revenir sur les marqueurs, la contrainte de marqueurs uniques (et possiblement non universels) chez les descendants de l'ancêtre offre une certaine sécurité. Autoriser des marqueurs répétés chez les descendants nécessiterait de localiser précisément l'origine de ces duplications dans la phylogénie des espèces (chapitre A) pour s'assurer du caractère unique de tel marqueur dans l'ancêtre. La contrainte d'unicité est cependant non-nécessaire chez les "groupe extérieurs" et il serait souhaitable de la relâcher pour ces génomes, en autorisant des marqueurs répétés provenant, en théorie, de la dispersion de segments ancestraux (au niveau de l'ancêtre commun de tous les génomes étudiés) dans les lignées spécifiques des groupe extérieurs [Ma 2006]. Cela suppose de modifier la définition des caractères conservés qui définissent la matrice $\mathcal{R}$ et d'utiliser une notion d'intervalle commun avec répétition ou même d'équipe de gènes, mais restreinte au cas où un génome de référence ne contient pas de marqueurs dupliqués, un problème non étudié jusqu'à présent.

D'un point de vue méthodologique, la Propriété des Uns Consécutifs (et ses variantes) offre une base solide pour la reconstruction de Régions Ancestrales Contiguës. Cependant il reste encore un gros travail d'exploration des propriétés combinatoires et algorithmiques de ce cadre, notamment car la plupart des variantes que nous avons explorées mènent à des

problèmes difficiles. En particulier, la notion de PQ-arbre n'a pas d'équivalent naturel, à ma connaissance, pour le plupart des variantes de la Propriété des Uns Consécutifs introduites récemment dans nos travaux. Or cette possibilité de représenter tous les génomes ancestraux possible de manière compacte est un aspect central de notre travail jusqu'ici. Il s'agit d'un problème qui devra être résolu d'une façon ou d'une autre. Cela pourrait nécessiter d'abandonner l'approche combinatoire pour explorer des méthodes probabiliste d'échantillonnage et des représentations alternatives d'ensembles de génomes utilisées dans d'autres contextes, comme des ordres partiels par exemple. De manière générale, les travaux décrits dans ce mémoire ont complètement ignoré les méthodes probabilistes. Je compte explorer ces méthodes dans un futur proche, en combinant des modèles d'évolution de caractères génomiques (adjacences, intervalles communs de petite taille) à des approches de type Voyageur de Commerce déjà utilisées dans problèmes de cartographie [Faraut 2007, Servin 2010] et qui sont naturellement reliées à la Propriété des Uns Consécutifs.

Le problème de combiner plus finement Régions Ancestrales Contiguës et Groupes Synténiques Ancestraux est complètement ouvert. La séparation nette entre ces deux notions est en effet une conséquence de la modélisation combinatoire, relativement arbitraire, plutôt qu'une nécessité réelle. Tous les marqueurs d'un Groupe Synténique Ancestral ont en effet vocation à être ordonnés linéairement le long de chromosome ancestraux, tout en respectant les Régions Ancestrales Contiguës auxquelles ils appartiennent. L'intégration d'une information de distance attendue entre marqueurs dans les Groupes Synténiques Ancestraux semble être une voie prometteuse pour combiner modèles de Régions Ancestrales Contiguës (PQ-arbres, ordres partiels, ou autre), méthodes de Voyageur de Commerce et notion de Bande Passante (utilisée dans le cadre de problèmes d'assemblage de génomes [Gao 2011]) en un cadre unique. Cette approche est sans doute plus difficile à mettre en oeuvre qu'à énoncer, mais doit être explorée pour arriver à une représentation combinatoire de génomes ancestraux en segments linéaires représentant mieux leur structure réelle.

Même dans le cadre de la Propriété des Uns Consécutifs classique, il reste un problème important à explorer, celui de la confiance que l'on peut avoir dans un PQ-arbre. Si la matrice binaire obtenue a presque cette propriété on peut alors avoir confiance dans le PQ-arbre obtenu après élimination des conflits. Si, au contraire, il faut supprimer de nombreuses lignes pour éliminer les conflits, on peut avoir des doutes sur la qualité du signal de contiguïté encodé dans cette matrice[4]. Une question similaire se pose naturellement pour toute structure combinatoire représentant un génome ancestral assemblé. Ces questions mènent au problème du développement d'une mesure de qualité du signal présent dans les données (la matrice binaire contenant le signal de contiguïté ou le graphe des Groupes Synténiques Ancestraux), cette notion de qualité du signal pouvant, par exemple, être exprimée en termes d'obstruction à l'ordonnancement linéaire des marqueurs (scores probabilistes, ensembles conflictuels minimaux, ...).

L'intégration des approches "sans réarrangements" (assemblage/cartographie) et des ap-

[4]Voir [Ouangraoua 2009b] par exemple, où on peut remarquer que l'utilisation combinée des Synténies Doublement Conservées et de la Propriété des Uns Consécutifs classique est une erreur dans la reconstruction du génome amniote ancestral.

proches basées sur des scénarios parcimonieux est une question importante. Les deux approches apportent des éléments de réponse de nature différente, mais qui peuvent être reliés entre eux. Le choix de ne présenter ici que des résultats de méthodes "sans réarrangements" n'est pas basé sur un jugement de valeur a priori sur les possibilités des deux approches, mais sur la volonté d'explorer cette voie en profondeur. On peut cependant remarquer que le cadre méthodologique décrit dans ce chapitre pour la reconstruction de Régions Ancestrales Contiguës comporte deux éléments : l'utilisation de la Propriété des Uns Consécutifs pour l'assemblage de caractères potentiellement ancestraux (reprenant ainsi une méthode de cartographie physique) et la définitions de ces caractères en termes de structures génomiques conservées (que l'on peut rapprocher des techniques utilisées en cytogénétique). Rien n'interdit cependant d'utiliser une approche moins stricte pour détecter des caractères potentiellement ancestraux, notamment des méthodes de calcul de scénarios parcimonieux d'évolution par réarrangements. Par exemple, l'utilisation des caractères communs à toutes les solutions optimales d'un problème de réarrangements est un premier pas dans cette direction. De plus les progrès spectaculaires des deux types d'approches militent pour leur intégration, par exemple dans un cadre de type "Expectation-Maximisation" qui verraient des génomes ancestraux limiter les solutions possibles des problèmes de réarrangement, qui à leur tour permettraient de raffiner ces ancêtres. Nous avons effectué quelques pas dans cette direction dans [Chauve 2009b], en montrant comment les réarrangements probables définis dans [Zhao 2009] permettent d'inférer des adjacences qui appartiennent à toutes les solutions optimales dans un problème de médian, et sont donc de potentielles adjacences ancestrales. Ces premiers résultats demandent à être étendus.

Finalement, comment comparer deux propositions de génome ancestral pour le même organisme ? Cette question n'a pour le moment été qu'effleurée, en se basant sur des caractères génomiques relativement grossiers (caryotype et associations synténiques). Ce problème nécessite un effort de développement méthodologique pour pouvoir descendre au niveau des Régions Ancestrales Contiguës, notamment pour prendre en compte le fait que les génomes descendants et jeux de marqueurs utilisés par différentes méthodes peuvent être différents. Une notion de distance, ou plutôt de dissimilarité, entre génomes ancestraux, possiblement utilisant un génome de descendant de référence, est une première approche naturelle à explorer.

Contributions. Introduction des PQ-arbres pour représenter des génomes ancestraux.

- [Bergeron 2004] A. Bergeron, M. Blanchette, A. Chateau et C. Chauve. *Reconstructing Ancestral Gene Orders Using Conserved Intervals.* Algorithms in Bioinformatics, 4th International Workshop, WABI 2004, pages 14–25. Springer, 2004.

Reconstruction de CARs.

- [Chauve 2008b] C. Chauve et E. Tannier. *A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes.* PLoS Comptutational Biology, vol. 4, no. 11, page e1000234, 2008.
- [Chauve 2009b] C. Chauve, H. Gavranovic, A. Ouangraoua et E. Tannier. *Yeast ancestral genome reconstructions : the possibilities of computational methods II.* Journal of Computational Biology, vol. 17, no. 9, pages 1097–1112, 2009.

– [Gavranovic 2011] H. Gavranovic, C. Chauve, J. Salse et E. Tannier. *Mapping ancestral genomes with massive gene loss : a matrix sandwich problem*. Bioinformatics, vol. 27, no. 13, 2011. Actes de ISMB/ECCB 2011, sous presse.

Le graphe des synténies ancestrales et les DCS chez les vertébrés.

– [Ouangraoua 2009b] A. Ouangraoua, F. Boyer, A. McPherson, E. Tannier et C. Chauve. *Prediction of Contiguous Regions in the Amniote Ancestral Genome*. Bioinformatics Research and Applications, 5th International Symposium, ISBRA 2009, pages 173–185. Springer, 2009.
– [Ouangraoua 2011b] A. Ouangraoua, E. Tannier et C. Chauve. Reconstructing the architecture of the ancestral amniote genome. Soumis à *Bioinformatics*, en cours de révision, 2011.

Étude combinatoire et algorithmes de la Propriété des Uns Consécutifs.

– [Chauve 2009c] C. Chauve, J. Manuch et M. Patterson. *On the Gapped Consecutive-Ones Property*. Electronic Notes in Discrete Mathematics, vol. 34, pages 121–125, 2009. Actes de EuroComb 2009, version étendue soumise à Discrete Applied Mathematics, en cours de révision.
– [Chauve 2010] C. Chauve, U.-U. Haus, T. Stephen et V. P. You. *Minimal Conflicting Sets for the Consecutive Ones Property in Ancestral Genome Reconstruction*. Journal of Computational Biology, vol. 17, no. 9, pages 1167–1181, 2010. Version préliminaire publiée dans les actes de RECOMB-CG 2009.
– [Chauve 2011] C. Chauve, J. Manuch, M. Patterson et R. Wittler. *Tractability results for the Consecutive-Ones Property with multiplicity*. Combinatorial Pattern Matching, 21st Annual Symposium, CPM 2011. Springer, 2011. Sous presse.
– [Chauve 2011b] C. Chauve, T. Stephen et M. Tamayo. *Output-sensitive enumeration of Tucker patterns*. Poster présenté à *IWOCA 2011*, 2011.

Conclusion et perspectives 6

Pour débuter cette conclusion, il me semble important de replacer dans un contexte général l'ensemble des travaux présentés dans les chapitres précédents. Ils ne forment en effet pas une collection plus ou moins disparate de résultats, mais suivent une ligne directrice qui s'est clarifiée au cours des années et est maintenant bien définie. Ils tendent vers le but ambitieux d'intégrer génomes séquencés et génomes ancestraux inférés dans un cadre méthodologique à la fois général et robuste pour la génomique évolutive [Muffato 2008]. Dans un cadre idéal, les génomes ancestraux seraient obtenus via des techniques d'assemblage ou de cartographie, indépendamment d'approches parcimonieuses, qui seraient utilisées pour l'étude de l'évolution le long de branches individuelles de la phylogénie des génomes étudiés.

Mes travaux des dix dernières années correspondent à une première phase de ce programme, centrée en grande partie sur l'utilisation des intervalles communs et PQ-arbres dans l'analyse de génomes encodés par permutations signées. Les questions abordées au cours des trois dernières années amorcent une seconde phase, plus générale car sortant du cadre des seules permutations signées.

6.1 Conclusion

La plupart des questions abordées dans ce mémoire trouvent leur source dans deux articles parus en 2001 et 2002. L'article [Bergeron 2002a] porte sur une question naturelle : peut-on calculer tous les scénarios parcimonieux par inversions entre deux permutations signées ? Sans totalement répondre à cette question, cet article montre que le nombre de tels scénarios peut être énorme. On se trouve donc face au problème classique de la multiplicités de solutions optimales, qui soulève la question de la confiance que l'on peut avoir en l'un de ces scénarios, obtenu arbitrairement par un algorithme parmi l'ensemble de toutes les solutions possibles. À peu près au même moment, Steffen Heber et Jens Stoye publient un article sur le calcul des intervalles communs entre permutations [Heber 2001], un concept qui apparaît vite comme important pour l'étude des réarrangements génomiques. Ces deux éléments sont à la base des résultats présentés dans ce mémoire.

La question des modèles combinatoires de conservation de structures génomiques est sans cesse présente. Les intervalles communs sont des objets très bien adaptés à l'étude des données

basées sur des marqueurs uniques et universels ; ils capturent un signal évolutif clair et important tout en ayant des propriétés combinatoires rigoureuses menant à des algorithmes efficaces. L'utilisation des intervalles communs dans le calcul de scénarios parfaits et la reconstruction de génomes ancestraux récents illustrent ces propriétés. Le débat de 2006 sur les divergences entre approches cytogénétiques et approches *in silico* pour la reconstruction de génomes ancestraux, combiné à mes premières discussions avec Éric Tannier, ont contribué à ré-orienter mes travaux vers les problèmes de reconstruction de génomes ancestraux. Nos premiers résultats reposent uniquement sur les intervalles communs et les PQ-arbres, concepts qui s'avèrent en fait suffisants pour étudier l'ancêtre des mammifères placentaires. Ces travaux closent en quelque sorte cette séquence centrée sur les intervalles communs. Nos efforts depuis 2008 pour la reconstruction du génome de l'ancêtre des amniotes ouvrent un second volet, centré sur l'exploration et le développement de concepts adaptés à l'inférence de génomes ancestraux plus anciens.

Je distingue deux grandes lignes dans ce travail sur les génomes ancestraux, que j'aimerais prendre le temps de décrire et de justifier : un effort de développement méthodologique et de modélisation, illustré sur des jeux de données réels et, en parallèle, l'exploration des propriétés mathématiques et algorithmiques de ces modèles.

Concernant le premier point, j'estime que notre travail méthodologique est important et nécessaire pour mieux comprendre, discuter, voire critiquer les nombreuses méthodes existantes. En particulier, je crois fermement, qu'en l'état actuel, il n'est pas réaliste d'évaluer des méthodes de reconstruction de génomes ancestraux par des simulations. La connaissance des propriétés stochastiques de l'évolution de génomes complets (fréquence des différents mécanismes de réarrangements, de leur longueur, des propriétés des points de cassure, ...) est trop préliminaire. De plus, à l'exception de quelques génomes ancestraux de mammifères ou de levures, il n'existe pas de standard auquel on peut comparer les résultats obtenus. Je défend une approche qui repose sur des méthodes de reconstruction rigoureuses, dont le but est de détecter du signal évolutif clairement défini, en se basant sur des principes méthodologiques clairement posés. En outre, l'obtention de génomes ancestraux devrait aider, en combinaison avec des méthodes d'analyse plus classiques de réarrangement de génomes, à préciser les propriétés stochastiques des processus d'évolution par réarrangements génomiques. Pour accompagner de tels travaux méthodologiques, la définition de modèles de génomes ancestraux adaptés à la nature des données disponibles est nécessaire. L'étude de la Propriété des Uns Consécutifs et de ses variantes, un sujet d'une grande richesse combinatoire et algorithmique, est un aspect fondamental de cet effort.

