ÉCOLE DOCTORALE DU GÉNOME AUX ORGANISMES

THÈSE DE DOCTORAT

Spécialité : Bio-informatique, Biologie structurale, et Génomique

Présentée par

Matthieu Muffato

Pour l'obtention du grade de

Docteur de l'université d'Évry Val d'Essonne

Reconstruction de génomes ancestraux chez les vertébrés

Soutenue le 15 décembre 2010 devant le jury compsé de :

M. Gilles FISCHER	Rapporteur
M. Christophe HITTE	Rapporteur
Pr. Florence D'ALCHÉ-BUC	Examinatrice
M. Thomas FARAUT	Examinateur
M. Javier HERRERO	Examinateur
M. Éric TANNIER	Examinateur
Pr. Franck DELAPLACE	Directeur de thèse
M. Hugues ROEST CROLLIUS	Codirecteur de thèse

Remerciements

Avant tout chose, je lance un clin d'œil à Steven et Michael, qui ont préparé le terrain de la vulgarisation de mon travail de thèse (avant même que je ne la commence, avant même que je ne découvre la bio-informatique) dans une épopée sauvage et reptilienne ... Jurassic Park.

J'adresse tout d'abord mes sincères salutations aux membres de mon jury de thèse : Gilles FISCHER, Christophe HITTE, Javier HERRERO, Thomas FARAUT, Éric TANNIER et Florence D'ALCHÉ-BUC. Un des accomplissements d'une carrière de chercheur est d'être reconnu par ses pairs, et je suis très heureux de m'approcher du «monde des grands» en soutenant ma thèse devant vous. En particulier, je salue Éric, Christophe, et Gilles pour nos diverses rencontres de travail, qui m'ont souvent motivé à aller plus loin dans mon travail, et poussé à remettre en cause certains acquis. D'autre part, nous conviendrons aisément que je n'ai pas réalisé ce travail seul. L'encadrement dont j'ai bénéficié (Franck DELAPLACE et Hugues ROEST CROLLIUS) m'a permis d'explorer toutes les idées qui me traversaient (les meilleures idées sont souvent les plus incongrues), sans qu'ils hésitent à me rappeler à l'ordre sur mes obligations ou sur la validité de mes pistes.

J'ai la chance d'avoir un sujet de thèse aisément compréhensible («Comment était le génome des espèces qui vivaient il y a des dizaines, des centaines, de millions d'années ?»), ce qui permet de ne pas passer pour un extraterrestre dans les conversations, et extraordinairement ludique (de mon point de vue, mon travail consiste à jouer au LEGO avec les génomes du chat, de la grenouille, de l'écureuil et du kangourou), ce qui m'a évité de me poser la question de l'utilité de mon travail. D'ailleurs, il m'a bien fallu deux ans pour que je commence à envisager de faire autre chose que m'amuser au travail, avant de réaliser encore un ou deux ans plus tard que ce que je faisais allait grandement servir à la communauté, et que donc je pouvais légitimement continuer à prendre du plaisir. D'autant plus, que je dispose d'un certain confort dans le jugement de mes résultats car les véritables génomes ancestraux ne seront probablement jamais connus ! (à moins d'inventer une machine à remonter le temps, mais ceci est le travail d'autres étudiants en thèse).

Merci infiniment à Hugues d'avoir cru en moi depuis le début, depuis ce premier stage auprès de Sarah, et de m'avoir subi en stage de M2 / en thèse durant 1779 jours d'affilée sans jamais -trop- broncher (pourtant, je le reconnais, j'ai des aspects très énervants : «C'est pour demain ? Bon, ok, je m'y mets», «Ah bon ? Il faut écrire un mémoire ? Et préparer une soutenance aussi ! ! !»). Mon unique crainte (et oui, il m'arrive de stresser) a été de te décevoir, et j'espère que ça n'a pas été le cas.

Le risque de la thèse est de se replier, et l'ambiance, et la cohésion de l'équipe et du laboratoire sont primordiaux pour assurer une bonne qualité de vie, et des résultats. En ce sens, j'ai été chanceux, d'avoir eu des co-bureaux si agréables à vivre : Sarah, Alexandra, Charles, Camille, et ma plante verte. Bien entendu, cela n'enlève rien aux qualités de ceux et celles avec qui je n'ai pas eu l'occasion de partager un bureau (d'autant plus, que moi j'ai la place à l'ombre tandis que vous, vous auriez cramé au premiers rayons de soleil !) :

David (même à 6000 km de distance, n'hésite pas à m'envoyer par e-mail tes blagues / commentaires graveleux / salaces), Stéphane, Guy, Magali, et Marlène.

La vie au laboratoire a été agrémentée régulièrement de *Monday Cakes*, de séances de Counter Strike, de ~~soirées pizza-film~~ séminaires étudiants, de sorties diverses (comme le cinéma plein-air à La Villette). Tout ceci s'est fait en compagnie de Laurent, Mathilde (à qui je joins des remerciements particuliers, pour m'avoir ouvert les yeux au monde de la pâtisserie et à une sensibilité culturelle), Baptiste, Sophie, Thomas, Yann, Sébastien, et je mesure le vide que je devrai combler une fois parti, ne serait-ce que tout simplement les rencontres quotidiennes.

Rendons à César ce qui est à César, cette thèse n'aurait pas eu la même saveur sans les desserts de la cantine (ce serait mentir que de dire que je ne les apprécie pas), les courgettes de la cantine (ce serait mentir de dire que je les apprécie, mais j'avoue qu'elles m'ont motivé comme jamais pour apprendre à -bien- cuisiner les courgettes), et les tétraodons de l'aquarium (Roger, pour les intimes) avec leurs coquilles de moules.

D'un point de vue plus professionnel, je tiens à féliciter le travail, la patience et l'écoute de l'équipe du service informatique menée par Pierre VINCENS : Édouard, Arnaud, Catherine, Clarisse et Jean-Pierre. Ils ont maintenu un environnement matériel et logiciel de travail propice à l'accomplissement de cette thèse, malgré nos rouspétances continuelles. Le développement des outils et serveurs décrits dans ce manuscrit n'auraient pu avoir lieu sans les moyens qui ont été mis à ma disposition.

Il est encore question de patience et d'écoute pour décrire les qualités de Brigitte, Anne, et Martine. Malheureusement pour vous, un chercheur reste un chercheur, et ne saura jamais faire les tâches administratives correctement et/ou à temps. Merci pour votre implication et votre compréhension.

Mes seuls regrets sur ces -presque- 5 années sont de ne pas avoir photographié chaque jour l'arbre en face de ma fenêtre (cela aurait fait un magnifique roman-photo), et de ne pas avoir fait de statistiques sur la ponctualité du RER D (on fait un concours de la meilleure excuse ? Moi j'ai eu un «En raison d'une biche sur la voie» une fois !), chose qui mérite amplement un sujet de thèse ...

Passons la porte du laboratoire et allons saluer les amis docteurs (Yun-Kang, Philippe, Valentin), ingénieurs (Ted, Vivien, Jean-Marie, Sylvain, Aline, Sami, Tanguy, Samuel, David, Mathieu) ou bienfaiteurs (Karine, Noémie, Rémy, Marie-Claire) et la famille (Martin, Maman, Papa). Vous avez parfaitement tenu votre rôle : être là pour me changer les idées, tout en me laissant m'investir autant que je le voulais (parfois à votre détriment, donc) dans un sujet qui m'a passionné.

Enfin, pour être complet en cette période de reconnaissance, je me dois de saluer l'impact bénéfique sur mon bien-être qu'ont eu deux groupes de personnes : Tijs, Armin, Menno, Markus, et Paul d'une part (*In trance we trust o/*), Adam, Randy, John, Shelton et Rob d'une autre (*PAR DEssus la troisième cooorde ! !*).

Sur ce, passons à la science !

Table des matières

Première partie

Introduction

Chapitre 1

Préambule

1.1 Problématique

Pourquoi reconstruire des génomes ancestraux ? D'un point de vue fondamental, les études des systèmes biologiques contemporains en utilisant, par exemple, des approches d'anatomie, de biochimie, de physiologie, et biologie moléculaire, sont sérieusement limitées par l'absence d'un cahier de laboratoire de l'Évolution qui décrirait et expliquerait leur mise en place, leur organisation et leur fonctionnement. L'objectif à long terme de posséder les génomes ancestraux est d'établir un vaste cadre d'étude de l'évolution pour corriger le manque cruel de données historiques (la molécule d'ADN ne se conserve guère plus d'une centaine de milliers d'années). Pour atteindre cet objectif, de nombreux développements algorithmiques sont nécessaires pour traiter de manière efficace et systématique les larges volumes de données disponibles, selon une méthodologie rigoureuse.

Les données génomiques suivent en général bien ce paradigme parce qu'elles ont une résolution très élevée (à la base près), sont très fiables (beaucoup de séquences de génomes contiennent moins d'une erreur toutes les 10000 bases), très abondantes (plus de 100 génomes eucaryotes séquencés, plus de 1000 procaryotes), et centralisées dans des bases de données publiques. Le génome fournit aussi des points d'entrées fondamentaux vers les propriétés fonctionnelles des organismes, comme la présence ou l'absence de gènes, l'expansion ou la régression des familles de gènes, la topologie des éléments cis-régulateurs, qui en retour nous informent sur la vraisemblance de certaines voies métaboliques ou de développement qui peuvent exister dans un organisme, et l'importance des fonctions spécifiques à chaque espèce. Les génomes représentent ainsi la fondation sur laquelle de nombreuses avancées peuvent être faites, et accéder à ces informations dans un génome ancestral fournit un large spectre de ces propriétés.

D'un point de vue plus pratique, compte tenu de la quantité astronomique de données génomiques apportées à la communauté, dont le rythme, vraisemblablement, s'accélèrera encore dans les prochaines années, il est critique de garder un degré d'organisation substantiel pour la distribution et la présentation des données. Les résultats de reconstructions de génomes ancestraux permettront de lier naturellement les séquences et les annotations des espèces modernes entre elles, et avec celles des espèces ancestrales, dans le sens de l'évolution, en suivant la phylogénie des espèces. Les génomes ancestraux serviront de points de référence unique pour comparer des génomes descendants, ce qui facilitera grandement l'identification de propriétés génomiques ancestrales, et donc les gains ou pertes lignées-spécifiques. Réciproquement, les résultats qui continueront à être obtenus avec les différents organismes modèles les enrichiront en retour. Pour résumer, les

génomes ancestraux représentent les fondations qui nous aideront à déchiffrer les différentes composantes moléculaires contribuant à l'évolution des espèces, et qui ont mené à une telle variété d'espèces et de systèmes biologiques.

Le but de cette thèse et de mettre au point les outils capables de reconstruire les ancêtres successifs dans plusieurs lignées, fournissant ainsi pour la première fois une vue dynamique de l'évolution des génomes. Tout comme les points temporels dans une expérience sont cruciaux pour comprendre la dynamique ou la physique d'un processus biologique, les génomes ancestraux représentent ces fameux points de mesure, au cours d'une expérience qui dure des millions d'années.

1.2 Résumé de la thèse

Ce travail de thèse décrit une nouvelle méthode, appelée AGORA (*Algorithms for Gene Order Reconstruction in Ancestors*), pour reconstruire de manière automatique et systématique l'ordre des gènes et les caryotypes de toutes les espèces ancestrales dans une phylogénie donnée. AGORA est capable de gérer les duplications de gènes, les délétions, et les gains, et interprète de manière réaliste des phylogénies complexes de gènes. Nous avons appliqué la méthode chez les vertébrés (en utilisant huit espèces outgroups supplémentaires) pour reconstruire des ordres de gènes ancestraux dans 43 génomes ancestraux à partir de 54 espèces modernes séquencées et annotées. Les performances d'AGORA ont été mesurées par des simulations de génomes de vertébrés, et par confrontation à des génomes ancestraux déjà connus. Les données, présentées graphiquement dans un serveur web nommé Genomicus [Muffato *et al.*, 2010] fournissent une nouvelle ressource pour l'étude de l'évolution des génomes dans un cadre dynamique. Les données couvrent près de 600 millions d'années d'évolution, et sont disponibles pour de nombreuses études biologiques comme l'évolution des gènes, ou les réarrangements dans les génomes.

1.3 Plan de la thèse

Le manuscrit est organisé en cinq parties. La première partie présente l'état de l'art dans le domaine de la reconstruction des génomes ancestraux, en décrivant les différentes techniques existantes, leurs avantages et leurs limites. La partie se clôt par la description des génomes de trois espèces ancestrales, que nous chercherons à reconstruire de manière automatisée et intégrée dans un pipeline commun.

Les deux parties suivantes présentent l'ensemble des outils développés au cours de la thèse. Tout d'abord, la Partie II décrit succinctement des méthodes génériques en vue de leur réutilisation à plusieurs endroits du processus de reconstruction (voyageur de commerce, clustering, calcul d'une moyenne sur un arbre).

La majeure partie du travail de la thèse, la description des méthodes AGORA pour la reconstruction de génomes, est concentré dans la Partie III. AGORA assemble graduellement les données issues de la comparaison des génomes, en partant des gènes comme le plus petit dénominateur commun des génomes. Des arbres phylogénétiques définissent le contenu en gènes des espèces ancestrales, et les liens d'homologie qui permettent, dans l'étape suivante, de comparer les différents génomes modernes. Par parcimonie, la conservation de l'ordre des gènes entre deux espèces est assignée à leur dernier ancêtre commun. Les algorithmes décrits dans les chapitres suivants permettent de combiner ces données (à différentes échelles allant de la paire de gènes au chromosome entier) en éliminant

les incohérences, inhérentes à des données biologiques. Un chapitre sera consacré à la description des méthodes tirant bénéfice d'un événement particulier des génomes : les duplications complètes.

La Partie IV détaille le paramétrage du processus de reconstruction, et sa validation. Une fois introduit un protocole de simulation qui permettra de mesurer quantitativement (et empiriquement) les performances d'AGORA, nous l'utilisons pour valider AGORA

En particulier, nous montrons comment dépasser les limites des données et les problèmes qu'elles imposent pour offrir, au final, une qualité de reconstruction encore plus élevée. AGORA est en plus validé par le protocole de simulation que nous avons mis en place : des performances équivalentes aux autres méthodes de reconstruction sur des jeux de données limités, et très satisfaisantes pour le jeu de données complet, réel, que seul AGORA est capable d'interpréter. Les résultats des reconstructions AGORA concordent avec les résultats déjà connus, ce qui valident, là encore, l'ensemble du pipeline. La partie conclut sur le serveur web, Genomicus [1], que nous avons mis en place pour mettre à disposition de la communauté l'ensemble des reconstructions. Ce serveur est en ligne depuis près de deux ans et permet à de nombreuses équipes de comparer et étudier les génomes dans le sens de l'évolution.

Pour finir, la Partie V conclut la thèse en discutant des avantages et des limites d'AGORA, et surtout de quelques analyses désormais disponibles, grâce aux données des génomes ancestraux.

1.4 Publications liées à la thèse

Cette thèse a donné lieu aux publications suivantes, publiées, à paraître, ou en préparation.

Le domaine de la reconstruction des génomes ancestraux (état de l'art, avancées et concordance des méthodes) a été décrit dans un article de revue :

– Muffato et Crollius [2008] Matthieu Muffato and Hugues Roest Crollius. **Paleogenomics in vertebrates, or the recovery of lost genomes from the mist of time**. *BioEssays*, 30(2) :122–134, Feb 2008. doi: 10.1002/bies.20707.

Le serveur et la base de données Genomicus (chapitre 15) ont été publiés dans une note d'application :

– Muffato *et al.* [2010] Matthieu Muffato, Alexandra Louis, Charles-Edouard Poisnel & Hugues Roest Crollius. **Genomicus : a database and a browser to study gene synteny in modern and ancestral genomes**. *Bioinformatics*, 26(8) :1119–1121, Apr 2010. doi: 10.1093/bioinformatics/btq079.

La méthode de reconstruction en elle-même (AGORA) et les analyses qui en découleront feront l'objet de trois articles.

– Matthieu Muffato & Hugues Roest Crollius. **Simulation of the evolution of gene content & order**. Note d'application soumise à *Bioinformatics* (équivalente au chapitre 12 de ce manuscrit).
– Matthieu Muffato, Alexandra Louis & Hugues Roest Crollius. **Automatic Reconstruction of Gene Order in Multiple Ancestral Vertebrate Genomes**. Article *en cours de rédaction* (équivalent à la Partie III de ce manuscrit).
– Matthieu Muffato, Alexandra Louis & Hugues Roest Crollius. **Genome-wide and Phylum-wide Framework for the Analysis of Vertebrates Evolution**. Article

1. http://www.dyogen.ens.fr/genomicus/

en préparation (titre provisoire) qui présentera les analyses rendues possibles par les génomes ancestraux (chapitre 17).

Les techniques de comparaison de génomes utilisées dans le cadre de cette thèse ont servi dans le cadre du projet international d'analyse du génome d'*Oikopleura dioica*. Ce travail ne traitant pas de reconstructions ancestrales, il ne sera pas abordé dans le manuscrit.

– Denoeud *et al.* [2010] France Denoeud, Simon Henriet, Sutada Mungpakdee, Jean-Marc Aury, Corinne Da Silva, Henner Brinkmann, Jana Mikhaleva, Lisbeth Charlotte Olsen, Claire Jubin, Cristian Cañestro, Jean-Marie Bouquet, Gemma Danks, Julie Poulain, Coen Campsteijn, Marcin Adamski, Ismael Cross, Fekadu Yadetie, Matthieu Muffato, Alexandra Louis, Stephen Butcher, Georgia Tsagkogeorga, Sarabdeep Singh Anke Konrad, Marit Flo Jensen, Evelyne Huynh Cong, Helen Eikeseth-Otteraa, Benjamin Noel, Véronique Anthouard, Betina M. Porcel, Rym Kachouri-Lafond, Atsuo Nishino, Matteo Ugolini, Pascal Chourrout, Hiroki Nishida, Rein Aasland, Snehalata Huzurbazar, Eric Westhof, Frédéric Delsuc, Hans Lehrach, Richard Reinhardt, Jean Weissenbach, Scott W. Roy, François Artiguenave, John H. Postlethwait, J. Robert Manak, Eric M. Thompson, Olivier Jaillon, Louis Du Pasquier, Pierre Boudinot, David A. Liberles, Jean-Nicolas Volff, Hervé Philippe, Boris Lenhard, Hugues Roest Crollius, Patrick Wincker & Daniel Chourrout. **Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate**. *Science*, Nov 2010. doi: 10.1126/science.1194167.

Chapitre 2

État de l'art des méthodes de reconstruction

Sommaire

Le problème de la reconstruction de génomes ancestraux a été adressé selon une gamme d'approches assez différentes, allant de l'optimisation combinatoire à des approches de cytogénétique en passant par des méthodes basées sur les conservations d'adjacences. La résolution des résultats varie selon les méthodes et va de la reconstruction du caryotype à celle d'un ordre de marqueurs (gènes, blocs de séquence conservée), et atteignent éventuellement la reconstruction de la séquence complète d'ADN. Chez les vertébrés, les reconstructions se sont concentrées sur *Boreoeutheria* (le dernier ancêtre commun des primates, des rongeurs et des carnivores) compte tenu de sa position avantageuse dans la phylogénie des espèces. Les duplications complètes de génomes fournissent également un formidable levier pour reconstruire le caryotype d'un ancêtre situé juste avant la duplication, et ces méthodes ont été appliquées avec succès chez les poissons, les vertébrés et les levures. Ce chapitre est consacré à la description des principes de toutes ces méthodes, tandis que leurs résultats en eux-mêmes seront évoqués dans le chapitre suivant.

La notion de synténie et son éventuelle conservation entre plusieurs espèces est cruciale pour les descriptions suivantes. Deux gènes d'une espèce donnée sont synténiques s'ils sont situés sur le même chromosome. Cette synténie est conservée si dans une seconde espèce, leurs orthologues sont aussi situés sur un même chromosome. Si des gènes synténiques sont contigus et que leurs orthologues le sont aussi dans un autre génome (c'est-à-dire dans le même ordre), alors les gènes sont dans un état de synténie conservée

et d'ordre conservé. Le langage courant utilise néanmoins le terme de «segment de synténie» pour désigner un bloc de gènes dont l'ordre est conservé entre deux espèces. Entre deux espèces, la synténie et l'ordre des gènes sont presque complètement conservés au moment de leur spéciation d'un ancêtre commun, et se dégradent progressivement au gré de l'évolution et des réarrangements du génome.

2.1 Cytogénétique

Les premières expériences de génomique comparative datent de plus de 30 ans, et ont ouvert la possibilité de comparer les chromosomes (morphologie et bandes chromosomiques) à partir de cellules en métaphase. Ces techniques ont fourni les premières données pour permettre de déterminer les réarrangements ancestraux de chromosomes chez les vertébrés [Rumpler et Dutrillaux, 1976, Yunis et Prakash, 1982]. Plus récemment, des expériences d'hybridation fluorescente in-situ (*FISH* : *fluorescence in situ hybridisation*) entre espèces ont été développées et ont permis d'améliorer fortement la précision et la portée de ces approches. Dans une expérience typique, l'ADN d'un chromosome donné d'une espèce de référence (souvent l'homme) est purifié, marqué par fluorescence, découpé, puis hybridé sur tous les chromosomes d'une autre espèce, cible. Il est même possible d'étudier la répartition de plusieurs chromosomes de l'espèce référence à la fois en utilisant plusieurs types de marqueurs fluorescents. Cette technique s'appelle *Chromosome painting* ou encore *Zoo-FISH* [Scherthan *et al.*, 1994, Wienberg *et al.*, 1990]. L'analyse des images au microscope (Figure 2.1) permet de découvrir des différences chromosomiques de grande échelle (de l'ordre du mégabase), comme les fusions, les fissions, ou les translocations d'une grande région. En revanche, les translocations de petites régions ne sont pas détectables, et les inversions (intra-chromosomique) sont, par principe de l'expérience, impossibles à repérer.

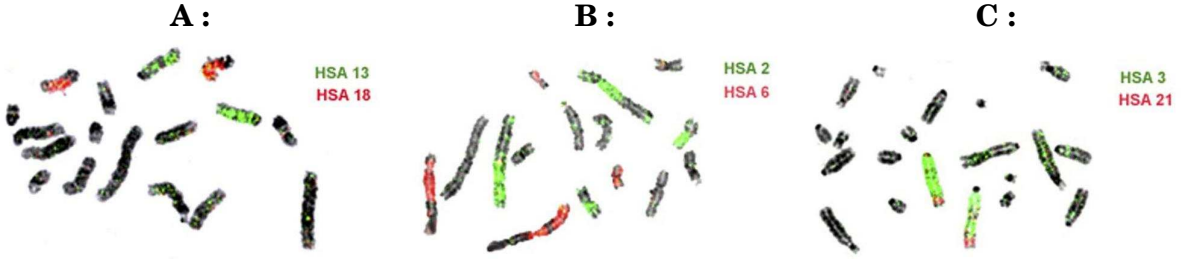


FIGURE 2.1 – Exemples d'études de cytogénétique (hybridation) tirés de Svartman *et al.* [2006]. Dans chaque expérience, deux chromosomes humains sont hybridés simultanément. **A :** Les chromosomes 13 et 18 humains sont chacun d'un seul tenant chez le tatou à neuf bandes, *Dasypus novemcinctus*, ce qui laisse supposer que c'était aussi le cas chez leur dernier ancêtre commun. **B :** Les chromosomes humains 2 et 6 sont fragmentés en deux morceaux chacun dans le génome du fourmilier à collier, *Tamandua tetradactyla*, ce qui révèle des réarrangements ancestraux de translocations, fusions ou fissions. **C :** Les chromosomes humains 3 et 21 s'hybrident sur le même chromosome du paresseux d'Hoffmann, *Choleopus Hoffmanii*, ce qui peut signifier une association ancestrale 3/21. Pour **B** et **C**, il faudra une troisième espèce pour décider de la configuration ancestrale.

Classiquement, les expériences révèlent le nombre de segments chromosomiques de l'espèce cible qui correspondent à un chromosome (entier) de l'espèce de référence. Si ce

nombre est 1, alors l'ancêtre commun de ces deux espèces devait certainement posséder ce chromosome d'un seul tenant (Figure 2.1.A). En revanche, si ce nombre est 2 ou plus (Figure 2.1.B), alors, il est nécessaire d'utiliser les données de comparaisons avec d'autres espèces (éventuellement des outgroups) pour décider quel état est ancestral (la version contiguë de l'espèce de référence ou la version fragmentée de l'espèce cible). Ensuite (Figure 2.2), on peut étudier des associations de chromosomes de l'espèce de référence dans l'espèce cible. Selon le même principe, si deux chromosomes de l'espèce de référence s'hybrident sur des chromosomes différents, alors ils devaient certainement être sur des chromosomes différents chez l'ancêtre. Si, au contraire, ils s'hybrident sur le même chromosome de l'espèce cible (Figure 2.1.C), alors, il est nécessaire d'utiliser d'autres données de comparaisons pour décider de l'état ancestral et de la position temporelle d'un réarrangement de translocation (voire une fusion ou une fission).

Le raisonnement sous-jacent pour déduire les réarrangements et l'état ancestral suit en général les principes de la cladistique [Dobigny *et al.*, 2004], et les décisions prises pour définir un caractère comme ancestral sont basées sur la parcimonie. Dans la plupart des cas, les données sont analysées à la main (les nombres d'espèces et de réarrangements impliqués sont généralement limités), mais certaines situations, rares, avec des données en grande quantité ou complexes, demandent une analyse informatique : le logiciel *PAUP* est alors utilisé [Müller *et al.*, 2003].

Le *Zoo-FISH* est extrêmement puissant car virtuellement, n'importe quelle espèce peut être étudiée, sans demander de technologie particulière (tel le séquençage) : la seule ressource nécessaire est un échantillon de tissu à partir duquel on peut faire évoluer des lignées cellulaires. Ces analyses ont cependant une limite physico-chimique due à la capacité des molécules d'ADN à s'hybrider. Deux espèces trop éloignées phylogénétiquement ont en général une trop grande divergence moléculaire de leurs chromosomes (mutations, insertions, délétions) et ne peuvent être comparées par cette technique. Ainsi, chez les mammifères, il est possible de comparer les euthériens (mammifères placentaires) entre eux (environ 100 millions d'années de divergence), mais difficilement les euthériens aux métathériens (marsupiaux), sauf pour le cas, unique, du chromosome X [Glas *et al.*, 1999, Wienberg, 2004].

La première reconstruction d'un caryotype ancestral en utilisant des données cytogénétiques était fondée sur du *Zoo-FISH* entre l'homme et sept espèces de mammifères [Chowdhary *et al.*, 1998]. Depuis lors, les résultats de nombreuses études ont permis de retrouver l'organisation ancestrale des chromosomes dans différents ancêtres, en s'attachant particulièrement aux quatre clades de mammifères placentaires [Richard *et al.*, 2003, Yang *et al.*, 2003, Froenicke, 2005, Yang *et al.*, 2006, Ferguson-Smith et Trifonov, 2007, Stanyon *et al.*, 2008, Westerman *et al.*, 2010]. La base de données *ChromHome* [1] [Nagarajan *et al.*, 2008] recense une partie des résultats des comparaisons cytogénétiques entre espèces de mammifères. On peut toutefois rappeler la possibilité d'études cytogénétiques en dehors des mammifères, comme Schneider *et al.* [2009] chez des espèces de scorpions, à condition que les génomes n'aient pas encore trop divergé.

Le *E-Painting* [Kemkemer *et al.*, 2006, 2009] est une méthode hybride entre les manipulations de cytogénétique et les données issues du séquençage des génomes. Les marqueurs conservés (typiquement des paires de gènes orthologues) permettent de simuler l'hybridation de sondes, et ainsi de comparer deux génomes. Il est donc possible, en appliquant les mêmes principes, d'atteindre une résolution jamais atteinte jusqu'alors. Cependant, le risque d'erreur est fortement accru car le nombre d'espèces séquencées (et donc

1. http://www.chromhome.org

Superorder/ order	Species	Syntenic associations																								
		3/21	4/8p	7/16	12/22	14/15	16/19	10p/12	19p/1	5/21	2/8/4	3/20	18/19	2/8	7/10	4/20	1q/10q	2/20	3/19p	18/22	5/19p	11/19	19p/q	1/10p	20/15	12/8
Afrotheria	*Aardvark*	•	•	•	•	•	•	•	•	•	•	•														
	Elephant shrew	•	•	•	•	•	•	•	•	•	•	•														
	African elephant	•		•	•	•	•	•	•	•			•													
	Golden mole	•	•	•	•	•	•	•	•	•	•															
	Manatee	•		•	•	•	•	•	•	•			•													
Xenarthra	*Two-toed sloth*	•	•	•	•	•								•	•											
	Anteater	•	•	•	•	•	•		•					•	•											
Eulipotyphla	*Shrew-hedgehog*	•	•	•	•	•		•	•					•		•										
	Common shrew	•	•	•	•	•	•									•			•							
Carnivora	*Cat*	•	•	•	•	•	•	•										•	•	•						
	Hyena	•	•	•	•	•	•	•										•	•	•						
	Dog	•	•	•	•	•											•		•							
	Mink	•	•	•	•	•	•	•										•	•	•						
	Giant panda	•	•		•		•												•							

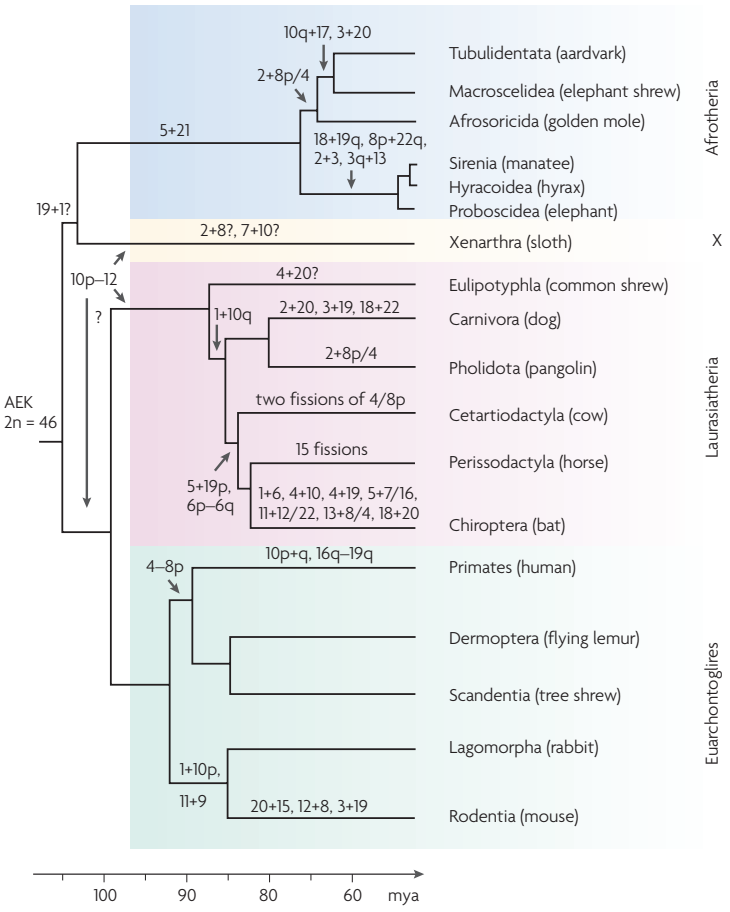


FIGURE 2.2 – Exemples d'études de cytogénétique (études d'associations ancestrales) tiré de Ferguson-Smith et Trifonov [2007]. Le tableau (tronqué) recense les associations de chromosomes humains vues dans 14 espèces de mammifères. On peut en déduire les réarrangements (fusions, translocations) qui ont marqué l'évolution des génomes des mammifères et les placer sur les branches de l'arbre phylogénétique.

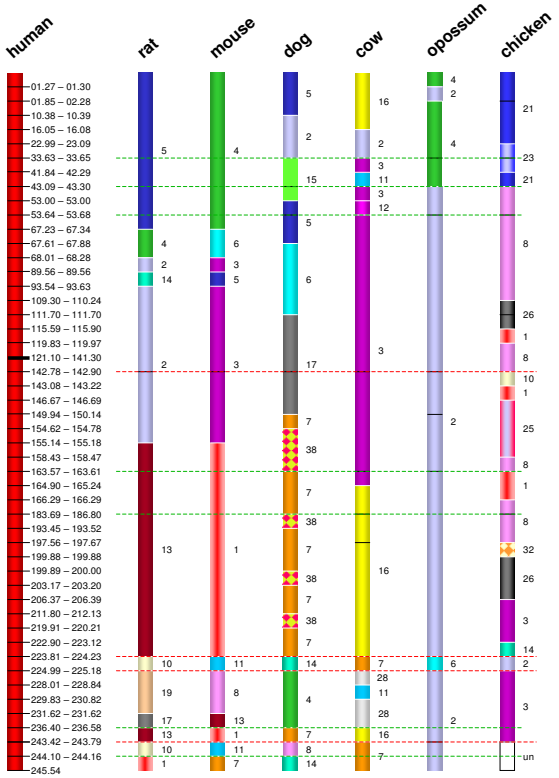


FIGURE 2.3 – Exemple de *E-painting* d'après Kemkemer *et al.* [2009]. Les gènes orthologues entre l'homme et six autres amniotes permettent de définir des blocs de synténie, de la même manière que la cytogénétique, mais avec une précision supplémentaire : la localisation sur le chromosome humain. Ainsi, on peut connaître précisément les régions du chromosome 1 humain qui se retrouvent sur le chromosome 4 de l'opossum, chose impossible en cytogénétique.

comparées) n'est pas aussi important que pour la cytogénétique «classique». Les branches espèce-spécifiques sont donc beaucoup plus longues, ce qui augmente la chance que les mêmes réarrangements aient eu lieu sur des lignées indépendantes (homoplasie), et les erreurs de reconstructions. Malgré ceci, les reconstructions les plus récentes [Kemkemer *et al.*, 2009] concordent nettement avec les prédictions de la cytogénétique.

2.2 Analyse des points de cassure

Alors que la cytogénétique permet d'analyser de nombreuses espèces à une (relativement) faible résolution, le séquençage des génomes a permis d'atteindre une précision inégalée sur (relativement) peu de génomes. Comparées aux méthodes de cytogénétique, les avantages des méthodes basées sur la séquence outrepassent les contraintes actuelles imposées par un échantillonnage limité des espèces : la possibilité de viser des niveaux différents de précision de reconstruction comme le caryotype, l'ordre des gènes (et ainsi résoudre les inversions intra-chromosomiques, indétectables en cytogénétique) et la séquence génomique.

Les séquences des génomes peuvent être converties assez intuitivement en ensembles d'objets discrets (des nucléotides, des gènes, des segments d'homologie). Dans ce domaine, les génomes sont fréquemment représentés sous la forme de permutations : un arrangement linéaire de nombre entiers représentant la position de ces objets dans le gé-

nome. Éventuellement, la permutation peut être signée pour refléter l'orientation des marqueurs (par exemple le sens de transcription des gènes), ce qui permet de manipuler un objet plus proche de la structure biologique sous-jacente.

Une analyse computationnelle compare alors les génomes modernes écrits selon ces principes pour calculer le génome ancestral. Parce que basé sur les séquences des génomes modernes, le génome ancestral possède alors des liens vers des séquences (nucléiques ou protéiques, selon les études), ce qui pose les fondements de la reconstruction de la séquence ancestrale.

Lorsque les génomes sont représentés sous forme de permutation, un défi classique est de trouver l'ordre optimal de réarrangements qui transforme un génome en un autre, et le nombre de tels événements est appelé une «distance». Cela correspond directement au problème de la reconstruction de génomes ancestraux, car l'ancêtre commun de deux génomes se situe quelque part sur le chemin (en termes de réarrangements) menant de l'un à l'autre.

Pour énumérer la liste des événements qui transforment un génome en un autre, la plupart des algorithmes identifient des points de cassure (*breakpoints*). Chaque point de cassure correspond à une paire de marqueurs voisins dans un génome, mais pas dans un autre (car séparés par un ou plusieurs marqueurs, ou présents sur des chromosomes différents). Les points de cassure symbolisent les frontières entre les segments de synténie créés par les réarrangements.

2.2.1 Optimisation combinatoire

Le problème de trouver la suite optimale de réarrangements pour transformer un génome en un autre a en premier été abordé avec les génomes composés d'un seul chromosome et a été appelé le problème du «tri par inversions» (*reversal sorting problem*), parce que seules les inversions y étaient alors autorisées. Ce sujet a été intensivement étudié au cours de la dernière décennie, et maintenant, la séquence complète des inversions peut être récupérée en une complexité [2] sub-quadratique [Tannier *et al.*, 2007] alors que le nombre d'inversions (*reversal distance*) est plus rapide à calculer (complexité linéaire, Bader *et al.* [2001]). Ces algorithmes ont ensuite été étendus pour gérer les génomes composés de plusieurs chromosomes et les réarrangements inter-chromosomiques (le problème du «tri génomique», *genomic sorting problem*). Là encore, une solution de complexité raisonnable (sub-quadratique) est disponible [Ozery-Flato et Shamir, 2006].

Ces développements sont restés essentiellement théoriques et peu d'applications pratiques à la reconstruction de génomes à grande échelle ont été disponibles jusqu'à leur implémentation dans les logiciels *GRIMM* et *MGR* (section suivante).

Il faut rajouter que ces analyses sont en passe d'être révolues par la notion de *Double Cut and Join* (DCJ) [Yancopoulos *et al.*, 2005]. Un DCJ désigne une opération générique d'édition d'un génome au cours de laquelle deux paires de marqueurs sont séparées et croisées. Là où les analyses classiques considéraient les inversions et les translocations comme deux réarrangements différents à modéliser simultanément, et les fusions et fissions de chromosomes comme des cas particuliers de translocation, le DCJ permet de modéliser les quatre à la fois, naturellement, comme quatre variantes de la même opération. La formulation des problèmes (distance et séquence optimale de réarrangements

2. La complexité d'un problème est une estimation (une borne supérieure, ou un équivalent) de la durée ou de l'espace mémoire nécessaire à sa résolution. On la note $O(f(n))$ où $f(n)$ est une fonction de n, la taille des données.

entre deux génomes) est nécessairement modifiée, mais, de fait, plus simple, et ceux-ci sont désormais de complexités linéaires. La nouveauté qui nourrit le plus d'espoir est la possibilité de tenir compte du contenu en gènes différents entre les génomes (pertes spécifiques de gènes, et duplications) [Yancopoulos et Friedberg, 2009]. Cela permettrait de faire enfin concorder l'évolution des génomes avec l'analyse combinatoire des réarrangements.

2.2.2 *GRIMM – MGR – GRIMM-Synteny*

GRIMM (*Genome Rearrangements In Man and Mouse*) contient une implémentation efficace des algorithmes de comparaison de deux génomes (en autorisant plusieurs chromosomes, avec des inversions, des translocations, des fusions et des fissions), disponible sur un serveur en ligne [Tesler, 2002]. GRIMM permet de trouver le scénario optimal (en termes de réarrangements) pour transformer un génome en un autre. Cependant, il n'est pas capable de définir le génome du dernier ancêtre commun, qui va se retrouver, en principe, quelque part entre les deux. Cette tâche demande en effet le génome d'une espèce outgroup, qui a divergé avant l'ancêtre que l'on cherche à reconstruire (ce qui équivaut à la recherche de la racine d'un arbre phylogénétique), et est implémentée dans *MGR* [Bourque et Pevzner, 2002].

Pour trois génomes, *MGR* sélectionne les réarrangements qui permettent à un génome de se rapprocher des deux autres à la fois et les applique. Au fur et à mesure, l'algorithme fait converger les trois génomes vers un même génome : le génome «médian», le génome de l'ancêtre commun des deux espèces les plus proches. Pour plus de trois génomes, l'algorithme fonctionne sur le même principe en calculant les «bons» réarrangements (ceux qui font se rapprocher un génome à tous les autres à la fois). Appliqué sur les deux génomes les plus proches (en nombre de réarrangements), ceci permet de les transformer en un génome, celui de leur dernier ancêtre commun, et de diminuer de 1 le nombre total de génomes. *MGR* continue de la sorte jusqu'à avoir reconstruit tous les ancêtres communs des espèces qu'il avait à analyser. Par ce processus, *MGR* fournit également un arbre phylogénétique de ces espèces établi sur des distances en nombre de réarrangements. *MGR* est donc capable d'analyser des nouveaux génomes, sans connaître leur phylogénie.

Même pour trois génomes, ce problème est NP-difficile [Caprara, 1999] (y compris dans la formulation *DCJ*), ce qui signifie que l'algorithme ne peut pas tester toutes les combinaisons de bons réarrangements possibles, et est amené, via des heuristiques, à faire des choix. En particulier, même si chaque étape intermédiaire (chaque bon réarrangement) est parcimonieuse, le résultat final n'est pas nécessairement la solution optimale. De plus, les auteurs montrent que souvent, l'algorithme fournit de nombreuses solutions, équivalentes entre elles (en nombre de réarrangements), et ne sait en choisir une.

Malheureusement, les résultats de *MGR* peuvent être perturbés par la présence de nombreux réarrangements de petite échelle [Carver et Stubbs, 1997, Puttagunta *et al.*, 2000], qui peuvent correspondre soit à de véritables événements, soit à des artefacts d'annotation ou d'assemblage. *GRIMM-Synteny* [Pevzner et Tesler, 2003b] est un algorithme qui aide à la définition de blocs de synténie (qui peuvent servir de marqueurs pour *GRIMM* et *MGR*). Pour ceci, le programme demande un ensemble de marqueurs supposés de très bonne qualité (des ancres) qui vont être filtrés en éliminant les ancres trop petites, ou trop éloignées de leurs voisines (les ancres problématiques, créées par des micro-réarrangements ou des artefacts, sont en effet généralement isolées). Les ancres restantes peuvent alors servir de marqueurs pour *MGR*, ce qui permet d'améliorer signi-

ficativement la qualité des reconstructions. Le trio *GRIMM-Synteny – GRIMM – MGR* définit une chaine de programmes qui permet efficacement de calculer un génome ancestral optimal, sur des données réelles de génomes séquencés.

MGR a d'abord été appliqué à des génomes mitochondriaux et aux virus de l'herpès [Bourque et Pevzner, 2002] avant d'être utilisé avec des génomes de mammifères. Comme sa puissance est équivalente à celle des marqueurs utilisés pour décrire les génomes modernes, *MGR* a au fil du temps été utilisé avec un nombre croissant de marqueurs et d'espèces de mammifères. *MGR* a d'abord été utilisé avec 114 marqueurs sur 3 espèces (l'homme, la souris, et le chat, Bourque et Pevzner [2002]), puis avec 391 marqueurs sur 3 espèces (l'homme, la souris et le rat Bourque *et al.* [2004]), 586 marqueurs sur 4 espèces (en rajoutant le poulet comme outgroup Bourque *et al.* [2005], Consortium [2004]), et enfin sur 1159 marqueurs sur 8 espèces (homme, souris, rat, chien, chat, taureau, et cochon, Murphy *et al.* [2005]). Ces différentes reconstructions sont parfois très différentes entre elles, ce qui reflète la part de variations dues à l'utilisation d'un nombre limité d'espèces,

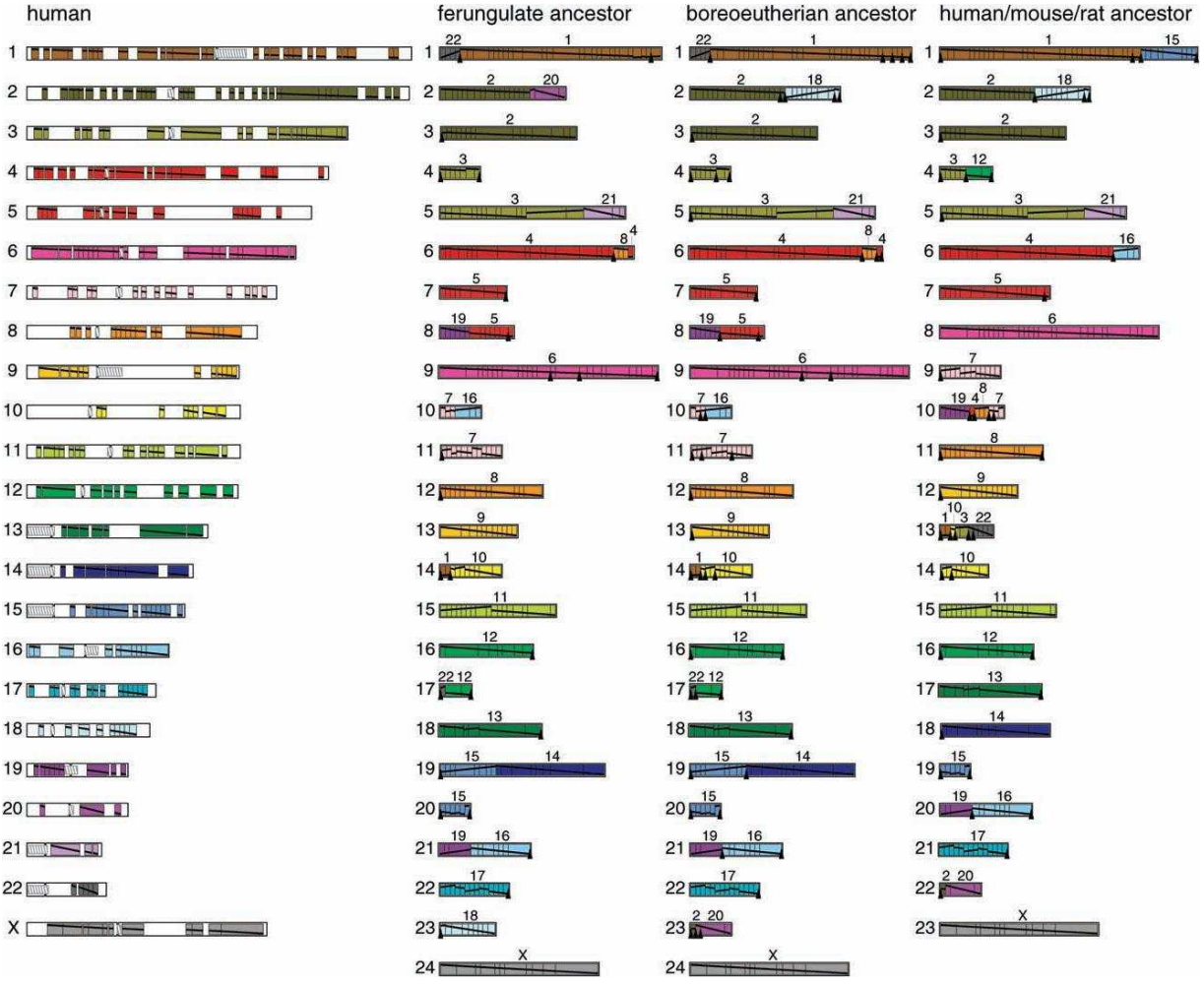


FIGURE 2.4 – Résultats de MGR dans Murphy *et al.* [2005]. L'étude porte sur 1159 marqueurs identifiés entre 8 espèces de mammifères (homme, souris, rat, chien, chat, taureau, et cochon). Les caryotypes reconstruits présentent des différences flagrantes avec les autres reconstructions, en particulier celles basées sur la cytogénétique (fusions des chromosomes humains 2 et 18, ou 1 et 22).

mais aussi les faiblesses des heuristiques.

MGR a depuis quelques concurrents comme, en particulier, *EMRAE* [Zhao et Bourque, 2009], et *MGRA* [Alekseyev et Pevzner, 2009]. *EMRAE* a pour but de retrouver les réarrangements avec une très haute spécificité, sans prédire le génome ancestral. Les auteurs affirment atteindre une spécificité de ≈85% quand *MGR* stagne à ≈45%, avec une sensibilité équivalente. De son côté, *MGRA* traite les données beaucoup plus efficacement que *MGR*, et limite le recours aux heuristiques, ce qui le rend plus efficace pour des génomes composés de plus de marqueurs.

2.3 Analyse des adjacences

Toujours dans la veine de la représentation d'un génome selon un ensemble d'objets discrets, deux méthodes abordent le problème de la reconstruction de génome ancestral sous un autre angle. Au lieu de raisonner sur les pertes d'adjacence (les points de cassure), ces méthodes repèrent et interprètent la conservation d'adjacence. Parce que toutes les adjacences ancestrales ne se retrouvent plus toujours dans les génomes modernes, le résultat de la reconstruction ne sera pas toujours un ensemble de chromosomes entiers, mais plutôt un ensemble de fragments de chromosomes. Ces fragments sont communément appelés *CARs* (*Contiguous Ancestral Regions*). De plus, ces méthodes reconstruisent directement le génome ancestral, en ne prédisant ni la suite de réarrangements qui les mènent aux génomes modernes, ni les génomes ancestraux intermédiaires (bien que ceux-ci puissent être inférés a posteriori). Les deux analyses bio-informatiques présentées ci-dessous sont les premières à montrer une grande concordance avec les résultats de la cytogénétique (pour la reconstruction de *Boreoeutheria*).

2.3.1 Ma et al., 2006

Tout d'abord, la méthode *inferCARs* [Ma *et al.*, 2006] utilise un alignement multiple de plusieurs espèces modernes. Ces alignements multiples définissent des blocs d'orthologie entre les génomes modernes, exempts de réarrangements de grande échelle. Les blocs sont traités selon un paramètre de résolution t : les blocs de taille inférieure à t sont supprimés, et la présence d'une insertion ou d'une délétion de taille supérieure à t dans un bloc coupe le bloc en deux. Les blocs sont ensuite fusionnés s'ils sont adjacents dans tous les génomes modernes. Le résultat est un ensemble de segments conservés entre les génomes modernes (maximaux car chaque extrémité est le théâtre au moins d'un réarrangement dans un génome moderne), définis à une résolution t, et représentent l'unité minimale de la reconstruction.

Les génomes modernes sont alors résumés à la séquence de ces segments conservés, ce qui permet de définir des adjacences modernes. *inferCARs* utilise le principe de parcimonie Fitch [1971], Hartigan [1973] pour faire remonter à un ancêtre cible l'information d'adjacence de ces segments, en suivant la phylogénie (supposée connue) des espèces entre elles. En cas d'ambiguïté (un segment ayant deux voisins possibles à une de ses extrémités), *inferCARs* calcule une probabilité d'adjacence ancestrale en fonction des espèces modernes qui valident l'adjacence, en tenant compte de la phylogénie des espèces. Les adjacences sont fixées chez l'ancêtre en fonction de cette probabilité, en favorisant les adjacences universelles (vues par toutes les espèces), puis les plus probables.

Dans l'article, la méthode a été appliquée chez les mammifères (homme, souris, rat, et chien) avec deux espèces outgroups (opossum et poulet), pour reconstruire le génome de

l'ancêtre *Boreoeutheria* avec les alignements multiples disponibles sur l'UCSC traités à une résolution $t = 50$ kb. 1338 segments conservés (chacun équivalent à six segments de chromosomes, un de chaque espèce moderne) ont été ainsi définis et représentent 94,31% du génome humain. Le résultat est un ensemble de 29 CARs, presque équivalents aux chromosomes de *Boreoeutheria* définis par la cytogénétique.

2.3.2 Chauve et Tannier, 2008

Dans l'autre analyse, les données proviennent là encore d'un alignement multiple des espèces modernes, avec un filtre sur les blocs d'orthologie qui y sont définis. Les blocs sont regroupés lorsqu'ils sont séparés de moins de *max_gap*, et les ensembles ainsi formés qui recouvrent au moins *min_len* de chaque génome définissent le jeu de marqueurs utilisé dans les reconstructions.

La méthode de reconstruction utilise la résolution du «problème des 1 consécutifs». Il s'agit, dans une matrice de 0 et de 1, d'ordonner les colonnes pour que sur chaque ligne, les 1 soient consécutifs. Dans le cas général, il n'existe pas toujours de solution à ce problème, et il faut donc supprimer le moins de lignes possibles pour rendre le problème soluble (ce qui est NP-difficile). Le résultat a une structure d'arbre PQ (un arbre dont tous les nœuds internes sont étiquetés par un P ou un Q). L'étiquette Q fixe l'ordre de ses fils entre eux, mais autorise à les parcourir dans un sens ou dans l'autre. Cette étiquette sert pour désigner des chromosomes (l'ordre de parcours d'un chromosome est en effet

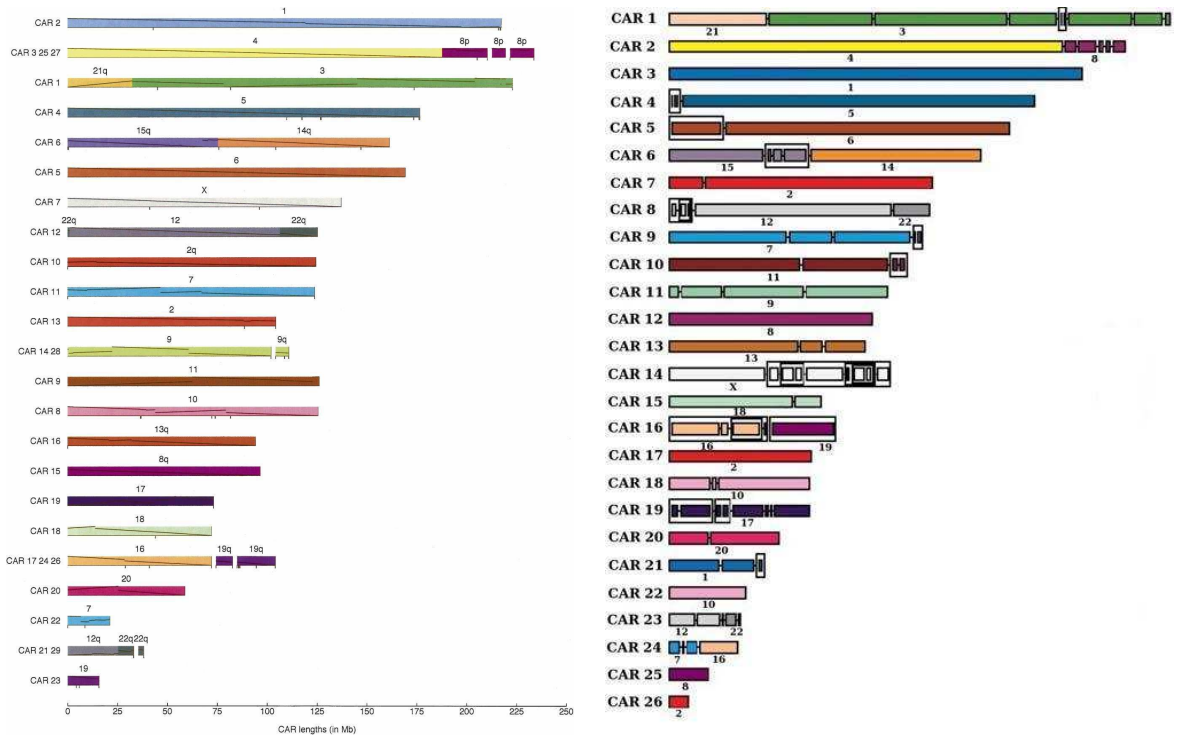


FIGURE 2.5 – Résultats d'études des analyses des adjacences. Chaque CAR désigne un (segment de) chromosome ancestral. Dans Ma *et al.* [2006] (à gauche), plusieurs CARs sont mis sur la même ligne quand les auteurs estiment qu'ils représentent le même chromosome ancestral. La ligne oblique indique la position sur le génome humain. Dans Chauve et Tannier [2008] (à droite), les encadrés montrent les étiquettes P de l'arbre de la solution, c'est-à-dire les zones dans lesquelles l'ordre n'est pas connu.

arbitraire) et pour des segments de chromosomes. L'étiquette P indique que l'ordre des fils n'est pas connu, et que n'importe quelle permutation est possible. Par exemple, si on sait qu'un chromosome est composé de deux segments (d'ordre interne connu), mais que l'ordre et l'orientation des segments entre eux ne sont pas connus, on pourra désigner ce chromosome par une étiquette P ayant deux fils d'étiquette Q. La structure d'arbre PQ est nécessaire pour décrire les différentes solutions (équivalentes) qui résultent de l'optimisation des 1 consécutifs.

Pour la reconstruction, les colonnes de la matrice désigneront les marqueurs, et chaque ligne correspondra à une adjacence conservée entre deux espèces modernes, ou à un «intervalle commun», c'est-à-dire un ensemble de marqueurs contigus dans deux génomes, mais sans conservation d'ordre. Comme dans Ma *et al.* [2006], chaque ligne (adjacence conservée ou intervalle commun) est associée à une probabilité ancestrale qui indique la conservation à travers plusieurs espèces, et qui permet de guider la suppression des lignes ambiguës (les lignes qui empêchent de résoudre le problème des 1 consécutifs). Les auteurs ont utilisé *min_len*=200kb et *max_gap*=100kb, ce qui définit 824 marqueurs, 1431 synténies ancestrales (lignes dans la matrice) dont seulement 14 doivent être éliminées pour permettre une résolution. Le résultat est un ensemble de 26 CARs, là encore, presque équivalents aux chromosomes de *Boreoeutheria*. De plus, les auteurs montrent en particulier la stabilité de leur approche selon les données (en faisant varier les paramètres *min_len* et *max_gap*).

2.4 Duplications complètes de génomes

Toutes les méthodes décrites ci-dessus nécessitent des correspondances un-à-un entre les génomes comparés pour identifier le contenu de leurs ancêtres communs, et ne savent tenir compte des duplications car celles-ci compliquent généralement le processus de reconstruction. La duplication complète d'un génome, aussi appelée tetraploïdisation, est un mécanisme assez courant d'évolution chez les plantes, moins chez les vertébrés, au cours duquel le contenu en chromosomes d'une espèce est doublé. Elle permet, de manière inattendue, de reconstruire relativement facilement des génomes ancestraux, mais uniquement celui qui précède immédiatement la duplication, ou celui qui la suit immédiatement.

2.4.1 Présentation

Chez les vertébrés, de nombreuses espèces modernes ont subi des polyploïdisations [Otto et Whitton, 2000] en plus de celle, bien documentée, qui a précédé la radiation des téléostes [Jaillon *et al.*, 2004]. Le principe qui est exploité pour reconstruire le génome pré-duplication est simple : si les chromosomes dupliqués peuvent être reconnus et appariés (éventuellement des fragments de chromosomes), alors ils désignent naturellement le contenu pré-duplication. Évidemment, plus une duplication est ancienne, plus le génome ancestral aura de l'intérêt, mais plus la reconstruction sera rendue difficile par les réarrangements chromosomiques, qui ont lentement remodelé le caryotype et effacé les traces de la polyploïdisation. La première étude, à l'échelle d'un génome complet, des effets d'une duplication complète chez les eucaryotes a été conduite chez les levures [Wolfe et Shields, 1997] et a montré que le réarrangement le plus important dans les génomes dupliqués était une perte massive de gènes qui ramenait progressivement le nombre de gènes au nombre initial (avant la duplication), processus appelé diploïdisation [Otto et Whitton,

2000]. Le terme diploïdisation ne fait ici référence qu'au contenu en gènes et non aux chromosomes, dont le nombre peut rester constant. Le processus de perte des gènes peut supprimer plus de 90% des copies surnuméraires, ce qui explique pourquoi, chez la paramécie, le nombre de gènes n'a été multiplié que par 2 au bout de trois duplications de génome successives (au lieu de $2^3 = 8$) [Aury *et al.*, 2006]. On qualifiera d'«ohnologues» deux gènes issus d'une duplication complète de génomes, et tous deux conservés.

Malgré la perte massive de gènes et de nombreux réarrangements supplémentaires, dont le rythme est d'ailleurs accéléré à cause de la duplication elle-même [Sémon et Wolfe, 2007], une duplication complète laisse deux marques distinctes dans les génomes.

La première (Figure 2.6), comme démontré chez les levures [Dietrich *et al.*, 2004, Kellis *et al.*, 2004, Gordon *et al.*, 2009] et les poissons téléostéens [Jaillon *et al.*, 2004, Kasahara *et al.*, 2007], consiste en un motif d'alternance qui s'observe entre un génome ayant subi la duplication complète et un autre ne l'ayant pas subie. Les gènes des chromosomes de l'espèce non-dupliquée auront leurs orthologues alternativement sur deux chromosomes de l'espèce dupliquée. La raison de ce phénomène, nommé synténie dédoublée, est que la diploïdisation ne conserve qu'une copie de chaque gène, aléatoirement, parmi les deux disponibles sur les deux chromosomes. La conservation de l'ordre des gènes au sein d'un bloc de synténie dédoublée existe tant que les réarrangements n'ont pas eu le temps de remodeler l'ordre des gènes sur les chromosomes. Aux deux extrêmes, les segments paralogues des génomes de levures sont quasiment colinéaires, tandis que ceux des poissons n'arborent aucune conservation d'ordre.

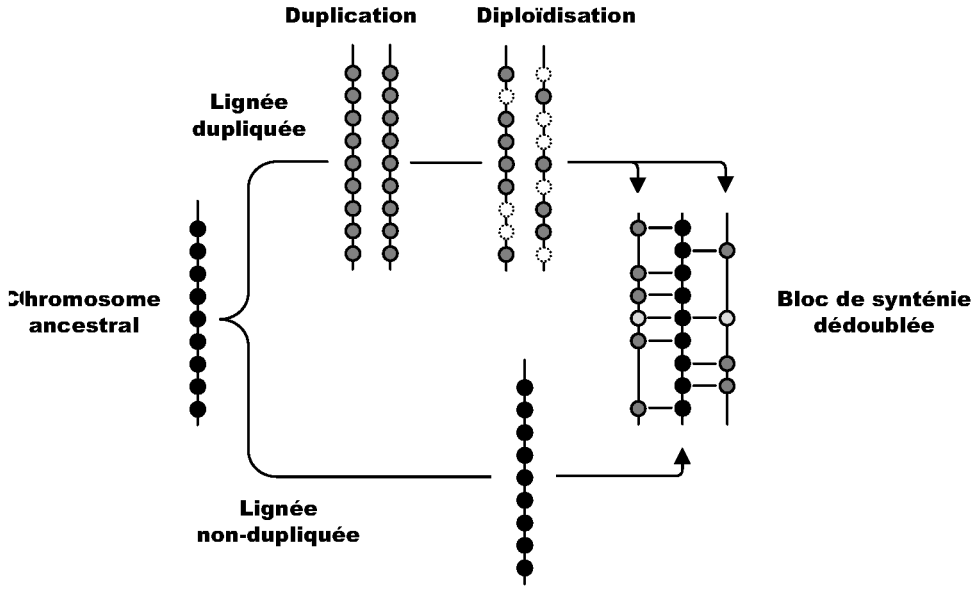


FIGURE 2.6 – Schéma d'un bloc de synténie dédoublée. La comparaison du génome d'une espèce ayant subie une duplication de génome à un autre génome ne l'ayant pas subie révèle un profil d'alternance, due au processus de diploïdisation.

La seconde signature est que, tant que les réarrangements n'ont pas encore trop remodelé le caryotype, et qu'il subsiste suffisamment d'ohnologues, ces derniers doivent être distribués en faisceaux liant les deux copies de chaque chromosome (éventuellement des segments s'il y a eu des réarrangements). Cette technique a principalement été utilisée chez les vertébrés [Nakatani *et al.*, 2007] et chez les plantes [Salse *et al.*, 2009].

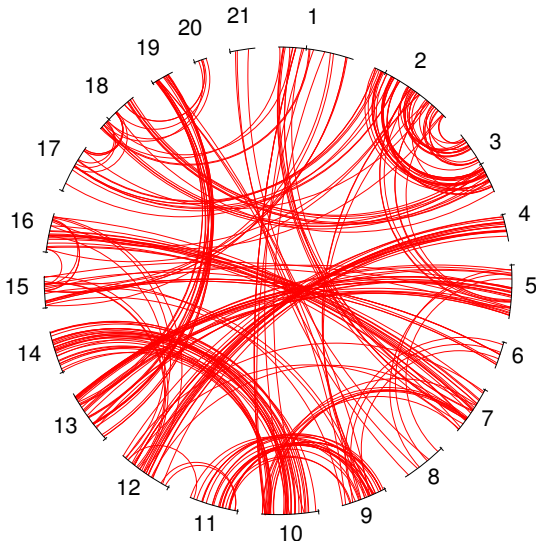


FIGURE 2.7 – Répartition dans le génome du tétraodon des ohnologues dus à la duplication complète de génome des poissons téléostéens, d'après Jaillon *et al.* [2004]. Les paires de chromosomes paralogues sont encore identifiables (9 et 11, 10 et 14, etc) et permettent de définir des chromosomes pré-duplication.

2.4.2 Reconstructions

Chez les poissons téléostéens, la première étude de l'événement de duplication complète qui leur est commun est Naruse *et al.* [2004]. Elle utilisait une carte génétique du medaka (*Oryzias latipes*) dans une analyse de sa synténie avec l'homme et le poisson-zèbre (*Danio rerio*), et a permis de prédire un génome ancestral pré-duplication de 12 chromosomes en suivant un raisonnement similaire à celui de la synténie dédoublée (compte tenu du faible nombre de gènes annotés). Les études suivantes – Jaillon *et al.* [2004] (*Tetraodon nigroviridis*), Woods *et al.* [2005] (poisson-zèbre), Kasahara *et al.* [2007] (medaka) – ont confirmé la majeure partie du génome ancestral, malgré des différences sur le nombre de chromosomes (11, 12, ou 13), qui s'expliquaient principalement par des quantités de données disponibles différentes (séquence des génomes, annotations des gènes, orthologie entre poissons).

Le principe de la synténie dédoublée a été également employé chez les levures, dans un premier temps selon un processus automatique [Kellis *et al.*, 2004]. Les auteurs ont extrait des segments de longueur maximale du génome de la levure non-dupliquée *Kluyveromyces waltii* dont les orthologues sur *Saccharomyces cerevisae* alternaient sur deux chromosomes à la fois. Ces segments ont ensuite été regroupés en chromosomes pré-duplication lorsqu'ils présentaient la même alternance. L'étude montrait que l'ordre des gènes à l'intérieur de chaque paire de chromosomes dupliqués était encore bien conservé, ce qui a permis plus tard une reconstruction manuelle du même ancêtre [Gordon *et al.*, 2009]. Dans cette étude, les auteurs ont comparé les génomes de 5 levures dupliquées et de 6 non-dupliquées, établi et vérifié tous les blocs de synténie dédoublée. Le résultat, un génome de 8 chromosomes, est donc considéré comme la référence, que les méthodes automatiques cherchent à atteindre (comme les reconstructions cytogénétiques chez les mammifères).

Avec la connaissance gagnée à travers le séquençage de plus en plus d'espèces de vertébrés, les frontières de ce qui était autrefois pensé comme impossible en termes de reconstruction de génomes ancestraux ont rapidement été repoussées. L'hypothèse for-

mulée il y a 40 ans par S. Ohno [Ohno *et al.*, 1968] selon laquelle les vertébrés auraient subi deux duplications complètes successives (*2R : 2 rounds of whole genome duplication*), intensivement débattue [Hughes *et al.*, 2001, Larhammar *et al.*, 2002, McLysaght *et al.*, 2002] est désormais soutenue par de nombreuses évidences [Dehal et Boore, 2005, Panopoulou et Poustka, 2005, de Peer *et al.*, 2010]. Une étude récente [Nakatani *et al.*, 2007] utilise cette information a priori pour reconstruire le génome ancestral pré-2R. La méthode identifie tout d'abord dans le génome humain toutes les copies des gènes issues des deux duplications. Lorsqu'on observe les positions de ces ohnologues, des réseaux de régions apparaissent, et chacun de ces réseaux est censé correspondre à un chromosome pré-duplication (de la même manière que sur la Figure 2.7, des paires de régions identifiaient les chromosomes pré-duplication). Au sein de chaque réseau, on est peut identifier 4 groupes (un par chromosome post-duplication) tels que les ohnologues sont répartis uniquement entre groupes, et jamais à l'intérieur d'un même groupe. Malheureusement, les réarrangements subis entre ou depuis les deux duplications brouillent ce signal : les groupes sont fragmentés et mélangés dans le génome humain.

Lorsque des génomes sont très éloignés, et que les réarrangements ont mélangé les gènes sur les chromosomes, le seul signal conservé est la synténie. La méthode décrite dans [Putnam *et al.*, 2008] permet de clusteriser deux génomes à la fois, et chaque cluster formé représente un chromosome de leur dernier ancêtre commun. Au-delà de l'ancêtre commun des vertébrés, l'ordre des gènes n'est plus assez conservé, et la synténie conservée devient (avec les 2R) la seule arme pour inférer un caryotype ancestral. Ainsi, le clustering mené sur les scaffolds d'amphioxus (un céphalochordé) et des segments humains a permis de reconstruire le dernier ancêtre commun des chordés, qui était proche phylogénétiquement du génome ancestral pré-2R.

Enfin, des duplications ont été identifiées et ont mené à des reconstructions de génomes ancestraux, grâce à des méthodes ad-hoc, chez la paramécie (*Paramecium tetraurelia*, Aury *et al.* [2006]) et les plantes (Abrouk *et al.* [2010], Murat *et al.* [2010])

2.5 Discussion

Nous pouvons conclure cet état de l'art en posant les prémisses d'une bonne méthode de reconstruction, dans l'objectif de servir de points temporels de référence à l'étude de l'évolution des génomes. Une méthode idéale pour reconstruire la structure génomique ancestrale (des blocs de séquences, des ordres de gènes, des caryotypes) est censée cibler plusieurs ancêtres à la fois pour fournir une vue dynamique de l'évolution des génomes ancestraux. Pour ceci, il est nécessaire d'avoir un jeu de marqueurs qui inclut une grande fraction des données disponibles sur chaque génome, et qui a une large couverture phylogénétique pour fournir des liens entre les ancêtres. Chez les vertébrés, seules les séquences protéiques peuvent être correctement et facilement alignées sur les espèces les plus éloignées (mammifères vs poissons, Miller *et al.* [2007]), ce qui limite le potentiel des méthodes basées sur des alignements multiples. Une limitation de certains programmes bio-informatiques est la nécessité que les marqueurs soient présents en une et une seule copie dans toutes les branches de l'arbre qui lient les ancêtres ciblés, ignorant de fait les duplications, les délétions, et les gains de marqueurs au fil de l'évolution, qui sont des propriétés incontournables de l'évolution des génomes. Les méthodes de cytogénétique, en plus de la résolution limitée, sont inappropriées à cause de l'impossibilité d'hybrider des espèces au-delà d'une limite de divergence. Les méthodes qui utilisent les adjacences

sont moins sujettes aux erreurs de reconstruction que celles basées sur l'optimisation des réarrangements.

L'apport d'une nouvelle méthode de reconstruction est par nature difficile à mesurer car la réponse exacte (le caryotype) ne sera jamais connue, et toute reconstruction restera une conjecture fondée sur des données disponibles. Cela est d'autant plus vrai que les méthodes de reconstruction sont rarement accompagnées d'une estimation de leurs qualités ou de leurs taux d'erreurs, même s'il faut toutefois noter que les méthodes bio-informatiques sont de plus en plus capables, au minimum, de pointer des zones de la reconstruction moins fiables, et au mieux de calculer théoriquement leur taux d'erreur. Nous pouvons esquisser trois moyens de mesurer la fiabilité d'une reconstruction.

1. La première est la comparaison de plusieurs reconstructions issues de méthodes complètement différentes. C'est par exemple ce qui est naturellement fait pour les méthodes basées sur la séquence des génomes : celles-ci se comparent systématiquement entre elles, et aux résultats des méthodes de cytogénétique (en particulier, ces dernières disposent d'un jeu de données beaucoup plus large). Toutefois, la comparaison ne peut dépasser la limite de résolution des méthodes comparées, par exemple lorsque le détail de la reconstruction dépasse la résolution du Zoo-FISH, ou quand l'ancêtre visé est trop éloigné pour être estimé par cytogénétique (une centaine de millions d'années au maximum, pour les mammifères).
2. Un autre critère de confiance en une reconstruction peut être établi en la comparant à un génome outgroup qui n'a pas servi à la reconstruction. Par exemple, une reconstruction de *Boreoeutheria* en utilisant uniquement des génomes de mammifères peut se comparer au génome du poulet, et est censée en être plus proche (en termes de réarrangements). Ce critère permettrait même de classer différentes reconstructions, la «meilleure« étant la plus proche du génome out-group.
3. Finalement, en particulier pour les méthodes bio-informatiques, une quantification du taux d'erreurs devrait toujours être fournie, qu'elle soit théorique ou empirique. En particulier, il est possible d'estimer les performances d'une méthode de reconstruction par des simulations. Le principe est dans ce cas de définir un génome ancestral virtuel constitué d'ensembles ordonnés (des chromosomes) de gènes ou de blocs d'orthologie. Un logiciel mime alors l'évolution en appliquant des réarrangements selon une phylogénie des espèces et des paramètres d'évolution (comme des taux de réarrangements, des tailles d'inversion, etc), en générant des génomes modernes simulés. Une méthode de reconstruction peut alors utiliser les génomes modernes simulés pour reconstruire un génome ancestral qui est comparé au génome ancestral initial de la simulation. Cette approche estimera quantitativement la précision de la méthode de reconstruction, étant donné les paramètres de la simulation.

Bien que les technologies évoluent vite, le séquençage des génomes ne peut bien entendu se justifier uniquement par la volonté de reconstruire les génomes ancestraux. Les méthodes de cytogénétique, parce qu'elles permettent de reconstruire rapidement le caryotype de nombreux ancêtres, et malgré leur «faible» résolution (relativement aux méthodes qui utilisent la séquence des génomes), ont donc encore des applications potentielles. Cependant, l'accélération du séquençage des génomes, en particulier due aux technologies de «nouvelle génération», permet d'anticiper sur la prépondérance des méthodes basées sur ce type d'information dans les années à venir.

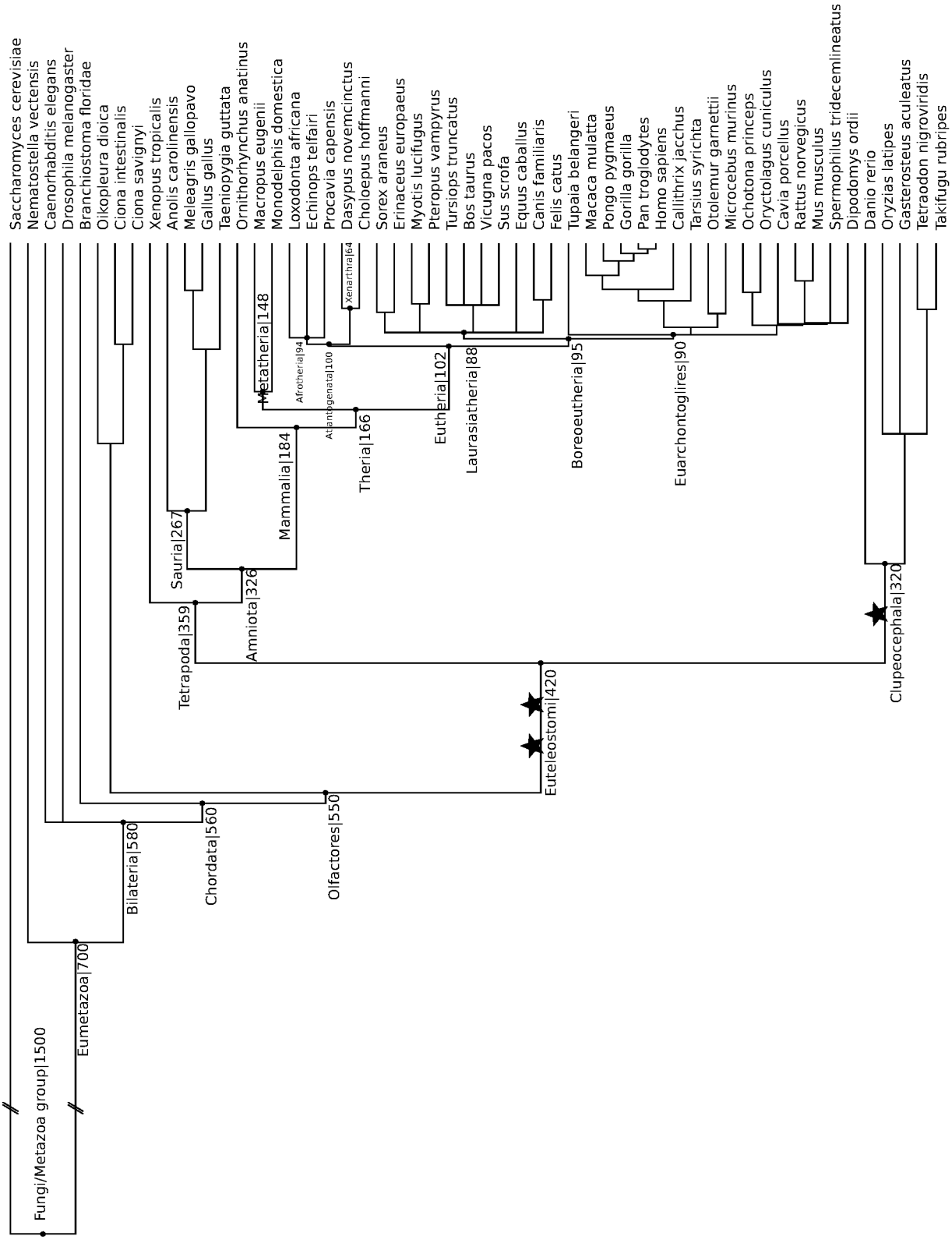


FIGURE 2.8 – Arbre phylogénétique des vertébrés utilisé au cours de cette thèse. Les dates de spéciation indiquées après certains ancêtres sont données en millions d'années et sont fournies par la base de données Ensembl, ou estimées d'après TimeTree [Hedges *et al.*, 2006]. Les étoiles indiquent les duplications connues et complètes du génome. Les branches issues de *Fungi/Metazoa group* sont raccourcies pour l'aisance de la lecture.