6.2 Perspectives

Ce regard rétrospectif sur mes travaux passés m'amène naturellement à esquisser leur prolongement. Dans un élan d'optimisme au moment d'écrire ces lignes, je pense que, d'ici quelques années, les techniques de reconstruction de l'organisation de génomes ancestraux auront progressé au point de produire des génomes ancestraux relativement bien définis et

acceptés par la communauté, pour les vertébrés tout au moins. L'essentiel de mes projets à court et moyen terme vont participer à cet effort.

Comme je l'indiquais plus haut, dans un tel cadre les distances et scénarios évolutifs seraient donc calculés branche par branche et non plus entre paires de feuilles. La notion de scénario parfait, notamment asymétrique [Braga 2009a], prend alors tout son sens du fait de l'élimination des problèmes d'évolution convergente. Par contre cela nécessitera des développements dans le calcul de distances et l'échantillonnage de scénarios évolutifs entre génomes ambigus, un sujet qui n'a été abordé que de manière très préliminaire jusqu'ici. L'arrivée prochaine d'un déluge de données, souvent de qualité discutable [Consortium 2009] et représentant en fait les génomes actuels avec une certaine dose d'ambigu¨é, rend cette problématique très importante. Les développements phénoménaux des vingt dernières années dans ce domaine [Fertin 2009] permettent cependant d'envisager ces questions avec optimisme.

En disposant de bons génomes ancestraux, on pourra aussi étudier plus directement des questions telles que le lien entre régulation et évolution des chromosomes [Mongin 2009], ou encore l'évolution de génomes après une duplication de génome complet [Sankoff 2010]. De plus, en utilisant des méthodes de reconstruction de séquences d'ADN ancestrales [Paten 2008b][1], on pourra certainement étudier plus précisément des questions comme l'évolution de la structure en isochore des génomes [Romiguier 2010] ou les caractéristiques des régions de cassure [Sankoff 2009]. Il s'agit là de questions appliquées qu'il me semble important d'attaquer pour illustrer le potentiel des approches générales que je viens de décrire.

Ce travail doit débuter par une remise à plat des méthodes de construction de marqueurs. Cette étape est trop importante dans tout travail de génomique comparée pour être occultée, voire même parfois négligée. Il s'agit là d'un effort de recherche que je compte entreprendre (poursuivre en fait), notamment en développant des méthodes de *comparaison* et *évaluation/validation* de jeux de marqueurs. De telles questions, pourtant basiques, n'ont pour le moment été traitées que superficiellement. Cet effort doit être dirigé à la fois vers le calcul de familles de blocs (pour les génomes de mammifères) et vers le calcul de familles de gènes (pour les génomes de vertébrés plus anciens). Idéalement, on aimerait conclure ce travail par la mise à la disposition de la communauté de jeux de marqueurs de référence.

De manière générale, la conclusion du chapitre 5 illustre l'étendue du travail restant à faire dans les méthodes de reconstruction de génomes ancestraux. Il faut en particulier se pencher sur le développement de modèles, possiblement hiérarchiques, capables à la fois d'utiliser les différents types de signaux évolutifs qui peuvent être détectés et de représenter les génomes ancestraux en structures les plus linéaires possibles. Pour le moment les modèles utilisés sont en général spécifiques à un ancêtre particulier. Le développement de modèles génériques permettant d'analyser l'évolution d'un large ensemble de génomes actuels et ancestraux est un but que je compte poursuivre. En particulier, il n'est ni réaliste, ni raisonnable en fait, d'espérer se couper complètement des approches parcimonieuses, notamment du fait du lien

[1] Cependant, pour les génomes de mammifères tout au moins, l'obtention de génomes ancestraux complets incluant les séquences d'ADN chromosomes demandera de mieux comprendre les nombreuses duplications, segmentales notamment, qu'ils contiennent et leur histoire évolutive.

théorique naturel entre structures conservées et scénarios parcimonieux. Par exemple, ignorer les résultats récents comme la définition d'adjacences présentes dans toutes les solutions d'un problème de médian ou les algorithmes efficaces d'échantillonnage de scénarios parcimonieux, serait sans doute une erreur. Une meilleure intégration des deux approches est une des questions que je compte aborder.

De même que les progrès des techniques de séquençage ne se sont pas arrêtés à la suite de l'obtention de la séquence du génome humain, je pense que l'assemblage de génomes ancestraux est un sujet qui ne sera jamais complètement résolu et demandera sans cesse de nouvelles avancées. Sans vouloir m'enfermer dans une thématique de recherche trop étroite, je suis persuadé de l'importance de ces questions et je compte consacrer la majeure partie de mon temps dans les prochaines années à les faire progresser.

On ne peut cependant pas passer sous silence d'autres domaines de recherche qui sont naturellement liés aux perspectives que je viens de décrire. Les aspects méthodologiques de la reconstruction de génomes ancestraux peuvent sans doute être adaptés à d'autres structures biologiques. À plus long terme, on peut donc espérer se diriger vers des travaux de biologie systémique au niveau des espèces éteintes, comme par exemple lier l'évolution des génomes et des réseaux d'interactions de protéines. Finalement on peut décrire ces problèmes en termes d'assemblage de génomes non disponibles directement. Ce type de problème se pose dans d'autre cadres, comme par exemple l'assemblage de génomes de tumeurs ou l'assemblage de métagénomes viraux. Il est donc tentant d'explorer ces problèmes et de comprendre quelles techniques ou idées développées pour les génomes ancestraux peuvent s'appliquer dans ces cadres sensiblement différents [Wittler 2011]. Étant donné le foisonnement d'activités autour de ces problèmes à Vancouver, je compte explorer quelques unes de ces pistes nouvelles dans un futur proche.

Bibliographie

[Adam 2007] Z. Adam, M. Turmel, C. Lemieux et D. Sankoff. *Common Intervals and Symmetric Difference in a Model-Free Phylogenomics, with an Application to Streptophyte Evolution*. Journal of Computational Biology, vol. 14, pages 436–445, 2007. 40, 41

[Akerborg 2009] O. Akerborg, B. Sennblad, L. Arvestad et J. Lagergren. *Simultaneous Bayesian gene tree reconstruction and reconciliation analysis*. Proceedings of the National Academy of Sciences of the United States of America, vol. 106, no. 14, pages 5714–5719, 2009. 87

[Alekseyev 2009] M.A. Alekseyev et P.A. Pevzner. *Breakpoint graphs and ancestral genome reconstructions*. Genome Research, vol. 19, no. 5, pages 943–957, 2009. 40, 41

[Alizadeh 1995] F. Alizadeh, R.M. Karp, D.K. Weisser et G. Zweig. *Physical mapping of chromosomes using unique probes*. Journal of Computational Biology, vol. 2, no. 2, pages 159–184, 1995. 7, 44

[Angibaud 2007] S. Angibaud, G. Fertin, I. Rusu et S. Vialette. *A General Framework for Computing Rearrangement Distances between Genomes with Duplicates*. Journal of Computational Biology, vol. 14, no. 4, pages 379–393, 2007. 22

[Angibaud 2008] S. Angibaud, G. Fertin, I. Rusu, A. Thévenin et S. Vialette. *Efficient Tools for Computing the Number of Breakpoints and the Number of Adjacencies between two Genomes with Duplicate Genes*. Journal of Computational Biology, vol. 15, no. 8, pages 1093–1115, 2008. 22

[Angibaud 2009] S. Angibaud, G. Fertin, I. Rusu, A. Thévenin et S. Vialette. *On the Approximability of Comparing Genomes with Duplicates*. Journal of Graph Algorithms and Applications, vol. 13, no. 1, pages 19–53, 2009. 22

[Arvestad 2004] L. Arvestad, A.-C. Berglund, J. Lagergren et B. Sennblad. *Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution*. In P.E. Bourne et D. Gusfield, éditeurs, Proceedings of the Eighth Annual International Conference on Computational Molecular Biology, 2004, San Diego, California, USA, March 27-31, 2004, pages 326–335. ACM, 2004. 86, 87

[Arvestad 2009] L. Arvestad, J. Lagergren et B. Sennblad. *The gene evolution model and computing its associated probabilities*. Journal of the ACM, vol. 56, no. 2, pages 1–44, 2009. 87

[Badr 2010] G. Badr, K. M. Swenson et D. Sankoff. *Listing All Parsimonious Reversal Sequences : New Algorithms and Perspectives*. In E. Tannier, éditeur, Comparative Genomics - International Workshop, RECOMB-CG 2010, Ottawa, Canada, October 9-11, 2010. Proceedings, volume 6398 of *Lecture Notes in Computer Science*, pages 39–49. Springer, 2010. 30, 34

[Bansal 2008] M.S. Bansal et O. Eulenstein. *An $\Omega(n^2/\log n)$ speed-up of TBR heuristics for the gene-duplication problem*. IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 5, no. 4, pages 514–524, 2008. 88

[Bansal 2009] M.S. Bansal, O. Eulenstein et A. Wehe. *The Gene-Duplication Problem : Near-Linear Time Algorithms for NNI Based Local Searches.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 6, no. 2, pages 221–231, 2009. 88

[Bansal 2011] M.S. Bansal et R. Shamir. *A Note on the Fixed Parameter Tractability of the Gene-Duplication Problem.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 8, no. 3, pages 848–850, 2011. 88

[Batada 2007] N. Batada et L.D. Hurst. *Evolution of chromosome organization driven by selection for reduced gene expression noise.* Nature Genetics, vol. 39, pages 945–949, 2007. 21

[Béal 2004] M.-P. Béal, A. Bergeron, S. Corteel et M. Raffinot. *An algorithmic view of gene teams.* Theoretical Computer Science, vol. 320, no. 2-3, pages 395–418, 2004. 20

[Belcaid 2007] M. Belcaid, A. Bergeron, A. Chateau, C. Chauve, Y. Gingras, G. Poisson et M. Vendette. *Exploring Genome Rearrangements using Virtual Hybridization.* In D. Sankoff, L. Wang et F. Chin, éditeurs, Proceedings of 5th Asia-Pacific Bioinformatics Conference, APBC 2007, 15-17 January 2007, Hong Kong, China, volume 5 of *Advances in Bioinformatics and Computational Biology*, pages 205–214. Imperial College Press, 2007. 12

[Bérard 2004] S. Bérard, A. Bergeron et C. Chauve. *Conservation of Combinatorial Structures in Evolution Scenarios.* In J. Lagergren, éditeur, Comparative Genomics, RECOMB 2004 International Workshop, RCG 2004, Bertinoro, Italy, October 16-19, 2004, Revised Selected Papers, volume 3388 of *Lecture Notes in Computer Science*, pages 1–14. Springer, 2004. 33

[Bérard 2007] S. Bérard, A. Bergeron, C. Chauve et C. Paul. *Perfect Sorting by Reversals Is Not Always Difficult.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 4, no. 1, pages 4–16, 2007. Version préliminaire publiée dans les actes de WABI 2005. 31, 32, 33, 36

[Bérard 2008] S. Bérard, C. Chauve et C. Paul. *A more efficient algorithm for perfect sorting by reversals.* Information Processing Letters, vol. 106, no. 3, pages 90–95, 2008. 33

[Bérard 2009] S. Bérard, A. Chateau, C. Chauve, C. Paul et E. Tannier. *Computation of Perfect DCJ Rearrangement Scenarios with Linear and Circular Chromosomes.* Journal of Computational Biology, vol. 16, no. 10, pages 1287–1309, 2009. Version préliminaire publiée dans les actes de RECOMB-CG 2008. 34, 35

[Bergeron 2002a] A. Bergeron, C. Chauve, T. Hartman et K. St-Onge. *On the properties of sequences of reversals that sort a signed permutation.* In Journées Ouvertes en Biologie, Informatique et Mathématiques, pages 99–108, 2002. 27, 30, 33, 59

[Bergeron 2002b] A. Bergeron, S. Heber et J. Stoye. *Common intervals and sorting by reversals : a marriage of necessity.* Bioinformatics, vol. 18, no. Suppl. 2, pages S54–S63, 2002. 21

[Bergeron 2004] A. Bergeron, M. Blanchette, A. Chateau et C. Chauve. *Reconstructing Ancestral Gene Orders Using Conserved Intervals.* In I. Jonassen et J. Kim, éditeurs,

Algorithms in Bioinformatics, 4th International Workshop, WABI 2004, Bergen, Norway, September 17-21, 2004, Proceedings, volume 3240 of *Lecture Notes in Computer Science*, pages 14–25. Springer, 2004. 41, 46, 47

[Bergeron 2005] A. Bergeron, J. Mixtacki et J. Stoye. Mathematics of evolution and phylogeny (édité par O. Gascuel), chapitre The inversion distance problem, pages 262–290. Oxford University Press, 2005. 21, 27, 28

[Bergeron 2006a] A. Bergeron, J. Mixtacki et J. Stoye. *A Unifying View of Genome Rearrangements*. In P.Bucher et B.M.E. Moret, éditeurs, Algorithms in Bioinformatics, 6th International Workshop, WABI 2006, Zurich, Switzerland, September 11-13, 2006, Proceedings, volume 4175 of *Lecture Notes in Computer Science*, pages 163–173. Springer, 2006. 29

[Bergeron 2006b] A. Bergeron et J. Stoye. *On the similarity of sets of permutations and its applications to genome comparison*. Journal of Computational Biology, vol. 13, no. 7, pages 1340–1354, 2006. Version préliminaire publiée dans les actes de COCOON 2003. 21, 22

[Bergeron 2008a] A. Bergeron, C. Chauve, F. de Montgolfier et M. Raffinot. *Computing Common Intervals of K Permutations, with Applications to Modular Decomposition of Graphs*. SIAM Journal on Discrete Mathematics, vol. 22, no. 3, pages 1022–1039, 2008. Version préliminaire publiée dans les actes de ESA 2005. 15, 19

[Bergeron 2008b] A. Bergeron, C. Chauve et Y. Gingras. Bioinformatics algorithms : Techniques and applications, chapitre Formal Models of Gene Clusters, pages 177–202. Wiley Series in Bioinformatics. Wiley Interscience, 2008. 15

[Bergeron 2008c] A. Bergeron, J. Mixtacki et J. Stoye. *On Computing the Breakpoint Reuse Rate in Rearrangement Scenarios*. In C.E. Nelson et S. Vialette, éditeurs, Comparative Genomics, International Workshop, RECOMB-CG 2008, Paris, France, October 13-15, 2008. Proceedings, volume 5267 of *Lecture Notes in Computer Science*, pages 226–240. Springer, 2008. 30