Chapitre 3

Structures connues des génomes ancestraux

Sommaire

Dans ce chapitre, nous allons voir les principaux résultats des méthodes de reconstruction présentées dans le chapitre précédent et appliquées aux vertébrés. Bien que de nombreuses reconstructions existent (en particulier chez les mammifères), nous nous concentrerons sur trois ancêtres cibles de beaucoup d'attentions (du point de vue des reconstructions) et / ou clés dans l'évolution des génomes des vertébrés : *Boreoeutheria*, *Teleostei* (avant la duplication complète de génome), et *Chordata* (avant et après les deux duplications complètes du génome). La Figure 3.5 montre sous la même figure l'évolution des caryotypes en partant de la reconstruction de Putnam *et al.* [2008].

3.1 *Boreoeutheria*

De tous les ancêtres, *Boreoeutheria* est celui sur qui le plus d'études se sont concentrées, à cause de son placement idéal dans l'arbre des mammifères : de nombreuses espèces disponibles, et une radiation assez rapide [Blanchette *et al.*, 2004a]. Malgré quelques différences notables entre les reconstructions, et un débat vain sur le bienfait de la cytogénétique par rapport aux méthodes bio-informatiques [Bourque *et al.*, 2006, Froenicke *et al.*, 2006, Robinson *et al.*, 2006], un consensus assez fort existe aujourd'hui sur la structure du caryotype.

La Figure 3.1 est basée sur le modèle établi par les méthodes cytogénétiques [Froenicke, 2005], et confirmé depuis par les méthodes basées sur les séquences des génomes [Ma *et al.*, 2006, Chauve et Tannier, 2008, Kemkemer *et al.*, 2009]. Une modification a néanmoins été apportée : la fusion entre les segments humains $10p$ et $12a - 22a$ n'y est pas reportée car elle est faiblement supportée dans les analyses cytogénétiques (présente uniquement chez les descendants *Afrotheria* et de *Carnivora*), et non retrouvée par les reconstructions de Ma *et al.* [2006] et Chauve et Tannier [2008]. Cette association ne peut en fait être vue par aucune méthode basée sur les génomes séquencés car parmi ces génomes [Ferguson-Smith et Trifonov, 2007], seul l'éléphant (outgroup de *Boreoeutheria*)

possède cette association. Elle ne fait donc pas partie de notre «objectif». Les résultats proposent un génome ancestral composé de 24 chromosomes :

- 10 toujours «intacts» chez l'homme (c'est-à-dire n'ayant subi aucun réarrangement inter-chromosomique) ;
- 2 s'étant cassé 1 fois chacun depuis ;
- 2 paires de chromosomes qui se sont depuis fusionnés ;
- 8 ayant échangé du matériel par translocations.

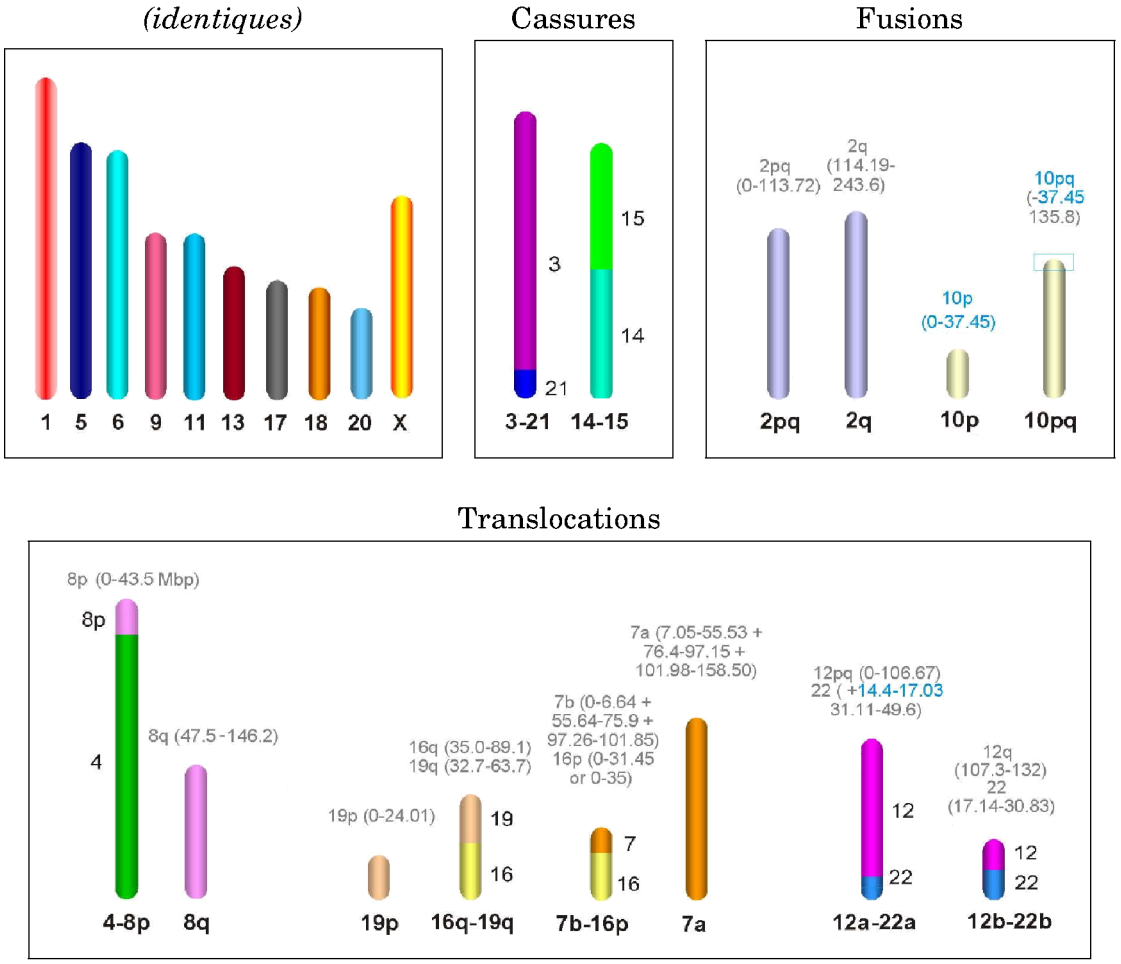


FIGURE 3.1 – Caryotype de l'ancêtre *Boreoeutheria*. Les 24 chromosomes sont représentés selon leurs chromosomes humains homologues. Les positions des segments humains (lorsque le chromosome ancestral s'est fragmenté) sont indiquées en Mb. Le caryotype est celui défini dans [Kemkemer *et al.*, 2009] mais avec le segment ancestral $10p$ placé en tant que chromosome entier.

3.2 *Teleostei* (pré-duplication)

L'ancêtre pré-duplication des poissons téléostéens est aussi le sujet de nombreuses attentions à cause des signaux exceptionnellement forts laissés par la duplication complète du génome, et encore présents après ~350 millions d'années d'évolution. Là encore, des petites variations existent selon les reconstructions, mais ne sauraient masquer le fait qu'elles s'accordent sur la majeure partie du caryotype. Les résultats des deux principales études [Jaillon *et al.*, 2004, Kasahara *et al.*, 2007] concordent presque entièrement,

la différence principale (13 chromosomes au lieu de 12) provenant d'une précision plus importante dans Kasahara *et al.* [2007] (utilisation de medaka en plus de tetraodon). Le génome post-duplication (Figure 3.2) a rapidement subi 8 réarrangements majeurs de chromosomes avant la divergence du poisson-zèbre (qui a subi 15 réarrangements supplémentaires) et des percomorphes (medaka et tétraodon). Medaka a ensuite gardé le même génome depuis (près de ~320 millions d'années), ce qui explique la clarté du signal et l'intérêt d'utiliser son génome pour les reconstructions ancestrales, tandis que tetraodon a subi 4 réarrangements supplémentaires.

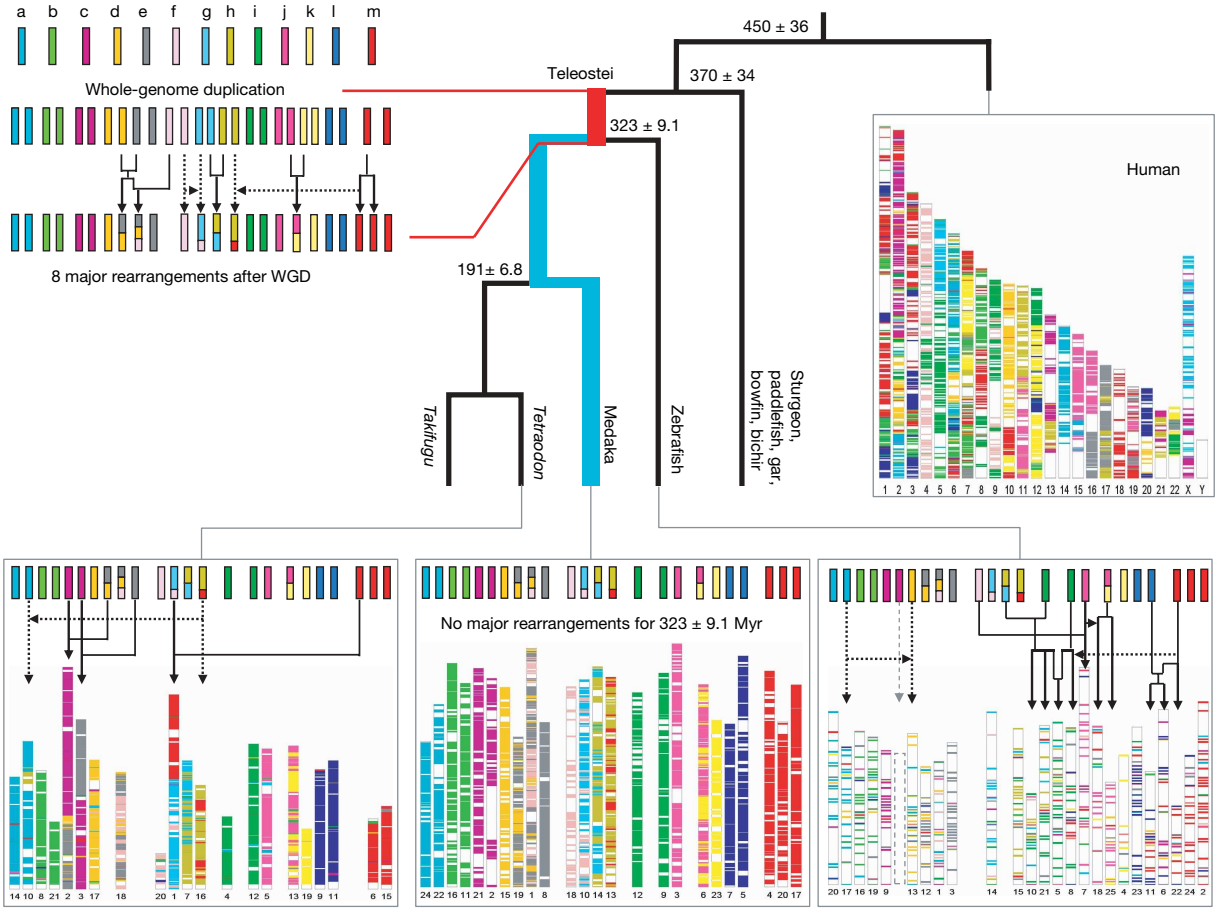


FIGURE 3.2 – Évolution du caryotype chez les téléostes depuis la duplication du génome (figure tirée de Kasahara *et al.* [2007]).

3.3 *Chordata* (pré/post-duplication)

Au delà de l'ancêtre des vertébrés, dans la branche menant de *Chordata* à *Euteleostomi*, se sont produites également deux duplications complètes de génomes, qui permettent, comme chez les téléostes, de reconstruire un caryotype pré-duplication. À une telle distance évolutive (près de 500 millions d'années), les nombreux réarrangements accumulés depuis gênent la découverte d'orthologie entre les gènes et le signal de synténie. Deux reconstructions existent, sur des ancêtres différents :

– celle de l'ancêtre pré-duplication Nakatani *et al.* [2007] (10 chromosomes, Figure 3.3) ;

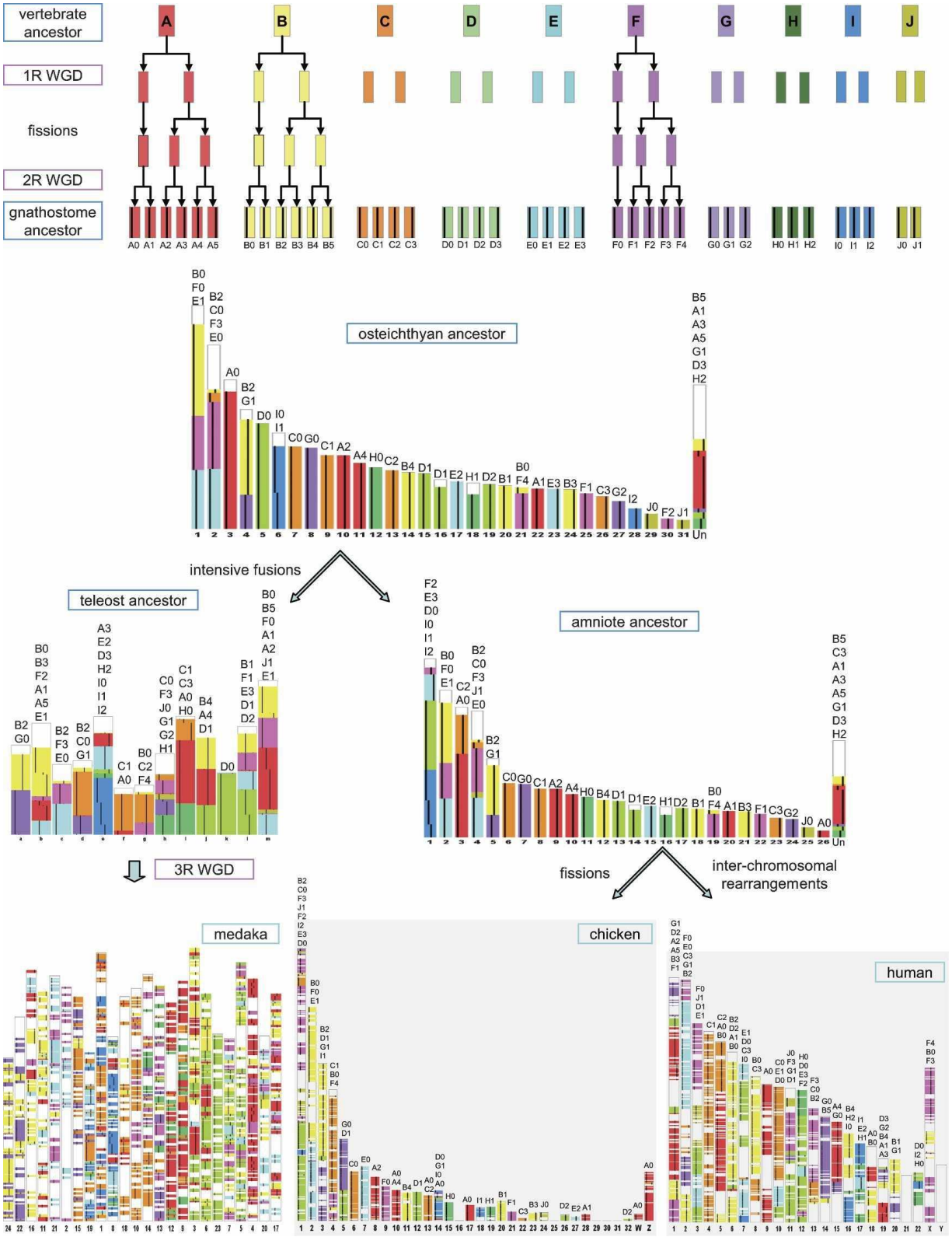


FIGURE 3.3 – Caryotypes pré- et post-*2R* d'après Nakatani *et al.* [2007]

– celle de du dernier ancêtre commun des chordés [Putnam *et al.*, 2007, 2008, Srivastava *et al.*, 2008] (17 chromosomes, Figure 3.4).

Des similarités (visibles sur la Figure 3.5) apparaissent, comme entre le chromosome 1 pré-2R, et les chromosomes 1 et 2 de chordés, ou les 5 et 8 pré-2R et le 16 de chordés, mais nous ne les avons pas toutes énumérées, ni cherché à établir si les différences de reconstruction étaient justifiées phylogénétiquement.

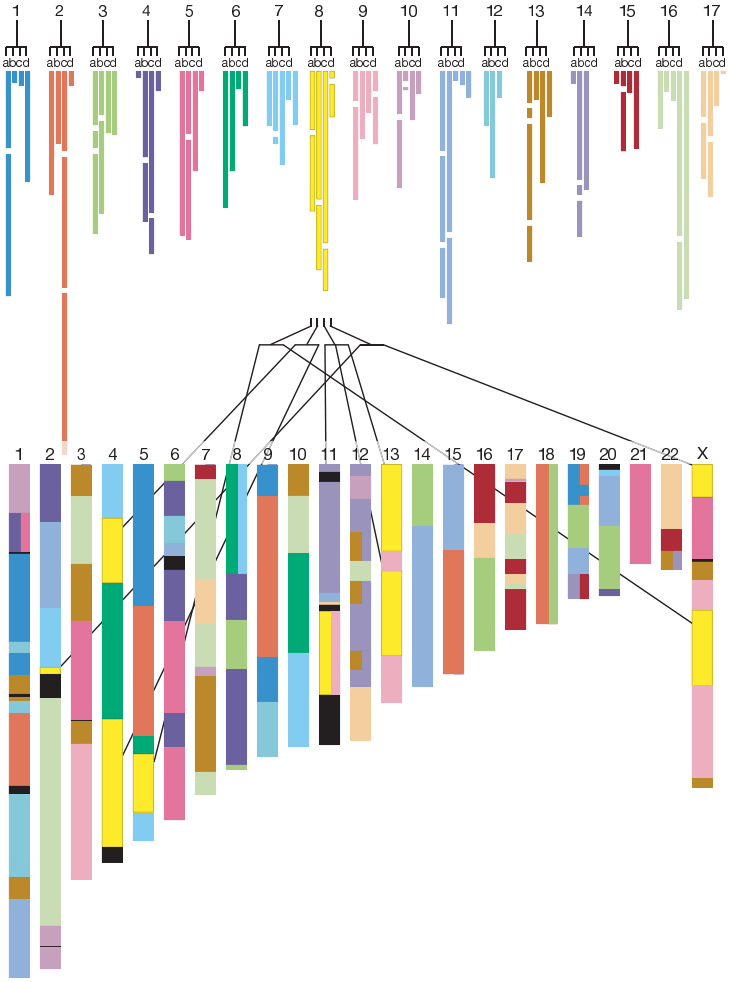


FIGURE 3.4 – Caryotype pré-*2R* d'après Putnam *et al.* [2008]

3.4 Résumé

La Figure 3.5 rassemble les caryotypes précédemment vus sous deux codes couleurs. Le premier utilise le génome humain en référence et permet de remonter jusqu'avant les *2R* selon les deux modèles existants et jusqu'avant la duplication complète chez les poissons. À cet instant, un nouveau code couleur basé sur les 13 chromosomes ancestraux prend le relais. Chez les mammifères, la représentation est enrichie des caryotypes de la souris et du gibbon (génome très réarrangé), et chez les poissons, de tétraodon et de medaka.

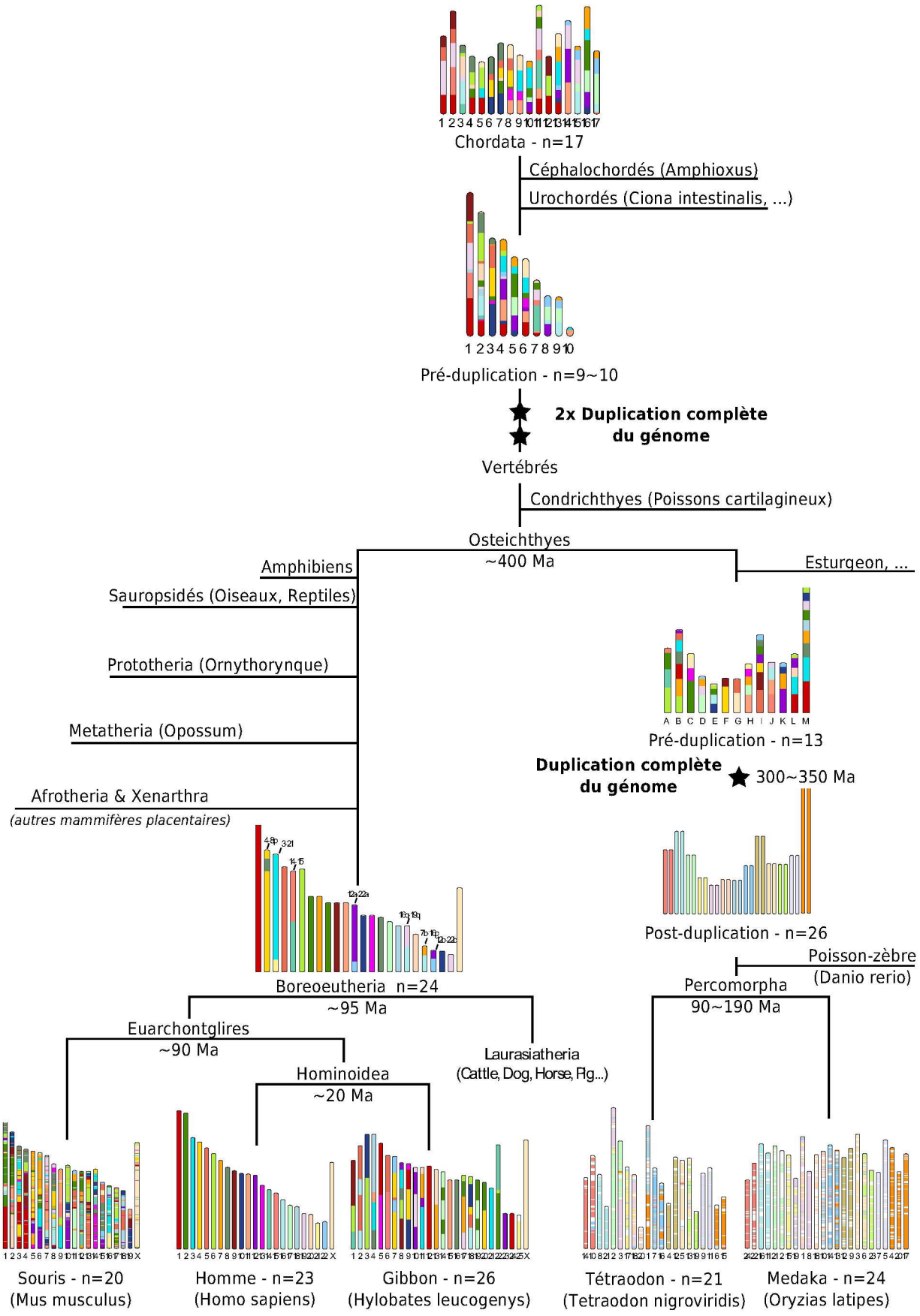


FIGURE 3.5 – Évolution du caryotype chez les vertébrés, tirée de Muffato et Crollius [2008], et mise à jour pour inclure les reconstructions des 2R.

Deuxième partie

Méthodes informatiques

Sommaire

Cette courte partie est consacrée à la présentation de problèmes généraux informatiques et mathématiques, dont les solutions seront utilisées à de nombreuses reprises dans le cadre des développements d'AGORA pour la reconstruction de génomes ancestraux.

Les deux premiers chapitres présentent des problèmes algorithmiques majeurs (le problème du voyageur de commerce, chapitre 4, et le regroupement hiérarchique, chapitre 5), bien étudiés, et pour lesquels des solutions existent. Dans le cadre de notre étude, ils n'ont donc pas nécessité le développement de nouvelles méthodes, mais d'un interfaçage avec des programmes de résolution existants.

Le chapitre 6 présente une méthode de calcul de moyenne qui tient compte d'un arbre phylogénétique. En effet, il sera fréquent de devoir calculer une valeur pour différentes espèces modernes (par exemple, un score), et de vouloir une unique valeur qui désignerait le score de l'ancêtre commun de ces espèces. En pratique, cette méthode a uniquement nécessité l'écriture d'équations, la résolution en elle-même étant déléguée à des librairies spécialisées.

Chapitre 4

Voyageur de commerce

Le problème du voyageur de commerce est un problème d'optimisation aux applications extrêmement concrètes. Comment un représentant de commerce peut-il visiter tous ses clients (une seule fois chacun), en minimisant son temps de déplacement ?

Formellement, le problème est modélisé sous la forme d'un graphe non-dirigé, connexe, et pondéré, dans lequel les sommets représentent les clients à visiter, les arêtes, les routes qu'il peut emprunter et les poids des arêtes, les distances des trajets entre chaque ville. La question est de trouver un chemin (ou un circuit si on veut revenir au point de départ) hamiltonien (c'est-à-dire qui passe une seule fois par tous les sommets) qui minimise la somme des poids des arêtes (appelée score).

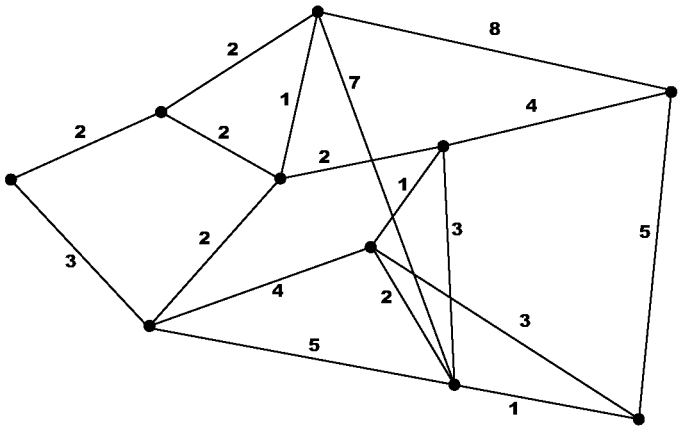
Il n'existe pas de méthode de résolution «rapide» de ce problème, car il est rangé dans la classe de complexité NP-complet. C'est-à-dire que les méthodes de résolution actuelles ne garantissent pas la découverte d'une solution exacte en un temps polynomial ($O(n^p)$, n représentant le nombre de villes, et p une constante). La complexité naïve de la résolution est $O(n!)$ et correspond au test de tous les chemins possibles. Malgré l'évolution des techniques de résolution, la complexité actuelle reste de l'ordre de $O(n^p 2^n)$ [Bellman, 1962] avec p entier. Cela signifie (si on se concentre sur le terme 2^n) que l'ajout d'une ville au problème multiplie son temps de résolution par 2 (dans le pire des cas).

L'implémentation de référence de la résolution du voyageur de commerce est le programme *concorde* [1]. Il a été écrit par David Applegate, Robert E. Bixby, Vašek Chvátal, et William J. Cook (auteurs du livre Applegate *et al.* [2006]).

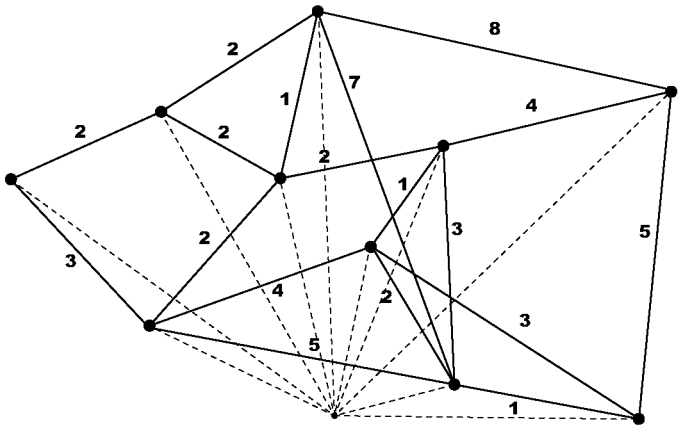
Par défaut, *concorde* cherche un circuit, c'est-à-dire que la solution permet de revenir de manière optimale en son point de départ. En pratique, nous chercherons à utiliser *concorde* sur des données génomiques de vertébrés pour définir des chemins qui seront assimilés à des chromosomes. Or ceux-ci (chez les vertébrés) sont linéaires, et non circulaires. Pour récupérer un chemin (et non un circuit) optimisé, il suffit de créer un nœud supplémentaire lié à tous les autres nœuds. Ainsi, toute solution sera obligée de passer par ce nœud. De plus, si ce nœud est lié à tous les autres avec exactement le même poids, on ne privilégiera aucun point d'insertion. Il suffit ensuite de couper le circuit solution au niveau de ce nœud pour retrouver un chemin.

1. http://www.tsp.gatech.edu/concorde. Le programme est un exécutable autonome, libre d'utilisation dans le milieu de la recherche académique, et demande à ce que le graphe soit donné en entrée sous un format textuel simple d'utilisation.

A :



B :



C :

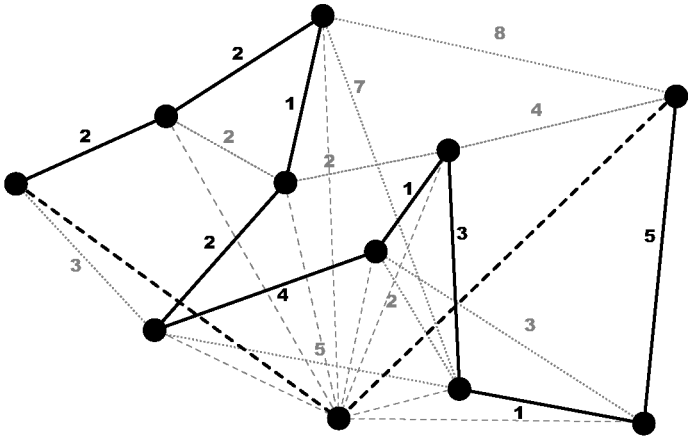


FIGURE 4.1 – Exemple d'application du voyageur de commerce. **A :** Graphe initial. **B :** Ajout d'un nœud supplémentaire lié à tous les autres (avec un poids constant). **C :** La solution optimale au problème du voyageur de commerce est un circuit passant par le nœud supplémentaire, et donc un chemin passant par les nœuds du graphe initial. Ici, le score de la solution vaut 18.

Chapitre 5

Regroupement hiérarchique

Le regroupement hiérarchique, aussi appelé *clustering*, a pour but d'ordonner des objets par similarité. Le résultat est un dendrogramme, c'est-à-dire une structure d'arbre dans lequel les objets similaires sont proches. Le dendrogramme contient la liste des fusions à faire entre les objets, pour les regrouper en classes d'objets similaires. Dans l'exemple de la Figure 5.1, le dendrogramme en (b) est le résultat du clustering des 36 sommets du graphe de (a), selon un critère de proximité.

À chaque dendrogramme est associé un ensemble naturel de partitions [1]. La partition de départ (la racine du dendrogramme) correspond à l'ensemble de tous les objets. Au fur et à mesure du parcours du dendrogramme, on peut diviser, à chaque nœud du dendrogramme, un des ensembles de la partition en deux, jusqu'au moment où chaque objet sera seul dans son ensemble. Toujours dans l'exemple de la Figure 5.1, le dendrogramme en (c) (équivalent, pour ce qui est de la structure de l'arbre, au dendrogramme de (b)) montre deux niveaux de coupure en pointillés bleus. Ces deux niveaux correspondent à deux partitions du graphe de (a).

Le programme *walktrap* [Pons et Latapy, 2005] utilise des «marches aléatoires» pour mesurer la similarité des objets à clusteriser, une fois ceux-ci placés dans un graphe. Partant d'un nœud donné du graphe, *walktrap* établit la liste de tous les chemins possibles d'une longueur donnée en paramètre (en général, 5 nœuds), éventuellement en passant deux fois par le même nœud, en associant à chaque chemin une probabilité : le produit des probabilités de choix des arêtes à chaque nœud traversé. Pour un nœud donné, si le graphe n'est pas pondéré, la probabilité de choix d'arête est uniforme ($1/n_a$ si il y a n_a arêtes sortantes), sinon, la probabilité de la i-ème arête est $\omega_i / \sum_j \omega_j$ où ω_j représente le poids de la j-ème arête (cette mesure est insensible à une multiplication de toutes les arêtes par un même facteur multiplicatif). Deux nœuds seront considérés comme d'autant plus proches qu'ils proposent d'aller aux mêmes nœuds du graphe avec des probabilités similaires. À partir de cette mesure de proximité des nœuds, *walktrap* fusionne au fur et à mesure les nœuds, en mettant à jour les mesures de proximités, et construit ainsi un dendrogramme.

Un dendrogramme, à la base, ne contient que les informations de similarité des objets clusterisés. *walktrap* propose un algorithme pour ajouter une «hauteur» aux nœuds d'un dendrogramme pour qu'ils soient au même niveau lorsqu'ils désignent des classes d'objets équivalentes (c'est la différence entre les dendrogrammes (b) et (c) de la Figure 5.1). De plus, *walktrap* renvoie un indice de pertinence des partitions que l'on pourrait tirer

1. Une partition d'un ensemble E est un ensemble d'ensembles $\{S_i\}_i$ tous non vides et disjoints, tels que leur union fait E.

du dendrogramme (courbe $R(\alpha)$ à droite de la Figure 5.1.c). *walktrap* peut donc proposer pour n'importe quel graphe une partition, sans que l'on doive spécifier le nombre de clusters. Les deux partitions en bleu correspondent ainsi aux deux partitions optimales (les maxima locaux de $R(\alpha)$) du graphe de (a).

On se servira de *walktrap* [2] dès que l'on a besoin d'un regroupement hiérarchique ou d'une partition d'un ensemble, en particulier si on ne connait pas à l'avance le nombre de clusters de la solution. Il faudra pour ceci définir le graphe qui contient les objets à clusteriser, ainsi que les poids sur les arêtes si on travaille avec des arêtes pondérées.

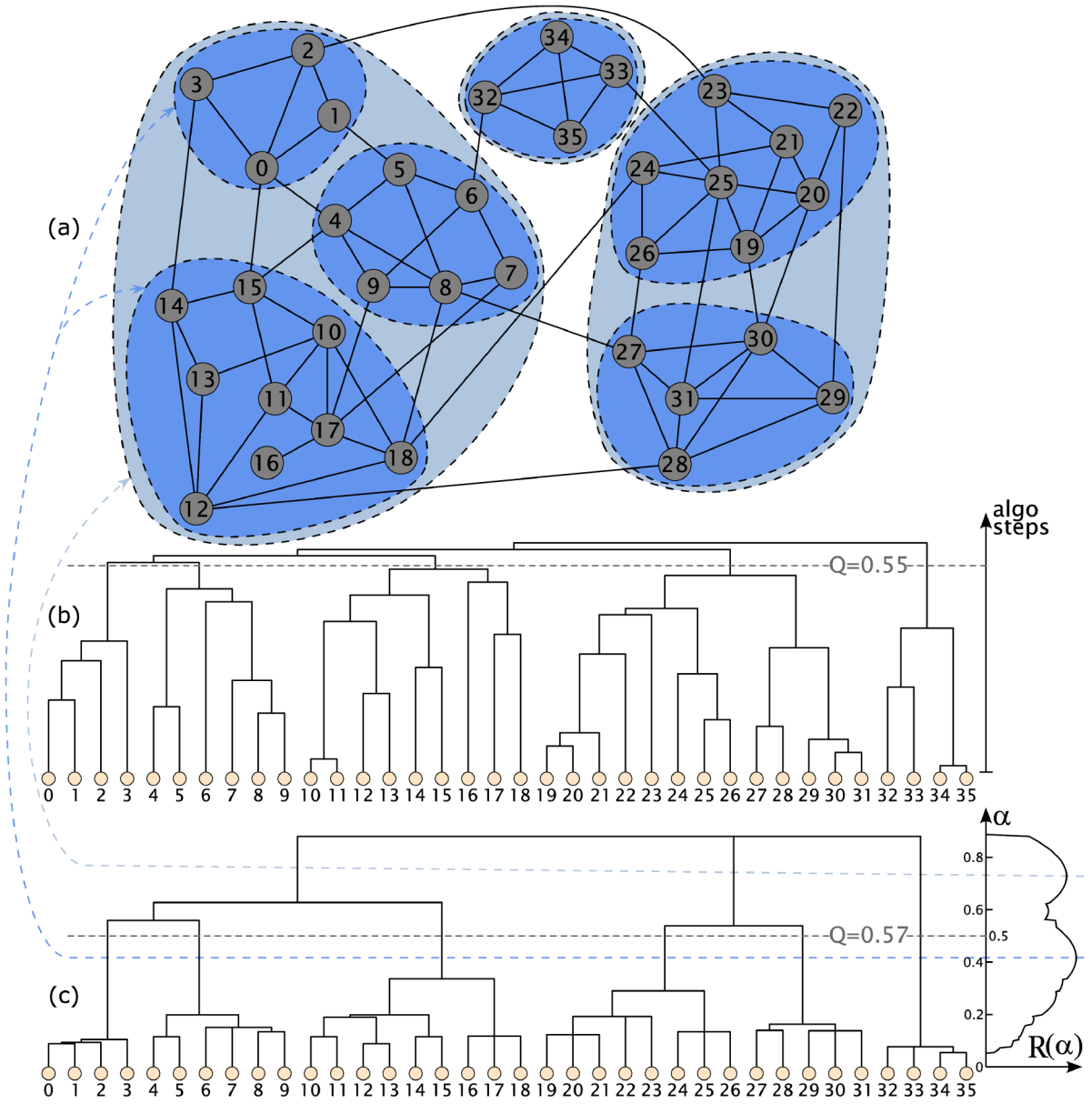


FIGURE 5.1 – Schéma du regroupement effectué par *walktrap* et de la sélection automatique de la meilleure partition. **A :** Graphe non-pondéré contenant 36 nœuds. **B :** Dendrogramme issu du regroupement par *walktrap*. **C :** Dendrogramme mis à l'échelle et courbe de pertinence des partitions du graphe.

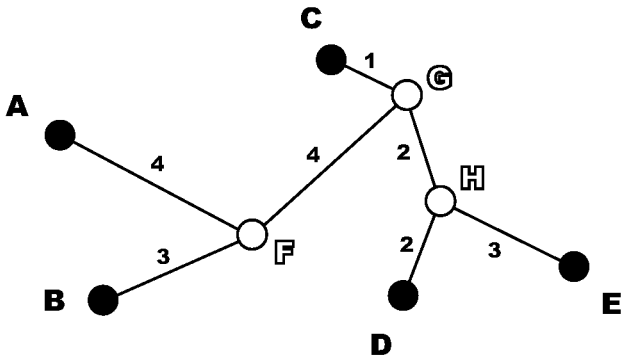
2. Distribué en tant que logiciel libre, et utilisant des données dans un format texte simple.