[Bergeron 2009] A. Bergeron, J. Mixtacki et J. Stoye. *A new linear time algorithm to compute the genomic distance via the double cut and join distance*. Theoretical Computer Science, vol. 410, no. 51, pages 5300–5316, 2009. 28

[Bernt 2006] M. Bernt, D. Merkle et M. Middendorf. *Genome rearrangement based on reversals that preserve conserved intervals*. IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 3, no. 3, pages 275–288, 2006. 37

[Bernt 2008] M. Bernt, D. Merkle et M. Middendorf. *Solving the preserving reversal median problem*. IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 5, no. 3, pages 332–347, 2008. 37

[Berry 2008] V. Berry. *On building and comparing trees : Application to supertrees in phylogenetics*. Habilitation à Diriger des Recherches, Université Montpellier II, 2008. 90

[Bertrand 2010] D. Bertrand, Y. Gagnon, M. Blanchette et N. El-Mabrouk. *Reconstruction of Ancestral Genome Subject to Whole Genome Duplication, Speciation, Rearrangement*

and Loss. In V. Moulton et M. Singh, éditeurs, Algorithms in Bioinformatics, 10th International Workshop, WABI 2010, Liverpool, UK, September 6-8, 2010. Proceedings, volume 6293 of *Lecture Notes in Computer Science*, pages 78–89. Springer, 2010. 41, 83

[Bhutkar 2007] A. Bhutkar, W.M. Gelbart et T.F. Smith. *Inferring genome-scale rearrangement phylogeny and ancestral gene order : a* Drosophilia *case study*. Genome Biology, vol. 8, page R236, 2007. 40

[Bie 2006] T. De Bie, N. Cristianini, J.P. Demuth et M.W. Hahn. *CAFE : a computational tool for the study of gene family evolution*. Bioinformatics, vol. 22, no. 10, pages 1269–1271, 2006. 48

[Bininda-Edmonds 2004] O.R.P Bininda-Edmonds, éditeur. Phylogenetic supertrees : Combining information to reveal the tree of life. Kluwer, 2004. 90

[Blanchette 2007] M. Blanchette. *Computation and analysis of genomic multi-sequence alignments*. Annual Review of Genomics and Human Genetics, vol. 8, pages 193–213, 2007. 12

[Blin 2004] G. Blin, C. Chauve et G. Fertin. *The breakpoint distance for signed sequences*. In K. S. Gimaraes et M.-F. Sagot, éditeurs, CompBioNets 2004 : Algorithms and Computational Methods for Biochemical and Evolutionary Networks, volume 3 of *Text in Algorithms*, pages 3–16. College Publications, 2004. 22

[Blin 2005] G. Blin, C. Chauve et G. Fertin. *Genes Order and Phylogenetic Reconstruction : Application to γ-Proteobacteria*. In A. McLysaght et D. H. Huson, éditeurs, Comparative Genomics, RECOMB 2005 International Workshop, RCG 2005, Dublin, Ireland, September 18-20, 2005, Proceedings, volume 3678 of *Lecture Notes in Computer Science*, pages 11–20. Springer, 2005. 22

[Blin 2006] G. Blin, A. Chateau, C. Chauve et Y. Gingras. *Inferring Positional Homologs with Common Intervals of Sequences*. In G. Bourque et N. El-Mabrouk, éditeurs, Comparative Genomics, RECOMB 2006 International Workshop, RCG 2006, Montreal, Canada, September 24-26, 2006, Proceedings, volume 4205 of *Lecture Notes in Computer Science*, pages 24–38. Springer, 2006. 21

[Blin 2007a] G. Blin, E. Blais, D. Hermelin, P. Guillon, M. Blanchette et N. El-Mabrouk. *Gene maps linearization using genomic rearrangement distances*. Journal of Computational Biology, vol. 14, no. 4, pages 394–407, 2007. 7

[Blin 2007b] G. Blin, C. Chauve, G. Fertin, R. Rizzi et S. Vialette. *Comparing Genomes with Duplications : A Computational Complexity Point of View*. IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 4, no. 4, pages 523–534, 2007. Version préliminaire publiée dans les actes de IWBRA 2006. 22

[Böcker 2009] S. Böcker, K. Jahn, J. Mixtacki et J. Stoye. *Computation of Median Gene Clusters*. Journal of Computational Biology, vol. 16, no. 8, pages 1085–1099, 2009. 20

[Bonizzoni 2005] P. Bonizzoni, G. Della Vedova et R. Dondi. *Reconciling a gene tree to a species tree under the duplication cost model*. Theoretetical Computer Science, vol. 347, no. 1-2, pages 36–53, 2005. 86

[Booth 1976] K.S. Booth et G. S. Lueker. *Testing for the Consecutive Ones Property, Interval Graphs, and Graph Planarity Using PQ-Tree Algorithms*. Journal of Computer Systems and Sciences, vol. 13, no. 3, pages 335–379, 1976. 16, 19, 43

[Bourque 2002] G. Bourque et P.A. Pevzner. *Genome-scale evolution : reconstructing gene orders in the ancestral species*. Genome Research, vol. 12, no. 1, pages 26–36, 2002. 36, 40, 41

[Bourque 2004a] G. Bourque, P.A. Pevzner et G. Tesler. *Reconstructing the genomic architecture of ancestral mammals : lessons from human, mouse and rat genomes*. Genome Research, vol. 14, no. 4, pages 507–516, 2004. 16

[Bourque 2004b] G. Bourque, P.A. Pevzner et G. Tesler. *Reconstructing the genomic architecture of ancestral mammals : lessons from human, mouse and rat genomes*. Genome Research, vol. 14, no. 4, pages 507–516, 2004. 33, 36, 40, 41

[Bourque 2005a] G. Bourque, Y. Yacef et N. El-Mabrouk. *Maximizing Synteny Blocks to Identify Ancestral Homologs*. In A. McLysaght et D.H. Huson, éditeurs, Comparative Genomics, RECOMB 2005 International Workshop, RCG 2005, Dublin, Ireland, September 18-20, 2005, Proceedings, volume 3678 of *Lecture Notes in Computer Science*, pages 21–34. Springer, 2005. 22

[Bourque 2005b] G. Bourque, E.M. Zdobnov, P. Bork, P.A. Pevzner et G. Tesler. *Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages.* Genome Research, vol. 15, no. 1, pages 98–110, 2005. 36, 40, 41

[Bourque 2006] G. Bourque, G. Tesler et P.A. Pevzner. *The convergence of cytogenetics and rearrangement-based models for ancestral genome reconstruction*. Genome Research, vol. 16, pages 311–313, 2006. 40, 41, 45

[Bouvel 2009] M. Bouvel, C. Chauve, M. Mishna et D. Rossin. *Average-Case Analysis of Perfect Sorting by Reversals*. In G. Kucherov et E. Ukkonen, éditeurs, Combinatorial Pattern Matching, 20th Annual Symposium, CPM 2009, Lille, France, June 22-24, 2009, Proceedings, volume 5577 of *Lecture Notes in Computer Science*, pages 314–325. Springer, 2009. Version étendue soumise à *Discrete Mathematics Algorithms and Applications*, en cours de révision. 33, 36

[Boyer 2005] F. Boyer, A. Morgat, L. Labarre, J. Pothier et A. Viari. *Syntons, metabolons and interactons : an exact graph-theoretical approach for exploring neighbourhood between genomic and functional data*. Bioinformatics, vol. 21, no. 23, 2005. 20

[Braga 2008a] M.D.V. Braga. *L'espace de solutions du tri par inversions et son utilisation dans l'analyse de réarrangements de génomes*. Thèse de doctorat, Université Claude Bernard Lyon I, 2008. 30, 34

[Braga 2008b] M.D.V. Braga, M.-F. Sagot, C. Scornavacca et E. Tannier. *Exploring the Solution Space of Sorting by Reversals, with Experiments and an Application to Evolution*. IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 5, no. 3, pages 348–356, 2008. 30, 34

[Braga 2009a] M.D.V. Braga, C. Gautier et M.F. Sagot. *An asymmetric approach to preserve common intervals while sorting by reversals.* Algorithms for Molecular Biology, vol. 4, page 16, 2009. 36, 61

[Braga 2009b] M.D.V. Braga et J. Stoye. *Counting All DCJ Sorting Scenarios.* In F. Ciccarelli et I. Miklós, éditeurs, Comparative Genomics, International Workshop, RECOMB-CG 2009, Budapest, Hungary, September 27-29, 2009. Proceedings, volume 5817 of *Lecture Notes in Computer Science*, pages 36–47. Springer, 2009. 30

[Brouwer 2008] R.W Brouwer. *The relative value of operon predictions.* Briefings in Bioinformatics, vol. 9, no. 5, pages 367–375, 2008. 21

[Bryant 1997] D. Bryant. *Hunting for trees, building trees and comparing trees : theory and methods in phylogenetic analysis.* Thèse de doctorat, Department of Mathematics, University of Canterbury, New Zealand, 1997. 88

[Bui-Xuan 2005] B.M. Bui-Xuan, M. Habib et C. Paul. *Revisiting T. Uno and M. Yagiura's Algorithm.* In X. Deng et D.-Z. Du, éditeurs, Algorithms and Computation, 16th International Symposium, ISAAC 2005, Sanya, Hainan, China, December 19-21, 2005, Proceedings, volume 3827 of *Lecture Notes in Computer Science*, pages 146–155. Springer, 2005. 15

[Bui-Xuan. 2008] B.-M. Bui-Xuan. *Tree-representation of set families in graph decompositions and efficient algorithms.* Thèse de doctorat, Université Montpellier II, 2008. 15

[Burgetz 2007] I.J. Burgetz, S. Shariff, A. Pang et E.R. Tillier. *Positional homology in bacterial genomes.* Evolutionary Bioinformatics Online, vol. 2, pages 77–90, 2007. 21

[Burleigh 2011] J. G. Burleigh, M. S. Bansal, O. Eulenstein, S. Hartmann, A. Wehe et T. J. Vision. *Genome-scale phylogenetics : Inferring the plant tree of life from 18,896 discordant gene trees.* Systematic Biology, vol. 60, no. 2, pages 117–125, 2011. 88

[Byrka 2010] J. Byrka, S. Guillemot et J. Jansson. *New results on optimizing rooted triplets consistency.* Discrete Applied Mathematics, vol. 158, no. 11, pages 1136–1147, 2010. 88

[Catchen 2008] J.M. Catchen, J.S. Conery et J.H. Postlethwait. *Inferring Ancestral Gene Order.* In J.M. Keith, éditeur, Bioinformatics, Volume I : Data, analysis, and Evolution, volume 452, pages 365–383. Humana Press, Springer, 2008. 40

[Chauve 2006] C. Chauve, Y. Diekmann, S. Heber, J. Mixtacki, S. Rahmann et J. Stoye. *On Common Intervals with Errors.* Rapport technique 2006-02, Technische Fakultät der Universität Bielefeld, 2006. 20

[Chauve 2008a] C. Chauve, J.-P. Doyon et N. El-Mabrouk. *Gene Family Evolution by Duplication, Speciation, and Loss.* Journal of Computational Biology, vol. 15, no. 8, pages 1043–1062, 2008. Version préliminaire publiée dans les actes de RECOMB-CG 2007. 89, 90

[Chauve 2008b] C. Chauve et E. Tannier. *A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes.* PLoS Comptutational Biology, vol. 4, no. 11, page e1000234, 2008. 43, 44, 45, 46, 48, 49, 50

[Chauve 2009a] C. Chauve et N. El-Mabrouk. *New Perspectives on Gene Family Evolution : Losses in Reconciliation and a Link with Supertrees.* In S. Batzoglou, éditeur, Research in Computational Molecular Biology, 13th Annual International Conference, RECOMB 2009, Tucson, AZ, USA, May 18-21, 2009. Proceedings, volume 5541 of *Lecture Notes in Computer Science*, pages 46–58. Springer, 2009. 86, 90

[Chauve 2009b] C. Chauve, H. Gavranovic, A. Ouangraoua et E. Tannier. *Yeast ancestral genome reconstructions : the possibilities of computational methods II.* Journal of Computational Biology, vol. 17, no. 9, pages 1097–1112, 2009. 49, 56

[Chauve 2009c] C. Chauve, J. Manuch et M. Patterson. *On the Gapped Consecutive-Ones Property.* Electronic Notes in Discrete Mathematics, vol. 34, pages 121–125, 2009. Actes de EuroComb 2009, version étendue soumise à Discrete Applied Mathematics, en cours de révision. 48

[Chauve 2010] C. Chauve, U.-U. Haus, T. Stephen et V. P. You. *Minimal Conflicting Sets for the Consecutive Ones Property in Ancestral Genome Reconstruction.* Journal of Computational Biology, vol. 17, no. 9, pages 1167–1181, 2010. Version préliminaire publiée dans les actes de RECOMB-CG 2009. 47

[Chauve 2011a] C. Chauve, J. Manuch, M. Patterson et R. Wittler. *Tractability results for the Consecutive-Ones Property with multiplicity.* In R. Giancarlo et G. Manzini, éditeurs, Combinatorial Pattern Matching, 21st Annual Symposium, CPM 2011, Palermo, Italy, June 27-29, 2011. Proceedings, Lecture Notes in Computer Science. Springer, 2011. Sous presse. 49

[Chauve 2011b] C. Chauve, T. Stephen et M. Tamayo. Output-sensitive enumeration of Tucker patterns. Poster présenté à *IWOCA 2011*, 2011. 47

[Chen 2005] X. Chen, J. Zheng, Z. Fu, P. Nan, Y. Zhong, S. Lonardi et T. Jiang. *Assignment of orthologous genes via genome rearrangement.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 2, no. 3, pages 302–315, 2005. 22

[Consortium 2009] Genome 10K Consortium. *Genome 10K : A proposal to obtain whole-genome sequence for 10 000 vertebrate species.* Journal of Heredity, vol. 100, no. 6, pages 659–674, 2009. 12, 61

[de Montgolfier 2003] F. de Montgolfier. *Décomposition modulaire des graphes. Théorie, extensions et algorithmes.* Thèse de doctorat, Université Montpellier II, 2003. 15

[Demuth 2009] J.P. Demuth et M.W. Hahn. *The life and death of gene families.* BioEssays, vol. 31, no. 1, pages 29–39, 2009. 83

[Dewey 2011] C.N. Dewey. *Positional orthology : putting genomic evolutionary relationships into context.* Briefings in Bioinformatics, 2011. Sous presse, tt doi : 10.1093/bib/bbr040. 21

[Didier 2007] G. Didier, T. Schmidt, J. Stoye et D. Tsur. *Character sets of strings.* Journal of Discrete Algorithms, vol. 5, no. 2, pages 330–340, 2007. 20