Chapitre 6

Interpolation linéaire d'une variable sur un arbre

Sommaire

Dans ce travail, il a été nécessaire à plusieurs reprises de pouvoir calculer la valeur moyenne d'une variable V chez des ancêtres, connaissant les valeurs de cette variable chez les espèces existantes. Cela est nécessaire dès que l'on a besoin d'une valeur unique, ancestrale, condensant les valeurs disponibles pour les espèces modernes (typiquement un score moyen ou une probabilité moyenne). Il est risqué d'utiliser la moyenne des valeurs des espèces modernes car cela favoriserait les branches de l'arbre qui contiennent beaucoup de génomes. Il est donc nécessaire de pondérer chaque espèce en fonction de la structure de l'arbre, et de définir ainsi une moyenne «phylogénétique». Cette méthode ne fait aucune supposition sur le mode d'évolution de V, interdisant en particulier de l'utiliser pour prédire des valeurs ayant un sens biologique (comme le taux de GC). Dans ce chapitre, nous allons définir une méthode pour calculer les valeurs de V chez les ancêtres et exhiber quelques-unes de ses propriétés.



$$\text{avec} \begin{cases} V_{\mathbf{A}} = 2,4 \\ V_{\mathbf{B}} = 1 \\ V_{\mathbf{C}} = 9,6 \\ V_{\mathbf{D}} = 5,8 \\ V_{\mathbf{E}} = 7,1 \end{cases}$$

FIGURE 6.1 – Exemple d'arbre phylogénétique avec valeurs associées aux espèces modernes. Dans l'arbre, les espèces existantes sont représentées par les lettres pleines, les espèces ancestrales par les lettres blanches, et les nombres indiquent les longueurs des branches. La variable V indique des valeurs associées à chaque espèce existante. L'objet de ce chapitre est de définir une méthode de calcul de $V_{\mathbb{F}}$, $V_{\mathbb{G}}$, et $V_{\mathbb{H}}$.

6.1 Résolution formelle

Dans un modèle linéaire, on utilise ici de manière récursive la notion de moyenne pondérée. Chaque nœud interne est lié à au moins 2 autres nœuds, et recevra la moyenne pondérée en fonction des longueurs des branches des valeurs associées à chacun de ces nœuds. Le problème se résume donc à la résolution d'un système d'équations linéaires. Pour deux nœuds i et j, on introduit $\omega_{i,j}$ qui vaut l'inverse du poids de l'arête liant ces nœuds si elle existe, et 0 sinon. Le système s'écrit de la manière suivante :

$$S = \begin{cases} x_i = V_i & \text{pour les nœuds terminaux ;} \\ x_i \sum_j \omega_{i,j} - \sum_j \omega_{i,j} x_j = 0 & \text{pour les nœuds internes.} \end{cases} \tag{6.1}$$

On notera que le nœud j apportant la plus grande contribution à un nœud i est celui le plus proche dans l'arbre.

Lemme 1. *Le système S admet une solution, et cette solution est unique.*

Démonstration. En notant $Ax = b$ l'équivalent matriciel du système S, la question est équivalente à l'inversibilité de A. Il faut donc montrer que toute solution x du système réduit (l'équivalent du système général avec $V = 0$) $Ax = 0$ est nulle. Supposons qu'il existe une telle solution x (il en existe au moins une : $x = 0$). Nous allons montrer que tous les $|x_i|$ sont bornés par 0 (et que donc $x = 0$) en utilisant la propriété de connexité d'un arbre phylogénétique, et le fait qu'au moins un des x_i est fixé à V_i.

Choisissons i_0 tel que $|x_{i_0}| = \max_i |x_i|$.

- Si i_0 est un nœud pour lequel x_{i_0} est fixé à V_{i_0}, alors $x_{i_0} = 0$ car on est dans le système réduit $V = 0$, ce qui clôt la démonstration.
- Sinon, on a l'équation $x_{i_0} \sum_j \omega_{i_0,j} = \sum_j \omega_{i_0,j} x_j$ (on note cette quantité S_{i_0}). Par passage à la valeur absolue : $|S_{i_0}| \leqslant \sum_j \omega_{i_0,j} |x_j| \leqslant \sum_j \omega_{i_0,j} |x_{i_0}| = |S_{i_0}|$. Les inégalités doivent donc être des égalités, et pour les j tel que $\omega_{i_0,j} \neq 0$, on doit avoir $|x_j| = |x_{i_0}|$. Autrement dit, tous les voisins de i_0 dans l'arbre partagent la même valeur pour x. Dès lors, le choix de i_0 peut se faire pour un de ses voisins, et en appliquant le même raisonnement de proche en proche dans l'arbre, on arrive par propriété de connexité sur un nœud pour lequel V est fixé (cas précédent), ce qui permet de conclure que $x_{i_0} = 0$, et que donc $x = 0$.

□

Lemme 2. *Les valeurs ancestrales sont soit toutes égales entre elles, soit strictement comprises entre les extrema des valeurs imposées.*

Démonstration. On sait que la moyenne pondérée par des coefficients compris entre 0 et 1 de p nombres est soit égale à tous ces nombres (s'ils sont tous égaux entre eux), soit différente de tous ces nombres (si au moins deux sont différents), et dans ce cas, comprise entre leurs extrema. Appliqué à un nœud interne i_0, on a dans le premier cas l'égalité locale de x (sur tous les voisins). On peut donc utiliser le même raisonnement successivement sur les voisins pour prouver la constance de x sur tout l'arbre, ce qui est un cas particulier du résultat voulu. Sinon, on peut trouver un nœud i_1 voisin tel que $x_{i_0} < x_{i_1}$. En l'appliquant à i_1 et à ses voisins, on peut construire une suite de nœuds tous différents $(i_j)_j$ telle que la suite $(x_{i_j})_j$ soit strictement croissante. Cette suite s'arrête nécessairement sur un nœud terminal i_p censé vérifier $V_{i_p} = x_{i_p} > x_{i_0}$, ce qui montre que x est borné par la valeur maximale de V. Le même raisonnement tient pour montrer que x est également borné par la valeur minimale de V en exhibant une suite décroissante. □

Lemme 3. *La valeur en un nœud ancestral est une combinaison linéaire des valeurs imposées, dont tous les coefficients sont compris entre 0 et 1, et dont la somme vaut 1.*

Démonstration. Puisque le système S est linéaire, la forme générale des solutions est $x_i = \sum_j \alpha_{i,j} V_j$. Pour un i_0 donné, si on pose $V_i = \begin{cases} 1 & \text{si } i = i_0 \\ 0 & \text{si } i \neq i_0 \end{cases}$. on a $x_i = \alpha_{i,i_0} \in [0,1]$ par application du résultat précédent. Tous les coefficients sont donc entre 0 et 1. D'autre part, si $\forall i,\ V_i = 1$, alors, $\forall i,\ x_i = \sum_j \alpha_{i,j} V_j = \sum_j \alpha_{i,j} \in [1,1]$. La somme des coefficients vaut donc 1. □

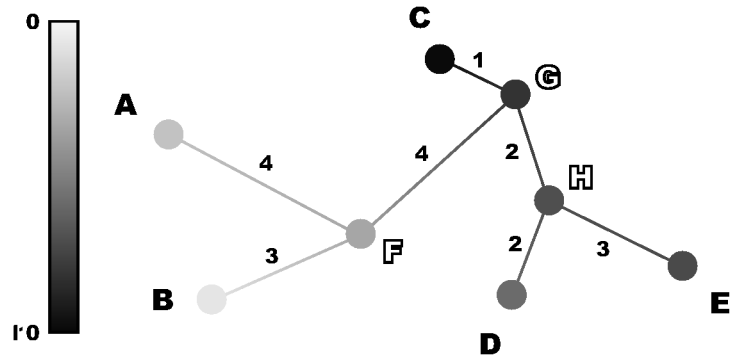
Ces trois lemmes prouvent que la résolution du système (Équation 6.1) est équivalente à l'attribution de coefficients aux nœuds terminaux en fonction de la topologie de l'arbre. Nous possédons donc désormais une méthode pour calculer une moyenne pondérée («phylogénétique») de valeurs «modernes».

6.2 Exemple de résolution

Le système associé à la Figure 6.1 est composé de deux sous-systèmes. Le premier concerne les espèces terminales et reprend les valeurs qui y sont imposées. Le deuxième est l'écriture des formules de moyenne pour tous les nœuds ancestraux, en fonction de la topologie de l'arbre est des longueurs des branches. La résolution du système linéaire de cette taille (moins de 100 nœuds chez les vertébrés) est, de nos jours, immédiate (ici, la librairie utilisée est LAPACK, via numpy).

$$\begin{cases} x_{\mathbf{A}} = V_{\mathbf{A}} \\ x_{\mathbf{B}} = V_{\mathbf{B}} \\ x_{\mathbf{C}} = V_{\mathbf{C}} \\ x_{\mathbf{D}} = V_{\mathbf{D}} \\ x_{\mathbf{E}} = V_{\mathbf{E}} \\ \left(\frac{1}{4}+\frac{1}{3}+\frac{1}{4}\right) x_{\mathbb{F}} = \frac{x_{\mathbf{A}}}{4} + \frac{x_{\mathbf{B}}}{3} + \frac{x_{\mathbb{G}}}{4} \\ \left(\frac{1}{4}+\frac{1}{1}+\frac{1}{2}\right) x_{\mathbb{G}} = \frac{x_{\mathbb{F}}}{4} + \frac{x_{\mathbf{C}}}{1} + \frac{x_{\mathbb{H}}}{2} \\ \left(\frac{1}{2}+\frac{1}{2}+\frac{1}{3}\right) x_{\mathbb{H}} = \frac{x_{\mathbb{G}}}{2} + \frac{x_{\mathbf{D}}}{2} + \frac{x_{\mathbf{E}}}{3} \end{cases}$$

$$\begin{cases} x_{\mathbb{F}} = (150V_{\mathbf{A}} + 200V_{\mathbf{B}} + 96V_{\mathbf{C}} + 18V_{\mathbf{D}} + 12V_{\mathbf{E}})/476 \\ x_{\mathbb{G}} = (24V_{\mathbf{A}} + 32V_{\mathbf{B}} + 320V_{\mathbf{C}} + 60V_{\mathbf{D}} + 40V_{\mathbf{E}})/476 \\ x_{\mathbb{H}} = (9V_{\mathbf{A}} + 12V_{\mathbf{B}} + 120V_{\mathbf{C}} + 201V_{\mathbf{D}} + 134V_{\mathbf{E}})/476 \end{cases}$$



et $\begin{cases} x_{\mathbf{A}} = 2{,}4 \\ x_{\mathbf{B}} = 1 \\ x_{\mathbb{F}} = 3{,}51 \\ x_{\mathbf{C}} = 9{,}6 \\ x_{\mathbb{G}} = 7{,}97 \\ x_{\mathbb{H}} = 6{,}94 \\ x_{\mathbf{D}} = 5{,}8 \\ x_{\mathbf{E}} = 7{,}1 \end{cases}$

FIGURE 6.2 – Résultats de l'interpolation sur l'arbre de la Figure 6.1. Les facteurs sont mis au même dénominateur, 476, pour faciliter la compréhension de la combinaison linéaire. Les valeurs sont dessinées selon un dégradé de gris sur les nœuds et les branches.

Pour la résolution de l'exemple de la Figure 6.1 (Figure 6.2), on remarque dans l'expression de $x_{\mathbb{F}}$ que $150+200+96+18+12=476$ et que le nœud tient sa valeur principalement de **B** et **A** (davantage **B** que **A** car **B** est plus proche dans l'arbre), puis un peu de **C**, et très légèrement de **D** et **E**.

6.3 Remarques

La méthode reste valable si V n'est pas défini pour tous les nœuds terminaux. En effet, l'hypothèse minimale est que le graphe soit connexe, et qu'il y ait au moins une valeur V_i de fixée. En particulier, sur les zones de l'arbre qui ne sont rattachées que par un seul nœud au sous-arbre qui contient les valeurs fixées, la solution x aura la valeur de ce nœud. Ainsi, dans l'arbre de la Figure 6.2, si $V_{\mathbf{A}}$ et $V_{\mathbf{B}}$ ne sont pas définis, l'interpolation pour les nœuds **A**, **B**, et $\mathbb{F}$ donnera la même valeur que pour $\mathbb{G}$.

En notant par 0 ou 1 l'absence ou la présence d'un caractère dans des espèces existantes, cette méthode de calcul renvoie des probabilités de présence de ce caractère dans les espèces ancestrales, sans prendre de décision, contrairement aux méthodes d'inférence de caractère. La méthode est néanmoins utile si on n'a pas besoin de décider de la présence (probabilité $> 0,5$), mais juste d'avoir une indication. D'autre part, la solution n'est pas nécessairement parcimonieuse, car la méthode ne cherche pas à optimiser des gains / pertes ou des hausses / baisses. Pour tout calcul sur des variables génomiques pour lesquelles un mode d'évolution est connu (taux de GC, séquence ancestrale), on préfèrera des techniques de maximum de vraisemblance ou de combinatoire (minimisation du nombre de pertes ou de duplications dans l'arbre).

Troisième partie

Développement d'outils bio-informatiques pour la reconstruction de génomes ancestraux

Sommaire

Cette partie contient l'essentiel des développements effectués dans le cours de cette thèse pour permettre la reconstruction des génomes ancestraux. Le processus de reconstruction suit d'une certaine manière le séquençage par shotgun du génome d'une nouvelle espèce. L'ensemble des algorithmes s'appelle AGORA, pour *Algorithms for Gene Order Reconstruction in Ancestors*.

Nous verrons tout d'abord (chapitre 7) comment définir les gènes que possédait chaque ancêtre reconstruit, ce qui va permettre de définir les relations d'orthlogie et de paralogie utiles pour comparer des génomes.

Ensuite (chapitre 8), nous définirons deux méthodes qui permettent de comparer des paires de génomes. Leur résultat est le liant qui va permettre réellement la reconstruction, sous le principe de parcimonie : ce qui est commun à deux génomes est hérité de leur dernier ancêtre commun. Les adjacences conservées entre les paires d'espèces comparées vont alimenter le processus de reconstruction comme les lectures le font pour l'assemblage d'un génome.

La combinaison de ces données *pairwise* est régie par les algorithmes décrits théoriquement dans le chapitre 9 (majoritairement des parcours de graphe), et appliqués dans le chapitre 10. Les méthodes de reconstruction AGORA permettent successivement de définir des contigs ancestraux, puis des scaffolds grâce à différents protocoles qui peuvent se combiner, ce qui procure à AGORA une haute adaptabilité.

Les duplications complètes de génomes offrent des leviers particuliers pour reconstruire directement l'ancêtre qui précède (ou suit) juste la duplication, et nécessitent des méthodes particulières présentées chapitre 11.

Chapitre 7

Définition des gènes ancestraux

Sommaire

Ce chapitre présente les données, brutes et formatées, utilisées par AGORA. Nous décrivons tout d'abord la source de ces données (Ensembl), puis quelques manipulations basiques (interprétation des positions des gènes, formatage, ajout d'espèces supplémentaires) avant de rentrer dans les algorithmes d'AGORA qui définissent le contenu en gènes des génomes ancestraux (éventuellement filtré) et les relations d'homologie entre les gènes.

7.1 Origine des données

7.1.1 Ensembl / EnsemblGenomes

AGORA nécessite deux types d'information pour reconstruire les génomes ancestraux : la position des gènes dans les génomes modernes, et les relations phylogénétiques entre ces gènes. De nombreuses ressources existent pour fournir ces données, et en particulier, la base de données Ensembl [1] [Flicek *et al.*, 2010] met à disposition de la communauté scientifique ces informations de manière intégrée et exhaustive. En effet, Ensembl dispose de processus automatiques pour annoter les gènes (éventuellement corrigés manuellement pour certaines espèces comme l'homme, Wilming *et al.* [2008]), et construire les phylogénies des gènes à partir d'alignements multiples. Ensembl met à jour ses données tous les deux mois environ, et dans la version 57 (utilisée au long de ce manuscrit) sont

1. http://www.ensembl.org

présentes 46 espèces de vertébrés et 5 espèces modèles de non-vertébrés. Les données d'Ensembl sont accessibles via un site internet, via Biomart (interface d'interrogation), via une interface de programmation en Perl, ou enfin en téléchargeant les données formatées afin d'être chargées dans les bases de données.

De plus, Ensembl, initialement focalisé sur les vertébrés, a initié un réseau de navigateurs et de bases de données nommé EnsemblGenomes [2] [Kersey *et al.*, 2010]. Il s'agit de versions d'Ensembl dédiées aux bactéries, aux protistes, aux plantes, aux champignons et aux métazoaires (vertébrés exclus). Ces navigateurs disposent exactement des mêmes outils qu'Ensembl et leurs données sont disponibles dans les mêmes formats. Les méthodes AGORA sont donc facilement applicables sur d'autres familles d'organismes que les vertébrés (voir perspectives).

7.1.2 Compara / TreeBest

Ensembl propose des reconstructions phylogénétiques exhaustives effectuées à partir de tous les gènes annotés de toutes les espèces de la base de données. Le processus d'inférence de ces phylogénies s'appelle Compara [Vilella *et al.*, 2009] et se compose des étapes suivantes :

1. aligner toutes les séquences protéiques entre elles avec *WUBlastp* et *Smith-Waterman* ;
2. clusteriser les gènes en familles, grâce au score de *WUBlastp*, avec *hcluster_g* ;
3. pour chaque cluster, aligner les séquences protéiques avec un méta-aligneur multiple (*MCoffee2*, Wallace *et al.* [2006]) ;
4. pour chaque cluster, construire un arbre phylogénétique réconcilié avec l'arbre des espèces avec *TreeBeST* (arbre consensus de 5 méthodes différentes).

Ces données sont réconciliées avec la phylogénie des espèces. Ainsi, pour chaque famille de gènes, son histoire évolutive (événements de spéciation, duplication, ou de perte) est connue et peut être lue dans les arbres phylogénétiques. L'exemple de la Figure 7.2 montre l'arbre phylogénétique de quelques espèces d'amniotes, ainsi qu'un arbre phylogénétique réconcilié pour une famille fictive de gènes amniotes. Le gène ancestral était présent en une seule copie à l'origine (chez *Amniota*), a été perdu dans la lignée du chien, et s'est dupliqué entre les nœuds *Boreoeutheria* et *Catarrhini*. L'algorithme exact de définition des gènes ancestraux sera décrit en sous-section 7.3.1.

La version 57 d'Ensembl contient 886 547 gènes (codant pour des protéines) pour 51 espèces modernes (soit 17 383 en moyenne). Ensembl Compara fournit 36 450 arbres phylogénétiques, qui incluent au total 857 861 gènes (96.8%). Leurs 686 997 nœuds de spéciation et 151 566 nœuds de duplication permettent de définir 832 509 gènes ancestraux dans 43 ancêtres (soit 19 361 en moyenne).

7.2 Gènes et arbres phylogénétiques

7.2.1 Annotation des gènes

Au cours de ce travail de thèse, nous nous sommes limités aux gènes codant les protéines car eux seuls sont utilisés par Ensembl Compara pour construire des alignements

2. http://www.ensemblgenomes.org

multiples et des arbres phylogénétiques, essentiels aux études réalisées. Depuis peu, Ensembl met néanmoins à disposition des phylogénies pour les ARNs non codants. Ces gènes n'ont pas encore été intégrés à AGORA, mais le seront bientôt (voir perspectives).

On dispose souvent (en particulier pour des espèces bien annotées comme l'homme ou la souris) de plusieurs transcrits par gène. Prises dans leur intégralité, ces données de transcrits définissent pour chaque gène une «étendue génomique», c'est-à-dire le plus petit intervalle du génome qui contienne tous les transcrits. Sur l'exemple de la Figure 7.1, le gène WDR63 (*WD repeat domain 63*) possède un transcrit avec un exon non traduit en 5', situé bien en amont des trois autres, et englobe le gène MCOLN3 (*mucolipin 3*). Cette topologie crée une situation ambiguë lorsqu'il est nécessaire d'ordonner les gènes les uns par rapport aux autres sur l'axe du chromosome. Si la première extrémité 5' de transcrit rencontrée est utilisée pour ordonner les gènes, alors WDR63 sera ordonné avant MCCLN3 dans le sens gauche - droite. Si la dernière extrémité 3' est utilisée, alors WDR63 sera placé après MCCLN3. Dans le présent travail, nous avons choisi d'ordonner les gènes par rapport aux coordonnées de l'extrémité 5' du plus court transcrit de chaque gène. Ainsi, dans l'exemple de WDR63, WDR63 sera noté comme situé après MCOLN3, et non avant.

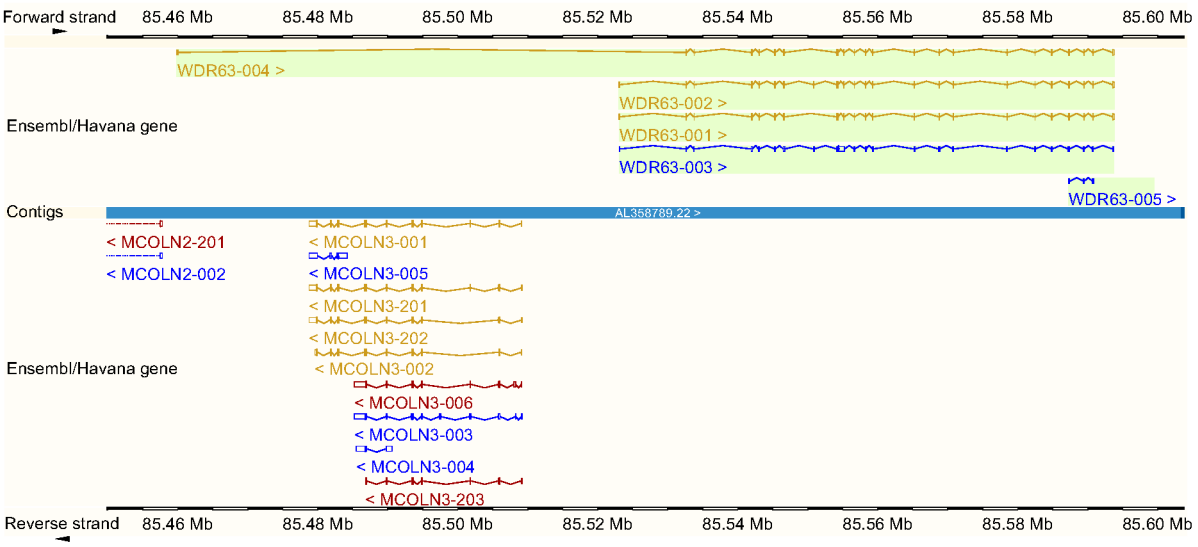


FIGURE 7.1 – Extrait de la base de données Ensembl montrant les annotations sur 150 kb autour de la position 85,5 Mb du chromosome 1 (p22.3) humain. Les transcrits d'un même gène sont notés avec le nom du gène suivi d'un suffixe numéroté. Le gène WDR63 possède un transcrit (WDR63-004) qui chevauche un autre gène : MCOLN3.

7.2.2 Nomenclature des noms d'espèces ancestrales

La plupart des nœuds ancestraux ont un nom connu tiré de la classification, plus ou moins consensuelle, disponible dans des bases de données de référence telle que le site du NCBI (Taxonomy)[3]. Ces noms sont utilisés dans les arbres phylogénétiques d'Ensembl et nous nous en servirons également. Cependant, certains nœuds de la phylogénie des vertébrés sont incertains, en particulier chez les mammifères. Le nœud d'origine des mammi-

3. http://www.ncbi.nlm.nih.gov/Taxonomy

fères placentaires, *Eutheria*, est parent des quatre ordres de mammifères : *Euarchontoglires* (primates, rongeurs, etc), *Laurasiatheria* (carnivores, insectivores, cétartiodactyles, etc), *Afrotheria* (éléphant, tenrecidés, etc), *Xenarthra* (paresseux, tatou, etc). Dans les arbres, l'ordre de spéciation des quatre ordres n'est pas résolu, alors que la littérature récente [Wildman *et al.*, 2007, Prasad *et al.*, 2008, Murphy *et al.*, 2007] montre qu'*Eutheria* est le nœud ancestral des *Atlantogenata* (qui groupe les *Afrotheria* et les *Xenarthra*) et des *Boreoeutheria* (qui groupe les *Euarchontoglires* et les *Laurasiatheria*).

7.2.3 Ajout d'une espèce

Nous avons ajouté trois espèces, absentes de la base de données Ensembl, qui nous semblaient importantes pour étudier l'évolution des génomes avant les vertébrés. En effet, comme nous le verrons plus tard, AGORA utilise les données de tous les génomes simultanément. Pour étudier une espèce et la comparer à celles déjà présentes, il est de fait préférable de l'inclure le plus tôt possible dans le processus de reconstruction. Ces trois espèces sont amphioxus (*Branchiostoma floridae*, un céphalochordé, Putnam *et al.* [2008]), l'anémone de mer (*Nematostella vectensis*, un cnidaire, Putnam *et al.* [2007]), et un tunicier (*Oikopleura dioica*, Denoeud *et al.* [2010]).

Deux solutions étaient possibles pour inclure ces espèces : reconstruire toutes les phylogénies en ré-exécutant le pipeline Compara ou conserver les phylogénies existantes et y rajouter les nouveaux gènes. Pour éviter une redondance des calculs avec Ensembl, c'est la deuxième approche qui a été retenue, bien que la première trouve un sens dans d'autres scénarios (voir perspectives).

Les annotations des gènes des nouvelles espèces ont été téléchargées (des sites du JGI et du Genoscope) et les séquences protéiques correspondantes ont été alignées avec BLASTp contre les protéines du nématode *Caenorabibdtis elegans*, de la drosophile *Drosophila melanogaster*, du poisson-zèbre, de la grenouille, du poulet, de l'opossum et de l'homme. On se sert des meilleurs hits réciproques (issus des résultats de BLASTp) pour sélectionner pour chaque gène d'une nouvelle espèce, les arbres d'Ensembl dans lesquels il devrait être inséré. Trois scénarios sont alors possibles.

- Un seul arbre est identifié. On insère le gène de la nouvelle espèce à sa position supposée d'après la phylogénie des espèces.
- Plusieurs arbres sont identifiés. On peut alors utiliser le nouveau gène uniquement s'il se positionne en espèce outgroup de chacun des arbres. On crée dans ce cas-là un ou plusieurs nœuds de duplication pour expliquer le nombre d'arbres.
- Aucun arbre n'est identifié. On crée alors un nouvel arbre pour inclure les gènes en meilleur hit réciproque, en copiant la phylogénie des espèces.

Espèce	Amphioxus	Anémone de mer	Oikopleura
Gènes annotés	28666	27273	18020
Gènes avec un meilleur hit	10295	9830	7504
Gènes insérés dans des arbres	8857	9549	6704
dont			
Gènes dans un nouvel arbre	290	421	352
Gènes dans un unique arbre, pré-existant	8442	8472	6318
Gènes dans plusieurs arbres, fusionnés	125	656	34

TABLE 7.1 – Statistiques sur les trois espèces supplémentaires insérées dans les données.

Les trois espèces sont rajoutées successivement, de préférence à partir de celles branchant la plus récemment (ici oikopleura) à celle branchant le plus anciennement (ici l'anémone de mer). Les décomptes totaux des gènes sont présentés dans le Tableau 7.1.

7.3 Liste des gènes ancestraux

Le plus petit élément qui va caractériser un génome ancestral reconstruit par AGORA est le gène. AGORA compare les génomes modernes sur la base de leur contenu en gènes et des données d'homologie. La première étape est donc de définir le contenu en gènes de chaque ancêtre et cette information est extraite des reconstructions phylogénétiques des gènes. De ces arbres phylogénétiques découlent des relations d'homologie : l'orthologie, utilisée pour la comparaison de deux génomes du point de vue de leur ancêtre commun, et la paralogie, utilisée pour la comparaison de deux génomes (souvent un génome par rapport à lui-même) du point de vue d'un événement de duplication. Nos études préliminaires (sous-section 13.2.2) ont montré que les contigs reconstruits étaient courts car le programme de reconstruction butait sur des gènes ancestraux peu annotés dans les génomes existants. Il a donc aussi été nécessaire de savoir établir un sous-ensemble des gènes, dits robustes, sur des critères de taille de famille.

7.3.1 Extraction à partir des arbres de protéines

Comme expliqué en sous-section 7.1.2, on dispose de reconstructions phylogénétiques pour tous les gènes codant pour des protéines. Le contenu en gènes des ancêtres se retrouve en parcourant chaque arbre depuis sa racine, en considérant un unique gène au départ. À chaque événement de duplication, une nouvelle copie du gène apparaît, sans qu'il soit possible de la distinguer de la copie d'origine (du moins, si on n'utilise que les données d'arbre phylogénétiques). À chaque événement de perte, le gène est supprimé dans toutes les espèces descendantes. Ce processus est appliqué récursivement depuis la racine de chaque arbre pour énumérer tous les gènes de tous les ancêtres.

Sur chaque nœud ancestral, on peut de la même manière définir des paires de gènes orthologues et paralogues. Chaque nœud de spéciation indique que toutes les paires de gènes pris dans deux sous-branches différentes sont orthologues, tandis qu'un nœud de duplication indique que toutes les paires de gènes ainsi considérées sont paralogues, qu'elles appartiennent ou non à la même espèce. Une paire de gènes orthologues représente un unique gène ancestral qui existait au dernier ancêtre commun, et peut donc être utilisée pour comparer deux génomes. Ce n'est pas le cas des paires de gènes paralogues, qui représentent des événements qui ont eu lieu sur des branches, et ne sont pas rattachées à l'existence d'un gène à un ancêtre donné, mais à une modification du contenu en gènes.

L'algorithme 7.1 permet d'extraire ces informations (liste des gènes ancestraux, paires de gènes orthologues et paralogues) d'un arbre, à partir de sa racine. Le point particulier de cet algorithme est qu'il faut tenir compte qu'une branche d'un arbre phylogénétique ne lie pas nécessairement deux ancêtres consécutifs à cause des pertes, et qu'il faut donc créer des gènes pour des ancêtres autres que ceux référencés par les nœuds internes (dans l'exemple de la Figure 7.2, il n'y a pas de nœud *Boreoeutheria*, mais cet ancêtre possédait bien un exemplaire du gène). Ainsi, l'algorithme crée récursivement tous les gènes désignés par l'arbre, puis à partir des listes de gènes de chacune des sous-branches (les L_i), crée les listes de paires de gènes homologues.

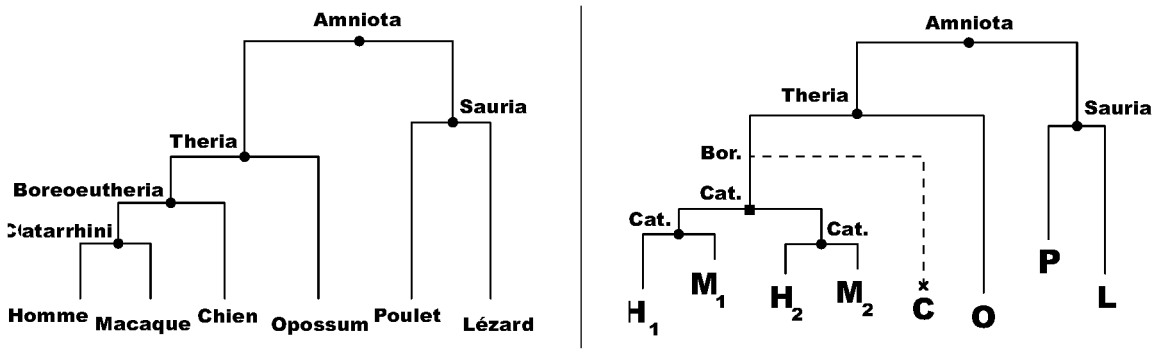


FIGURE 7.2 – Exemple d'arbre phylogénétique réconcilié. Les ronds indiquent des spéciations, les carrés des duplications. Traditionnellement, les pertes n'apparaissent pas directement dans la phylogénie du gène, mais sont révélées par comparaison avec la phylogénie des espèces. Ici, la perte du gène chez le chien est indiquée par une branche en tirets terminée par une croix.

Ancêtre	Gènes ancestraux
Amniota	$A_{\text{Amn}} = \{P, L, O, H_1, H_2, M_1, M_2\}$
Sauria	$A_{\text{Saur}} = \{P, L\}$
Theria	$A_{\text{Ther}} = \{O, H_1, H_2, M_1, M_2\}$
Boreoeutheria	$A_{\text{Boreo}} = \{O, H_1, H_2, M_1, M_2\}$
Catarrhini	$A_{\text{Cat}_1} = \{H_1, M_1\}$ $A_{\text{Cat}_2} = \{H_2, M_2\}$

TABLE 7.2 – Liste des gènes ancestraux définis à partir de l'arbre phylogénétique de la Figure 7.2.

Ancêtre	Paires de gènes orthologues	Paires de gènes paralogues
Amniota	$\{P, L\} \times \{O, H_1, H_2, M_1, M_2\}$	$\emptyset$
Sauria	$\{C\} \times \{L\}$	$\emptyset$
Theria	$\emptyset$	$\emptyset$
Boreoeutheria	$\{O\} \times \{H_1, H_2, M_1, M_2\}$	$\emptyset$
Catarrhini	$\{H_1\} \times \{M_1\}$ $\{H_2\} \times \{M_2\}$	$\{H_1, M_1\} \times \{H_2, M_2\}$

TABLE 7.3 – Liste des paires de gènes homologues identifiées pour chaque ancêtre, à partir de la Figure 7.2.

Algorithme 7.1 Extraction des gènes ancestraux et des paires de gènes homologues d'un arbre

Entrées: $\mathscr{A}$: un arbre phylogénétique réconcilié. N : position (nœud) courante dans $\mathscr{A}$ (initialement, sa racine)

Variables globales: $\mathscr{L}_{\text{para}}^{\text{Anc}}, \mathscr{L}_{\text{ortho}}^{\text{Anc}}, \mathscr{L}_{\text{gènes}}^{\text{Anc}}$: les listes pour chaque ancêtre Anc des paires de gènes paralogues, de celles orthologues, et des gènes ancestraux.

pour tout Fils F_i du nœud N de $\mathscr{A}$ **faire**
  // Liste des gènes créés dans la branche du fils F_i
  Appel récursif sur le nœud F_i
  Stocker la liste des gènes renvoyés par la fonction dans L_i
  pour tout Ancêtre Anc_j intermédiaire entre les nœuds N et F_i **faire**
    Définir un gène G_{F_i,Anc_j} pour l'ancêtre Anc_j, à partir de l'ensemble des gènes modernes présents dans $\mathscr{A}$ sous le nœud F_i
    // Ajout du nouveau gène ancestral
    Rajouter G_{F_i,Anc_j} à $\mathscr{L}_{\text{gènes}}^{\text{Anc}_j}$, et $\{G_{F_i,\text{Anc}_j}\} \times L_i$ à $\mathscr{L}_{\text{ortho}}^{\text{Anc}_j}$
    Rajouter G_{F_i,Anc_j} à L_i
  fin pour
fin pour
si le nœud N est un nœud de spéciation **alors**
  // N désigne donc un gène ancestral
  Définir un gène G_N, à partir de l'ensemble des gènes modernes présents sous le nœud N
  Rajouter G_N à $\mathscr{L}_{\text{gènes}}^{\text{Anc}_N}$
  pour tout Fils F_i **faire**
    Rajouter $\{G_N\} \times L_i$ à $\mathscr{L}_{\text{ortho}}^{\text{Anc}_N}$
  fin pour
fin si
// Paires d'homologues formées par le nœud N
pour tout Paire de fils (F_{i_1}, F_{i_2}) **faire**
  Rajouter $L_{i_1} \times L_{i_2}$ à $\mathscr{L}_{\text{para}}^{\text{Anc}_N}$ ou $\mathscr{L}_{\text{ortho}}^{\text{Anc}_N}$ selon si le nœud N est une duplication ou non
fin pour
renvoyer Liste de tous les gènes créés : $\bigcup_i L_i$ (en ajoutant G_N si on l'a défini)

7.3.2 Filtre sur le nombre d'événements dans les familles

L'expérience (sous-section 13.2.2) nous a montré que les reconstructions phylogénétiques, dont dépendent les reconstructions de génomes, sont de qualité hétérogène. Il est donc devenu opportun d'établir un classement des familles de gènes sur la base de la topologie des arbres phylogénétiques, afin de pouvoir travailler avec un sous-ensemble de familles respectant certains critères de qualité quantifiés. Deux méthodes ont été définies, qui pourront être utilisées alternativement selon les données disponibles et à traiter.

La première méthode (baptisée *events*) consiste en la sélection d'un sous-ensemble d'arbres, parmi l'ensemble des arbres définis par Ensembl. La méthode (algorithme 7.2) trie d'abord les arbres selon le nombre de gènes manquants (par rapport à un arbre de même topologie qui ne contiendrait aucun événement de perte), le nombre d'événements de duplications, et enfin le nombre d'événements de pertes. L'algorithme sélectionne alors dans l'ordre les arbres (en minimisant ces trois critères) jusqu'à inclure une proportion p_g de l'ensemble des gènes modernes référencés par l'ensemble des arbres, où p_g est un un paramètre du filtre indiquant une proportion.

Il faut noter que la sélection d'une proportion p_g d'arbres (à la place du nombre de gènes) n'est pas utilisée car cette solution aurait réduit le nombre de gènes ancestraux de manière incontrôlable (la proportion de gènes restants aurait été difficilement prévisible à partir de p_g). Cela s'explique par le fait que les arbres éliminés sont des familles ayant perdu beaucoup de gènes, mais en contenant aussi beaucoup (par exemple des familles contenant des duplications très anciennes, et s'étant rapidement perdues dans des clades entiers). Les plus grandes familles ainsi supprimées réduisaient drastiquement le nombre de gènes des ancêtres. C'est pourquoi la proportion p_g est appliquée sur le nombre de gènes désignés par les arbres et pas sur le nombre d'arbres en lui même. En procédant de cette manière, le taux de gènes conservés est proche de p_g pour tous les ancêtres.

Algorithme 7.2 Filtrage d'un ensemble d'arbres selon une proportion de gènes à conserver

Entrées: p_g : le paramètre de proportion de gènes ($0 \leqslant p_g \leqslant 1$). $(\mathscr{A}_i)$: l'ensemble des arbres phylogénétiques réconciliés.