[Diekmann 2007] Y. Diekmann, M.-F. Sagot et E. Tannier. *Evolution under Reversals : Parsimony and Conservation of Common Intervals.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 4, no. 2, pages 301–309, 2007. Version

préliminaire par E. Tannier et M.-F. Sagot publiée dans les actes de COCOON 2005. 33

[Dittmar 2010] K. Dittmar et D. Liberles, éditeurs. Evolution after gene duplication. Wiley-Blackwell, 2010. 85

[Dobzhansky 1938] T. Dobzhansky et A.T. Sturtevant. *Inversions in the Chromosomes of* Drosophila pseudoobscura. Genetics, vol. 23, pages 28–64, 1938. 8

[Dom 2008] M. Dom. *Recognition, Generation, and Application of Binary Matrices with the Consecutive-Ones Property.* Thèse de doctorat, Institut für Informatik, Friedrich-Schiller-Universität Jena, 2008. 44

[Doyon 2009] J.-P. Doyon, C. Chauve et Sylvie Hamel. *Space of Gene/Species Trees Reconciliations and Parsimonious Models.* Journal of Computational Biology, vol. 16, no. 10, pages 1399–1418, 2009. Version préliminaire publiée dans les actes de RECOMB-CG 2008. 87

[Doyon 2010] J.P. Doyon. *Algorithmes pour la réconciliation d'un arbre de gènes avec un arbre d'espèces.* Thèse de doctorat, DIRO, Université de Montréal, 2010. 83

[Doyon 2011a] J.-P. Doyon et C. Chauve. Software tools and algorithms for biological systems, volume 696 of *Advances in Experimental Medicine and Biology*, chapitre Branch-and-Bound Approach for Parsimonious Inference of a Species Tree From a Set of Gene Family Trees, pages 287–295. Springer, 2011. 88, 89

[Doyon 2011b] J.-P. Doyon, C. Chauve et S. Hamel. *An Efficient Method for Exploring the Space of Gene Tree/Species Tree Reconciliations in a Probabilistic Framework.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2011. Sous presse. 87

[Dubchak 2009] I. Dubchak, A. Poliakov, A. Kislyuk et M. Brudno. *Multiple whole-genome alignments without a reference organism.* Genome Research, vol. 19, no. 4, pages 682–689, 2009. 12

[Durand 2003] D. Durand et D. Sankoff. *Tests for gene clustering.* Journal of Computational Biology, vol. 10, no. 3–4, pages 453–482, 2003. 22, 50

[Eriksen 2007] N. Eriksen. *Reversal and transposition medians.* Theoretical Computer Science, vol. 374, no. 1-3, pages 111–126, 2007. 37

[Faraut 2007] T. Faraut, S. de Givry, P. Chabrier, T. Derrien, F. Galibert, C. Hitte et T. Schiex. *A comparative genome approach to marker ordering.* Bioinformatics, vol. 23, pages e50–e56, 2007. 55

[Faraut 2008] T. Faraut. *Addressing chromosome evolution in the whole-genome sequence era.* Chromosome Research, vol. 16, pages 5–16, 2008. 40

[Felsenstein 2004] J. Felsenstein. Inferring phylogenies. Sinauer Associates, 2004. 19, 40, 89

[Ferguson-Smith 2007] M.A. Ferguson-Smith et V. Trifonov. *Mammalian karyotype evolution.* Nature Reviews on Genetics, vol. 8, pages 950–962, 2007. 40

[Ferretti 1996] V. Ferretti, J. H. Nadeau et D. Sankoff. *Original Synteny.* In D.S. Hirschberg et E.W. Myers, éditeurs, Combinatorial Pattern Matching, 7th Annual Symposium,

CPM 96, Laguna Beach, California, USA, June 10-12, 1996, Proceedings, volume 1075 of *Lecture Notes in Computer Science*, pages 159–167. Springer, 1996. 7

[Fertin 2009] G. Fertin, A. Labarre, I. Rusu, E. Tannier et S. Vialette. Combinatorics of genome rearrangements. The MIT Press, 2009. 10, 22, 25, 61

[Figeac 2004] M. Figeac et J.-S. Varré. *Sorting by Reversals with Common Intervals*. In I. Jonassen et J. Kim, éditeurs, Algorithms in Bioinformatics, 4th International Workshop, WABI 2004, Bergen, Norway, September 17-21, 2004, Proceedings, volume 3240 of *Lecture Notes in Computer Science*, pages 26–37. Springer, 2004. 30, 31

[Flajolet 2009] P. Flajolet et R. Sedgewick. Analytic combinatorics. Cambridge University Press, 2009. 36

[Fostier 2011] J. Fostier, S. Proost, B. Dhoedt, Y. Saeys, P. Demeester an Y. Van de Peer et K. Vandepoele. *A greedy, graph-based algorithm for the alignment of multiple homologous gene lists*. Bioinformatics, vol. 27, no. 6, pages 749–756, 2011. 12

[Froenicke 2006] L. Froenicke, M. Garcia Caldés, A. Graphodatsky, S. Mueller, L.A. Lyons, T.J. Robinson, M. Volleth, F. Yang et J. Wienberg. *Are molecular cytogenetics and bioinformatics suggesting diverging models of ancestral mammalian genomes ?* Genome Research, vol. 16, pages 306–310, 2006. 40, 41, 42, 45

[Fu 2007] Z. Fu, X. Chen, V. Vacic, P. Nan, Y. Zhong et T. Jiang. *MSOAR : A High-Throughput Ortholog Assignment System Based on Genome Rearrangement*. Journal of Computational Biology, vol. 14, no. 9, pages 1160–1175, 2007. 22

[Fu 2008] Z. Fu et T. Jiang. *Clustering of Main Orthologs for Multiple Genomes*. Journal of Bioinformatics and Computational Biology, vol. 6, no. 3, pages 573–584, 2008. 22

[Fujishige 2005] S. Fujishige. Submodular functions and optimization, volume 58 of *Annals of Discrete Mathematics*. Elsevier, seconde édition, 2005. 91

[Gao 2011] S. Gao, N. Nagarajan et W.-K. Sung. *Opera : Reconstructing Optimal Genomic Scaffolds with High-Throughput Paired-End Sequences*. In V. Bafna et S.C. Sahinalp, éditeurs, Research in Computational Molecular Biology - 15th Annual International Conference, RECOMB 2011, Vancouver, BC, Canada, March 28-31, 2011. Proceedings, volume 6577 of *Lecture Notes in Computer Science*, pages 437–451. Springer, 2011. 55

[Gaul 2006] É. Gaul et M. Blanchette. *Ordering Partially Assembled Genomes Using Gene Arrangements*. In G. Bourque et N. El-Mabrouk, éditeurs, Comparative Genomics, RECOMB 2006 International Workshop, RCG 2006, Montreal, Canada, September 24-26, 2006, Proceedings, volume 4205 of *Lecture Notes in Computer Science*, pages 113–128. Springer, 2006. 53

[Gavranovic 2010] H. Gavranovic et E. Tannier. *Guided genome halving : provably optimal solutions provide good insights into the preduplication ancestral genome of* Saccharomyces cerevisiae. Pacific Symposium on Biocomputing, pages 21–30, 2010. 41, 49

[Gavranovic 2011] H. Gavranovic, C. Chauve, J. Salse et E. Tannier. *Mapping ancestral genomes with massive gene loss : a matrix sandwich problem*. Bioinformatics, vol. 27, no. 13, 2011. Actes de ISMB/ECCB 2011, sous presse. 47

[Gebhart 2008] E. Gebhart. *Ring chromosomes in human neoplasias.* Cytogenetics Genome Research, vol. 121, no. 3–4, pages 149–173, 2008. 29

[Goodman 1979] M. Goodman, J. Czelusniak, G.W. Moore, A.E. Romero-Herrera et G. Matsuda. *Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences.* Systematic Zoology, vol. 28, pages 132–163, 1979. 85

[Gordon 2009] J.L. Gordon, K.P. Byrne et K.H. Wolfe. *Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern Saccharomyces cerevisiae genome.* PLoS Genetics, vol. 5, no. 5, page e1000485, 2009. 40, 49

[Górecki 2006] P. Górecki et J. Tiuryn. *DLS-trees : a model of evolutionary scenarios.* Theoretetical Computer Science, vol. 359, no. 1, pages 378–399, 2006. 86

[Graur 2000] D. Graur et W.-H. Li. Fundamentals of molecular evolution. Sinauer Associates, seconde édition, 2000. 5, 85

[Habib 2000] M. Habib, R.M. McConnell, C. Paul et L. Viennot. *Lex-BFS and partition refinement, with applications to transitive orientation, interval graph recognition and consecutive ones testing.* Theor. Comput. Sci., vol. 234, no. 1-2, pages 59–84, 2000. 43

[Habib 2004] M. Habib, C. Paul et M. Raffinot. *Maximal Common Connected Sets of Interval Graphs.* In S.C. Sahinalp, S. Muthukrishnan et U. Dogrusöz, éditeurs, Combinatorial Pattern Matching, 15th Annual Symposium, CPM 2004, Istanbul,Turkey, July 5-7, 2004, Proceedings, volume 3109 of *Lecture Notes in Computer Science*, pages 359–372. Springer, 2004. 20

[Hachiya 2009] T. Hachiya, Y. Osana, K. Popendorf et Y. Sakakibara. *Accurate identification of orthologous segments among multiple genomes.* Bioinformatics, vol. 25, no. 7, pages 853–860, 2009. 12

[Hahn 2007] M.W. Hahn. *Bias in phylogenetic tree reconciliation methods : implications for vertebrate genome evolution.* Genome Biology, vol. 8, page R141, 2007. 86, 92

[Hampson 2005] S.E. Hampson, B.S. Gaut et P. Baldi. *Statistical detection of chromosomal homology using shared-gene density alone.* Bioinformatics, vol. 21, no. 8, pages 1339–1348, 2005. 12

[Hannenhalli 1995a] S. Hannenhalli et P.A. Pevzner. *Transforming cabbage into turnip : polynomial algorithm for sorting signed permutations by reversals.* In Proceedings of the Twenty-Seventh Annual ACM Symposium on Theory of Computing, 29 May-1 June 1995, Las Vegas, Nevada, USA, pages 178–189. ACM, 1995. 25

[Hannenhalli 1995b] S. Hannenhalli et P.A. Pevzner. *Transforming Men into Mice (Polynomial Algorithm for Genomic Distance Problem).* In 36th Annual Symposium on Foundations of Computer Science, Milwaukee, Wisconsin, 23-25 October 1995., pages 581–592. IEEE Computer Society Press, 1995. 25, 28

[He 2005] X. He et M.H. Goldwasser. *Identifying Conserved Gene Clusters in the Presence of Homology Families.* Journal of Computational Biology, vol. 12, no. 6, pages 638–656, 2005. 20

[Heber 2001] S. Heber et J. Stoye. *Finding All Common Intervals of k Permutations.* In A. Amir et G. M. Landau, éditeurs, Combinatorial Pattern Matching, 12th Annual Symposium, CPM 2001 Jerusalem, Israel, July 1-4, 2001 Proceedings, volume 2089 of *Lecture Notes in Computer Science*, pages 207–218. Springer, 2001. 18, 59

[Hoberman 2005a] R. Hoberman et D. Durand. *The Incompatible Desiderata of Gene Cluster Properties.* In A. McLysaght et D. H. Huson, éditeurs, Comparative Genomics, RECOMB 2005 International Workshop, RCG 2005, Dublin, Ireland, September 18-20, 2005, Proceedings, volume 3678 of *Lecture Notes in Computer Science*, pages 73–87. Springer, 2005. 22

[Hoberman 2005b] R. Hoberman, D. Sankoff et D. Durand. *The Statistical Analysis of Spatially Clustered Genes under the Maximum Gap Criterion.* Journal of Computational Biology, vol. 12, no. 8, pages 1083–1102, 2005. 22

[Hufton 2009] A.L. Hufton et G. Panopoulou. *Polyploidy and genome restructuring : a variety of outcomes.* Current Opinions in Genetics & Develeopment, vol. 19, pages 600–606, 2009. 51

[Iwata 2009] S. Iwata et J.B. Orlin. *A simple combinatorial algorithm for submodular function minimization.* In C. Mathieu, éditeur, Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2009, New York, NY, USA, January 4-6, 2009, pages 1230–1237. SIAM, 2009. 91

[Jahn 2010] K. Jahn. *Efficient Computation of Approximate Gene Clusters Based on Reference Occurrences.* In E. Tannier, éditeur, Comparative Genomics - International Workshop, RECOMB-CG 2010, Ottawa, Canada, October 9-11, 2010. Proceedings, volume 6398 of *Lecture Notes in Computer Science*, pages 264–277. Springer, 2010. 20

[Jaillon 2004] O. Jaillon et *et al. Genome duplication in the teleost fish* Tetraodon nigroviridis *reveals the early vertebrate proto-karyoty pe.* Nature, vol. 431, pages 946–957, 2004. 51

[Jean 2007] G. Jean et M. Nikolski. *Genome rearrangements : a correct algorithm for optimal capping.* Information Processing Letters, vol. 104, no. 1, pages 14–20, 2007. 28

[Jean 2009] G. Jean, D.J. Sherman et M. Nikolski. *Mining the semantics of genome superblocks to infer ancestral architectures.* Journal of Computational Biology, vol. 16, pages 1267–1284, 2009. 40, 49

[Kellis 2004] M. Kellis, B.W. Birren et E. S. Lander. *Proof and evolutionary analysis of ancient genome duplication in the yeast* Saccharomyces cerevisiae. Nature, vol. 428, pages 617–624, 2004. 49

[Kemkemer 2006] C. Kemkemer, M. Kohn, H. Kehrer-Sawatzki, P. Minich, J. Högel, L. Froenicke et H. Hameister. *Reconstruction of the ancestral ferungulate karyotype by electronic chromosome painting (E-painting).* Chromosome Research, vol. 14, pages 899–907, 2006. 40

[Kemkemer 2009] C. Kemkemer, M. Kohn, D.N. Cooper, L. Froenicke, H. Hameister et H. Kehrer-Sawatzki. *Gene synteny comparison between different vertebrates provide new insights into breakage and fusion events during mammalian karyotype evolution.* BMC Evolutionary Biology, vol. 9, page 84, 2009. 40

[Kohn 2006] M. Kohn, J. Högel, W. Vogel, P. Minich, H. Kehrer-Sawatzki, J.A. Graves et H. Hameister. *Reconstruction of a 450My-old ancestral vertebrate protokaryotype.* Trends in Genetics, vol. 22, pages 203–210, 2006. 40, 53

[Kovác 2010] J. Kovác, M.D.V. Braga et J. Stoye. *The Problem of Chromosome Reincorporation in DCJ Sorting and Halving.* In E. Tannier, éditeur, Comparative Genomics - International Workshop, RECOMB-CG 2010, Ottawa, Canada, October 9-11, 2010. Proceedings, volume 6398 of *Lecture Notes in Computer Science*, pages 13–24. Springer, 2010. 30

[Kumar 2001] S Kumar, S R Gadagkar, A Filipski et X Gu. *Determination of the number of conserved chromosomal segments between species.* Genetics, vol. 157, pages 1387–1395, 2001. 50, 51

[Landau 2005] G.M. Landau, L. Parida et O. Weimann. *Gene proximity analysis across whole genomes via PQ trees.* Journal of Computational Biology, vol. 12, no. 10, pages 1289–1306, 2005. 19, 21

[Larget 2005] B. Larget, J.B. Kadane et D.L. Simon. *A Bayesian approach to the estimation of ancestral genome arrangements.* Molecular Phylogenetics and Evolution, vol. 36, pages 214–223, 2005. 36

[Lemaitre 2008a] C. Lemaitre. *Réarrangements chromosomiques dans les génomes de mammifères : caractérisation des points de cassure.* Thèse de doctorat, Université Claude Bernard Lyon I, 2008. 7

[Lemaitre 2008b] C. Lemaitre et M.-F. Sagot. *A Small Trip in the Untranquil World of Genomes : A survey on the detection and analysis of genome rearrangement breakpoints.* Theoretical Computer Science, vol. 395, no. 2-3, pages 171–192, 2008. 7

[Lin 2010] Y. Lin, V. Rajan, K.M. Swenson et B.M. Moret. *Estimating true evolutionary distances under rearrangements, duplications, and losses.* BMC Bioinformatics, vol. 11, no. Suppl 1, page S54, 2010. 36

[Ling 2009] X. Ling, X. He et D. Xin. *Detecting gene clusters under evolutionary constraint in a large number of genomes.* Bioinformatics, vol. 25, no. 5, pages 571–577, 2009. 20

[Luc 2003] N. Luc, J.-L. Risler, A. Bergeron et M. Raffinot. *Gene teams : a new formalization of gene clusters for comparative genomics.* Computational Biology and Chemistry, vol. 27, no. 1, pages 59–67, 2003. 20

[Ma 2000] B. Ma, M. Li et L. Zhang. *From gene trees to species trees.* SIAM Journal on Computing, vol. 30, pages 729–752, 2000. 88

[Ma 2006] J. Ma, L. Zhang, B.B. Suh, B.J. Rany, R.C. Burhans, W.J. Kent, M. Blanchette, D. Haussler et W. Miller. *Reconstructing contiguous regions of an ancestral genome.* Genome Research, vol. 16, pages 1557–1565, 2006. 12, 40, 41, 42, 43, 45, 54

[Ma 2008] J. Ma, A. Ratan, B.J. Raney, B.B. Suh, L. Zhang, W. Miller et D. Haussler. *DUPCAR : Reconstructing contiguous ancestral regions with duplications.* Journal of Computational Biology, vol. 15, pages 1007–1027, 2008. 83

[Mahmoody-Ghaidary 2011] A. Mahmoody-Ghaidary, C. Chauve et L. Stacho. Tractability results for the double-cut-and-join multichromosomal median problem. Poster présenté à *IWOCA 2011*, 2011. 37

[Manuch 2010] J. Manuch et M. Patterson. *The Complexity of the Gapped Consecutive-Ones Property Problem for Matrices of Bounded Maximum Degree*. In E. Tannier, éditeur, Comparative Genomics - International Workshop, RECOMB-CG 2010, Ottawa, Canada, October 9-11, 2010. Proceedings, volume 6398 of *Lecture Notes in Computer Science*, pages 278–289. Springer, 2010. 48

[Marota 2002] I. Marota et F. Rollo. *Molecular paleontology.* Cellular and Molecular Life Sciences, vol. 59, no. 1, pages 97–111, 2002. 39

[McConnell 2004] R.M. McConnell. *A certifying algorithm for the consecutive-ones property.* In J.I. Munro, éditeur, Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2004, New Orleans, Louisiana, USA, January 11-14, 2004, pages 768–777. SIAM, 2004. 43, 44

[Miklós 2010] I. Miklós et E. Tannier. *Bayesian sampling of genomic rearrangement scenarios via double cut and join.* Bioinformatics, vol. 26, no. 24, pages 3012–3019, 2010. 36

[Mongin 2009] E. Mongin, K. Dewar et M. Blanchette. *Long-range regulation is a major driving force in maintaining genome integrity.* BMC Evolutionary Biology, vol. 9, page 203, 2009. 61

[Muffato 2008] M. Muffato et H. Roest-Crollius. *Paleogenomics, or the recovery of lost genomes from the mist of times.* BioEssays, vol. 30, pages 122–134, 2008. 40, 59

[Muffato 2010] M. Muffato. *Reconstruction de génomes ancestraux chez les vertébrés.* Thèse de doctorat, Université d'Évry Val d'Essonne, 2010. 40

[Munoz 2010] A. Munoz, C. Zheng, Q. Zhu, V.A. Albert, S. Rounsley et D. Sankoff. *Scaffold filling, contig fusion and comparative gene order inference.* BMC Bioinformatics, vol. 11, page 304, 2010. 53

[Murat 2010] F. Murat, J.H. Xu, E. Tannier, M. Abrouk, N. Guilhot, C. Pont, J. Messing et J. Salsé. *Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution.* Genome Research, vol. 20, no. 11, pages 1545–1557, 2010. 12, 40, 47, 53

[Murphy 2001] W.J. Murphy, R. Stanyon et S.J. O'Brien. *Evolution of mammalian genome organization inferred from comparative gene mapping.* Genome Biology, vol. 2, pages reviews0005–reviews0005.8, 2001. 40

[Murphy 2005] W.J. Murphy, D.M. Larkin, A. Everts van der Wind, G. Bourque, G. Tesler, L. Auvil, J.E. Beever, B.P. Chowdhary, F. Galibert et L. Gatzke et al. *Dynamics of Mammalian Chromosome Evolution Inferred from Multispecies Comparative Maps.* Science, vol. 309, pages 613–617, 2005. 36, 37, 40

[Nakatani 2007] Y. Nakatani, H. Takeda et S. Morishita. *Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates.* Genome Research, vol. 17, pages 1254–1265, 2007. 40, 53

[Ohno 1970] S. Ohno. Evolution by gene duplication. Springer-Verlag, 1970. 8

[Ouangraoua 2009a] A. Ouangraoua et A. Bergeron. *Parking Functions, Labeled Trees and DCJ Sorting Scenarios.* In F. Ciccarelli et I. Miklós, éditeurs, Comparative Genomics,

International Workshop, RECOMB-CG 2009, Budapest, Hungary, September 27-29, 2009. Proceedings, volume 5817 of *Lecture Notes in Computer Science*, pages 24–35. Springer, 2009. 30

[Ouangraoua 2009b] A. Ouangraoua, F. Boyer, A. McPherson, E. Tannier et C. Chauve. *Prediction of Contiguous Regions in the Amniote Ancestral Genome.* In I. I. Mandoiu, G. Narasimhan et Y. Zhang, éditeurs, Bioinformatics Research and Applications, 5th International Symposium, ISBRA 2009, Fort Lauderdale, FL, USA, May 13-16, 2009, Proceedings, volume 5542 of *Lecture Notes in Computer Science*, pages 173–185. Springer, 2009. 55

[Ouangraoua 2010a] A. Ouangraoua, A. Bergeron et K.M. Swenson. *Ultra-Perfect Sorting Scenarios.* In E. Tannier, éditeur, Comparative Genomics - International Workshop, RECOMB-CG 2010, Ottawa, Canada, October 9-11, 2010. Proceedings, volume 6398 of *Lecture Notes in Computer Science*, pages 50–61. Springer, 2010. 36

[Ouangraoua 2010b] A. Ouangraoua, K. M. Swenson et C. Chauve. *An Approximation Algorithm for Computing a Parsimonious First Speciation in the Gene Duplication Model.* In E. Tannier, éditeur, Comparative Genomics - International Workshop, RECOMB-CG 2010, Ottawa, Canada, October 9-11, 2010. Proceedings, volume 6398 of *Lecture Notes in Computer Science*, pages 290–301. Springer, 2010. Version étendue à paraître dans Journal of Computational Biology. 91

[Ouangraoua 2011] A. Ouangraoua, E. Tannier et C. Chauve. Reconstructing the architecture of the ancestral amniote genome. Soumis à *Bioinformatics*, en cours de révision, 2011. 49, 51, 52, 53, 54

[Overbeek 1999] R. Overbeek, M. Fonstein, M. D'Souza, G.D. Pusch et N. Maltsev. *The use of gene clusters to infer functional coupling.* Proceedings of the National Academy of Sciences of the United States of America, vol. 96, pages 2896–2901, 1999. 21

[Page 1994] R.D. Page. *Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas.* Systematic Biology, vol. 43, pages 58–77, 1994. 85

[Pasek 2005] S. Pasek, A. Bergeron, J.L. Risler, A. Louis, E. Ollivier et M. Raffinot. *Identification of genomic features using microsyntenies of domains : domain teams.* Genome Research, vol. 15, no. 6, pages 867–874, 2005. 20, 21

[Passarge 1999] E. Passarge, B. Horsthemke et R.A. Farber. *Incorrect use of the term synteny.* Nature Genetics, vol. 23, page 387, 1999. 9

[Paten 2008a] B. Paten, J. Herrero, K. Beal, S. Fitzgerald et E. Birney. *Enredo and Pecan : genome-wide mammalian consistency-based multiple alignment with paralogs.* Genome Research, vol. 18, no. 11, pages 1814–1828, 2008. 12, 48

[Paten 2008b] B. Paten, J. Herrero, S. Fitzgerald, K. Beal, P. Flicek, I. Holmes et E. Birney. *Genome-wide nucleotide-level mammalian ancestor reconstruction.* Genome Research, vol. 18, no. 11, pages 1829–1843, 2008. 61

[Paul 2006] C. Paul. *Aspects algorithmiques de la décomposition modulaire.* Habilitation à Diriger des Recherches, Université Montpellier II, 2006. 15, 20, 33

[Pevzner 2000] P.A. Pevzner, éditeur. Computational molecular biology : An algorithmic approach. The MIT Press, 2000. 27

[Pinney 2007] J.W. Pinney, G.R. Amoutzias, M. Rattray et D.L. Robertson. *Reconstruction of ancestral protein interaction networks for the bZIP transcription factors.* Proceedings of the National Academy of Sciences of the United States of America, vol. 104, no. 51, pages 20449–20453, 2007. 10

[Ponty 2006] Y. Ponty, M. Termier et A. Denise. *GenRGenS : Software for Generating Random Genomic Sequences and Structures.* Bioinformatics, vol. 22, no. 12, pages 1534–1535, 2006. 36

[Putnam 2007] N.H. Putnam et *et al. Sea Anemone Genome Reveals Ancestral Eumetazoan Gene Repertoire and Genomic Organization.* Science, vol. 317, pages 86–94, 2007. 40

[Putnam 2008] N.H. Putnam et *et al. The amphioxus genome and the evolution of the chordate karyotype.* Nature, vol. 453, pages 1064–1071, 2008. 40

[Raghupathy 2008] N. Raghupathy, R. Hoberman et D. Durand. *Two Plus Two Does not Equal Three : Statistical Tests for Multiple genome Comparison.* Journal of Bioinformatics and Computational Biology, vol. 6, no. 1, pages 1–22, 2008. 22

[Raghupathy 2009] N. Raghupathy et D. Durand. *Gene Cluster Statistics with Gene Families.* Molecular Biology and Evolution, vol. 26, no. 5, pages 957–968, 2009. 22

[Rascol 2007] V. Lopez Rascol, P. Pontarotti et A. Levasseur. *Ancestral animal genomes reconstruction.* Current Opinions in Immunology, vol. 19, pages 542–546, 2007. 40

[Richard 2003] F. Richard, M. Lombard et B. Dutrillaux. *Reconstruction of the ancestral karyotype of eutherian mammals.* Chromosome Research, vol. 11, pages 605–618, 2003. 40

[Robinson 2008] T.J Robinson et A. Ruiz-Herrera. *Defining the ancestral eutherian karyotype : a cladistic interpretation of chromosome painting and genome sequence assembly data.* Chromosome Research, vol. 16, no. 8, pages 1133–1141, 2008. 40

[Rocchi 2006] M. Rocchi, N. Archidiacono et R. Stanyon. *Ancestral genome reconstruction : An integrated, multi-disciplinary approach is needed.* Genome Research, vol. 16, pages 1441–1444, 2006. 40, 42, 45

[Romiguier 2010] J. Romiguier, V. Ranwez, E.J. Douzery et N. Galtier. *Contrasting GC-content dynamics across 33 mammalian genomes : relationship with life-history traits and chromosome sizes.* Genome Research, vol. 20, no. 8, pages 1001–1009, 2010. 61

[Salsé 2009a] J. Salsé, M. Abrouk, S. Bolot, N. Guilhot, E. Courcelle, T. Faraut, R. Waugh, T.J. Close, J. Messing et C. Feuillet. *Reconstruction of monocotelydoneous protochromosomes reveals faster evolution in plants than in animals.* Proceedings of the National Academy of Sciences of the United States of America, vol. 106, pages 14908–14913, 2009. 40

[Salsé 2009b] J. Salsé, M. Abrouk, F. Murat, U.M. Quraishi et C. Feuillet. *Improved criteria and comparative genomics tool provide new insights into grass paleogenomics.* Briefings in Bioinformatics, vol. 10, pages 619–630, 2009. 40

[Sanderson 2007] M. Sanderson et M. McMahon. *Inferring angiosperm phylogeny from EST data with widespread gene duplication*. BMC Evolutionary Biology, vol. 7, no. Suppl 1, 2007. 88

[Sankoff 1992a] D. Sankoff. *Edit Distances for Genome Comparisons Based on Non-Local Operations*. In A. Apostolico, M. Crochemore, Z. Galil et U. Manber, éditeurs, Combinatorial Pattern Matching, Third Annual Symposium, CPM 92, Tucson, Arizona, USA, April 29 - May 1, 1992, Proceedings, volume 644 of *Lecture Notes in Computer Science*, pages 121–135. Springer, 1992. 25, 26

[Sankoff 1992b] D. Sankoff, G. Leduc, N. Antoine, B. Paquin, B.F. Lang et R. Cedergren. *Gene order comparisons for phylogenetic inference : Evolution of the mitochondrial genome*. Proceedings of the National Academy of Sciences of the United States of America, vol. 89, pages 6575–6579, 1992. 25, 26, 28