$n_g \leftarrow$ nombre total de gènes référencés dans les arbres
$n_{d,i} \leftarrow$ nombre d'événements de duplication dans l'arbre $\mathscr{A}_i$
$n_{p,i} \leftarrow$ nombre d'événements de perte de gènes dans l'arbre $\mathscr{A}_i$
$n_{g,i} \leftarrow$ nombre de gènes perdus dans l'arbre $\mathscr{A}_i$
$c \leftarrow 0$
pour tout $\mathscr{A}_i$, dans l'ordre lexicographique croissant du triplet $(n_{p,i}, n_{d,i}, n_{p,i})$ **faire**
    $c \leftarrow c +$ nombre de gènes dans l'arbre $\mathscr{A}_i$
    si $c \leqslant p_g n_g$ **alors**
        Conserver l'arbre $\mathscr{A}_i$
    sinon
        Supprimer l'arbre $\mathscr{A}_i$
    fin si
fin pour

7.3.3 Filtre sur la taille des familles

L'autre solution (nommée *size*) pour sélectionner des gènes ancestraux permet une sélection plus fine sur chaque arbre. Le principe de la méthode (algorithme 7.3) est de sélectionner dans un arbre les sous-arbres dont la taille relative T (le rapport entre le nombre de gènes qu'il contient et le nombre d'espèces qu'il inclut) est comprise entre deux paramètres $T_{\min}$ et $T_{\max}$.

Par exemple, une famille n'ayant subi aucun événement (perte, duplication) a une taille relative de $T = 1$, une famille s'étant dupliquée dès sa racine une taille relative de $T = 2$, et une famille ayant perdu la moitié de ses gènes une taille relative de $T = 0,5$. Des valeurs typiques des paramètres sont $T_{\min} = 0,9$ et $T_{\max} = 1,1$ pour autoriser 10% de tolérance sur la taille d'une famille par rapport au nombre d'espèces attendues.

Une famille est conservée si elle vérifie ce critère de taille. Si ce n'est pas le cas, la famille est remplacée par les sous-familles qui en découlent, et on continue récursivement. Dans le pire des cas, une famille est donc éclatée en des sous-familles contenant chacune 1 gène. Dans le cas général, à cause du découpage, une famille perdra ses liens d'homologie anciens, et les ancêtres récents conserveront donc davantage de familles que les ancêtres plus anciens.

Il faut noter que l'algorithme ne vérifie la cohérence de l'arbre que dans sa globalité, et pas à tous les nœuds de l'arbre. Ainsi, dans les arbres sélectionnés pour des paramètres $T_{\min} = T_{\max} = 1$, une famille possèdera *à sa racine* autant de gènes que d'espèces. Cela inclut donc les arbres contenant des pertes de gènes dans certaines espèces, compensées par autant de duplications dans d'autres.

Algorithme 7.3 Découpage d'un arbre selon des critères de taille de familles

Entrées: $T_{\min}$ et $T_{\max}$: les paramètres de tailles ($0 \leqslant T_{\min} \leqslant 1 \leqslant T_{\max}$). $\mathscr{A}$: un arbre phylogénétique réconcilié.

1: $n_g \leftarrow$ nombre de gènes dans l'arbre $\mathscr{A}$
2: $n_e \leftarrow$ nombre d'espèces sous l'ancêtre désigné par la racine de l'arbre
3: **si** $T_{\min} \leqslant \frac{n_g}{n_e} \leqslant T_{\max}$ **alors**
4: **fin**
5: **sinon**
6: Supprimer l'arbre $\mathscr{A}$ et le remplacer par ses sous-arbres $\mathscr{A}_1, \mathscr{A}_2 \ldots \mathscr{A}_p$
7: Appeler récursivement la procédure pour chaque $\mathscr{A}_i$
8: **fin si**

Chapitre 8

Comparaison de deux génomes

Sommaire

L'information élémentaire et fondamentale dont a besoin AGORA pour la reconstruction réside dans la comparaison des génomes d'espèces modernes. Nous avons donc défini deux méthodes qui identifient des objets conservés entre deux génomes : une qui opère à bas niveau, la conservation de paires de gènes, et une qui opère à niveau supérieur, capable de reconstituer des segments conservés.

8.1 Paires conservées

Le postulat suivant est l'élément de base sur lequel s'appuie l'ordre ancestral reconstruit par AGORA, et ce postulat revient à une simple application du principe de parcimonie.

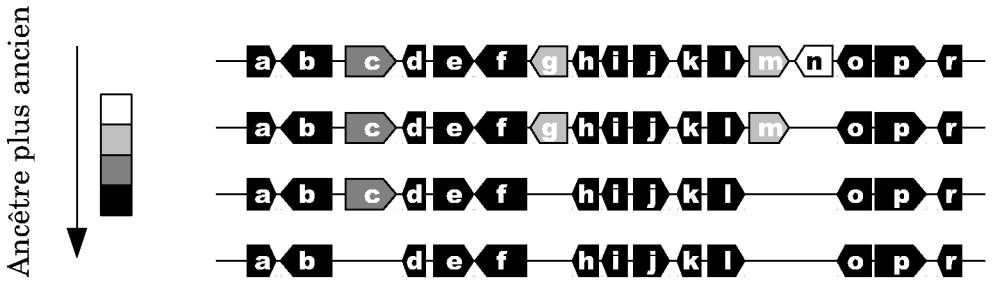
Deux gènes a_1 et b_1 situés dans le même génome ont deux gènes orthologues respectivement a_2 et b_2 dans un autre génome. Si les paires a_1 et b_1 d'une part et a_2 et b_2 d'autres part sont voisines et dans la même orientation transcriptionnelle dans leurs génomes respectifs alors cette configuration est considérée comme ancestrale.

Ceci, en dehors d'être validé empiriquement (voir chapitre 13), résulte du fait que deux génomes n'ont jamais évolué strictement indépendamment, mais partagent un ancêtre commun. Ils conservent donc encore un patrimoine de ce génome ancestral qui peut se retrouver, selon l'intervalle de temps, dans l'ordre des gènes (décrit dans ce chapitre), ou dans la synténie (section 10.3). Ce postulat doit être confronté au point de vue combinatoire qui nous permet d'estimer que la probabilité d'obtenir deux nombre consécutifs égaux entre deux permutations de $[1,n]$, avec n de l'ordre de 20000, est quasi-nulle.

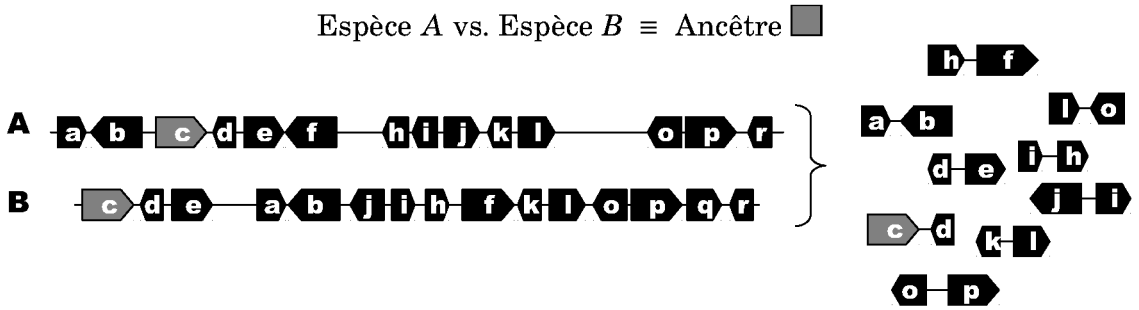
Pour la comparaison de deux génomes (algorithme 8.1 et Figure 8.1), AGORA filtre d'abord ces deux génomes pour ne retenir que les gènes présents dans leur dernier ancêtre commun. Puis il intersecte les deux ensembles de paires de gènes consécutifs (en tenant compte de l'orientation), et rajoute ces ensembles de paires de gènes conservées à tous les ancêtres définis entre les deux espèces comparées. Sur cette dernière étape, il faut faire attention à ne pas prédire de paire de gènes chez un ancêtre, si on sait qu'un gène s'est, entre temps, inséré entre les deux gènes d'intérêt.

A :

B :



C :



D :

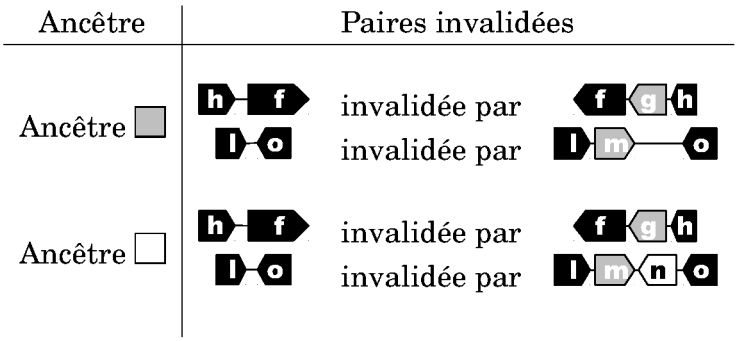


FIGURE 8.1 – Extraction des paires de gènes conservées entre 2 génomes. **A :** Extrait d'un chromosome du génome de l'espèce A, les gènes sont numérotés de a à r. **B :** La nuance de gris d'un gène indique l'âge du plus vieil ancêtre chez lequel il est retrouvé. Du point de vue d'un ancêtre donné, le segment de chromosome de A doit être filtré pour ne conserver que les gènes qui existaient à ce moment. Les lignes montrent le segment de chromosome de A filtré selon le contenu en gènes de chaque ancêtre, pris du plus jeune au plus ancien. **C :** La comparaison de deux génomes consiste en l'intersection des paires de gènes (orientés) consécutifs, une fois les génomes filtrés par rapport au contenu en gènes de leur dernier ancêtre commun. **D :** Les paires conservées sont propagées à tous les ancêtres intermédiaires où elles existaient (en reprenant le contenu dicté en **B**). Certaines paires ne peuvent cependant pas être propagées car l'apparition d'un gène (comme g entre f et h) dans un ancêtre récent peut interrompre l'adjacence observée dans un ancêtre plus ancien.

Algorithme 8.1 Compare deux génomes et extrait la liste des paires de gènes conservés

Entrées: $\mathscr{G}_A$ et $\mathscr{G}_B$: deux génomes à comparer.

Anc$_0$ ← ancêtre commun de A et de B
// Filtrage des génomes $\mathscr{G}_A$ et $\mathscr{G}_B$ (Figure 8.1.A)
pour tout Espèce ancestrale Anc entre A et Anc$_0$ **faire**
    $\mathscr{G}_A^{\text{Anc}} \leftarrow \mathscr{G}_A$ filtré pour ne conserver que les gènes présents dans $\mathscr{L}_{\text{gènes}}^{\text{Anc}}$
    $P_A^{\text{Anc}} \leftarrow \bigcup_{g_1,g_2 \text{ consécutifs dans } \mathscr{G}_A^{\text{Anc}}} \{(g_1,g_2),(\overline{g_2},\overline{g_1})\}$
fin pour
pour tout Espèce ancestrale Anc entre B et Anc$_0$ **faire**
    $\mathscr{G}_B^{\text{Anc}} \leftarrow \mathscr{G}_B$ filtré pour ne conserver que les gènes présents dans $\mathscr{L}_{\text{gènes}}^{\text{Anc}}$
    $P_B^{\text{Anc}} \leftarrow \bigcup_{g_1,g_2 \text{ consécutifs dans } \mathscr{G}_B^{\text{Anc}}} \{(g_1,g_2),(\overline{g_2},\overline{g_1})\}$
fin pour
// Intersection des paires de gènes (Figure 8.1.C)
$\mathscr{C}_{\text{Anc}_0} \leftarrow P_A^{\text{Anc}_0} \cap P_B^{\text{Anc}_0}$
// Propagation des paires conservées aux ancêtres intermédiaires (Figure 8.1.D)
pour tout Espèce ancestrale Anc entre A et Anc$_0$ **faire**
    $\mathscr{C}_{\text{Anc}} \leftarrow \mathscr{C}_{\text{Anc}_0} \cap P_A^{\text{Anc}}$
fin pour
pour tout Espèce ancestrale Anc entre B et Anc$_0$ **faire**
    $\mathscr{C}_{\text{Anc}} \leftarrow \mathscr{C}_{\text{Anc}_0} \cap P_B^{\text{Anc}}$
fin pour
renvoyer $\mathscr{C}$

8.2 Segments conservés

L'algorithme précédent travaille à une échelle très locale (au niveau des paires de gènes) et manque de vue globale sur la conservation à grande échelle de deux génomes. Or, il a été nécessaire de savoir comparer deux génomes en les alignant (par rapport à l'ordre des gènes) pour établir des régions (des segments) d'ordre conservé. En appliquant le même principe de parcimonie que précédemment, une région d'ordre conservé révèle l'ordre des gènes sur des segments de chromosomes ancestraux, et on appellera ces régions des «segments conservés».

L'algorithme décrit ici travaille comme une couche supplémentaire à l'extraction de paires de gènes conservées. La procédure est de définir des paires conservées, puis de les fusionner pour former des segments conservés, le tout régi par deux paramètres.

Le premier, R, indique sur quelle référence de gènes effectuer la comparaison des génomes et extraire les paires conservées. Là où l'algorithme précédent filtrait systématiquement les génomes comparés selon leur dernier ancêtre commun, nous introduisons trois valeurs possibles (trois niveaux de sélectivité) qui permettent de moduler la stringence des segments conservés :

- R_{tous}, la totalité des gènes : aucun filtre n'est appliqué ;
- $R_{\text{anc-comm}}$, la liste des gènes de l'ancêtre commun : on ne tient pas compte des gains de gènes espèce-spécifiques (paramètre implicitement utilisé dans l'algorithme 8.1) ;
- R_{inters}, l'intersection des gènes des deux espèces : on ne tient compte ni des gains, ni des pertes de gènes espèce-spécifiques.

$R_{\text{anc-comm}}$ permet de reconstruire des segments de chromosomes ancestraux, qu'on retrouve dans les deux génomes comparés avec éventuellement des insertions de gènes

A :

B :

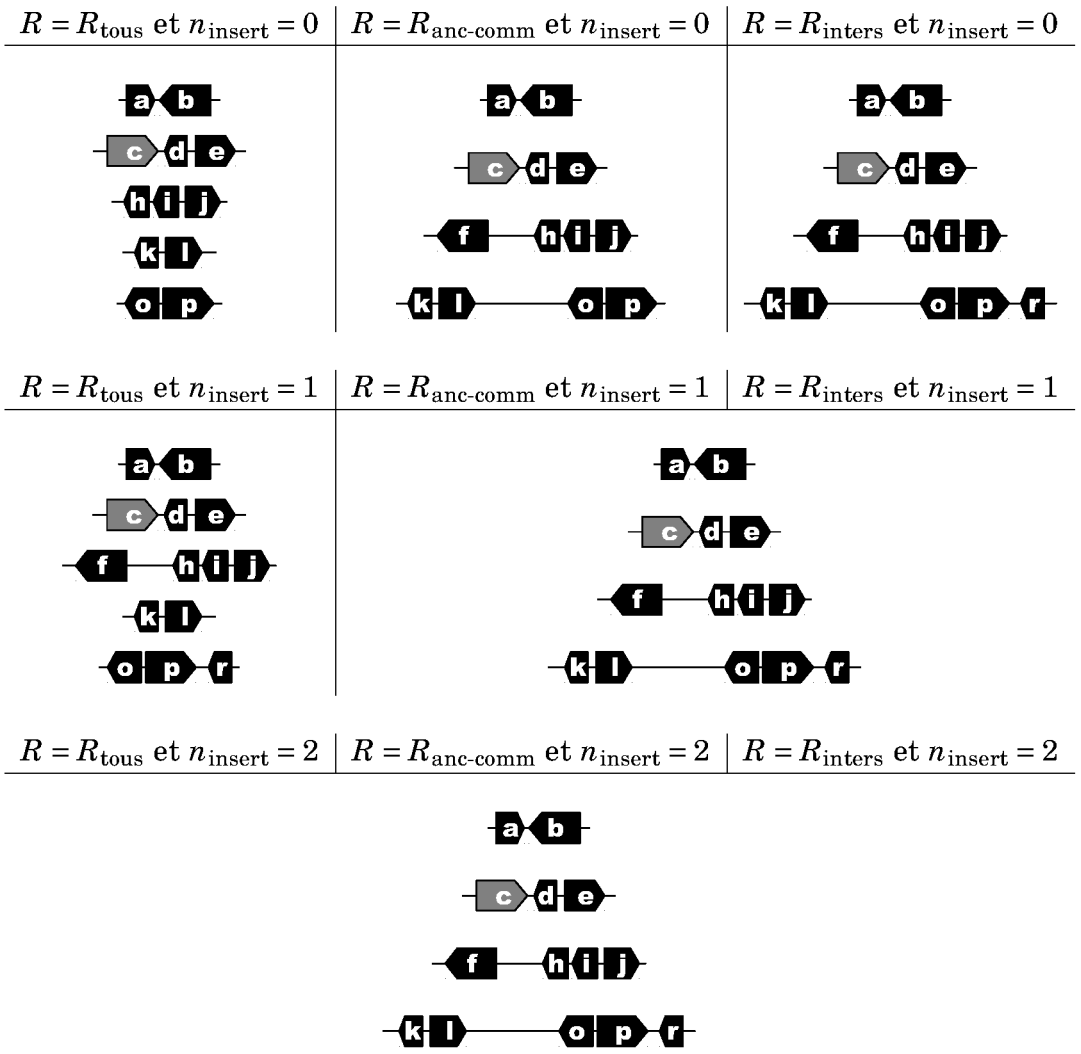


FIGURE 8.2 – Extraction des blocs de gènes conservés entre 2 génomes. **A :** Extraits de deux génomes. Les gènes sont marqués d'une lettre si ils sont dans la liste des gènes ancestraux, et en blanc si ils sont spécifiques de l'espèce. **B :** Segments conservés extraits selon les trois valeurs de R, avec n_{insert} valant 0, 1, ou 2. Pour $R = R_{\text{tous}}$, les segments conservés sont les mêmes quelque soit la valeur de n_{insert} : les configurations à la droite du tableau sont identiques.

espèce-spécifiques. Avec R_{tous}, les segments conservés désignent toujours des segments ancestraux, mais qui, cette fois, sont restés strictement identiques dans les deux génomes. Enfin, avec R_{inters}, les segments conservés ne désignent plus des segments ancestraux car les gènes ancestraux qui ont été perdus indépendamment dans les deux génomes comparés ne seront pas présents.

Le deuxième paramètre, n_{insert}, indique le nombre maximal d'insertions consécutives de gènes dans chaque génome. On sait ainsi que les gènes d'un bloc se suivent tous avec au maximum n_{insert} gènes insérés entre eux.

Par exemple (f, h, i, et j dans la Figure 8.2), avec des paramètres $R = R_{\text{tous}}$ et $n_{\text{insert}} = 0$, un segment de gènes identique dans deux génomes A et B mais avec une insertion de gène spécifique dans A apparaît en deux morceaux. Avec les paramètres $R = R_{\text{tous}}$ et $n_{\text{insert}} = 1$, ou $R = R_{\text{anc-comm}}$ et $n_{\text{insert}} = 0$, alors le segment apparaît un seul morceau.

L'algorithme 8.2 définit des segments initiaux à partir de paires conservées et de gènes orthologues. Ces segments vont progressivement être fusionnés lorsqu'ils sont séparés de moins de n_{insert} gènes.

Algorithme 8.2 Compare deux génomes et extrait des segments de gènes conservés

Entrées: $\mathscr{G}_A$ et $\mathscr{G}_B$: deux génomes à comparer. R : paramètre indiquant quelle référence de gènes utiliser. n_{insert} : nombre maximal d'insertion de gènes autorisés.

Construire G : la liste de gènes de référence selon R
// Filtrage des génomes $\mathscr{G}_A$ et $\mathscr{G}_B$
$\mathscr{G}'_A \leftarrow \mathscr{G}_A \cap G$
$\mathscr{G}'_B \leftarrow \mathscr{G}_B \cap G$
// Segments conservées initiaux entre $\mathscr{G}_A$ et $\mathscr{G}_B$
$P \leftarrow$ ensemble des paires (p_A, p_B) conservées entre $\mathscr{G}'_A$ et $\mathscr{G}'_B$ // Segments de longueur 2
$O \leftarrow$ ensemble des paires de gènes orthologues (g_A, g_B) entre $\mathscr{G}'_A$ et $\mathscr{G}'_B$, non inclus dans des paires conservées // Segments de longueur 1
// Fusion des segments conservés selon n_{insert}
$B = P \cup O$
tant que existent deux blocs (b_1, b_2) dans B, consécutifs dans $\mathscr{G}'_A$ et $\mathscr{G}'_B$ **faire**
  si $\text{dist}_{A'}(b_1, b_2) \leqslant n_{\text{insert}}$ et $\text{dist}_{B'}(b_1, b_2) \leqslant n_{\text{insert}}$ **alors**
    Fusionner les deux blocs et mettre à jour B
  fin si
fin tant que
renvoyer B

Chapitre 9

Choix d'un ordre de marqueurs ancestral

Sommaire

Les adjacences extraites à partir de l'algorithme 8.1 du chapitre précédent définissent des ensembles de paires conservées censées appartenir aux ancêtres. Pour un ancêtre donné, dans le cas idéal (Figure 9.1.A), toutes les paires qu'il contient sont cohérentes entre elles, et peuvent se fusionner pour former directement des segments ancestraux. Dans ce scénario, la reconstruction ancestrale est immédiate. Malheureusement, les graphes d'adjacence réels (Figure 9.1.B) sont des graphes au sens général (avec des cycles et des bifurcations). Ils sont alors éloignés de la structure biologique des chromosomes de vertébrés. Il y a donc nécessité de définir des algorithmes de parcours de graphes pour sélectionner un ensemble de chemins indépendants et sans cycles, qui pourront être assimilés à des chromosomes, Ces algorithmes seront aidés par une pondération des adjacences par le nombre de comparaisons qui les soutiennent.

Après quelques définitions, le chapitre est composé de deux sections qui proposent des algorithmes développés dans le but d'identifier des ordres de marqueurs ancestraux. La première opère dans le cas le plus général (section 9.2), lorsque la seule source d'information est le résultat de la comparaison de génomes modernes. Il existe cependant des cas particuliers où l'on dispose d'informations sur l'ordre de certains marqueurs, que l'on souhaite fixer dès le départ. Ces contraintes imposent des modifications du cas général, décrites dans la section suivante (section 9.3). Enfin, nous proposons une méthode de repli lorsqu'aucune information d'adjacence n'est disponible (section 9.4).

Hormis l'orientation des marqueurs, les techniques de reconstruction d'ordre ancestral suivantes sont formulées dans le cadre de graphes généraux, afin de se détacher de particularités génomiques. Pour faciliter la compréhension, on pourra assimiler les marqueurs à des gènes, et les chemins à des segments de chromosomes.

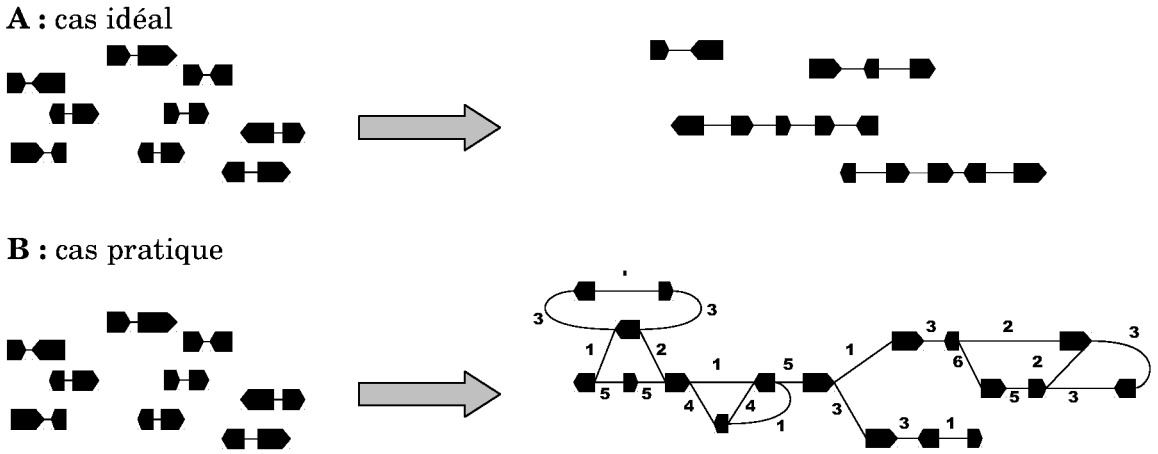


FIGURE 9.1 – Nécessité des algorithmes de parcours de graphe. Pour un ancêtre, le graphe d'adjacence issu de l'ensemble des paires conservées. (**B**) correspond rarement directement à des segments de chromosomes (**A**).

9.1 Définitions

On note M un ensemble de marqueurs. Chaque marqueur m peut être orienté, et possède dès lors deux extrémités : une entrée m^- et une sortie m^+. $\overrightarrow{m}$ représente alors le marqueur dans une orientation canonique (m^-, m^+), et $\overleftarrow{m}$ le même marqueur pris dans l'autre sens (m^+, m^-). L'opérateur $m \mapsto \overline{m}$ permet d'inverser l'orientation d'un marqueur orienté. L'ensemble des extrémités des marqueurs est $M^{\pm} = \bigcup_{m \in M} \{m^-, m^+\}$. L'ensemble des marqueurs, en distinguant les deux orientations possibles, est $\overleftrightarrow{M} = \bigcup_{m \in M} \{\overrightarrow{m}, \overleftarrow{m}\}$. Un intervalle est une paire de marqueurs orientés $(m_1, m_2) \in \overleftrightarrow{M}^2$. Il représente la jonction (non orientée) entre la sortie de m_1 et l'entrée de m_2. Une telle paire est équivalente à $(\overline{m_2}, \overline{m_1})$.

On définit ensuite un graphe non-orienté G sur $\overleftrightarrow{M}$ comme un ensemble d'arêtes $A \subset \overleftrightarrow{M} \times \overleftrightarrow{M}$. Ce graphe est pondéré si on ajoute une fonction de valuation $v : A \mapsto \mathbb{R}$.

Un chemin c de longueur l_c dans le graphe G est une suite de marqueurs orientés $c = (c_1 \dots c_{l_c})$ telle que toutes les paires (c_i, c_{i+1}) sont présentes dans A. Un chemin de longueur 1 est appelé singleton. Le chemin $(c_1 \dots c_{l_c})$ est équivalent au chemin $(\overline{c_{l_c}} \dots \overline{c_1})$. Un chemin c' est un sous-chemin de c s'il reprend un sous-ensemble des marqueurs de c, dans le même ordre que dans c. Un chemin c' est un fragment de c s'il reprend un sous-ensemble de marqueurs consécutifs de c, dans le même ordre.

Un chemin c est génomique s'il ne passe pas deux fois par le même marqueur (tout singleton est naturellement génomique). Un ensemble de chemins C est lui-même génomique si chaque chemin l'est, et si aucun marqueur n'est présent dans deux chemins. Enfin, un ensemble génomique de chemins est une partition génomique si chaque marqueur de M est présent dans un (unique) chemin. On a alors la propriété $\sum_{c \in C} l_c = |M|$. Une partition génomique est considérée comme triviale si elle n'implique que des singletons.

Ces définitions reprennent le fait que les chromosomes de vertébrés sont non-circulaires et qu'évidemment, un gène n'est présent qu'à une seule position dans un génome. On pourra donc assimiler un chemin génomique à un chromosome, un ensemble génomique à un ensemble de chromosomes et une partition génomique à un génome complet. Les algorithmes présentés dans les sections suivantes travaillent dans de tels graphes et permettent d'extraire des ensembles génomiques de chemins, voire des parti-

tions génomiques. La fonction de valuation v sera systématiquement utilisée pour choisir les «meilleurs» chemins (la définition de meilleur dépendant du contexte).

La notion de «meilleur» chemin réside dans la définition d'une fonction que l'on cherche à optimiser. Dans nos graphes, les fonctions linéaires (comme la somme des poids des arêtes) peuvent en général être résolues par des algorithmes de recherche de flot de coût minimal, et donc en complexité polynomiale. Cependant, rien ne permet pas d'affirmer que ces fonctions sont les plus crédibles biologiquement parlant. Le cœur et l'objet du travail ne sont donc pas la recherche d'un algorithme de complexité optimale, mais d'une méthode de résolution compatible avec nos attentes des génomes ancestraux (plus une adjacence est observée dans les espèces modernes, plus elle a de chances d'être ancestrale). Nous avons donc en général préféré des solutions gloutonnes qui suivent le raisonnement qu'aurait un décideur humain, et décrit les méthodes de résolution correspondantes.

9.2 Dans un graphe sans contraintes

Ce premier algorithme est utilisé lorsqu'on n'a aucun a priori sur la structure du résultat. En reprenant la formalisation introduite, le problème est simplement de définir une partition génomique à partir du graphe d'adjacences. Le nombre de chemins (d'au moins deux marqueurs) est alors un résultat annexe de cet algorithme, car il représente le nombre de chromosomes reconstruits. Le reste (les singletons) sont des marqueurs n'ayant pu être reconstruit et restant sans voisins.

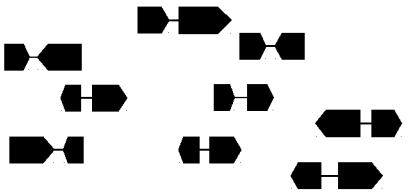
Le graphe d'adjacences n'est (en général) pas directement une partition génomique (Figure 9.1), il faut donc enlever des arêtes, et de préférence le moins possible, ou celles avec les poids les plus faibles. L'algorithme 9.1 (illustré Figure 9.2) indexe les arêtes en $A = \{a_i\}_{1 \leqslant i \leqslant n_a}$ selon les poids décroissants (la suite $(v(a_i))_{1 \leqslant i \leqslant n_a}$ est décroissante). En cas d'égalité de poids, l'ordre des a_i est arbitraire. Partant d'une partition triviale C de G (uniquement des singletons), l'algorithme rajoute chaque arête a_i à C, uniquement si elle lie deux extrémités de chemins, en fusionnant les chemins correspondant. Comme on ne fait qu'ajouter des liens entre extrémités de chemins, on a l'assurance que les marqueurs ne sont présents qu'une et une seule fois chacun, et que tous les chemins de C sont génomiques. C désigne donc continuellement une partition génomique de M. L'algorithme s'arrête lorsqu'il a essayé d'insérer toutes les arêtes.

Algorithme 9.1 Extraction d'une partition génomique dans un graphe de marqueurs orientés, sans contraintes

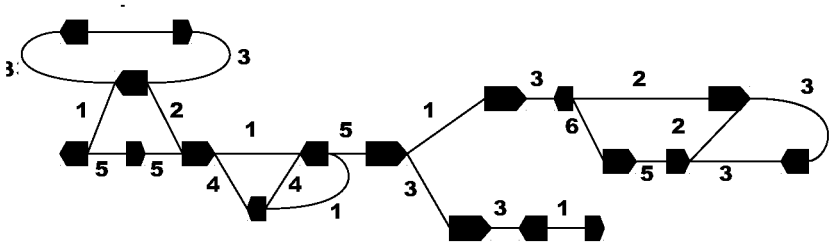
Entrées: $G = (\overleftrightarrow{M}, A, v)$: un graphe pondéré.

1: $C \leftarrow$ partition triviale sur M
2: $(a_i)_{1 \leqslant i \leqslant n_a} \leftarrow$ arêtes de A, triées par poids décroissant selon v
3: **pour tout** arête $a_i = (m_1, m_2)$ **faire**
4: **si** l'extrémité sortante de m_1 et l'extrémité entrante de m_2 sont libres dans C, et qu'ils sont utilisés dans des chemins différents **alors**
5: Remplacer les deux chemins désignés par leur concaténation selon a_i
6: **fin si**
7: **fin pour**
8: **renvoyer** C

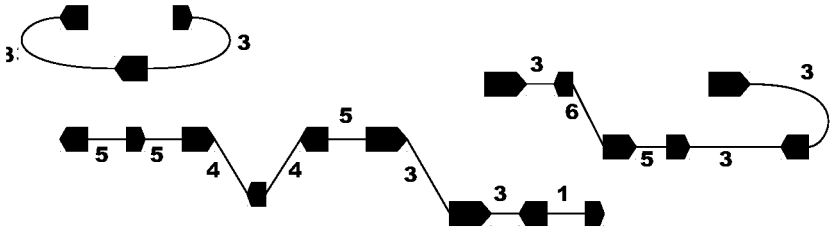
A :



B :



C :



D :

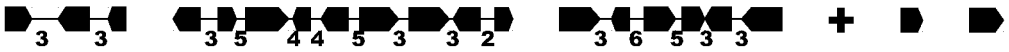


FIGURE 9.2 – Extraction d'une partition génomique dans un graphe de marqueurs orientés, sans contraintes. **A :** L'entrée de l'algorithme est un ensemble pondéré de paires de marqueurs orientés. **B :** Les paires de marqueurs sont placées dans un graphe non orienté, pondéré. Ce graphe contient généralement des cycles, et les degrés entrants et sortants peuvent être supérieurs à 2. **C :** L'algorithme teste successivement toutes les arêtes par ordre décroissant de poids, et conserve celles qui sont cohérentes avec les arêtes précédemment sélectionnées. **D :** Le résultat est une partition génomique : une liste de chemins maximaux et non chevauchants, qui représentent un ordre optimal de marqueurs (dont une liste de singletons).

Notre solution répond aux critères suivants.

1. Il n'existe plus aucune arête liant deux extrémités de chemins différents (les chemins sont maximaux).
2. Chaque arête choisie a le plus haut poids possible, compte tenu des autres arêtes. C'est-à-dire que si une arête ne peut être sélectionnée, c'est parce qu'une autre arête de poids supérieur (ou égal) a déjà été sélectionnée.

Dans la solution actuelle, en cas d'égalité de poids, l'ordre de sélection des arêtes se fait au hasard, alors que cela peut avoir des conséquences sur le choix des futures arêtes (Figure 9.3). Il serait plus rigoureux de tester la meilleure combinaison d'arêtes.

L'optimal que l'on cherche à obtenir est le maximum (selon l'ordre lexicographique) de la liste des poids des arêtes choisies triées par ordre décroissant, soit par exemple $(5,5,4,3) > (5,5,3,3,2)$.

L'algorithme 9.2 classe les arêtes par poids décroissants, puis les traite par ensembles d'arêtes de même poids. Pour chaque poids, l'algorithme essaie d'insérer le maximum d'arêtes possibles, ce qui se traduit par la recherche de cliques [1] de taille maximale (problème NP-complet) dans un graphe liant les arêtes compatibles entre elles.

1. Une clique est un sous-graphe complet, c'est-à-dire dans lequel tout sommet est lié à tous les autres.

Algorithme 9.2 Extraction d'une partition génomique dans un graphe de marqueurs orientés, sans contraintes (traitement optimal des égalités de v)

Entrées: $G = (\overleftrightarrow{M}, A, v)$: un graphe pondéré

1: // L_C contient l'ensemble des solutions possibles
2: $L_C \leftarrow \{C_0\}$ (C_0 est une partition génomique triviale de M)
3: $(v_i)_{1 \leqslant i \leqslant n_v} \leftarrow$ valeurs accessibles par v, triées par ordre décroissant
4: **pour tout** poids v_i **faire**
5: // Liste des insertions possibles d'arêtes de poids v_i
6: **pour tout** $C \in L_C$ **faire**
7: $A \leftarrow$ le sous-ensemble d'arêtes $a = (m_1, m_2)$ de A telles que $v(a) = v_i$, l'extrémité sortante de m_1 et l'extrémité entrante de m_2 sont libres dans C, et qu'ils sont utilisés dans des chemins différents
8: Définir des liens sur les paires d'arêtes de A qui ont des entrées différentes, et des sorties différentes // Ces liens lient les arêtes de A qui peuvent être sélectionnées conjointement
9: $C_C \leftarrow$ les cliques de A de taille maximale selon ces liens
10: **fin pour**
11: Filtrer L_C pour ne garder que les partitions qui permettent d'atteindre les plus grands cliques (c'est-à-dire. d'insérer le maximum d'arêtes de poids v_i)
12: // Construction des nouvelles solutions, en vue de la prochaine itération
13: $L'_C \leftarrow \emptyset$
14: **pour tout** $C \in L_C$ et clique $c \in C_C$ **faire**
15: Ajouter à L'_C une copie de C dans laquelle les chemins correspondants à des arêtes de c ont été concaténés
16: **fin pour**
17: $L_C \leftarrow L'_C$
18: **fin pour**
19: **renvoyer** L_C l'ensemble des partitions génomiques possibles sur G

Au fur et à mesure de l'algorithme, le nombre de solutions augmente, si plusieurs cliques de taille maximale existent, et diminue, si une des solutions de l'itération précédente ne permet pas d'inclure autant d'arêtes que les autres. Cependant, la résolution pratique se heurte à des problèmes combinatoires (recherche de cliques dans des graphes de plusieurs milliers de nœuds, et explosion du nombre de solutions). L'algorithme ne pourra être utilisé que sur des graphes d'adjacences de taille restreinte.

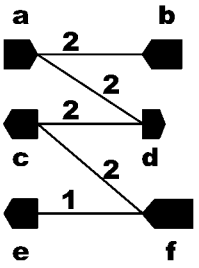


FIGURE 9.3 – Exemple de gestion litigieuse des égalités d'arêtes. Selon l'ordre choisi pour traiter les arêtes de poids 2, l'algorithme 9.1 peut choisir les arêtes (a,d) et (c,f), ce qui empêche de sélectionner ensuite (e,f), ou choisir (a,b) et (c,d) et pouvoir alors prendre (e,f). L'algorithme 9.2, en traitant les arêtes de poids 2, garde en mémoire les deux options, puis choisit la deuxième solution lors de l'étude des arêtes de poids 1.

9.3 Dans un graphe avec contraintes

Cette section décrit les variations de l'algorithme précédent, pour tenir compte d'informations supplémentaires sur la structure de la solution (arêtes fixées, ou relations à longue distance).

9.3.1 Arêtes fixées

Ici, on dispose déjà de l'ordre de certains marqueurs organisés en chemins, et on cherche à rajouter les autres marqueurs *autour* de ces chemins, en les étendant, mais sans jamais modifier leur intérieur. On se servira de cette méthode quand on aura reconstruit l'ordre d'un sous-ensemble de gènes que l'on sait contigus, et que les gènes restant doivent être ordonnés, indépendamment ou en fusionnant des chemins existants.

On dispose donc d'un ensemble génomique de chemins C_0, donné en référence, et d'un ensemble d'arêtes pondérées, et on va construire une partition génomique qui contient tous les marqueurs. En pratique, il s'agit du même algorithme de résolution que dans la version sans contraintes, mais en utilisant comme partition de départ C_0 augmentée de tous les marqueurs non référencés, et disposés en singletons. Comme l'algorithme ne fait que rajouter des arêtes, on a l'assurance que les chemins de la partition finale contiendront toujours les chemins de C_0 intacts.

9.3.2 Règles de précédence

Ici, on connaît l'architecture globale des chemins (on connaît l'ordre relatif d'un sous-jeu de gènes, pas nécessairement contigus), et on cherche à rajouter les autres marqueurs *à l'intérieur* de ces chemins, en les rendant plus précis, mais sans jamais modifier l'ordre

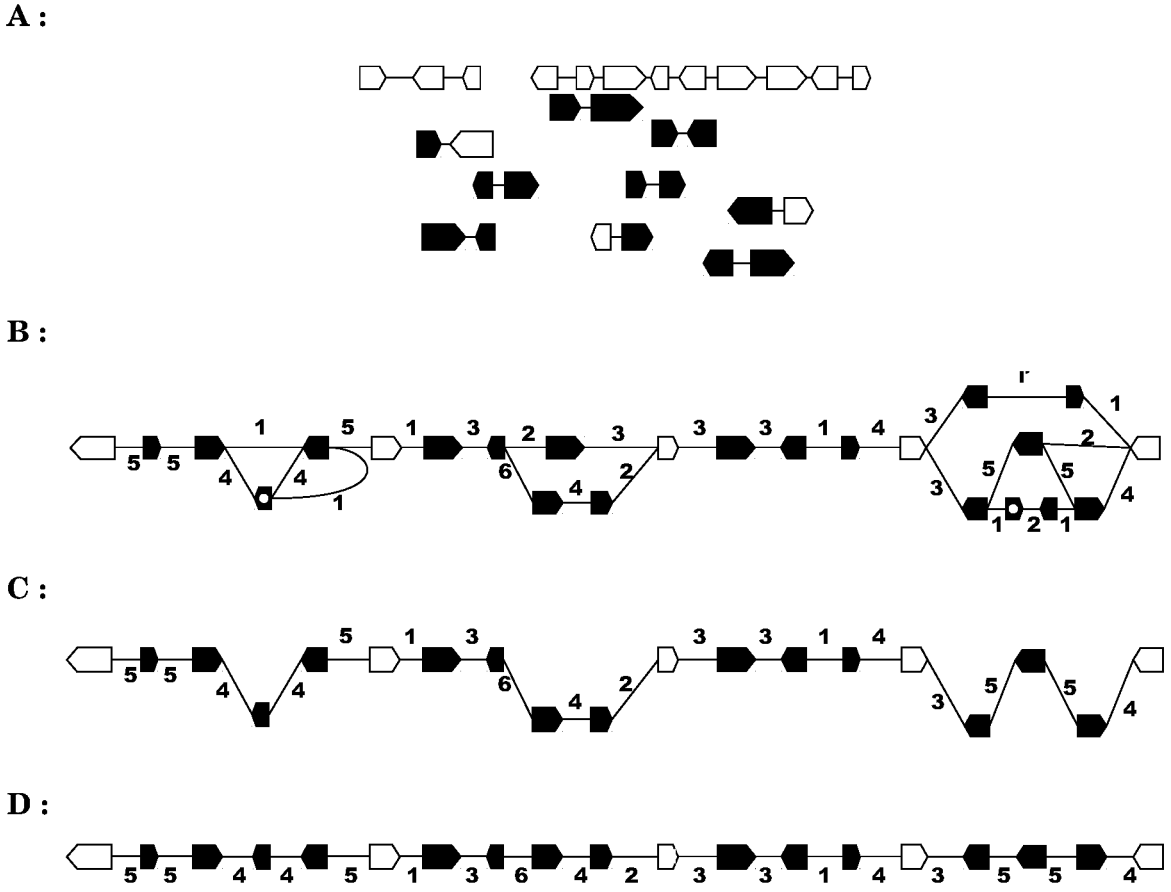


FIGURE 9.4 – Reconstruction de l'ordre ancestral des gènes avec règles de précédence. **A :** L'entrée de l'algorithme est un ensemble de chemins qui serviront de référence (en blanc), et un ensemble pondéré de paires de marqueurs orientés. **B :** Les paires de marqueurs sont placées dans un graphe dans lequel on n'étudiera que les liens entre deux marqueurs consécutifs dans un chemin de référence. La référence sert ainsi d'ossature qui va supporter l'ajout de nouveaux marqueurs, sans être elle-même remise en cause. Les marqueurs sont ici tous différents, sauf un, présent en double, marqué avec un rond blanc en surimpression. **C :** L'algorithme choisit pour chaque intervalle de référence le plus long chemin possible. En cas de conflit (marqueur devant être sélectionné dans deux intervalles), on conserve le chemin qui utilise les meilleurs poids autour de ce marqueur. **D :** Le résultat est un ordre de marqueurs s'appuyant sur une ossature robuste mais restreinte, étendue de façon à couvrir une plus grande fraction du génome ancestral.

Algorithme 9.3 Extraction d'une partition génomique dans un graphe de marqueurs orientés, avec des règles de précédence

Entrées: $G = (\overleftrightarrow{M}, A, v)$: un graphe pondéré. C_0 : partition génomique de référence. f : une fonction de sélection de chemins dans un graphe

$I \leftarrow$ Ensemble des intervalles des chemins de C_0
$C \leftarrow$ partition triviale sur M
tant que $I \neq \emptyset$ **faire**
  // Choix des meilleurs chemins avec la fonction de sélection f
  pour tout Intervalle $i_i = (m_1, m_2)$ dans I **faire**
    $C_i \leftarrow$ le chemin sélectionné par f parmi tous les chemins de G reliant m_1 à m_2 et passant uniquement par des singletons de C (si un tel chemin existe, sinon, on considère le chemin $(m_1 m_2)$)
  fin pour
  Compter, pour chaque singleton m de C, N_m : le nombre de chemins parmi les C_i passant par m (sans tenir compte des extrémités)
  // Inclure tous les chemins passant par uniquement des marqueurs n'ayant qu'un seul point d'insertion possible
  pour tout Intervalle i_i tel que $\{N_m\}_{m \in C_i} = \{1\}$ **faire**
    Ajouter C_i à C
    Retirer i_i de I
  fin pour
  // Choisir un chemin parmi ceux contenant des marqueurs en commun
  Choisir un chemin C_{i_0} contenant un marqueur en conflit
  $L \leftarrow \{C_{i_0}\}$
  Ajouter à L tous les chemins C_j qui partagent un marqueur avec un chemin C_i de L (et répéter cette mise à jour de L tant que possible)
  Associer à chaque chemin $C_i \in L$ la somme des poids des arêtes entrante et sortante des marqueurs en conflit parmi les chemins de L
  Ajouter à C le chemin C_{i_1} qui maximise ce score
  Retirer i_{i_1} de I
fin tant que
renvoyer C

relatif des gènes qu'ils contiennent. On se servira de cette méthode quand on aura reconstruit un ordre fiable sur un sous-ensemble de gènes, et que les gènes restant (moins fiables) doivent être inclus à l'intérieur des chemins déjà reconstruits, mais sans les remettre en cause. On dispose donc d'une partition génomique C_0 et on cherche à construire une partition génomique C telle que :

1. tout chemin de C possède un unique sous-chemin dans C_0 de mêmes extrémités ;
2. réciproquement, chaque chemin (non singleton) de C_0 est sous-chemin d'un unique chemin de C, de mêmes extrémités.

Autrement dit, ce sont les singletons de C_0 qui vont être insérés dans les chemins (de taille $\geqslant 2$) de C_0. Nous avons voulu mimer le comportement humain en travaillant, d'abord, indépendamment sur tous les intervalles, avant de résoudre d'éventuelles incompatibilités.

Notre solution (algorithme 9.3, illustré Figure 9.4) dépend principalement de la définition d'une fonction f de sélection de chemins. Cette fonction est appelée pour chaque intervalle de marqueurs de référence, et sélectionne un chemin, parmi tous les chemins

possibles de singletons, liant les deux marqueurs de l'intervalle. Cependant, comme cette fonction de sélection est appelée indépendamment pour chaque intervalle, elle peut faire intervenir le même marqueur dans deux chemins différents (on dit que ce marqueur est en conflit). Il est donc nécessaire de rajouter une protection qui va choisir quel est le chemin le plus probable pour chaque marqueur en conflit. Là encore, nous utilisons les poids des arêtes en calculant pour chaque chemin la moyenne du poids de l'arête entrante au marqueur en conflit, et de son arête sortante. Le chemin qui maximise ce score est sélectionné, et l'autre chemin est recalculé (nouvel appel de f, qui ne pourra donc plus utiliser le marqueur précédemment en conflit).

Cinq implémentations de f sont proposées, et permettent de paramétrer la méthode suivant les données disponibles :

1. prendre le chemin le plus long ;
2. prendre le chemin le plus court ;
3. maximiser la somme des poids des arêtes ;
4. maximiser la moyenne des poids des arêtes ;
5. maximiser les poids des arêtes choisies (comme dans la section 9.2).

9.3.3 Arêtes fixées et règles de précédence

Ici, on connaît à la fois l'ordre de certains marqueurs (contigus) et l'ordre d'autres marqueurs (non forcément contigus). Il faut *fusionner* les deux sources de données pour avoir un unique ensemble de chemins. On se servira de cette méthode quand on aura reconstruit l'ordre de sous-ensembles de gènes (séparément) et qu'on doit insérer un ensemble dans l'autre.

On dispose donc de deux ensembles génomiques de chemins, C_1 et C_2, C_1 ne contenant pas de singletons, et $C_1 \cup C_2$ devant être une partition génomique. On veut insérer les chemins de C_2 dans C_1 pour créer une partition génomique C. La solution répondra aux deux critères suivants.

1. Les chemins de C_1 sont des sous-chemins des chemins C.
2. Les chemins de C_2 sont des fragments des chemins C.

Le problème est donc de trouver un couplage entre toutes les extrémités des marqueurs de C_1, et les extrémités des chemins de C_2, ce qui définira les points d'insertion des chemins de C_2 dans C_1.

L'algorithme 9.4 (illustré Figure 9.5) associe à chaque marqueur d'un chemin de C_1, et pour chacune de ses extrémités, les chemins de C_2 auxquels il pourrait être lié. La contrainte est qu'évidemment, on ne peut inclure qu'un seul chemin à chaque extrémité de marqueur, et qu'un chemin ne peut être inséré qu'à un seul endroit. La liste des insertions possibles est considérée par ordre décroissant des poids des arêtes, et on effectue les inclusions qui sont compatibles avec les deux règles précédentes (les autres ayant donc été supplantées par d'autres insertions, de poids supérieur).

9.4 Sans information d'adjacence

Cette dernière méthode est utilisée lorsqu'on a épuisé toutes les informations des graphes. Ceci servira par exemple si on a identifié des groupes de chemins, chaque groupe

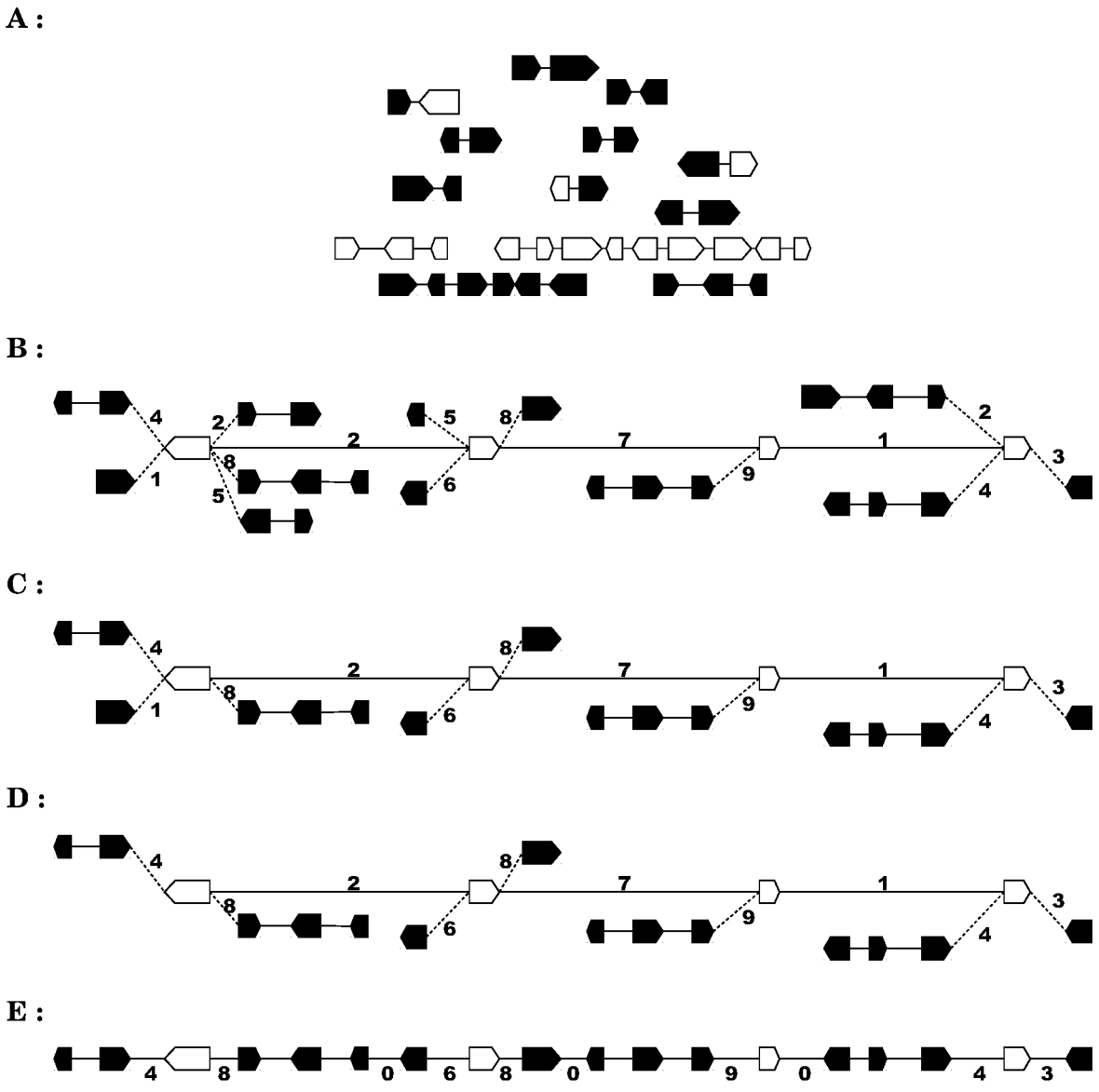


FIGURE 9.5 – Extraction d'une partition génomique dans un graphe de marqueurs orientés, avec règles de précédence et arêtes fixées. **A :** L'entrée de l'algorithme est un ensemble pondéré de paires de marqueurs orientés, et deux ensembles génomiques de chemins (tous différents) qui serviront de référence (en blanc et noir). **B :** L'algorithme construit la liste de jonctions de marqueurs blancs avec des chemins noirs, grâce aux paires de marqueurs (dessinées en pointillés). **C :** L'algorithme choisit alors successivement les jonctions internes les mieux supportées en éliminant celles correspondant à un chemin noir déjà utilisé. **D :** Enfin, la même procédure est appliquée aux jonctions aux extrémités de chromosomes. **E :** Le résultat est un ordre de marqueurs plus précis que l'ordre initial. On notera que certaines jonctions prédites peuvent n'avoir été observées dans aucune paire de marqueurs (leur poids vaut 0)

Algorithme 9.4 Extraction d'une partition génomique dans un graphe de marqueurs orientés, avec règles de précédence et arêtes fixées

Entrées: $G = (\overleftrightarrow{M}, A, v)$: un graphe pondéré. C_1 et C_2 : un ensemble de chemins génomiques de référence tels que $C_1 \cup C_2$ est une partition génomique.

$C \leftarrow C_1$ // représente la partition génomique que l'on construit
$R_m \leftarrow \emptyset$ // contient les chemins de C_2 insérés dans C
$R_c \leftarrow \emptyset$ // contient les extrémtiés des chemins de C liées à des chemins de C_2
$A \leftarrow$ arêtes (m_1, m_2) de G telles que m_1 est dans un chemin de C_1 et m_2 est une extrémité de chemin de C_2
// Test des jonctions possibles
pour tout arête $(m_1, m_2) \in A$, choisies par poids décroissant **faire**
  $c_2 \leftarrow$ chemin de C_2 commençant en m_2
  si $c_2 \notin R_c$ et $m_1 \notin R_m$ **alors**
    Insérer c_2 à côté de m_1 dans le chemin correspondant dans C
    Ajouter c_2 à R_c
    Ajouter m_1 à R_m
  fin si
fin pour
renvoyer C

correspondant à un chromosome, mais sans que l'ordre des chemins dans les groupes ne soit connu. Le but est donc d'ordonner chaque groupe de chemins.

Formellement (pour un chromosome), on dispose d'une partition génomique $C = \{c_i\}$, et on veut construire un chemin génomique c tel que tous les c_i en seront des fragments. Comme plus aucune information d'adjacence n'est disponible, nous demandons la définition d'une fonction de distance d entre les chemins, pour pouvoir invoquer une méthode de résolution (globale) par le voyageur de commerce (chapitre 4).

Dans l'instanciation du voyageur de commerce, les villes correspondront aux chemins de C et seront donc orientées, ce qui modifie légèrement la formulation du problème. En fait, il suffit de décomposer chaque chemin c_i en entrée c_i^- et sortie c_i^+, et de demander que la solution fasse intervenir les deux extrémités de chaque segment à la suite. Par exemple, pour trois chemins $\overrightarrow{c_1}$, $\overrightarrow{c_2}$, et $\overrightarrow{c_3}$, il faut que la solution soit du type $c_1^- c_1^+ c_2^- c_2^+ c_3^- c_3^+ = \overrightarrow{c_1}\overrightarrow{c_2}\overrightarrow{c_3}$ et non du type $c_1^- c_2^- c_3^+ c_1^+ c_3^- c_2^+$.

Le point crucial est donc de définir une fonction $d : C^{\pm} \mapsto \mathbb{R}^+$ qui estimera la distance entre deux extrémités de chemins, et qui favorise fortement les jonctions (c_i^-, c_i^+). Pour cela, on pose sur la fonction de distance d les conditions suivantes [2] :

$$\forall\ i,\ d(c_i^+, c_i^-) = d_0 \tag{9.1a}$$

$$\forall\ i_1 \neq i_2,\ d_0 < d(c_{i_1}^+, c_{i_2}^-) \tag{9.1b}$$

$$\forall\ i_1 \neq i_2,\ d(c_{i_1}^+, c_{i_2}^-) + d_0 < \begin{cases} d(c_{i_1}^+, c_{i_2}^+) \\ d(c_{i_1}^-, c_{i_2}^-) \end{cases} < d(c_{i_1}^-, c_{i_2}^+) - d_0 \tag{9.1c}$$

où d_0 est une constante qui sert à étalonner d (qui est bien évidemment symétrique). La dernière équation est une inégalité triangulaire appliquée à l'orientation des chemins (de

2. La deuxième équation est écrite dans le cas où la configuration optimale est $\overrightarrow{c_{i_1}}\overrightarrow{c_{i_2}}$.

manière générale, d ne doit pas vérifier l'inégalité triangulaire pour que l'on puisse appliquer le voyageur de commerce). On veillera donc à bien définir d selon ces contraintes, pour que les chemins soient correctement orientés dans la solution.

Chapitre 10

Reconstruction de l'ordre ancestral des gènes

Sommaire

Par opposition avec le chapitre 8, on pourrait appeler ce chapitre «comparaison de n génomes», puisqu'ici on va combiner le résultat des comparaisons obtenues sur des paires d'espèces pour produire un unique résultat intégré reflétant la structure du génome ancestral. Ce processus d'intégration se déroule en deux étapes principales. L'ordre ancestral des gènes est d'abord reconstruit en appelant successivement les méthodes décrites au chapitre précédent avec comme matériel d'entrée une liste de paires de gènes conservées. Cette étape génère des contigs de gènes adjacents chez l'ancêtre, et ceux-ci sont ensuite assemblés en «scaffolds» en utilisant une définition de l'adjacence plus relâchée. Enfin, nous présentons deux méthodes qui ne se fondent pas sur la notion d'adjacence, mais sur la synténie (appartenance au même chromosome) et l'éloignement pour (respectivement) regrouper des contigs en chromosomes, et ordonner des contigs dans un chromosome. Les algorithmes présentés ici travaillent indépendamment sur chaque ancêtre, évitant ainsi de propager des erreurs entre les nœuds de l'arbre. De plus, les valeurs des paramètres des programmes ne sont pas encore spécifiées, car elles feront l'objet d'une optimisation dans un chapitre suivant (chapitre 13).

10.1 Ordre ancestral des gènes – Assemblage en contigs

Dans cette première étape, on extrait d'abord l'ensemble des paires de gènes conservées entre toutes les paires de génomes. Les résultats de chaque ancêtre sont stockés dans un multi-ensemble. Un multi-ensemble permet d'associer à chaque paire de gènes conservée dans un ancêtre donné le nombre de comparaisons d'espèces modernes dans lesquelles cette paire a été observée. Pour un ancêtre A donné, en notant n_0 le nombre d'espèces extérieures (outgroups) et n_i ($i \geqslant 1$) le nombre d'espèces qui descendent de A (dans chaque

sous-arbre), le nombre de comparaisons est compris entre 0 et $\prod_{i \neq j} n_i n_j$ sauf si la paire de gène se duplique dans un génome, auquel cas ce nombre peut augmenter.

On peut appliquer les algorithmes d'extraction de chemins génomiques dans des graphes (présentés au chapitre 9) en utilisant les gènes comme marqueurs, et les paires de gènes conservées comme arêtes. Nous avons toutefois défini deux protocoles standards pour la reconstruction.

Le premier est d'utiliser l'ensemble des paires de gènes conservées en une seule passe, pour produire un jeu de contigs [1] et un ensemble de singletons. C'est l'approche la plus simple, qui fonctionne très bien en théorie mais affiche quelques limitations en pratique à cause de certains gènes (chapitre 13). Le deuxième protocole utilise un jeu de gènes dits robustes (voir sous-section 7.3.3 et sous-section 7.3.2) pour une reconstruction tout d'abord à basse résolution des chromosomes ancestraux, avant de rajouter le reste des gènes. Cette version est donc appelée multi-passes et produit une classe particulière de contigs, supportés par les gènes robustes : les supercontigs.

Plus précisément, le premier protocole («protocole 1-passe») consiste en l'application de l'algorithme 9.1 (appelé ici *de-novo*) sur l'ensemble des paires de gènes conservées (algorithme 8.1). Le deuxième protocole («protocole multi-passes») se décompose en 4 étapes (Figure 10.1) nommées *de-novo*, *affinage*, *fusion* et *insertion*. *de-novo* correspond en fait à un protocole 1-passe (algorithme 9.1), mais cette fois effectué sur un sous-jeu de gènes, dits robustes, identifiés par un des algorithmes 7.3 ou 7.2. On dispose alors d'une architecture de supercontigs basée uniquement sur des familles robustes, et dans lesquels il faut insérer les gènes non-robustes (le rôle des trois étapes restantes). Tout d'abord, l'étape d'*affinage* insère les gènes non-robustes lorsqu'ils permettent de créer des chemins d'adjacence entre les gènes robustes (algorithme 9.3). Puis les gènes non-robustes restants sont fusionnés avec, encore, l'algorithme 9.1, mais cette fois restreint à l'ensemble des gènes exclus des contigs. Enfin, ces contigs de gènes non-robustes sont réinsérés dans les supercontigs grâce à l'algorithme 9.4. C'est uniquement à la fin de cette quatrième étape que toutes les paires reconstruites dans la version 1-passe le sont aussi (Figure 10.2).

Une des limites de la méthode de reconstruction AGORA est qu'elle utilise des informations d'adjacence stricte. Ainsi, un chromosome quasi identique entre deux espèces dans lequel il manque une paire de gènes conservée, sera reconstruit en deux contigs dans leur dernier ancêtre commun. Les familles robustes sélectionnées désignent un sous-ensemble des gènes pour lesquels l'ordre (et sa conservation) est moins sensible aux artefacts d'annotation et aux micro-réarrangements affectant une seule paire de gène. Ainsi, le protocole multi-passes offre une chance de reconnaître la conservation de l'ordre à une plus grande échelle, et donc de fusionner des contigs du protocole 1-passe. En revanche, il n'a aucun effet sur les contigs composés uniquement de familles non-robustes qui restent identiques dans les deux protocoles.

On remarquera que la première étape du protocole multi-passes est en fait un protocole 1-passe appliqué sur les gènes robustes. On peut donc imaginer imbriquer plusieurs protocoles multi-passes, sur plusieurs niveaux de robustesse. L'existence des deux protocoles permet à AGORA de s'adapter aux marqueurs (leur qualité et leur nombre) utilisés pour la comparaison des génomes. Pour les reconstructions AGORA chez les vertébrés (chapitre 13), nous avons testé dans des simulations une reconstruction 1-passe sur l'ensemble des gènes, et des reconstructions multi-passes en utilisant à chaque fois un seul sous-ensemble de gènes robustes, pour définir les reconstructions optimales.

1. Le terme *contig* est utilisé par analogie avec le séquençage d'un génome, où un contig désigne un segment de séquence contigu dans le génome.

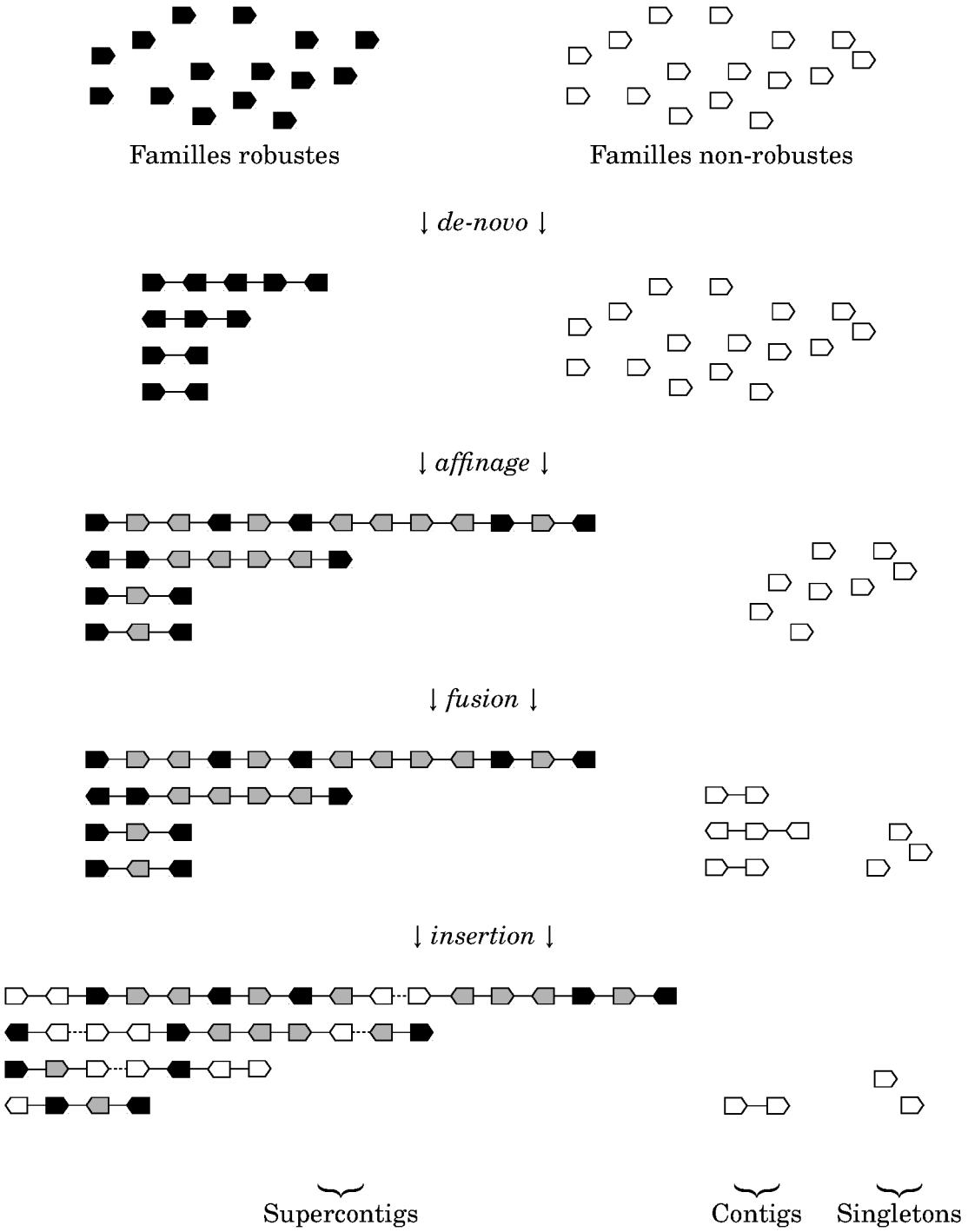


FIGURE 10.1 – Schéma récapitulatif du protocole de reconstruction multi-passes.

10.2 Adjacence de contigs – Assemblage en scaffolds

De la même manière que les génomes modernes ont été considérés comme des suites de gènes, et qu'on a reconstruit l'ordre ancestral de ces gènes, on va désormais décrire les génomes modernes comme des suites de contigs et reconstruire l'ordre ancestral de ces contigs. L'adjacence mesurée entre les contigs est de fait plus relâchée que celle entre les gènes, car elle peut s'observer, selon les espèces, entre des paires différentes de gènes des extrémités : des paires, elles, qui peuvent éventuellement ne jamais être en situation d'adjacence conservée.

Puisque reconstruire l'ordre ancestral de marqueurs a fait l'objet du paragraphe précédent, l'objet de ce paragraphe est de décrire comment reconnaître l'adjacence des contigs dans les génomes modernes. La difficulté tient dans le fait qu'un contig reconstruit n'est pas nécessairement continu dans chaque génome moderne, mais peut-être interrompu

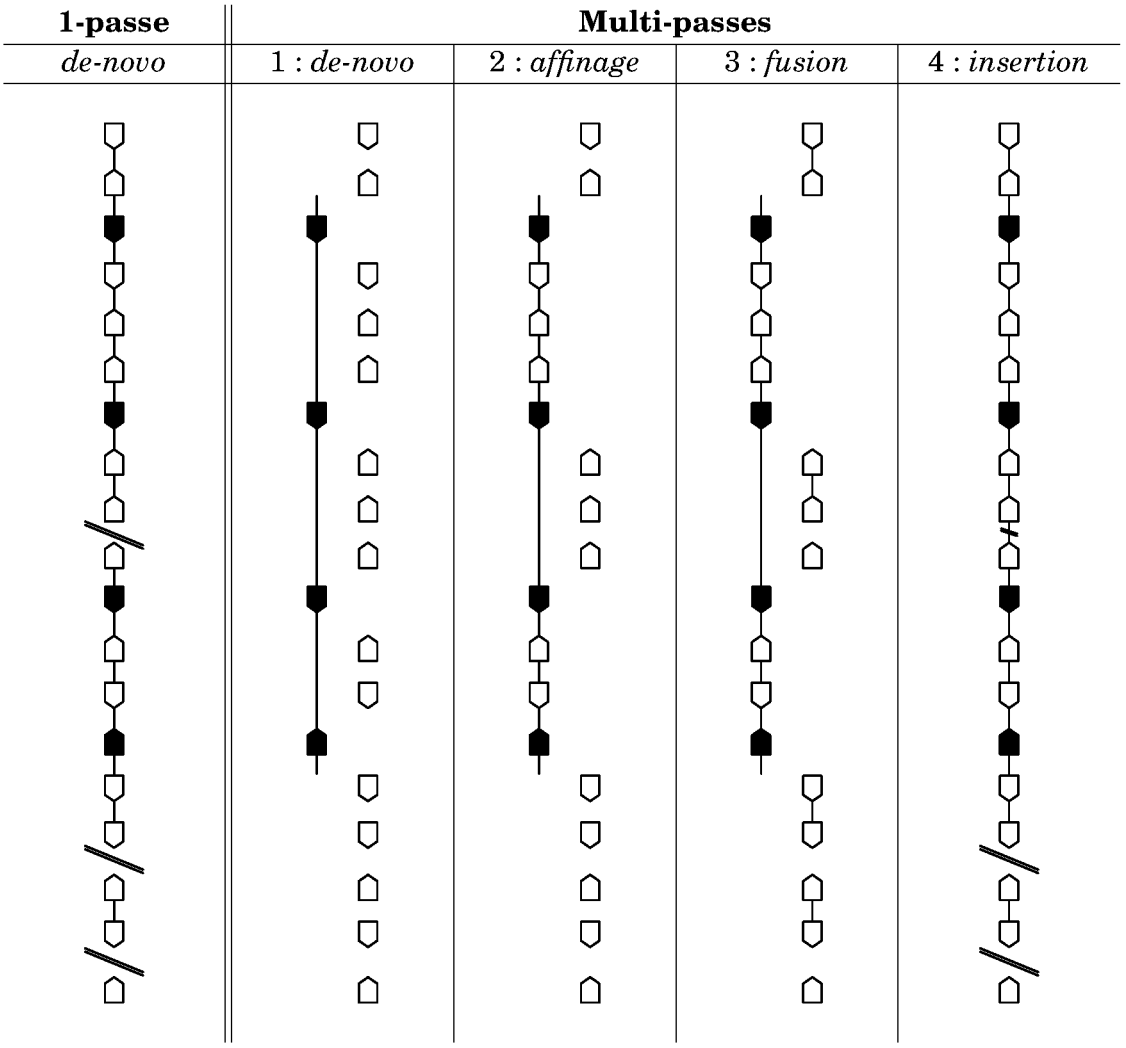


FIGURE 10.2 – Comparaison des protocoles de reconstruction 1-passe / multi-passes. Les gènes robustes sont en noir, et les non-robustes sont en blanc. Dans cet exemple, la reconstruction 1- passe (en utilisant tous les gènes en même temps) produit 18 gènes en 3 contigs, soit 15 paires, et 1 singleton. L'utilisation des familles robustes permet de fusionner les deux premiers contigs dès l'étape *de-novo*. En revanche, il faut attendre l'étape dite *insertion* pour que toutes les 15 paires se retrouvent dans la version multi-passes. Le résultat est à ce moment là 1 supercontig de 16 gènes, 1 contig de 2 gènes, et 1 singleton.

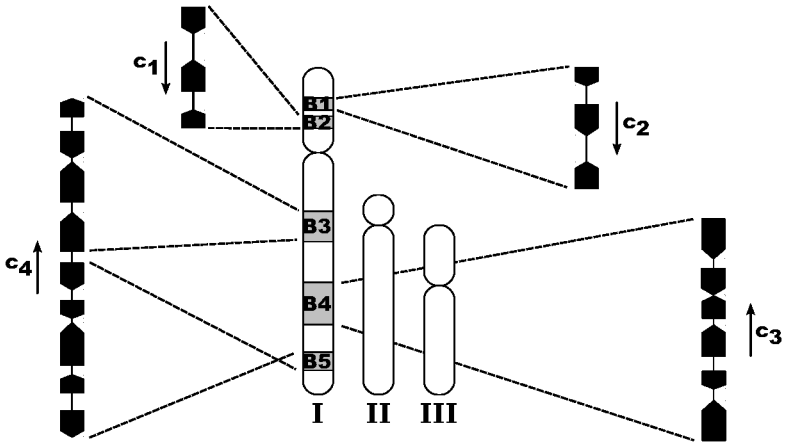
par des réarrangements. En effet, puisque les contigs sont le résultat d'une étape d'intégration de toutes les paires conservées entre tous les génomes, un contig peut inclure deux régions de chromosomes différents d'une espèce moderne.

Pour identifier des adjacences de contigs, il est nécessaire d'identifier la position des extrémités des contigs sur les génomes modernes, et d'extraire les cas où deux positions sont voisines. L'algorithme 8.2 permet d'aligner deux génomes, et donc en particulier, les contigs d'un génome ancestral sur les chromosomes d'un génome moderne. On utilise les paramètres $R = R_{\text{inters}}$ et $n_{\text{insert}} = 0$ pour identifier des ancres (des blocs de gènes contigus dans les deux génomes) tout en s'affranchissant des pertes et gains de gènes.

Les ancres sont des segments d'au moins l_a gènes colinéaires entre les contigs et les génomes modernes, qui servent à les repositionner. Par conséquent, les contigs qui pourront être utilisés contiendront au minimum l_a gènes (en pratique, $l_a = 2$ gènes). Cela exclut, de fait, les singletons de ce niveau de reconstruction (la sous-section 13.2.2 décrira des raisons de se séparer des singletons), et permet là-encore de considérer des adjacences plus relâchées qu'au niveau des gènes.

L'extraction des adjacences de contigs (Figure 10.3) dans un génome moderne se fait avec l'algorithme 10.1. L'information d'adjacence des contigs est équivalente à l'information d'adjacence des gènes. On peut donc reprendre exactement les mêmes techniques

A :



B :

Chromosome I

(B_1, B_2)	$(\overrightarrow{\text{contig}_2} - \overrightarrow{\text{contig}_1})$
(B_2, B_3)	$(\overrightarrow{\text{contig}_1} - \overleftarrow{\text{contig}_4})$
(B_3, B_4)	
(B_4, B_5)	$(\overleftarrow{\text{contig}_3} - \overrightarrow{\text{contig}_4})$

FIGURE 10.3 – Extraction des adjacences de contigs. **A :** Pour chaque génome moderne, on repositionne les contigs sur chaque chromosome. Ici est représenté un génome fictif d'espèce moderne à trois chromosomes. Quatre contigs sont alignés sur 5 segments (numérotés de B_1 à B_5) de son chromosome I. **B :** On parcourt les blocs d'alignement consécutifs en listant ceux qui correspondent à des extrémités de contigs, ce qui définit les adjacences des contigs dans cette espèce moderne.

décrites avec les gènes : intersecter les ensembles d'adjacences des génomes modernes pour extraire des adjacences conservées, et utiliser ces adjacences conservées pour reconstruire des contigs de contigs (des scaffolds[2]). Là encore, le protocole multi-passes peut apporter un avantage en échafaudant d'abord une ossature de reconstruction. Pour cela, nous avons défini deux méthodes de sélection de contigs «robustes» sur des concepts similaires aux gènes robustes.

length Étant donné une taille de contigs limite $l_{\min}$, on sélectionne pour chaque ancêtre les contigs de taille supérieure ou égale à $l_{\min}$.

proportion Étant donné une proportion p ($0 < p < 100$), on sélectionne les contigs par ordre décroissant de taille, jusqu'à englober $p\%$ des gènes ancestraux.