[Sankoff 1997] D. Sankoff, V. Ferretti et J.H. Nadeau. *Conserved segment identification*. Journal of Computational Biology, vol. 395, no. 2–3, pages 171–192, 1997. 15

[Sankoff 2005] D. Sankoff. Mathematics of evolution and phylogeny (édité par O. Gascuel), chapitre Conserved segment statistics and rearrangement inferences in comparative genomics, pages 236–261. Oxford University Press, 2005. 15

[Sankoff 2009] D. Sankoff. *The where and wherefore of evolutionary breakpoints*. Journal of Biology, vol. 8, no. 7, page 66, 2009. 8, 61

[Sankoff 2010] D. Sankoff, C. Zheng et Q. Zhu. *The collapse of gene complement following whole genome duplication*. BMC Genomics, vol. 11, page 313, 2010. 51, 61

[Schmidt 2004] T. Schmidt et J. Stoye. *Quadratic Time Algorithms for Finding Common Intervals in Two and More Sequences*. In S.C. Sahinalp, S. Muthukrishnan et U. Dogrusöz, éditeurs, Combinatorial Pattern Matching, 15th Annual Symposium, CPM 2004, Istanbul,Turkey, July 5-7, 2004, Proceedings, volume 3109 of *Lecture Notes in Computer Science*, pages 347–358. Springer, 2004. 20

[Schmidt 2007] T. Schmidt et J. Stoye. *Gecko and GhostFam - Rigorous and Efficient Gene Cluster Detection in Prokaryotic Genomes*. In N.H. Bergman, éditeur, Comparative Genomics, Volume 2., volume 396 of *Methods in Molecular Biology*, pages 165–182. Humana Press, 2007. 21

[Scornavacca 2011] C. Scornavacca, V. Berry et V. Ranwez. *Building species trees from larger parts of phylogenomic databases*. Information and Computation, vol. 209, no. 3, pages 590–605, 2011. 91

[Sémon 2007a] M. Sémon et K. H. Wolfe. *Reciprocal gene loss between Tetraodon and zebrafish after whole genome duplication in their ancestor*. Trends in Genetics, vol. 23, no. 3, pages 108 – 112, 2007. 51

[Sémon 2007b] M. Sémon et K.H. Wolfe. *Consequences of genome duplication*. Current Opinions in Genetics & Development, vol. 17, pages 505–512, 2007. 51

[Semple 2000] C. Semple et M. Steel. *A supertree method for rooted trees*. Discrete Applied Mathematics, vol. 105, no. 1–3, pages 147–158, 2000. 90

[Sennblad 2009] B. Sennblad et J. Lagergren. *Probabilistic Orthology Analysis*. Systematic Biology, vol. 58, no. 4, pages 411–424, 2009. 87

[Servin 2010] B. Servin, S. de Givry et T. Faraut. *Statistical confidence measures for genome maps : application to the validation of genome assemblies*. Bioinformatics, vol. 26, no. 24, pages 3035–3042, 2010. 55

[Simillion 2008] C. Simillion, K. Janssens, L. Sterck et Y. Van de Peer. *i-ADHoRe 2.0 : an improved tool to detect degenerated genomic homology using genomic profiles*. Bioinformatics, vol. 24, no. 1, pages 127–128, 2008. 12

[St-Onge 2005] K. St-Onge, A. Bergeron et C. Chauve. *Fast identification of gene clusters in prokaryotic genomes*. In M.-F. Sagot et K. S. Gimaraes, éditeurs, CompBioNets 2005 : Algorithms and Computational Methods for Biochemical and Evolutionary Networks, volume 5 of *Text in Algorithms*. College Publications, 2005. 22

[Stege 1999] U. Stege. *Gene Trees and Species Trees : The Gene-Duplication Problem in Fixed-Parameter Tractable*. In F.K.H.A. Dehne, A. Gupta, J-R. Sack et R. Tamassia, éditeurs, Algorithms and Data Structures, 6th International Workshop, WADS '99, Vancouver, British Columbia, Canada, August 11-14, 1999, Proceedings, volume 1663 of *Lecture Notes in Computer Science*, pages 288–293. Springer, 1999. 88

[Stoye 2009] J. Stoye et R. Wittler. *A Unified Approach for Reconstructing Ancient Gene Clusters*. IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 6, no. 3, pages 387–400, 2009. 41, 47

[Swenson 2008] K. M. Swenson, Y. Lin, V. Rajan et B.M.E. Moret. *Hurdles Hardly Have to Be Heeded*. In C.E. Nelson et S. Vialette, éditeurs, Comparative Genomics, International Workshop, RECOMB-CG 2008, Paris, France, October 13-15, 2008. Proceedings, volume 5267 of *Lecture Notes in Computer Science*, pages 241–251. Springer, 2008. 27

[Swidan 2006a] F. Swidan, E.P. Rocha, M. Shmoish et R.Y. Pinter. *An integrative method for accurate comparative genome mapping*. PLoS Computational Biology, vol. 2, no. 8, page e75, 2006. 21

[Swidan 2006b] F. Swidan, M. Ziv-Ukelson et R.Y. Pinter. *On the repeat-annotated phylogenetic tree reconstruction problem*. Journal of Computational Biology, vol. 13, no. 8, pages 1397–1418, 2006. 35

[Tannier 2007] E. Tannier, A. Bergeron et M.-F. Sagot. *Advances on sorting by reversals*. Discrete Applied Mathematics, vol. 155, no. 6–7, pages 881–888, 2007. Version préliminaire par E. Tannier et M.-F. Sagot publiée dans les actes de CPM 2004. 28

[Tannier 2009] E. Tannier, C. Zheng et D. Sankoff. *Multichromosomal median and halving problems under different genomic distances*. BMC Bioinformatics, vol. 10, page 120, 2009. 37

[Tucker 1972] A. Tucker. *A structure theorem for the consecutive 1's property*. Journal of Combinatorial Theory, Series B, vol. 12, no. 2, pages 153–162, 1972. 47

[Uno 2000] T. Uno et M. Yagiura. *Fast Algorithms to Enumerate All Common Intervals of Two Permutations*. Algorithmica, vol. 26, no. 2, pages 290–309, 2000. 18

[Vilella 2009] A.J. Vilella, J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin et E. Birney. *Ensembl Compara GeneTrees : Complete, duplication-aware phylogenetic trees in vertebrates.* Genome Research, vol. 19, pages 327–335, 2009. 10, 51

[Wapinski 2007] I. Wapinski, A. Pfeffer, N. Friedman et A. Regev. *Natural history and evolutionary principles of gene duplication in fungi.* Nature, vol. 449, pages 54–61, 2007. 87

[Wehe 2008] A. Wehe, M.S. Bansal, J.G. Burleigh et O. Eulenstein. *DupTree : a program for large-scale phylogenetic analyses using gene tree parsimony.* Bioinformatics, vol. 24, pages 1540–1541, 2008. 88

[Wienberg 2004] J. Wienberg. *The evolution of eutherian chromosomes.* Current Opinions in Genetics & Development, vol. 14, pages 657–666, 2004. 40

[Wittler 2010] R. Wittler et J. Stoye. *Consistency of Sequence-Based Gene Clusters.* In E. Tannier, éditeur, Comparative Genomics - International Workshop, RECOMB-CG 2010, Ottawa, Canada, October 9-11, 2010. Proceedings, volume 6398 of *Lecture Notes in Computer Science*, pages 252–263. Springer, 2010. 48

[Wittler 2011] R. Wittler et C. Chauve. Conflicting deletion call in matched tumor/normal genomes. En révision pour BMC Bioinformatics, 2011. 62

[Wolfe 1997] K.H. Wolfe et D.C. Shield. *Molecular evidence for an ancient duplication of the entire yeast genome.* Nature, vol. 387, pages 708–713, 1997. 49

[Woods 2005] I.G. Woods, C. Wilson, B. Friedlander, P. Chang, D.K. Reyes, R. Nix, P.D. Kelly, F. Chu, J.H. Postlethwait et W.S. Talbot. *The zebrafish gene map defines ancestral vertebrates chromosomes.* Genome Research, vol. 15, pages 1307–1314, 2005. 40

[Xu 2008a] A.W. Xu et D. Sankoff. *Decompositions of Multiple Breakpoint Graphs and Rapid Exact Solutions to the Median Problem.* In K. A. Crandall et J. Lagergren, éditeurs, Algorithms in Bioinformatics, 8th International Workshop, WABI 2008, Karlsruhe, Germany, September 15-19, 2008. Proceedings, volume 5251 of *Lecture Notes in Computer Science*, pages 25–37. Springer, 2008. 37

[Xu 2008b] W. Xu. *The distribution of distances between randomly constructed genomes : generating function, expectation, variance and limits.* Journal of Bioinformatics and Computational Biology, vol. 6, no. 1, pages 23–36, 2008. 36

[Xu 2008c] W. Xu, B. Alain et D. Sankoff. *Poisson adjacency distributions in genome comparison : multichromosomal, circular, signed and unsigned cases.* Bioinformatics, vol. 24, no. 16, pages i146–i52, 2008. 36

[Xu 2009a] A.W. Xu. *DCJ Median Problems on Linear Multichromosomal Genomes : Graph Representation and Fast Exact Solutions.* In F. Ciccarelli et I. Miklós, éditeurs, Comparative Genomics, International Workshop, RECOMB-CG 2009, Budapest, Hungary, September 27-29, 2009. Proceedings, volume 5817 of *Lecture Notes in Computer Science*, pages 70–83. Springer, 2009. 37

[Xu 2009b] A.W. Xu. *A Fast and Exact Algorithm for the Median of three Problem : a Graph Decomposition Approach.* Journal of computational biology, vol. 16, no. 10, pages 1–13, 2009. 37

[Yancopoulos 2005] S. Yancopoulos, O. Attie et R. Friedberg. *Efficient sorting of genomic permutations by translocation, inversion and block interchange.* Bioinformatics, vol. 21, no. 16, pages 3340–3346, 2005. 29, 30

[Yang 2003] F. Yang, E.Z. Alkalaeva, P.L. Perelman, A.T. Pardini, W.R. Harrison, P.C.M. O'Brien, B. Fu, A.S. Graphodatsky, M.A. Ferguson-Smith et T.J. Robinson. *Reciprocal chromsome painting among human, aardvark, and elephant (superorder Afrotheria) reveals the likely eutherian ancestral karyotype.* Proceedings of the National Academy of Sciences of the United States of America, vol. 100, pages 1062–1066, 2003. 40

[Zhang 2009] M. Zhang, W. Arndt et J. Tang. *An exact solver for the DCJ median problem.* Pacific Symposium on Biocomputing, pages 138–149, 2009. 37

[Zhao 2009] H. Zhao et G. Bourque. *Recovering genome rearrangements in the mammalian phylogeny.* Genome Research, vol. 19, pages 934–942, 2009. 12, 16, 40, 56

[Zheng 2005a] C. Zheng, A. Lenert et D. Sankoff. *Reversal distance for partially ordered genomes.* Bioinformatics, vol. 21, no. Suppl. 1, pages i502–i508, 2005. 7

[Zheng 2005b] Y. Zheng, B.P. Anton, R.J. Roberts et S. Kasif. *Phylogenetic detection of conserved gene clusters in microbial genomes.* BMC Bioinformatics, vol. 6, page 243, 2005. 22

[Zheng 2006] C. Zheng et D. Sankoff. *Genome rearrangements with partially ordered chromosomes.* J. Comb. Optim., vol. 11, no. 2, pages 133–144, 2006. 7

[Zheng 2008] C. Zheng, Q. Zhu, Z. Adam et D. Sankoff. *Guided genome halving : hardness, heuristics and the history of the Hemiascomycetes.* Bioinformatics, vol. 24, no. 13, pages i96–i104, 2008. 40, 41

[Zheng 2011] C. Zheng et D. Sankoff. *On the PATHGROUPS approach to rapid small phylogeny.* BMC Bioinformatics, vol. 12, no. Suppl 1, page S4, 2011. 37, 41

Familles de gènes : Réconciliations et Phylogénomique

A

On s'intéresse ici à l'évolution de *familles de gènes*. Formellement, cette question ne concerne pas l'évolution de génomes mais seulement de gènes, ce qui explique sa "relégation" en appendice. Cependant une des motivations de ces travaux est la reconstruction du contenu en gènes d'un génome ancestral, étape préliminaire indispensable dans le cas de marqueurs définis par des gènes [Ma 2008, Bertrand 2010]. Cet aspect sera décrit dans un premier temps. Des résultats dans le domaine de la *phylogénomique* seront décrits dans un second temps. Les travaux présentés dans ce chapitre ont été effectués en collaboration avec Nadia El-Mabrouk, Sylvie Hamel et Jean-Philippe Doyon dans le cadre de sa thèse [Doyon 2010].

A.1 Évolution de familles de gènes

Une famille de gènes homologues est un ensemble de gènes qui descendent d'un gène ancestral commun par des événements de spéciations et de duplications, auxquels il faut ajouter les pertes de gènes, un mécanisme que nous avons passé sous silence dans le Chapitre 2 [Demuth 2009].

Une duplication, tout comme une spéciation, agit sur un gène ancestral et résulte en *deux copies* de ce gène : ces deux copies sont dans le même génome pour une duplication et dans deux génomes distincts pour une spéciation. La Figure A.1 illustre ce processus d'évolution, qui résulte en un *arbre binaire* dont les feuilles représentent des gènes actuels et les sommets internes des gènes ancestraux et des événements de spéciations ou de duplications. Par définition, les pertes de gènes ne sont pas observées. Cet arbre est appelé un *arbre de gènes*. Il contient possiblement des occurrences multiples d'une même étiquette x, indiquant un gène présent en plusieurs copies dans le génome x (on représente ici chaque génome par un entier distinct). Cette dernière caractéristique distingue un arbre de gènes d'un *arbre d'espèces*. Un arbre d'espèce est aussi un arbre aux *étiquettes uniques* : chaque étiquette représente un génome actuel et les sommets internes représentent les événements de spéciation qui ont

jalonné leur évolution. Un arbre d'espèces inféré peut ne pas être binaire si certains groupes de spéciations n'ont pas été résolus.