Algorithme 10.1 Extraction des adjacences de contigs ancestraux

Entrées: $\mathscr{G}$: le génome d'une espèce moderne. C : l'ensemble des contigs d'un ancêtre. l_a : la taille des ancres

1: // Alignement des contigs sur le génome de l'espèce moderne
2: $B \leftarrow$ blocs conservés entre $\mathscr{G}$ et C, identifiés avec l'algorithme 8.2 avec les paramètres $R = R_{\text{inters}}$ et $n_{\text{insert}} = 0$, et de taille $\geqslant l_a$.
3: **pour tout** contig c de C **faire**
4: Trouver B_{c^-} et B_{c^+} (parmi $\overleftrightarrow{B}$) les blocs les plus proches de chaque extrémité de c (éventuellement le même si c a été aligné à un seul endroit).
5: Définir $B_{\overline{\mathsf{c}}^+} = B_{\mathsf{c}^-}$ et $B_{\overline{\mathsf{c}}^-} = B_{\mathsf{c}^+}$
6: **fin pour**
7: // On parcourt les blocs consécutifs alignés pour détecter les adjacences
8: $\mathscr{A} \leftarrow \emptyset$
9: **pour tout** paire de blocs (b_1, b_2) de B, consécutifs dans $\mathscr{G}$ **faire**
10: // L'ensemble rajouté à $\mathscr{A}$ contient au plus 1 élément
11: $\mathscr{A} \leftarrow \mathscr{A} \cup \left\{(\mathsf{c}_1, \mathsf{c}_2) \in \overleftrightarrow{\mathsf{C}}^2 \text{ tels que } b_1 = B_{\mathsf{c}_1^+} \text{ et } b_2 = B_{\mathsf{c}_2^-}\right\}$
12: **fin pour**
13: **renvoyer** $\mathscr{A}$

10.3 Synténie ancestrale – Assemblage en chromosomes

Une fois les scaffolds obtenus, l'information d'adjacence (stricte ou relâchée) entre gènes (ou contigs) est nécessairement épuisée. Pour établir un ordre entre les scaffolds, il est donc indispensable de faire intervenir une mesure plus relâchée que la notion d'ordre strictement conservé. Cette mesure tient compte de la conservation de synténie (liaison physique sur un même chromosome) entre scaffolds, qui permet à son tour d'estimer la synténie ancestrale. Comme cet algorithme ne s'applique pas uniquement sur des scaffolds, mais aussi sur des contigs, et que les scaffolds sont des contigs de contigs, nous utiliserons le terme contigs dans la suite du texte.

La conservation de synténie (voir introduction) ne reflète pas nécessairement une conservation de l'ordre des marqueurs dont la synténie est conservé, mais simplement leur regroupement – dans l'espace du génome – sur un même chromosome. Cette mesure

2. Là encore, le terme *scaffold* fait référence au séquençage d'un génome, où un scaffold est un ensemble de contigs ordonnés, orientés, mais non contigus.

peut donc être utilisée pour regrouper des marqueurs, ici des contigs, sur la base de leur conservation de synténie. Nous utilisons l'algorithme *walktrap* (décrit au chapitre 5) pour calculer les regroupements optimaux entre groupes de marqueurs.

Pour utiliser *walktrap*, il faut construire un graphe dont les nœuds sont les contigs et dont les arêtes sont pondérées par les probabilités de synténies ancestrales. Comme indiqué sur la Figure 10.4, pour deux contigs donnés, des 1 et des 0 sont imposés à chaque espèce moderne selon si les contigs sont synténiques ou non. On calcule alors la probabilité de synténie ancestrale entre ces deux contigs avec l'algorithme d'interpolation (Équation 6.1) et on utilise cette valeur comme poids de l'arête les liant dans le graphe (si elle est définie, c'est-à-dire si l'information de synténie est disponible pour au moins deux sous-arbres). La Figure 10.5 montre un exemple fictif d'utilisation de *walktrap*. À partir de l'ensemble non-trié des contigs, *walktrap* permet de définir les chromosomes ancestraux grâce au calcul des probabilités de synténie ancestrale.

L'objectif de l'algorithme 10.2 est de détecter la présence de synténie dans les espèces modernes, pour ensuite définir une probabilité de synténie ancestrale grâce à l'algorithme d'interpolation (Équation 6.1), et pouvoir utiliser *walktrap*. Pour cela, il faut associer à chaque contig, et pour chaque espèce moderne, la liste (minimale) des chromosomes de cette espèce qui contiennent chacun au moins un gène du contig et ensemble au moins $p\%$ des gènes du contig (typiquement, $p = 90\%$). Pour rappel, un contig est un segment continu ancestral, mais qui peut être fragmenté dans le génome d'une espèce moderne, ce qui explique pourquoi il faut chercher, a priori, plusieurs chromosomes. Ainsi, pour un contig de 10 gènes dont 9 gènes serait sur le chromosome 1, et 1 gène sur le chromosome 2, on ne retiendra que le chromosome 1.

Il est important de souligner que le problème de la synténie ancestrale ressemble davantage à un problème d'inférence de caractère qu'à l'interpolation d'une variable conti-

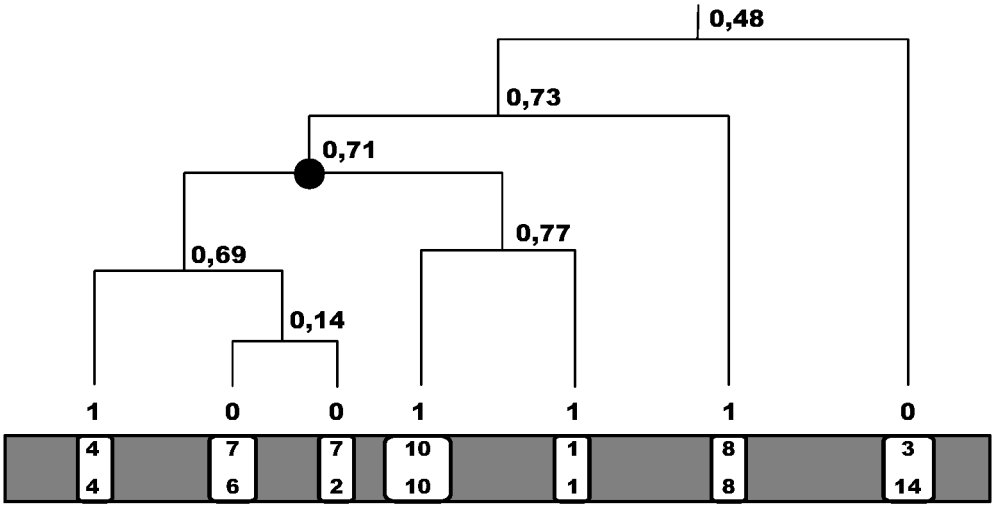
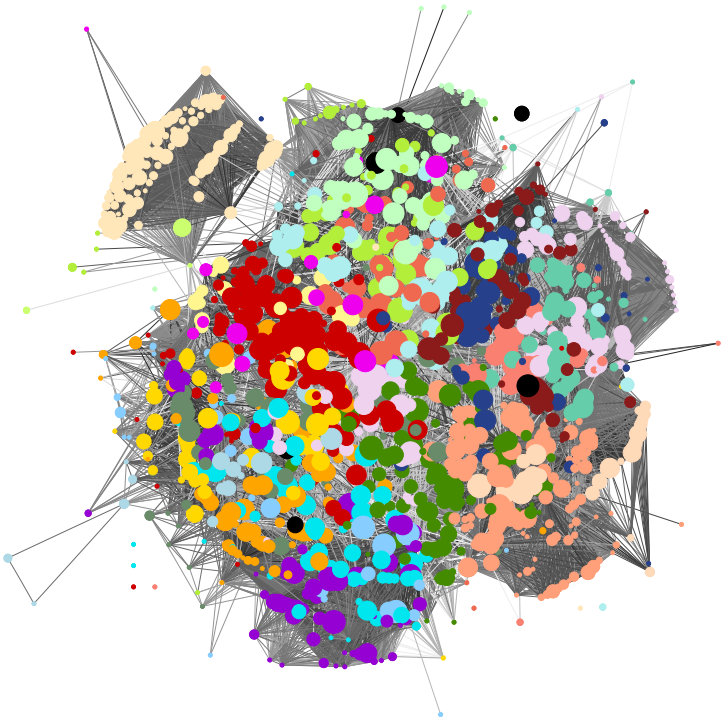


FIGURE 10.4 – Calcul de la probabilité de synténie ancestrale. Pour deux contigs donnés, on compare les chromosomes sur lesquels ils sont présents dans les génomes modernes (lignes en gris foncé) et on positionne des 1 et des 0 sur les génomes modernes selon si la synténie y existe ou non. Par exemple, dans l'espèce située à l'extrémité gauche de l'arbre, les deux contigs sont sur le chromosome 4 de cette espèce, et cette situation est reflétée par un 1. Dans les deux espèces voisines, les contigs sont sur des chromosomes différents (7 et 6, et 7 et 2), symbolisé par des 0. La probabilité de synténie ancestrale est donnée par interpolation selon l'Équation 6.1.

A :



B :

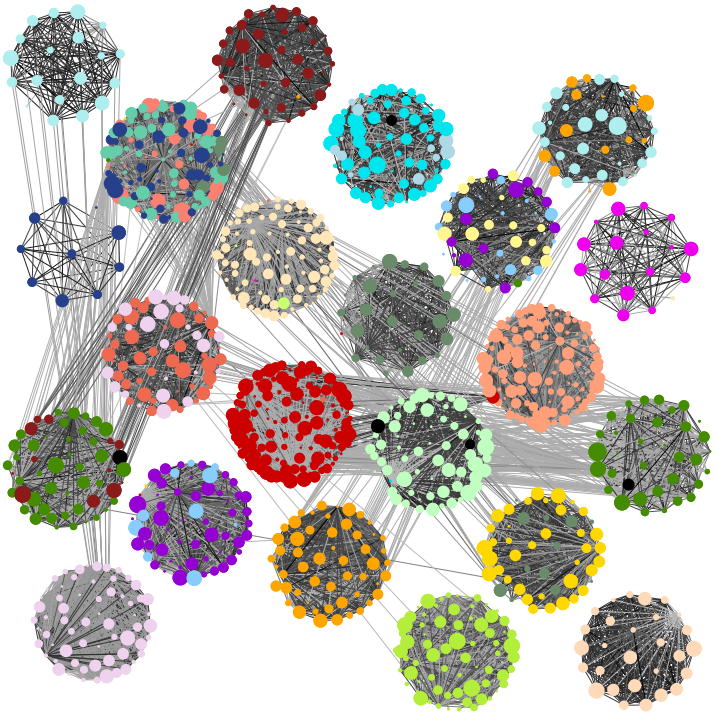


FIGURE 10.5 – Schéma de fonctionnement de *walktrap*. **A :** Données brutes de liens entre contigs. Les cercles indiquent les contigs qui doivent être clusterisés, colorés par un code indiquant le chromosome principal (d'une espèce moderne de référence) dont ils sont issus. Les liens relient les contigs ayant une probabilité de synténie ancestrale plus grande qu'un seuil arbitraire. **B :** Clusters définis par *walktrap*, représentés par les 22 regroupements circulaires, chacun équivalent à un chromosome ancestral. Les arêtes sont désormais concentrées à l'intérieur des chromosomes.

Algorithme 10.2 Regroupement d'un ensemble de contigs en chromosomes ancestraux

Entrées: $\mathscr{A}$: arbre phylogénétique des espèces. $\{\mathscr{G}_e\}$: l'ensemble des génomes de chaque espèce moderne, notée e. Anc, l'ancêtre sur lequel on travaille. C : l'ensemble des contigs de Anc à regrouper en chromosomes. p : seuil de sensibilité ($0 < p \leqslant 1$).

// Associations contig ↔ chromosomes d'espèces modernes
pour tout $\mathsf{c} \in \mathsf{C}$ **faire**
  pour tout espèce moderne e **faire**
    $l \leftarrow$ liste des chromosomes de e, qui contiennent chacun au moins 2 gènes de c
    $n \leftarrow$ nombre de gènes de c présents sur un chromosome de l
    $L_{\mathsf{c},e} \leftarrow$ sous-ensemble minimal de chromosomes de l, qui contient pn gènes de c (éventuellement $\emptyset$)
  fin pour
fin pour
// Graphe contenant les probabilités de synténie ancestrale
Définir un graphe $\mathsf{G} = (\mathsf{C}, \mathsf{A}, \mathsf{v})$ non orienté, pondéré, et initialement vide ($\mathsf{A} = \emptyset$)
pour tout $(\mathsf{c}_1, \mathsf{c}_2) \in \mathsf{C}^2,\ \mathsf{c}_1 \neq \mathsf{c}_2$ **faire**
  // Calcul de la probabilité de synténie ancestrale
  pour tout espèce moderne e **faire**
    si $L_{\mathsf{c}_1,e} \neq \emptyset$ et $L_{\mathsf{c}_2,e} \neq \emptyset$ **alors**
      Définir $S_e = \begin{cases} 1 & \text{si } L_{\mathsf{c}_1,e} \cap L_{\mathsf{c}_2,e} \neq \emptyset \\ 0 & \text{sinon} \end{cases}$
    fin si
  fin pour
  $S' \leftarrow$ interpolation (Équation 6.1) des S_e sur $\mathscr{A}$
  si S'_{Anc} est défini **alors**
    Rajouter l'arête $\{\mathsf{c}_1, \mathsf{c}_2\}$ à A
    Définir $\mathsf{v}(\{\mathsf{c}_1, \mathsf{c}_2\}) = S'_{\text{Anc}}$
  fin si
fin pour
renvoyer Regroupement effectué par *walktrap* du graphe G

nue. Nous préférons cependant garder des valeurs continues car *walktrap* utilise les poids des arêtes pour guider son choix, sans toutefois le forcer. Ainsi, même si on mettait directement le résultat de l'inférence de caractère dans le graphe donné à *walktrap* (des 1 et des 0), rien ne garantirait que les ensembles donnés en résultat seraient totalement compatibles avec (il faudrait pour cela que le graphe soit une union disjointe de graphes complets). D'autre part, avec des 1 et des 0, on perd toute information de confiance dans les synténies ancestrales (information que l'on retrouve dans la probabilité ancestrale interpolée). Néanmoins, il est envisageable d'utiliser l'information de confiance renvoyée par un programme d'inférence de caractère et cela fait partie des développements possibles de la méthode AGORA (voir discussion).

10.4 Ordre des contigs sur un chromosome

Dans le résultat de *walktrap*, chaque chromosome ancestral est un ensemble de contigs. Chaque chromosome est donc désigné par les contigs qu'il contient, sans information sur l'ordre de ces contigs à l'intérieur du chromosome. La méthode de la section 9.4

permet d'ordonner des contigs que l'on sait appartenir à un même chromosome, en modélisant un problème de voyageur de commerce, à condition de définir une distance d entre les contigs. On rappelle que, de plus, d doit vérifier les inégalités exprimées dans l'Équation 9.1 pour que *concorde* puisse orienter les contigs.

Algorithme 10.3 Définition dans un génome moderne d'une distance d entre contigs

Entrées: $\mathscr{G}$: le génome d'une espèce moderne. C : l'ensemble des contigs à unifier dans un seul chemin. l_a : la taille des ancres.

$B \leftarrow$ blocs conservés entre $\mathscr{G}$ et C, identifiés avec l'algorithme 8.2 avec les paramètres $R = R_{\text{inters}}$ et $n_{\text{insert}} = 0$, et de taille $\geqslant l_a$.
pour tout contig c de C **faire**
    Trouver B_{c^-} et B_{c^+} (parmi $\overleftrightarrow{B}$) les blocs les plus proches de chaque extrémité de c (éventuellement le même).
fin pour
$\mathscr{G}' \leftarrow \mathscr{G}$ écrit uniquement à partir de $\{B_{c^-}, B_{c^+}\}_{c \in C}$
pour tout $(c_1, c_2) \in C^2$, $c_1 \neq c_2$ **faire**
    Extraire les positions dans $\mathscr{G}'$ de $B_{c_1{}^+}$, $B_{c_1{}^-}$, $B_{c_2{}^+}$, $B_{c_2{}^-}$
    pour tout paire d'extrémités (c_1^*, c_2^*) **faire**
        Définir $d_{c_1^*, c_2^*} = \begin{cases} \infty \text{ si les extrémités sont sur deux chromosomes différents} \\ \text{(nombre d'intervalles les séparant) sinon} \end{cases}$
    fin pour
    si les $d_{c_1^*, c_2^*}$ ne vérifient pas l'Équation 9.1 **alors**
        Remplir les $d(c_1^*, c_2^*)$ avec ∞
    fin si
fin pour
renvoyer d

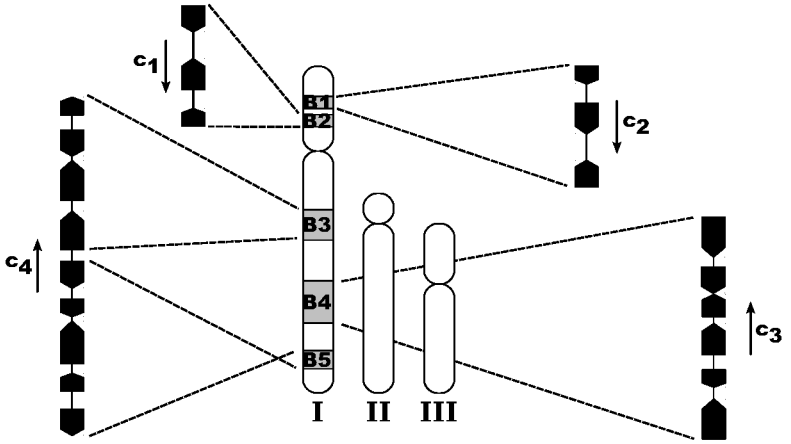
Algorithme 10.4 Ordre ancestral de contigs selon la méthode du voyageur de commerce

Entrées: $\mathscr{A}$: arbre phylogénétique des espèces. $\{\mathscr{G}_e\}$: le génome de chaque espèce moderne e. Anc, l'ancêtre sur lequel on travaille. C : l'ensemble des contigs de Anc à unifier dans un seul chemin. l_a : la taille des ancres.

pour tout espèce moderne e **faire**
    Définir d_e la fonction de distance spécifique à e, avec l'algorithme 10.3 avec $(\mathscr{G}_e, C, l_e)$ en paramètre.
fin pour
Définir $d : C^{\pm 2} \mapsto \mathbb{R}^+$ par :
pour tout $c \in C$ **faire**
    $d(c^-, c^+) = d(c^+, c^-) = 0$
fin pour
pour tout $(c_1^*, c_2^*) \in C^{\pm 2}$, $c_1 \neq c_2$ **faire**
    $d \leftarrow$ interpolation (Équation 6.1) des $d_e(c_1^*, c_2^*)$ sur $\mathscr{A}$
    Définir $d(c_1^*, c_2^*)$
fin pour
renvoyer le résultat de la résolution du voyageur de commerce (par *concorde*)

L'algorithme 10.3 (Figure 10.6) permet de définir une distance entre les contigs pour une espèce moderne e donnée. Il est sur le début très similaire à l'algorithme 10.1 car

A :



B :

Chromosome I : $\underset{C_{2^-}}{B_1}\ \underset{C_{2^+}}{}\quad \underset{C_{1^-}}{B_2}\ \underset{C_{1^+}}{}\quad \underset{C_{4^+}}{B_3}\quad \underset{C_{3^+}}{B_4}\ \underset{C_{3^-}}{}\quad \underset{C_{4^-}}{B_5}$

C :

		−+	+−	++	−−
C1	C_2	1	3	2	2
	C_3	3	3	2	4
	C_4	2	4	1	5
C_2	C_3	5	5	4	6
	C_4	4	6	3	7
C_3	C_4	∞	∞	∞	∞

D :

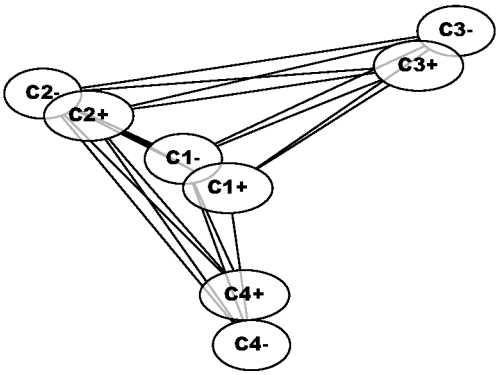


FIGURE 10.6 – Calcul des distances pour *concorde*. **A :** Positionnement de quatre contigs sur le chromosome I d'une espèce moderne (même figure qu'en Figure 10.3). **B :** Transformation de l'écriture du chromosome comme une suite de blocs alignés en une suite d'extrémités de contigs. **C :** Matrice de distance déduite de **B**. Les deux premières colonnes indiquent les contigs, et les quatre autres les extrémités entre lesquelles on mesure la distance. Pour C_3 et C_4, les distances lues ne vérifient pas l'Équation 9.1, il n'y aura donc pas de distance (sur cette espèce moderne). **D :** Dessin du graphe (avec l'outil *graphviz*) issu de la matrice de distance de cette espèce. On retrouve que C_1 et C_2 se suivent, et que C_1 est lié à C_3 ou à C_4, mais que la décision n'est pas encore possible.

il cherche aussi à positionner les extrémités de chaque contig de C. Après, on réduit le génome de l'espèce moderne à une suite d'extrémités de contigs et on définit une distance d_e entre deux extrémités en comptant le nombre d'intervalles qui les séparent. On ne tient pas compte des paires de contigs pour lesquels les distances calculées ne satisfont pas l'Équation 9.1, car cela signifie que les contigs correspondants sont mélangés et qu'il est impossible sur ce génome moderne de définir des distances. On espère alors que dans d'autres génomes, on sera capable de les comparer. Ensuite, avec l'algorithme 10.4, il suffit pour chaque paire d'extrémités de contigs d'interpoler les distances d_e pour avoir une estimation de la distance ancestrale entre ces deux extrémités. *Concorde* peut alors être appelé pour ordonner les extrémités de contigs, et donc les contigs eux-mêmes, si d a été définie correctement, pour donner l'ordre ancestral des contigs au sein d'un chromosome ancestral.

Chapitre 11

Duplications complètes de génomes

Sommaire

Les duplications de génomes (introduites en section 2.4) laissent des marques indélébiles dans les génomes et sont identifiables grâce aux relations de paralogie entre les gènes. Dans un génome ayant subi une duplication complète, toute région possède une région paralogue. Identifier de telles paires de régions permet de décrire une unique région pré-duplication, et donc de reconstruire des fragments d'un génome ancestral (l'ordre des gènes n'est généralement pas prédictible). Deux approches permettent ceci, à des niveaux de résolution différents :

- l'identification directe de paires de régions liées par des paires de gènes ohnologues (les fragments ancestraux reconstruits ne contiennent alors que des paires de gènes ohnologues) ;
- l'identification de synténie dédoublée grâce à un génome n'ayant pas subi la duplication (on peut ici inclure dans la version ancestrale des gènes conservés en une seule copie).

Il n'existe pas d'implémentation générique de ces deux méthodes. Nous avons donc développé de nouveaux programmes pour les inclure dans AGORA, en suivant les raisonnements déjà décrits dans des articles, et en les adaptant pour les rendre compatibles avec les scénarios d'utilisation d' AGORA.

Comme un événement de duplication se produit sur une branche de l'arbre phylogénétique, les paralogues issus d'une duplication sont identifiés en comparant les listes de gènes ancestraux du nœud précédant la duplication au nœud qui la suit. Si sur la branche considérée se sont produites plusieurs duplications complètes successivement, on ne pourra alors pas les distinguer. De plus, en raison des difficultés à identifier précisément les nœuds de duplication (sous-section 13.2.1), on préfèrera utiliser les nœuds terminaux à la place du nœud suivant la duplication. Enfin, par abus de langage, un génome non-dupliqué désignera un génome qui n'a pas subi la (les) duplication complète que l'on étudie.

11.1 Sans espèce non-dupliquée

La Figure 11.1 (tirée de Nakatani *et al.* [2007]) montre le processus de reconstruction que nous allons mettre en place (la figure se place dans le contexte des deux duplications complètes de génome chez les vertébrés). En **A** est représentée la comparaison de 8 chromosomes post-duplication (issus de 2 chromosomes pré-duplication). Initialement, les gènes sont toujours en quatre copies, et dans le même ordre le long des chromosomes. On peut facilement grouper les chromosomes par paquets de 4, chaque paquet indiquant un chromosome pré-duplication. Au fur et à mesure du temps (de **B** à **D**), les réarrangements remodèlent l'ordre des gènes dans les chromosomes, puis entre les chromosomes, jusqu'à arriver à un mélange de fragments des chromosomes initiaux post-duplication (amplifié par une numérotation imprévisible des chromosomes de l'espèce moderne). À ce moment, les paquets initiaux de 4 chromosomes sont fragmentés en de nombreux segments (plus que 4, en général) éparpillés sur le génome. La procédure de reconstruction doit remonter le temps tout d'abord (**E**) en identifiant ces segments humains liés par des ohnologues, et donc fragments de chromosomes ancestraux. Cela implique une étape de découpage du génome humain en segments non-réarrangés (découpage du chromosome 1 en $1a$, $1b$, et $1c$ dans l'exemple de **E**). Un clustering (**F**) peut regrouper les segments humains liés par des ohnologues, ce qui va permettre de reformer les paquets assimilables aux chromosomes pré-duplication. Enfin (**G**), il est possible, dans chaque paquet, de répartir les nombreux segments qui le composent en 4 sous-paquets, chacun d'entre eux formant un chromosome post-duplication, sous le postulat qu'un chromosome post-duplication ne possède pas d'ohnologues avec lui-même, mais avec les 3 autres.

La méthode que nous voulons mettre en place doit être capable de gérer d duplications complètes successives, en utilisant un génome dupliqué non nécessairement moderne (pour les reconstructions liées aux 2R, nous allons utiliser l'ancêtre *Amniota* comme base), et potentiellement fort fragmenté. La procédure de reconstruction est donc composé des trois étapes suivantes, chacune décrite dans une sous-section.

1. Découper le génome de l'espèce dupliquée en fragments non-réarrangés, selon la propriété qu'un segment non-réarrangé est lié à des ohnologues uniformément sur sa longueur, et qu'à l'inverse, la juxtaposition de deux segments d'origines pré-duplication différentes aura une distribution des ohnologues différentes selon les côtés.
2. Clusteriser avec *walktrap* les segments de l'espèce dupliquée liés par des ohnologues pour reformer les chromosomes pré-duplication.
3. Au sein de chaque cluster, reformer 2^d sous-paquets, correspondant aux chromosomes post-duplication en maximisant pour chaque sous-paquet le nombre d'ohnologues avec les $2^d - 1$ autres sous-paquets, et en minimisant le nombre d'ohnologues intra-sous-paquet. La méthode de Nakatani *et al.* [2007] consiste en un test de toutes les partitions possibles en 2^d ensembles, pour identifier la configuration optimale. Cette approche est en complexité exponentielle, et fonctionnait dans leur cas car le nombre de segments par cluster était d'au maximum 20. Nous aurons à traiter plusieurs centaines de fragments, et avons donc du modifier l'implémentation de cette étape de la reconstruction.

11.1.1 Découpage d'un génome dupliqué

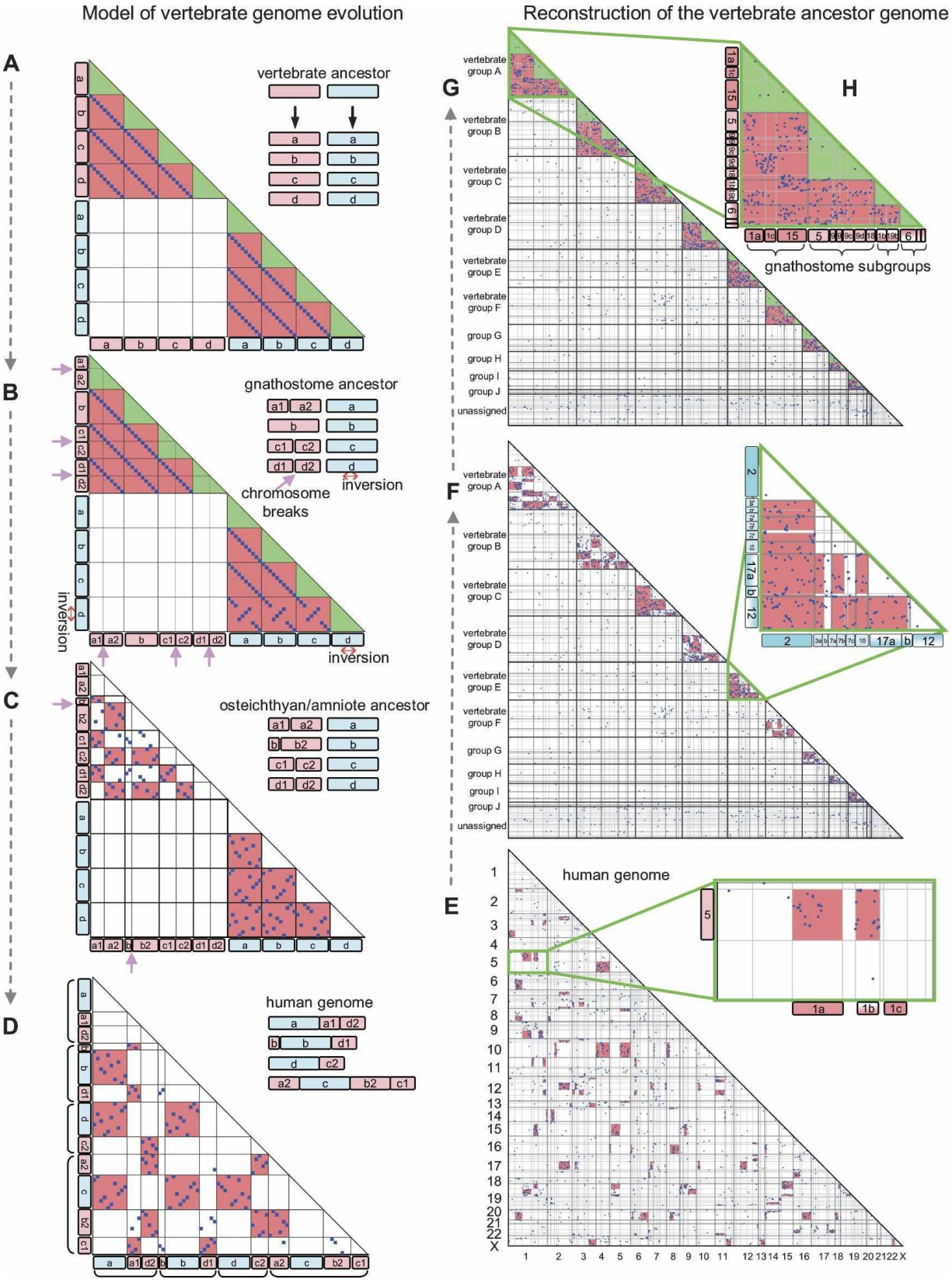


FIGURE 11.1 – Reconstruction de génomes pré- et post-duplication uniquement sur la base d'ohnologues (figure tirée de Nakatani *et al.* [2007]). La colonne de gauche (de haut en bas) décrit l'évolution de deux chromosomes après deux duplications complètes. La colonne de droite (de bas en haut) reprend les mêmes étapes, mais en remontant le temps à partir du génome humain, jusqu'à reconstruire les génomes ancestraux pré- et post-duplication. Les cases coloriées en rouge lient des segments de chromosomes liés par significativement beaucoup d'ohnologues. (voir texte pour des explications détaillées)

Algorithme 11.1 Découpage d'un génome en ohnologons

Entrées: $\mathscr{G}^{\mathscr{D}}$: le génome de l'espèce à découper. $\mathscr{L}_{\text{para}}$: la liste des paires de gènes ohnologues de $\mathscr{G}^{\mathscr{D}}$. $p_{\min}$: le seuil de probabilité en dessous duquel on considère que la distribution des gènes ohnologues est différente.

$R \leftarrow \emptyset$ // Contient la liste des intervalles découpés au maximum
pour tout chromosome c de $\mathscr{G}^{\mathscr{D}}$ **faire**
  $S \leftarrow \{(1, |\mathscr{G}^{\mathscr{D}}_c|)\}$ // Contient la liste des intervalles à essayer de découper
  tant que $S \neq \emptyset$ **faire**
    Retirer un élément de (x_1, x_2) de S
    pour tout $x_1 \leqslant x < x_2$ **faire**
      pour tout chromosome c' de $\mathscr{G}^{\mathscr{D}}$ **faire**
        $n_{1,c'} \leftarrow$ nombre de gènes de $\mathscr{G}^{\mathscr{D}}_{c,x_1 \to x}$ ayant un ohnologue sur c' d'après $\mathscr{L}_{\text{para}}$
        $n_{1,c'} \leftarrow$ nombre de gènes de $\mathscr{G}^{\mathscr{D}}_{c,x+1 \to x_2}$ ayant un ohnologue sur c' d'après $\mathscr{L}_{\text{para}}$
      fin pour
      $s_1 \leftarrow \sum_{c'} n_{1,c'}$
      $s_2 \leftarrow \sum_{c'} n_{2,c'}$
      // Critère de Cochran
      $C \leftarrow$ chromosomes c' tels que $(n_{1,c'} + n_{2,c'}) \min(s_1, s_2)/(s_1 + s_2) > 5$
      $p_x \leftarrow$ probabilité d'indépendance des échantillons $n_{1,c'}$ et $n_{2,c'}$ (pour $c' \in C$) selon le test du χ^2
    fin pour
    si $\min_x\{p_x\} < p_{\min}$ **alors**
      // Le minimum est atteint en x^*
      $S \leftarrow S \cup \{(x_1, x^*), (x^* + 1, x_2)\}$
    sinon
      // Aucune coupure n'est pertinente
      $R \leftarrow R \cup \{(c, x_1, x_2)\}$
    fin si
  fin tant que
fin pour
renvoyer R

La première étape de la méthode permet de démêler les réarrangements récents en utilisant la distribution des gènes ohnologues.

En effet, d'après la Figure 11.1.D, les segments humains d'origine pré-duplication «bleue» sont liés entre eux par des ohnologues, tout comme les segments d'origine «rose», mais aucun ohnologue ne relie un segment rose à un segment bleu. Ainsi, un chromosome humain qui contient deux segments d'origine rose et bleue a ses gènes ohnologues qui se distribuent de manière différentielle selon les deux segments, et nous pouvons utiliser ce signal pour les délimiter.

L'algorithme 11.1 scanne chaque chromosome de l'espèce dupliquée est essaie de le découper en deux, en utilisant un test de χ^2 (associé à un seuil de significativité $p_{\min}$) pour identifier une différence de répartition des gènes ohnologues. Si un chromosome est découpé, l'algorithme recommence la procédure sur chacun des deux intervalles créés, au cas où il faudrait découper le chromosome en plus de deux segments. Le résultat est une partition de chaque chromosome en intervalles.

11.1.2 Appariement en chromosomes pré-duplication

Algorithme 11.2 Groupement de segments de chromosomes modernes en chromosomes pré-duplication

Entrées: $\mathscr{G}$: le génome de l'espèce à découper. $\mathscr{L}_{\text{para}}$: la liste des paires de gènes ohnologues de $\mathscr{G}$.

1: $R \leftarrow$ Résultat de l'algorithme 11.1
2: Définir un graphe complet $G = (R, A, v)$ non orienté, pondéré, de poids non spécifiés
3: $n_g \leftarrow$ nombre de gènes de $\mathscr{G}$ ayant un ohnologue dans $\mathscr{L}_{\text{para}}$
4: $p \leftarrow |\mathscr{L}_{\text{para}}|/(n_g(n_g-1))$
5: **pour tout** $(s_1, s_2) \in R^2,\ s_1 \neq s_2$ **faire**
6: $\quad n_{1,2} \leftarrow$ nombre de ohnologues entre les segments s_1 et s_2 de $\mathscr{G}$, d'après $\mathscr{L}_{\text{para}}$
7: $\quad$ Définir $v(\{s_1, s_2\}) =$ probabilité d'après une loi binomiale $\mathscr{B}(|s_1||s_2|; p)$ d'avoir plus de $n_{1,2}$ ohnologues entre s_1 et s_2
8: **fin pour**
9: Appeler *walktrap* sur G
10: **renvoyer** Regroupement effectué par *walktrap* du graphe G

Dans un deuxième temps donc, il faut rassembler les segments de chromosomes ohnologues. Pour cela, il suffit de définir une mesure de similarité entre segments (partagent-ils beaucoup de gènes ohnologues ?) et de les clusteriser en fonction de cette mesure.

L'algorithme 11.2 calcule d'abord le taux moyen de paires d'ohnologues entre deux segments. Le reste de l'algorithme consiste à comparer pour chaque segment le nombre d'ohnologues qui les relient au nombre attendu selon une distribution uniforme des paires, pour établir une mesure qui permettra, grâce à *walktrap*, de les clusteriser en chromosomes pré-duplication. Nous avons choisi comme mesure $-\log p$, où p est la probabilité issue d'un test de conformité à une loi binomiale (paramétrée par le taux moyen d'ohnologues).

11.1.3 Séparation en chromosomes post-duplication

Grâce à la section précédente, nous disposons de chromosomes pré-duplication, et il reste à les séparer en 2^d paquets pour disposer des chromosomes post-duplication. Le but est donc de trouver (Figure 11.1.H.), pour chaque chromosome pré-duplication, la partition en 2^d ensembles qui maximise une répartition des ohnologues entre les ensembles (zone rouge sur la figure), et minimise une répartition interne des ohnologues dans les ensembles (zone verte sur la figure).

Le sélection de la meilleure partition se fait avec un test exact de Fisher. Pour chaque partition $(P_1 \dots P_k)$, nous appliquons ce test statistique sur une table de contigence 2x2 définie comme :

$$\begin{pmatrix} \sum_{i \neq j} n_{\text{para}}(P_i, P_j) & \sum_{i \neq j} |P_i||P_j| \\ \sum_i n_{\text{para}}(P_i, P_i) & \sum_i |P_i|^2 \end{pmatrix}$$

où $|P_i|$ désigne somme des tailles des segments inclus dans P_i, et n_{para} renvoie le nombre d'ohnologues entre deux ensembles.

Une probabilité proche de 0 indique que les paralogues sont répartis très différemment d'une distribution uniforme, et que donc les P_i contiennent peu d'ohnologues et les paires (P_i, P_j) en contiennent beaucoup. Plus proche de 0 est la probabilité, plus la répartition des ohnologues est baisée, et plus la partition est considérée comme optimale.

L'algorithme doit tester les partitions possibles des n segments en 2^d ensembles. Le nombre de partitions possibles de n objets en k ensembles ($n \geqslant k$) est appelé nombre de Stirling de deuxième espèce et est noté $S(n,k)$ ou $\left\{ {n \atop k} \right\}$.

Il est défini par la relation de récurrence $\left\{ {n \atop k} \right\} = \left\{ {n-1 \atop k-1} \right\} + k \left\{ {n-1 \atop k} \right\}$ avec $\left\{ {n \atop n} \right\} = \left\{ {n \atop 1} \right\} = 1$

La formule générale donne $\left\{ {n \atop 2} \right\} = 2^{n-1} - 1$ et $\left\{ {n \atop 4} \right\} = \frac{1}{6}(4^{n-1} - 1) - \frac{1}{2}(2^{n-1} - 3^{n-1})$

Ces nombres augmentent très rapidement avec n : $\left\{ {50 \atop 2} \right\} \approx 5 \cdot 10^{14}$ et $\left\{ {50 \atop 4} \right\} \approx 5 \cdot 10^{28}$

Dans le cas général, un test explicite de toutes les combinaisons n'est donc pas possible et nous avons choisi une heuristique gloutonne pour parcourir une partie de l'ensemble des solutions. L'algorithme choisit les meilleures solutions possibles avec les fragments les plus longs (ceux-ci donnent le contenu principal de chaque chromosome post-duplication) puis rajoute par vagues les fragments suivants par ordre décroissant de taille.