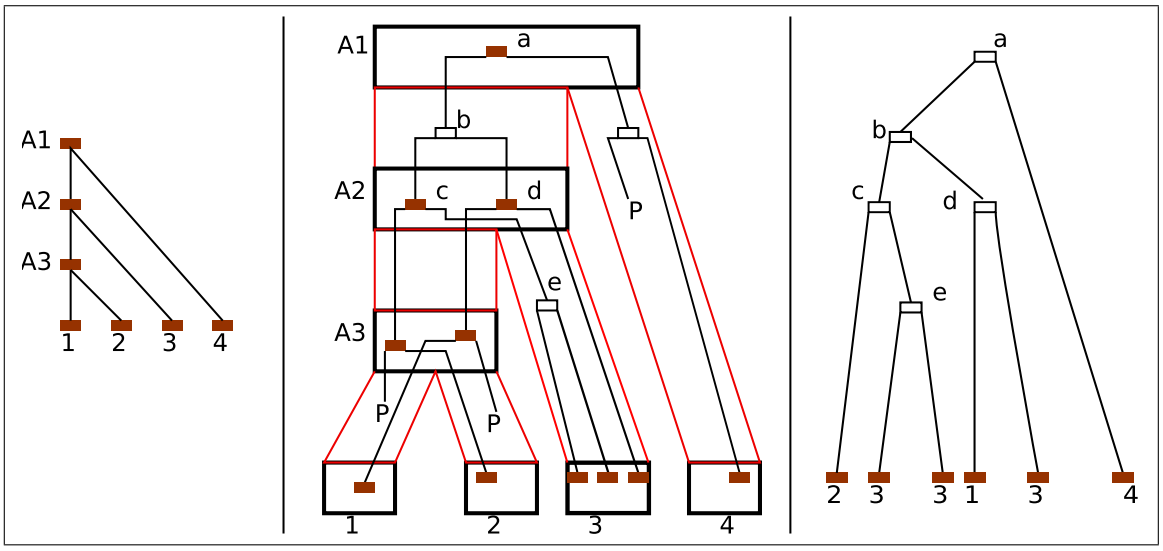

FIG. A.1 – (Gauche) Un arbre d'espèces S comportant trois événements de spéciation, trois espèces ancestrales ($A1$, $A2$, $A3$) et quatre espèces actuelles, représentées par les entiers de 1 à 4. (Centre) Évolution d'une famille de gènes le long de S. Les sommets creux, présents uniquement le long des branches de S, représentent des duplications, qui se produisent donc dans un génome entre deux spéciations. Les sommets pleins représentent des spéciations. Les pertes sont indiquées par la lettre P. Cette famille comporte une copie dans chacun des génomes 1, 2, 4 et trois copies dans le génome 3. (Droite) L'information disponible si l'évolution des gènes a été correctement inférée : un arbre de gènes T pour lequel on ne connaît pas les sommets internes indiquant des duplications et ceux indiquant des spéciations.

On peut distinguer essentiellement deux questions concernant l'évolution de familles de gènes. La première consiste, étant donnés un arbre de gènes T et un arbre d'espèce S, tous deux aux feuilles étiquetées sur un ensemble de génomes $\{1, 2, \ldots, m\}$, à décider quels sommets internes de T sont des événements de spéciations et lesquels sont des duplications. Il s'agit du problème de la *réconciliation entre un arbre de gènes et un arbre d'espèces*. La seconde prend en entrée un ensemble $\mathcal{T} = \{T_1, \ldots, T_k\}$ d'arbres de gènes, dont les gènes appartiennent aux génomes de m espèces, et cherche à inférer un arbre d'espèces (pour ces m espèces) qui minimise le nombre d'événements évolutifs dans les arbres $\mathcal{T}$. Il s'agit d'un problème d'inférence *phylogénomique*.

Notation et terminologie. Pour un sommet u d'un arbre A, on dénote par A_u le sous-arbre de A enraciné en u et par $L_A(u)$ l'ensemble des feuilles de A_u. Une *cerise* dans un arbre est un sous-arbre composé d'un sommet interne et de deux feuilles.

A.2 Réconciliations

Réconcilier un arbre de gènes T et un arbre d'espèces S est un concept classique [Goodman 1979, Page 1994], dont la principale application est la détermination de relations d'orthologie : deux gènes de T sont *orthologues* si leur plus proche ancêtre commun dans T est un sommet indiquant une spéciation. Dans le cas contraire, les deux gènes sont *paralogues*.

La notion de gènes homologues/orthologues est centrale en génomique évolutive et fonctionnelle[1]. En effet, un principe important de la génomique comparées est que deux gènes homologues sont plus susceptibles d'avoir la même fonction biologique si ils sont orthologues. La logique sous-jacente à ce principe est la suivante : après une duplication de gène, un génome contient deux copies du gène dupliqué, ce qui permet, entre autres, d'utiliser une de ces copies pour développer une nouvelle fonction biologique (la *néo-fonctionnalisation*) ou de spécialiser chacune des deux copies en une sous-fonction de la fonction initiale (la *sous-fonctionnalisation*) [Graur 2000, Dittmar 2010]. Si cette évolution de la fonction des gènes se produit avant une spéciation, ce qui est une hypothèse naturelle, alors les descendants des deux copies vont évoluer en développant des fonctions similaires si ils sont orthologues et différentes dans le cas contraire.

Du point de vue formel, le problème de la réconciliation est simple à énoncer : on veut séparer les sommets internes de T en deux classes, les sommets de spéciation et les sommets de duplication. En d'autres termes, on veut inférer une histoire évolutive à la fois compatible avec la topologie de T et avec la topologie de S (qui indique les spéciations). On peut déjà distinguer un cas simple.

Propriété 12 Soit u un sommet interne de T, et u_1 et u_2 ses deux enfants. Si il existe un descendant de u_1 de même étiquette qu'un descendant de u_2, alors aucune histoire évolutive par spéciations, duplications et pertes ne peut expliquer T sans avoir une duplication en u, qui est appelé une *duplication forcée*,

Par exemple, dans la Figure A.1, le génome 3 est le seul à avoir de multiples copies du gène, et tout sommet interne de l'arbre de gènes qui est le plus proche ancêtre commun de deux feuilles d'étiquette 3 est une duplication forcée. Les deux duplications de T (sommets internes b et e) sont donc forcées.

La réconciliation LCA. L'algorithme de réconciliation le plus répandu est basé sur une notion de correspondance entre les sommets de S et ceux de T en termes de plus proche ancêtre commun. On l'appelle la correspondance LCA : Pour un sommet u de T, $LCA(u)$ est le plus bas sommet x de S tel que $L_T(u) \subseteq L_S(x)$. On sépare alors les sommets de T en spéciations et duplications comme suit : un sommet u de T, dont les enfants sont u_1 et u_2,

[1] L'étude des données génomiques du point de vue de la fonction biologique des gènes et génomes.

est une duplication si et seulement si $LCA(u) = LCA(u_1)$ ou $LCA(u) = LCA(u_2)$. On note $d(T, S)$ le nombre de duplications dans T définies ainsi, étant donné S.

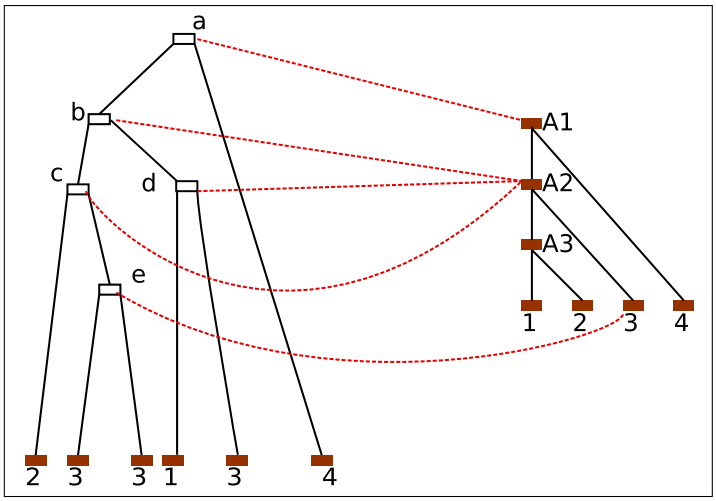

FIG. A.2 – Réconciliation LCA entre l'arbre de gènes T (gauche) et l'arbre d'espèces S de la Figure A.1. Seules les correspondances LCA des sommets internes sont montrées, les feuilles de T étant associées aux feuilles de S. Les deux duplications forcées de T satisfont la définition de duplication et sont les seuls sommets dans ce cas.

Cette réconciliation (dite LCA) entre T et S a des propriétés intéressantes. Il est bien connu qu'il s'agit d'une réconciliation parcimonieuse pour le nombre de duplications (i.e. qui minimise le nombre de duplications nécessaire pour expliquer T en fonction de S). Il peut cependant exister plusieurs réconciliations qui minimisent le nombre de duplications. De plus, les duplications ne sont pas les seuls événements que l'on peut prendre en compte. Ce n'est que récemment que cet aspect a été étudié.

Théorème 13
1. La réconciliation LCA est l'unique réconciliation qui minimise le nombre cumulé de duplications et de pertes [Górecki 2006].
2. La réconciliation LCA est l'unique réconciliation qui minimise le nombre de pertes [Chauve 2009a].

Explorer l'espace des réconciliations. La réconciliation LCA est donc une réconciliation parcimonieuse pour les trois critères combinatoires naturels associés à l'évolution d'une familles de gènes et, pour cette raison, est la plus couramment employée pour réconcilier un arbre de gènes et un arbre d'espèces. On peut cependant noter qu'il existe possiblement plusieurs réconciliations parcimonieuses pour les duplications seules. De plus une étude de Matthew Hahn [Hahn 2007] a illustré certains biais liés à l'utilisation de cette réconciliation. Cette constatation a motivé l'introduction de nouveaux modèles, plus généraux, de réconciliation [Arvestad 2004, Bonizzoni 2005, Górecki 2006], et à la question plus générale de l'exploration de l'espace de *toutes* les réconciliations possibles. On

note $\Phi(T, S)$ l'ensemble des réconciliations possibles entre T et S. Cette question est aussi motivée par l'introduction récentes de méthodes probabilistes (Monte-Carlo notamment) pour l'analyse de familles de gènes, basées sur un modèle classique de "Naissance et Mort" : une duplication correspond à la naissance d'un gène et une perte à la mort d'un gène [Arvestad 2004, Arvestad 2009, Akerborg 2009, Sennblad 2009].

Dans [Doyon 2009], nous avons simplifié la définition de réconciliation introduite dans [Arvestad 2004], qui englobe toutes les autres définitions. Nous avons décrit une réconciliation comme une correspondance entre les sommets de T et les sommets et arêtes de S qui respecte certaines conditions (simples et naturelles) de cohérence évolutive. Un sommet de T associé à une arête de S représente une duplication, et une spéciation sinon. Cette définition d'une réconciliation peut se traduire très naturellement en un algorithme de programmation dynamique qui permet de calculer $|\Phi(T, S)|$ et, combiné à des techniques classiques de génération aléatoire de structures combinatoires, de pouvoir engendrer des réconciliations sous la distribution uniforme [Doyon 2009].

Nous avons aussi défini deux opérateurs combinatoires simples qui permettent de transformer une réconciliation en une réconciliation voisine. Ces opérateurs permettent ainsi d'explorer tout ou partie de l'ensemble $\Phi(T, S)$. Ils permettent en effet de plonger $\Phi(T, S)$ dans une structure de graphe dont les sommets sont les réconciliations et les arêtes sont définies par les voisinage en termes de ces opérateurs. Nous avons ensuite défini un arbre couvrant de ce graphe dont le parcours permet d'explorer $\Phi(T, S)$ en temps constant amorti.

Théorème 14 [Doyon 2009]. L'ensemble des réconciliations de $\Phi(T, S)$ peut être exploré (avec calcul du nombre de duplications et pertes pour chacune) en temps $O(|\Phi(T, S)|)$ et en espace $O(nm)$, où n est la taille de T et m la taille de S.

Nous avons aussi adapté notre algorithme pour explorer, toujours en temps amorti constant, le sous-espace de $\Phi(T, S)$ contenant toutes les réconciliations ayant un nombre borné de duplications (resp. pertes, duplications+pertes). Ces algorithmes efficaces ont été utilisés sur des données simulées et les résultats de ces expériences suggèrent clairement qu'avec des taux de duplication et pertes raisonnables, la réconciliation représentant la vraie histoire d'une famille de gènes est très souvent la réconciliation LCA ou une des ses voisines proches.

Dans un second temps, nous avons étendu cette approche au cas de critères probabilistes introduits dans [Arvestad 2004, Arvestad 2009]. Dans [Doyon 2011b], nous avons montré comment mettre à jour efficacement la vraisemblance d'une réconciliation après avoir appliqué un de nos opérateurs. Ce résultats nous a permis d'explorer des espaces de réconciliations de grande taille et de montrer, sur des données simulées et réelles tirées de [Wapinski 2007], que la majeure partie de la masse des probabilités d'un ensemble de réconciliations est capturée par la réconciliation LCA et ses proches voisines.

A.3 Phylogénomique

Si on accepte que la duplication de gènes est un événement relativement rare, il est alors naturel de l'utiliser dans un cadre d'inférence phylogénétique parcimonieux. On peut alors définir le problème d'optimisation suivant, que l'on appelle le *Problème de Minimisation des Duplications* : étant donnés k arbres de gènes $\mathcal{T} = \{T_1, \ldots, T_k\}$, calculer un arbre d'espèces S qui minimise la somme $\sum_{i=1}^{k} d(T_i, S)$. On note $d(\mathcal{T})$ le nombre de duplications correspondant à un arbre d'espèces optimal.

Complexité. Comme la plupart des problèmes d'inférence d'un arbre d'espèces, le Problème de Minimisation des Duplications (MD) est NP-complet. Ce résultat de complexité a été prouvé explicitement dans [Ma 2000], mais était déjà présent implicitement dans la thèse de David Bryant [Bryant 1997] comme un problème d'*incohérences minimale de super-arbres* (le *Minimum Rooted Triplets Inconsistency*, *MRTI*), sur lequel nous reviendrons. La différence entre les deux preuves de NP-complétude est cependant importante. La preuve présentée dans [Ma 2000] repose sur des structures combinatoires relativement sophistiquées et peu réalistes en termes d'arbres de gènes issus de données réelles, alors que la preuve que l'on peut déduire de [Bryant 1997] est valide pour des triplets aux étiquettes uniques, c'est-à-dire des arbres ayant trois feuilles distinctes et deux sommets internes. De plus, le problème MRTI n'admet pas d'algorithme d'approximation de ratio meilleur que $\Omega(\log(n))$ ou d'algorithme paramétré par $d(\mathcal{T})$ [Byrka 2010]. Ces deux résultats s'appliquent aussi au problème MD, invalidant ainsi un résultat de complexité paramétrée annoncé dans [Stege 1999], une remarque faite, indépendamment, par Bansal et Shamir [Bansal 2011].

Algorithmes existants. Le problème MD est donc difficile à résoudre exactement. Il a été appliqué initialement sur des jeux de données comportant relativement peu de génomes [Sanderson 2007] en utilisant des méthodes de génération exhaustive des arbres d'espèces possibles. Dans sa thèse de doctorat, Jean-Philippe Doyon a développé un algorithme de type Branch-and-Bound [Doyon 2011a], que nous avons appliqué à des jeux de données couvrant 29 génomes de vertébrés. Cependant, même avec un tel nombre (relativement modeste) de génomes, cette technique ne permet pas de résoudre le problème MD exactement. Les résultats récents les plus intéressants sont basés sur des algorithmes rapides de recherche locale qui permettent d'analyser de grands jeux de données, sans garantie d'obtenir un résultat exact cependant [Wehe 2008, Bansal 2008, Bansal 2009, Burleigh 2011].

Mes principales contributions à ce problème de génomique concernent la question suivante : peut-on caractériser les instances $\mathcal{T}$ pour lesquelles résoudre le problème MD est faisable en temps polynomial ?

Instances explicables sans pertes. Un premier résultat trivial est que si tous les sommets internes de $\mathcal{T}$, sauf les cerises, sont des duplications forcées, alors résoudre le problème MD est facile : $d(\mathcal{T})$ est exactement le nombre de sommets internes de $\mathcal{T}$, hors cerises, et tout

arbre d'espèces est optimal. En d'autres termes, il n'y a plus de signal dans $\mathcal{T}$. Intuitivement, on préfère avoir à faire à un jeu de données contenant un signal clair (i.e. admettant peu de solutions optimales) et pour lequel on peut calculer une (ou mieux toutes les) solution(s) optimale(s). Dans [Chauve 2008a], nous avons caractérisé une famille d'instances ayant ces propriétés.

Théorème 15 [Chauve 2008a]

1. Si $\mathcal{T}$ peut être expliqué sans *perte de gènes* (i.e. en utilisant uniquement des duplications et spéciations), alors il existe un unique arbre d'espèces S optimal pour le problème MD.
2. Décider si $\mathcal{T}$ peut être expliqué sans perte de gènes et calculer S peut se faire en temps et espace $O(n)$, où n est le nombre de sommets de $\mathcal{T}$.

La motivation pour introduire ce problème est la suivante : pour comprendre l'histoire évolutive qui a mené aux arbres de $\mathcal{T}$, il suffit de replacer les arbres correspondants aux pertes de gènes. En effet, chaque perte correspond à la disparition d'un arbre complet (l'évolution du gène perdu). En replaçant ces sous-arbres (ou plutôt en représentant chacun par un ensemble de feuilles, correspondant aux descendants de l'espèce dans laquelle la perte a eu lieu) dans $\mathcal{T}$, on obtient un ensemble d'arbres ayant évolué sans perte (par construction), et donc un unique arbre d'espèces. Vu sous cet angle, le problème de reconstruire l'histoire d'une famille de gènes est donc plus un problème de détection des pertes de gènes que de minimisation de duplications.

L'algorithme de décision prouvant le point 2 du Théorème 15 est très simple et repose sur le principe simple que dans un arbre de gènes T explicable sans perte les cerises sont soit disjointes soit égales. Cette propriété est aisé à vérifier et permet de contracter les cerises, qui indiquent donc les dernières spéciations de S, avant d'utiliser le même principe sur l'arbre T ainsi réduit.

Plus généralement, dans [Chauve 2008a], nous avons introduit le Problème de Minimisation des Pertes, l'analogue du problème MD mais utilisant le nombre de pertes de gènes comme critère d'optimisation. La complexité de ce problème est encore inconnue, même dans le cas simple où $\mathcal{T}$ est un ensemble de cerises et de feuilles. Mais dans ce cas simple, il apparaît en fait comme une variante d'un problème classique de phylogénie (la parcimonie de Camin-Sokal [Felsenstein 2004]) et nous conjecturons qu'il est NP-complet. Nous avons développé une heuristique pour ce problème qui a donné de bons résultats sur des données simulées. Dans [Doyon 2011a], nous avons aussi défini un algorithme de Branch-and-Bound, dont le comportement sur des données réelles de vertébrés est intéressant par deux aspects. Premièrement, il semble que le nombre de pertes et le nombre de duplications sont très corrélés, ce qui n'est pas surprenant si on suppose que le devenir normal, attendu plutôt, d'une paire de gènes dupliqués est de voir une copie disparaître. Deuxièmement, l'algorithme de Branch-and-Bound est sensiblement plus rapide si on utilise les pertes de gènes comme critère d'optimisation, car les pertes sont plus faciles à localiser tôt et permettent de couper plus vite l'espace des arbres d'espèces à explorer. Ces propriétés restent toutefois à explorer plus en détail.

Duplications forcées. Dans un deuxième temps, nous avons étendu nos résultats de [Chauve 2008a] au cas des instances $\mathcal{T}$ que l'on peut expliquer en n'utilisant que leurs duplications forcées : $d(\mathcal{T})$ est le nombre de sommets duplications forcées dans $\mathcal{T}$. Il n'est pas difficile de voir que si une instance $\mathcal{T}$ peut être expliqué sans pertes, alors elle appartient à cette classe d'instances. L'exemple donné plus haut où tous les sommets de $\mathcal{T}$ hors cerises sont des duplications forcées montre que cette seule propriété n'est pas suffisante pour obtenir un résultat utile. Il faut donc aussi prendre en compte le nombre de solutions optimales. Nous avons étudié cette question dans [Chauve 2009a].

Dans un premier temps, nous avons montré que le problème MD est en fait équivalent à un problème d'incohérence minimale de super-arbres. Intuitivement, un problème d'incohérence minimale de super-arbres prend en entrée un ensemble d'arbres aux étiquettes uniques[2], repose sur une notion d'incohérence entre deux arbres aux étiquettes uniques, et cherche à inférer un arbre d'espèces qui minimise le nombre d'arbres donnés en entrée avec lesquels il est incohérent (voir [Bininda-Edmonds 2004, Berry 2008] pour des textes généraux sur les super-arbres). Si un arbre de gènes est un triplet aux étiquettes uniques, il n'est pas difficile de voir que seule la racine peut-être une duplication[3], ce qui arrive si et seulement si ce triplet est incohérent avec l'arbre d'espèces choisi. Si tous les arbres donnés en entrée sont des triplets aux étiquettes uniques, minimiser le nombre d'incohérences est donc équivalent à minimiser le nombre de duplications.

Pour étendre ce lien à des arbres plus généraux que les triplets aux étiquettes uniques, nous avons introduit la notion de *bipartition*. Soit u un sommet interne d'un arbre T de $\mathcal{T}$, dont les enfants sont u_1 et u_2. La bipartition associée à u est l'arbre composé d'une racine r, ayant deux enfants r_1 et r_2 tels que r_1 (resp. r_2) a pour enfants les feuilles $L_T(u_1)$ (resp. $L_T(u_2)$). Soit $\mathcal{B}$ le multi-ensemble des bipartitions associées à tous les sommets de $\mathcal{T}$ qui ne sont *pas* des duplications apparentes.

Théorème 16 [Chauve 2009a] Soient f le nombre de duplications forcées dans $\mathcal{T}$ et S un arbre d'espèces. Si $i(\mathcal{B}, S)$ est le nombre de bipartitions de $\mathcal{B}$ qui sont incohérentes avec S alors $d(\mathcal{T}, S) = f + i(\mathcal{B}, S)$.

Ce résultat montre donc qu'il suffit de résoudre un problème d'incohérence minimale de super-arbres avec des arbres de gènes spéciaux (les bipartitions) pour résoudre le problème MD. De plus, en utilisant des techniques classiques de la théorie des super-arbres (les approches de type *Coupe Minimum* [Semple 2000]) pour décider si, pour un ensemble $\mathcal{U}$ d'arbres aux étiquettes uniques, il existe au moins un arbre d'espèces qui est cohérent avec tous les arbres de $\mathcal{U}$, nous avons montré le résultat suivant.

Théorème 17 [Chauve 2009a] Décider si $\mathcal{T}$ peut être expliqué en utilisant seulement ses duplications forcées peut se faire en temps et espace polynomial.

[2] Typiquement, chacun représente l'histoire évolutive d'un ensemble de gènes orthologues deux-à-deux.
[3] Une cerise ne peut jamais être une duplication, sauf si ses deux feuilles ont même étiquette.

Notre preuve de ce résultat est constructive. Nous décrivons en effet un algorithme, basé sur le principe suivant : (1) on définit, à partir de $\mathcal{T}$ un graphe dont les sommets sont $\{1, 2, \ldots, m\}$ et dont les arêtes sont définies en termes de duplications apparentes de $\mathcal{T}$, (2) si ce graphe est connexe, alors il n'existe pas d'arbre d'espèces expliquant $\mathcal{T}$ uniquement avec ses duplications forcées, et (3) sinon, on peut supprimer certaines arêtes du graphe et appliquer le même principe récursivement aux différentes composantes connexes. Si ce processus se conclut par un graphe sans arête, alors chaque étape (3) indique une (ou plusieurs si le nombre de composantes connexes est supérieur à 2) spéciations et on obtient donc un arbre $\mathcal{S}$, possiblement non-binaire mais aux étiquettes uniques. Tout arbre d'espèces S qui permet d'expliquer $\mathcal{T}$ sans autre duplication que les duplications forcées est un raffinement binaire de $\mathcal{S}$ (l'énoncé inverse n'est cependant pas vrai en général). La structure de $\mathcal{S}$ permet ainsi d'estimer (mais pas de compter exactement) le nombre d'arbres d'espèces qui expliquent $\mathcal{T}$ en utilisant uniquement les duplications forcées. Ces résultats ont été étendus dans le travail récent [Scornavacca 2011].

Approximation pour la première spéciation. L'utilisation de l'approche par Coupe Minimale dans le problème précédent soulève naturellement la question de son applicabilité en dehors du cas des instances "faciles" que nous venons de décrire plus haut. Dans [Ouangraoua 2010b], nous avons montré qu'en utilisant une généralisation classique de la notion de Coupe Minimale, la *Minimisation de Fonction Sous-Modulaire* [Iwata 2009, Fujishige 2005], on peut obtenir en temps polynomial une première spéciation (i.e. une partition de l'ensemble des espèces en deux ensembles) qui induit une nombre de duplications la précèdant (i.e. les sommets de $\mathcal{T}$ associés par la correspondance LCA à la racine de l'arbre d'espèces) qui est au plus deux fois le nombre minimum de telles duplications.

Théorème 18 [Ouangraoua 2010b] On peut calculer une 2-approximation d'une spéciation optimale en temps et espace $O(kn)$, où k est le nombre d'arbres dans $\mathcal{T}$ et n le nombre de sommets dans $\mathcal{T}$.

Ce résultat est basé sur l'utilisation d'une variante de la Coupe Minimale d'un graphe, aux arêtes étiquetées de manière non-unique, dans lequel le coût d'une coupe est le nombre d'étiquettes sur les arêtes de cette coupe. Il est intéressant d'un point de vue théorique. En effet l'encodage naturel du problème de calcul d'une première spéciation optimale résulte en un graphe aux arêtes étiquetées pour lequel calculer une Coupe Minimale Étiquetée n'est pas faisable. Cependant, une modification astucieuse de ce graphe, qui consiste à ajouter des arêtes liées aux duplications forcées de $\mathcal{T}$, résulte en un graphe dont une Coupe Minimale Étiquetée peut être calculée par une Minimisation de Fonction Sous-Modulaire, ce qui est une technique inédite à notre connaissance. Des résultats expérimentaux préliminaires suggèrent que cette approche résulte souvent en une première spéciation parcimonieuse.

A.4 Conclusion

Les résultats présentés dans ce chapitre portant sur la notion de réconciliation sont importants dans l'optique de la reconstruction de génomes ancestraux utilisant les gènes comme marqueurs. Ils suggèrent en effet que la vraie réconciliation est proche de la réconciliation LCA. Ces résultats reposent sur des simulations qui ignorent un problème fondamental : les erreurs, souvent importantes, dans les arbres de gènes [Hahn 2007]. Une avenue naturelle pour attaquer ce problème pourrait être de détecter les branches et sommets des arbres de gènes qui présentent un signal douteux. Les algorithmes que nous proposons pour explorer rapidement un ensemble de réconciliations devraient s'avérer utile dans cette optique.

Concernant la reconstruction d'un arbre d'espèces à partir d'un ensemble d'arbres de gènes, les quelques résultats que nous avons obtenus ouvrent des pistes intéressantes. Par exemple, est-il facile de calculer tous les arbres d'espèces optimaux si $\mathcal{T}$ peut être expliqué en utilisant uniquement les duplications apparentes ? Et comment généraliser ce résultat ? On sait qu'il n'existe pas d'algorithme de complexité paramétrée par le nombre minimal de duplication, mais cela n'exclut pas d'essayer de définir un paramètre basé sur les duplications non-forcées qui doivent être utilisées. Finalement, nous avons montré que le calcul d'une première spéciation de coût au plus deux fois le coût optimal est faisable. Peut-on étendre ce résultat à un ensemble d'un nombre borné de spéciations ?

Contributions. Exploration de l'espace des réconciliations.

- [Doyon 2009] J.-P. Doyon, C. Chauve et Sylvie Hamel. *Space of Gene/Species Trees Reconciliations and Parsimonious Models*. Journal of Computational Biology, vol. 16, no. 10, pages 1399–1418, 2009. Version préliminaire publiée dans les actes de RECOMB-CG 2008.
- [Doyon 2011b] J.-P. Doyon, C. Chauve et S. Hamel. *An Efficient Method for Exploring the Space of Gene Tree/Species Tree Reconciliations in a Probabilistic Framework*. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2011. Sous presse.

Phylogénomique.

- [Chauve 2008a] C. Chauve, J.-P. Doyon et N. El-Mabrouk. *Gene Family Evolution by Duplication, Speciation, and Loss*. Journal of Computational Biology, vol. 15, no. 8, pages 1043–1062, 2008. Version préliminaire publiée dans les actes de RECOMB-CG 2007.
- [Chauve 2009a] C. Chauve et N. El-Mabrouk. *New Perspectives on Gene Family Evolution : Losses in Reconciliation and a Link with Supertrees*. Computational Molecular Biology, 13th Annual International Conference, RECOMB 2009, pages 46–58. Springer, 2009.
- [Ouangraoua 2010] A. Ouangraoua, K. M. Swenson et C. Chauve. *An Approximation Algorithm for Computing a Parsimonious First Speciation in the Gene Duplication Model*. Comparative Genomics - International Workshop, RECOMB-CG 2010, pages 290–301. Springer, 2010. Version étendue à paraître dans Journal of Computational Biology.
- [Doyon 2011a] J.-P. Doyon et C. Chauve. Software tools and algorithms for biological systems, volume 696 of *Advances in Experimental Medicine and Biology*, chapitre Branch-and-Bound Approach for Parsimonious Inference of a Species Tree From a Set of Gene Family Trees, pages 287–295. Springer, 2011.