L'approche gloutonne sélectionne les m_1 fragments de chromosomes les plus longs, en teste toutes les combinaisons (il y en a $\left\{ {m_1 \atop 2^d} \right\}$), et conserve les s meilleures. Puis, des paquets de m_2 fragments sont ajoutés successivement, en testant les $(2^d)^{m_2}$ combinaisons possibles, et en sélectionnant les s meilleures, et ainsi de suite, jusqu'à avoir ajouté tous les fragments. Le résultat est donc une partition des segments en chromosomes post-

Algorithme 11.3 Groupement de segments de chromosomes modernes en chromosomes post-duplication

Entrées: G : groupes (pré-duplication) de segments de chromosomes. m_1 : nombre de segments à combiner parmi les plus longs. m_2 : nombre de segments à rajouter à chaque itération. s : nombre de partitions à garder d'une itération à l'autre. d : nombre de duplications à parcourir.

Retirer les m_1 segments les plus longs de G et les stocker dans L
pour tout Partition en $P = (P_1 \dots P_{2^d})$ de L **faire**
    $p_P \leftarrow$ probabilité d'indépendance de la table de contigence associée à P selon un test exact de Fisher
fin pour
$S \leftarrow s$ meilleures partitions selon p
tant que $G \neq \emptyset$ **faire**
    Réinitialiser p
    Retirer les m_2 segments les plus longs de G et les stocker dans L
    pour tout $I \in [1, 2^d]^{m_2}$ et $P \in S$ **faire**
        pour tout $1 \leqslant i \leqslant m_2$ **faire**
            Rajouter le i-ème segment de L à P_{I_i}
        fin pour
        $p_P \leftarrow$ probabilité d'indépendance de la table de contigence associée à P selon un test exact de Fisher
    fin pour
    $S \leftarrow s$ meilleures partitions selon p
fin tant que
renvoyer Meilleure partition de S

duplication.

11.2 Avec espèce non-dupliquée

L'usage de la synténie dédoublée permet d'inclure dans la reconstruction des gènes ayant perdu leur deuxième copie après la duplication (puisqu'on ne se limite plus aux ohnologues). En revanche, cela implique d'utiliser le génome d'une espèce non-dupliquée, avec laquelle la synténie est suffisamment conservée. Chez les vertébrés, cette approche est possible avec la duplication spécifique des poissons, mais pas avec le double épisode de duplication qui a précédé la radiation des vertébrés.

L'implémentation décrite dans cette section généralise la définition couramment faite de la synténie dédoublée. Compte-tenu des réarrangements dans l'espèce dupliquée utilisée, et de la présence potentielle de plus d'une duplication complète, un segment de chromosome de l'espèce non-dupliquée alternera, en général, entre k segments de l'espèce dupliquée ($k \geqslant 2$). Cette alternance multiple sera justifiée par la présence d'ohnologues entre les segments.

L'approche que nous avons choisie consiste à scanner chaque génome non dupliqué et de le découper en segments issus d'un unique chromosome pré-duplication. On considère qu'un tel segment est issu d'un unique chromosome pre-duplication si les orthologues de ses gènes alternent entre des segments différents des chromosomes de l'espèce dupliquée et si cette alternance est soutenue par l'existence de gènes ohnologues. Ensuite, les blocs de synténie dédoublée seront clusterisés pour que les blocs de même alternance soient groupés dans le même chromosome pré-duplication.

11.2.1 Comparaison d'un génome dupliqué à un génome non-dupliqué

Les blocs de synténie dédoublée sont construits en parcourant les gènes d'un génome non dupliqué (algorithme 11.4 et Figure 11.2). Un gène est rajouté au bloc du gène qui le précède si son orthologue vérifie une des trois conditions suivantes :

1. être sur le même chromosome que l'orthologue du dernier gène (on considère que la synténie conservée est une preuve suffisante) ;
2. être proche de l'orthologue d'un gène du bloc courant (permet de revenir sur une région déjà visitée) ;
3. être proche d'un gène ohnologue d'un gène proche d'un gène du bloc courant (permet de créer l'alternance proprement dite, et de gérer une alternance sur k régions à la fois).

Si le gène ne vérifie pas ces conditions, il crée un nouveau bloc d'alternance qu'on va chercher à étendre en continuant le parcours du génome non-dupliqué. Deux gènes sont «proches» s'ils sont séparés d'au plus d gènes. Le test de proximité est réalisé en construisant en parallèle du bloc dans l'espèce non-dupliquée, une liste de segments de l'espèce dupliquée. Cette liste est l'union de tous les voisinages (selon une distance $\pm d$ gènes) de tous les orthologues des gènes inclus dans le bloc courant. Cette liste s'agrandit au fur et à mesure pour tenir compte de l'ajout de gènes dans le bloc. Un bloc de synténie dédoublée est au final, dans notre procédure, désigné par un segment du génome non-dupliqué et un ensemble de segments du génome dupliqué, sur lesquels on observe l'alternance. Cet ensemble est, en général, spécifique de chaque bloc.

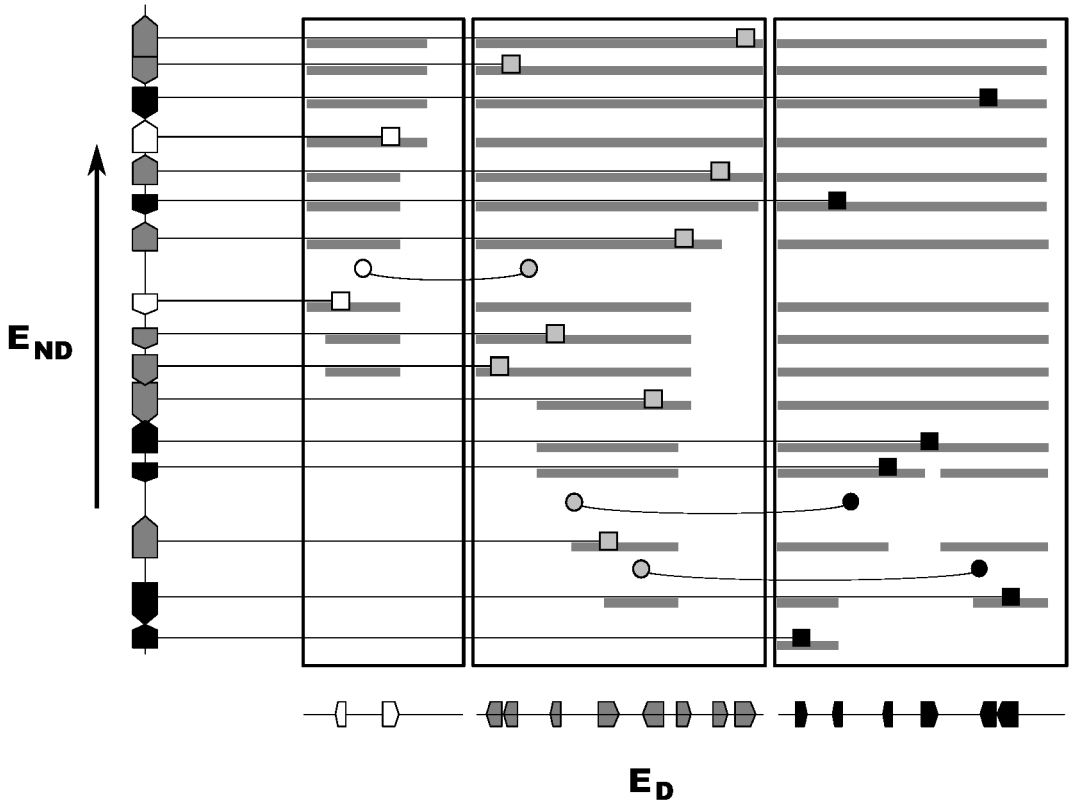


FIGURE 11.2 – Extraction d'un bloc de synténie dédoublée. Le chromosome de l'espèce non-dupliquée est disposé verticalement, et ceux de l'espèce dupliquée horizontalement. Les orthologues entre les deux génomes sont notés avec un carré, les ohnologues de l'espèce dupliquée avec des cercles. On parcourt les gènes de l'espèce non-dupliquée (de bas en haut) en construisant une liste de régions de l'espèce dupliquée (les barres grises) sur laquelle on va «autoriser» la synténie. Cette liste de régions est initialisée au voisinage de l'orthologue du premier gène du bloc (en bas). Lors du parcours de l'espèce non-dupliquée, un gène est rajouté au bloc courant si son orthologue est sur le même chromosome que celui du gène précédent (condition 1), ou s'il est sur une des régions autorisées (condition 2 ou 3). À chaque extension du bloc, la liste des régions autorisées est mise à jour en rajoutant le voisinage du dernier gène et les voisinages des ohnologues. Ici, le deuxième gène est rajouté car son orthologue est sur le même chromosome que celui du premier gène. Le troisième gène est rajouté car son orthologue est dans le voisinage d'un gène ohnologue d'un gène du voisinage du deuxième gène.

11.2.2 Combinaison des blocs de synténie dédoublée

Pour l'instant, chaque gène ancestral pré-duplication est inclus dans un bloc de synténie dédoublée pour chaque paire d'espèces non-dupliquée / dupliquée. Il faut donc introduire une étape de combinaison des reconstructions pour établir un unique jeu de chromosomes pré-duplication.

L'objectif est d'utiliser *walktrap* pour clusteriser des blocs en chromosomes pré-duplication. Nous avons donc défini (Figure 11.3) un score d'alternance pour un bloc de synténie dédoublée, et une mesure de similarité d'alternance entre deux blocs. Comme le clustering doit rassembler dans chaque chromosome pré-duplication les blocs de synténie identifiés avec toutes les espèces dupliquées, il est nécessaire que les objets à clusteriser soient transversaux du point de vue des espèces dupliquées. Nous combinons donc les blocs de synténie dont les segments sur les génomes non-dupliqués se chevauchent.

Cela nous amène à définir le score d'alternance d'un groupe de blocs de synténie

---

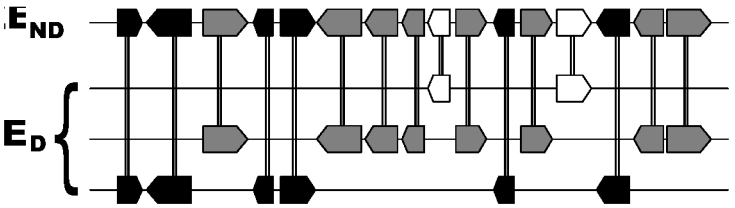
Algorithme 11.4 Comparaison d'un génome dupliqué à un génome non-dupliqué et identification de régions de synténie dédoublée

---

Entrées: $\mathscr{G}^{\mathscr{N}}$: le génome de l'espèce non-dupliquée. $\mathscr{G}^{\mathscr{D}}$: le génome de l'espèce dupliquée. d : la distance en nombre de gènes, maximale pour que deux gènes soient considérés comme proches.

1: $L_B \leftarrow \emptyset$ // contiendra la liste des blocs de synténie dédoublée
2: $l_c \leftarrow \emptyset$ // contiendra la liste des chromosomes de $\mathscr{G}^{\mathscr{D}}$ autorisés pour poursuivre le bloc courant
3: $l_g \leftarrow \emptyset$ // contiendra les gènes de $\mathscr{G}^{\mathscr{D}}$ autorisés pour poursuivre le bloc courant (selon le critère de proximité)
4: **pour tout** chromosome c de $\mathscr{G}^{\mathscr{N}}$ **faire**
5: $B \leftarrow \emptyset$ // le bloc courant
6: **pour tout** gène g de $\mathscr{G}^{\mathscr{N}}_c$ **faire**
7: $l \leftarrow$ Ensemble des orthologues de g dans $\mathscr{G}^{\mathscr{D}}$ qui vérifient «le chromosome de $g' \in l_c$», ou «$g' \in l_g$»
8: **si** $l \neq \emptyset$ **alors**
9: // Extension du bloc courant
10: Rajouter g à B
11: **pour tout** $g' \in l$ **faire**
12: $v \leftarrow$ voisinage de g' dans $\mathscr{G}^{\mathscr{D}}$ (à $\pm\, d$ gènes de distance)
13: Mettre à jour l_g avec v
14: **pour tout** gène $g'' \in v$ ayant un ohnologue **faire**
15: Rajouter à l_g les voisinage des ohnologues de g'' dans $\mathscr{G}^{\mathscr{D}}$ (à $\pm\, d$ gènes de distance)
16: **fin pour**
17: **fin pour**
18: **sinon**
19: Rajouter B à L_B
20: Réinitialiser B et l_g à $\emptyset$
21: **fin si**
22: **fin pour**
23: **fin pour**
24: **renvoyer** L_B

A :



B :

■/□ : 5 (3 × 1 + 1 × 1 + 1 × 1)
□/■ : 1 (1 × 1)
■/■ : 14 (2 × 1 + 1 × 2 + 2 × 3 + 1 × 1 + 1 × 1 + 1 × 2)

C :

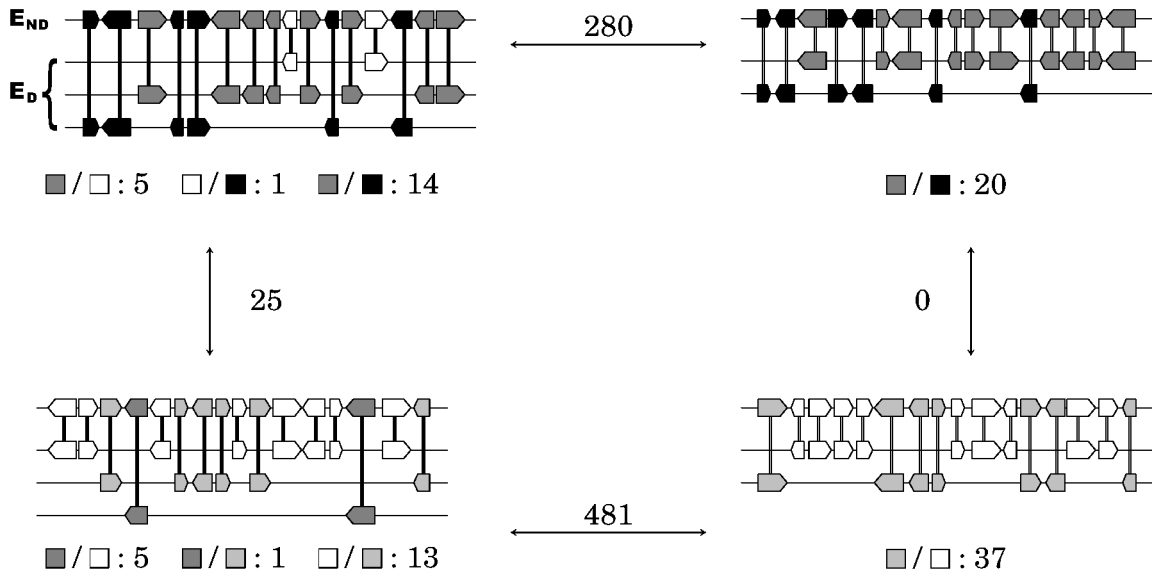


FIGURE 11.3 – Calcul d'un score d'alternance pour un bloc de synténie dédoublée. **A-B :** Pour chaque paire de chromosomes de l'espèce dupliquée, on fait la liste de tous les segments de gènes qui leur correspondent et on multiplie les longueurs de tels segments consécutifs. La somme correspond au score d'alternance de cette paire de chromosomes et est assimilable au nombre de paires de gènes qui ont validé le passage d'un chromosome à l'autre. **C :** Lorsque l'on compare deux blocs de synténie dédoublée, on fait le produit scalaire (dans l'ensemble des paires de chromosomes) des scores d'alternance de chaque bloc.

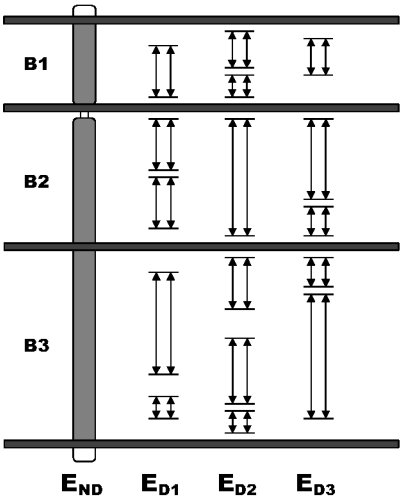


FIGURE 11.4 – Intégration des blocs de synténie dédoublée, pour une espèce non dupliquée. Les blocs de synténie dédoublée (identifiés sur trois génomes dupliqués E_{D1}, E_{D2}, E_{D3}) sont représentés par des doubles flèches. Ils sont regroupés lorsque les segments correspondants sur le génome non dupliqué E_{ND} se chevauchent, pour définir des blocs intégrés (B_1, B_2 et B_3).

comme la somme des scores d'alternance de ses blocs. La mesure de similarité doit combiner les scores d'alternance pour chaque espèce dupliquée en tenant compte de la phylogénie (méthode d'interpolation, Équation 6.1) pour définir la mesure de similarité qui est utilisée par *walktrap*.

Ces clusters pré-duplications contiennent tous les blocs de synténie dédoublée identifiés avec tous les génomes non-dupliqués. Rien ne garantit que pour un gène ancestral pré-duplication donné, ses descendants dans les différentes espèces non-dupliquées ne soient pas affectés, via leurs blocs de synténie respectifs, à des clusters pré-duplication différents.

Il est donc nécessaire de procéder à un dernier vote. Pour chaque gène ancestral, on interpole par l'Équation 6.1 son appartenance à chaque cluster (1 ou 0) chez les espèces non-dupliquées, et on choisit le cluster qui maximise la probabilité ancestrale.

La procédure de combinaison de tous les blocs de synténie dédoublée est donc :

1. mesurer pour chaque bloc de synténie son score d'alternance pour chaque paire de chromosomes des espèces dupliquées ;
2. pour chaque espèce non-dupliquée E_{ND}, grouper les blocs de synténie avec toutes les espèces dupliquées tant que les segments sur E_{ND} se chevauchent ;
3. clusteriser ces groupes avec *walktrap* en utilisant l'interpolation des produits scalaires sur chaque espèce dupliquée comme mesure de similarité ;
4. choisir pour chaque gène ancestral son cluster le plus probable (en interpolant la probabilité d'appartenance).

Quatrième partie

Résultats

Sommaire

Nous avons présenté dans la partie précédente les méthodes AGORA de reconstruction des génomes ancestraux. Cette partie montre l'utilisation pratique d'AGORA, ainsi que l'optmisation de certains paramètres en utilisant des simulations pour détecter la meilleure combinaison de valeurs.

Tout d'abord, le chapitre 12 présente une méthode de validation d'AGORA. Il s'agit d'un cadre logiciel pour la simulation de génomes de vertébrés. Il permettra de tester AGORA, de mesurer quantitativement ses performances, et de paramétrer les reconstructions.

Le chapitre 13 décrit la définition et l'optimisation des paramètres pour les reconstructions chez les vertébrés (quels protocoles appliquer, et avec quelles valeurs de paramètres).

Le chapitre 14 montre les reconstructions proprement dites et les compare aux reconstructions de référence présentées dans l'introduction, ou entre elles pour assurer leur cohérence.

Enfin, le chapitre 15 dévoile un serveur web, Genomicus, que nous avons mis en place pour comparer les génomes modernes et ancestraux dans la même interface. Le site sert de vitrine aux travaux du groupe sur les génomes ancestraux et est destiné à être enrichi de toutes les analyses qui suivront l'obtention des génomes ancestraux.

Chapitre 12

Simulation de l'évolution d'un génome

Sommaire

Comme évoqué dans l'introduction, les reconstructions de génomes ancestraux souffrent toujours du même problème : la difficulté de vérifier l'exactitude des résultats. Pour le résoudre, une des approches consiste à s'appuyer sur des simulations, et de pouvoir ainsi estimer empiriquement les performances d'une méthode de reconstruction. Nous avons développé un programme de simulation qui établit un ordre aléatoire de gènes dans l'ancêtre le plus vieux de l'arbre phylogénétique et le fait évoluer indépendamment sur chaque branche de l'arbre en suivant un jeu de règles prédéfinies. On dispose ainsi de génomes pour les espèces ancestrales, ainsi que pour les espèces modernes sur lesquels il est possible de tester une méthode de reconstruction. Le résultat de cette reconstruction est alors comparé aux génomes ancestraux créés lors de la simulation, ce qui permet d'établir les statistiques de performance de la méthode de reconstruction (Figure 12.1).

12.1 Liste et fréquences de référence des événements modélisés

Le réalisme des simulations tient dans la liste des événements modélisés, ainsi qu'à leur fréquence d'apparition. Le Tableau 12.1 montre les types d'événements considérés dans ce travail. Les duplications complètes de génome ne sont pas encore prises en compte, de par la difficulté d'étalonner la diploïdisation qui la suit, et l'augmentation des taux de réarrangements [Sémon et Wolfe, 2007].

Brièvement, une inversion est le réarrangement par lequel un segment d'un chromosome s'inverse sur place. L'inversion permet de modéliser les changements intra-chromosomiques de l'ordre des gènes. Une transposition est l'événement par lequel un segment d'un chromosome se déplace dans un autre chromosome. La transposition permet de modéliser les changements inter-chromosomiques de l'ordre des gènes. La fission

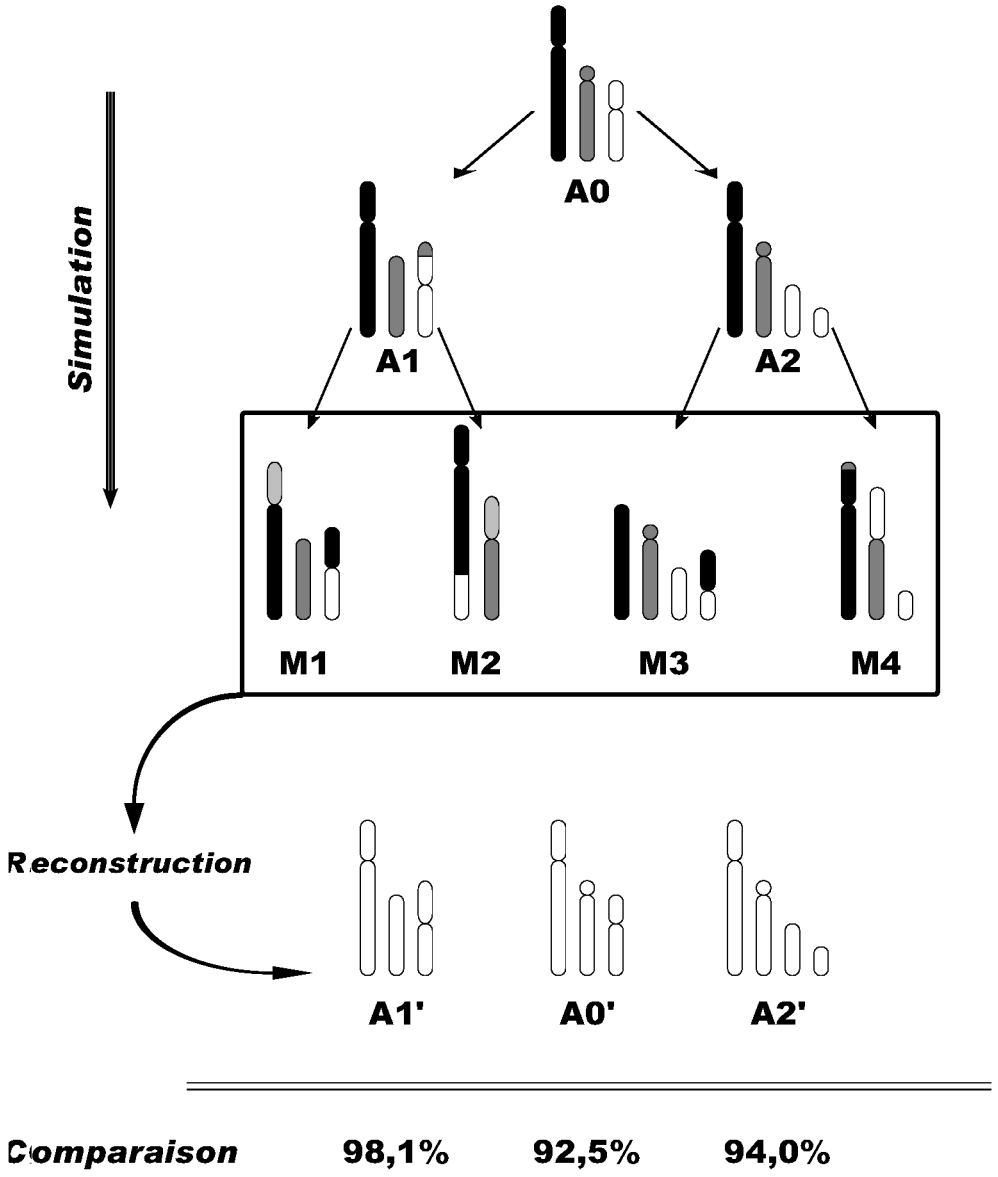


FIGURE 12.1 – Utilisation des simulations en tant qu'outil de validation. Le programme de simulation fait évoluer selon un ensemble de règles réalistes un génome ancestral aléatoire **A0** en des génomes ancestraux intermédiaires **A1** et **A2**, et des génomes modernes **M1** à **M4**. Une méthode de reconstruction peut alors reconstruire des génomes ancestraux **A0'**, **A1'** et **A2'** à partir des génomes modernes simulés et la comparaison de ces génomes ancestraux reconstruits à ceux attendus (donnés par le programme de simulation) fournit les mesures de performance de la méthode de reconstruction.

Évolution des familles de gènes	Réarrangements des chromosomes
Apparition d'une famille	Inversion
Perte d'un gène	Transposition
Duplication d'un gène	Fusion
ρ-groupage	Fission

TABLE 12.1 – Catégories d'événements modélisés

est le réarrangement qui sépare un chromosome en deux chromosomes, et la fusion en est l'opération inverse. Fusion et fission permettent de modéliser les changements de nombre de chromosomes.

Les duplications sont le mécanisme le plus courant de création de nouveaux gènes. En soi, l'apparition d'une famille de gène totalement nouvelle et sans lien avec un gène pré-existant est très rare [Knowles et McLysaght, 2009]. Dans nos simulations, chaque famille est systématiquement associée à un événement d'apparition au plus vieux nœud dans l'arbre au niveau duquel on a détecté une homologie. Dans la plupart des cas il s'agit probablement en réalité de la duplication d'un gène d'une autre famille, mais les séquences des deux familles de paralogues sont trop divergentes pour pouvoir être identifiées comme homologues. Enfin, le ρ-groupage n'est pas un événement en soi, mais une propriété simulée des génomes, paramétrée par un nombre ρ compris entre 0 et 1 qui permet de colocaliser des gènes en fonction de réarrangements (duplication, délétion) qu'ils subiront simultanément sur la même branche (voir section 12.4).

Nous avons estimé les fréquences (temporelles) des événements liés aux familles de gènes dans les données réelles en comptant le nombre de ces événements sur toutes les branches de l'arbre, et en divisant par la somme des longueurs des branches (l'algorithme de parcours des arbres et d'extraction des événements géniques sera expliqué en sous-section 7.3.1). Les comptages ont été faits en tenant compte uniquement des branches de l'arbre des vertébrés liant les espèces séquencées à haute couverture [1] car les génomes séquencées à faible couverture sont encore partiels : environ 2/3 de ces génomes est réellement séquencé (Figure 12.4), et donc environ 1/3 des gènes seront considérés comme perdus, ce qui biaiserait les statistiques. Nous avons mesuré 22915 apparitions de familles, 73023 nouvelles familles par duplication, et 94953 pertes de gènes, sur un total de 3877 millions d'années d'évolution. Les fréquences des réarrangements de chromosomes sont, elles, tirées de la littérature [Alekseyev et Pevzner, 2009]. Ces valeurs ont été calculées

1. Le taux de couverture d'un génome correspond au nombre de fois, en moyenne, qu'une base a été séquencée.

Réarrangement	Nombre par million d'années	Pourcentage
Apparition d'une famille	5,9	12,0 %
Perte d'un gène	24,5	49,8 %
Duplication d'un gène	18,8	38,2 %
Inversion	1,0	71,4 %
Transposition	0,3	21,4 %
Fusion	0,05	3,6 %
Fission	0,05	3,6 %

TABLE 12.2 – Taux moyens de modifications des familles de gènes et de réarrangements chromosomiques

chez les mammifères, et seront extrapolées à l'ensemble des espèces.

Afin de tester la méthode de reconstruction dans les conditions les plus réalistes possibles, les génomes ancestraux simulés pourront contenir les listes exactes des gènes ancestraux telles que lues dans les arbres phylogénétiques, au lieu d'une liste fictive modifiée au cours du temps par des pertes, gains et duplications aléatoires. Dès lors, seuls les réarrangements chromosomiques sont aléatoires et tout biais éventuel présent dans les familles réelles de gènes (sous-section 13.2.2) se retrouvera dans les génomes simulés.

Enfin, pour ce qui est des réarrangements chromosomiques, sur certaines branches de l'arbre, un facteur multiplicatif a été appliqué pour tenir compte des propriétés de certains génomes.

- L'opossum ne possède que 8 chromosomes, le taux de fusion de chromosomes sur la branche qui y mène a été multiplié par 2 tandis que celui de fission de chromosomes a été divisé par 2.
- Le poulet a, d'après Bourque *et al.* [2005] et Nakatani *et al.* [2007], un génome assez proche du génome ancestral des amniotes : les taux de réarrangements de chromosomes ont tous été divisés par 2 sur la branche qui y mène.
- Les génomes des rongeurs ont subi plus de réarrangements que ceux des autres mammifères [Zhao et Bourque, 2009] : les taux de réarrangements de chromosomes ont tous été multipliés par 2.

12.2 Nombres d'événements appliqués

Dans une simulation, un génome évolue le long d'une branche par un certains nombre d'événements. Nous avons modélisé le nombre effectif d'événements de type t par $n_t =$

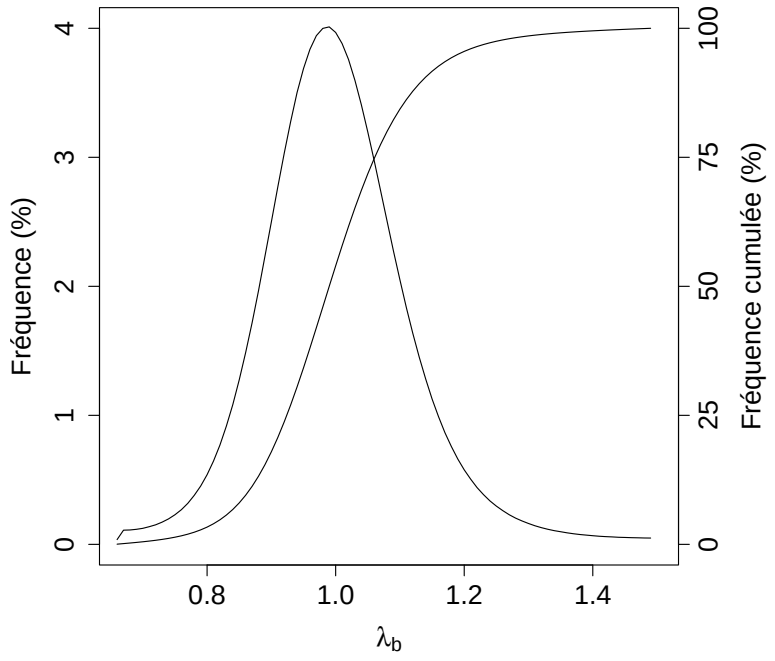


FIGURE 12.2 – Densité de probabilité (par pas de 0,01) et fonction de répartition des taux de réarrangements spécifiques de chaque branche.

$\lfloor l f_t \lambda_g \lambda_b \rfloor$ où l désigne la longueur de la branche, f_t la fréquence du type d'événement considéré, λ_g un facteur multiplicatif global appliqué à toutes les branches de l'arbre, et λ_b un facteur multiplicatif aléatoire, spécifique de chaque branche. λ_g sert à générer des instances de génomes aléatoires ayant tous subi λ_g fois plus de réarrangements, ce qui équivaut à considérer que toutes les branches sont λ_g plus longues. Cela permet d'appréhender le comportement de la méthode de reconstruction sur de grandes distances évolutives ou sur des génomes extrêmement réarrangés. λ_b est, lui, spécifique de chaque branche, et sert à donner de la variabilité inter-simulation au nombre de réarrangements appliqués à chaque branche. λ_b est calculé à partir d'une loi de distribution von Mises [2] de paramètre $\mu = 0$. Pour un x tiré selon cette loi, λ_b vaudra $\lambda_b^{*x/\pi}$, où λ_b^* est la valeur maximale possible pour λ_b (et $1/\lambda_b^*$ la valeur minimale). En pratique, la valeur 1,5 a été utilisée pour λ_b^*, et 2 pour κ (le paramètre de concentration de la distribution), ce qui implique que 2/3 des λ_b seront dans l'intervalle $[0{,}9\ ;\ 1{,}1]$. La distribution des λ_b calculés selon cette formule et ces paramètres est représentée sur la Figure 12.2.

12.3 Réarrangements de chromosomes

Aucun modèle n'existe encore pour simuler la répartition des points de cassure dans les génomes, ni les longueurs des segments réarrangés. Les réarrangements sont décrits dans le Tableau 12.3 comme des suites de fonctions de choix basiques. Pour appliquer les réarrangements, le protocole de simulation implémente donc les quatre fonctions de choix de la manière suivante.

- Le choix d'un chromosome se fait proportionnellement à sa taille.
- Sur un chromosome donné, les positions sont choisies uniformément.
- Les extrémités des chromosomes sont choisies uniformément.
- Sur un chromosome donné, un segment est d'abord défini par sa longueur (modélisée par une loi de distribution von Mises adaptée, voir paragraphe ci-dessous), puis par une position aléatoire sur le chromosome.

Réarrangement	Choix nécessaires
Inversion	Chromosome, segment de ce chromosome
Transposition	Deux chromosomes (source et cible), segment du chromosome source, position sur le chromosome cible
Fusion	Deux extrémités de chromosomes différents
Fission	Chromosome, position sur ce chromosome

TABLE 12.3 – Description des protocoles d'application de chaque type de réarrangements

La longueur des segments réarrangés doit être modélisée car aucune étude n'a défini leur longueur. Dans leur protocole de simulation, Ma *et al.* [2006] exprimaient leurs génomes simulés comme des arrangements de 6000 «blocs». Ils utilisaient une loi gamma $\Gamma(k = 0{,}7, \theta = 500)$ bornée à une limite de 50 blocs maximum. La distribution de leurs segments avait donc pour moyenne 20 blocs et pour médiane 18 blocs, par rapport à des chromosomes composés de 240 blocs en moyenne. Nous avons estimé que leur distribution

2. La loi de distribution von Mises peut être considérée comme l'adaptation d'une distribution normale à un intervalle de longueur 2π centrée sur une moyenne (et médiane) μ, et dépendante d'un paramètre de concentration κ, assimilable à l'inverse de la variance.

favorisait beaucoup trop les petites inversions et avons modélisé nous même une nouvelle distribution de longueurs de segments à partir de la distribution von Mises. Nous définissons un objectif μ^* qui désigne la valeur médiane de la longueur des segments, exprimée en proportion de la longueur totale du chromosome sur lequel le segment est sélectionné.

Dans le cas où $\mu^* < 1/2$, on tire un nombre dans la distribution von Mises et on applique la transformation affine $x \mapsto \mu^* + x\frac{1-\mu^*}{\pi}$. On dispose alors d'une valeur dans l'intervalle $[2\mu^* - 1,\ 1]$ et on revient à $[0,\ 1]$ par passage à la valeur absolue. Dans le cas où $\mu^* \geq 1/2$, la transformation affine est $x \mapsto \mu^* + x\frac{\mu^*}{\pi}$ et on doit alors transformer l'intervalle $[0,\ 2\mu^*]$ en $[0,\ 1]$ selon un procédé similaire à la valeur absolue, mais centré sur 1. En pratique, on fixe la longueur médiane $\mu^* = 0,2$, et on prend les paramètres $\mu = 0$ et $\kappa = 2$ pour la distribution von Mises. La distribution des longueurs de segment est présentée en Figure 12.3. Ainsi, la moitié des segments choisis mesurent moins de 20%, et les tailles inférieures à 20% sont à peu près uniformément choisies. Au delà de 20%, les tailles sont choisies avec une probabilité décroissante. Cela permet de faire apparaître que les longueurs des inversions sont en générales petites par rapport à la taille du chromosome qui les supporte.

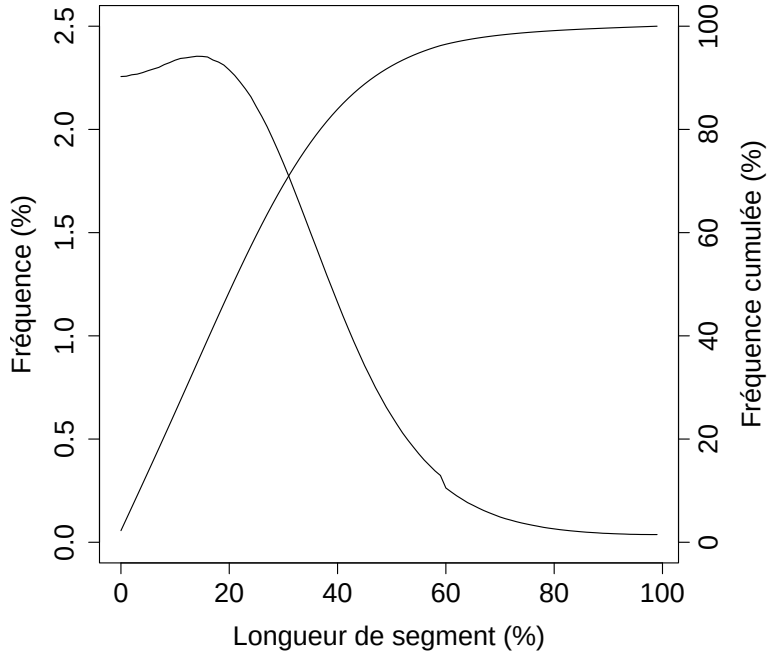


FIGURE 12.3 – Densité de probabilité (par pas de 1%) et fonction de répartition des longueurs des segments réarrangés

12.4 Regroupement spatial et temporel des gènes

Dans le génome, la perte de plusieurs gènes peut être causée par un unique événement de délétion qui aurait supprimé une importante région du génome. De même, la duplication d'un segment de chromosome implique la duplication de tous les gènes qui y sont présents. Ainsi, plusieurs modifications des familles de gènes peuvent avoir lieu

simultanément en raison d'un seul événement, si les gènes représentant ces familles sont contigus dans le génome. Pour rendre le protocole de simulation plus réaliste, nous avons cherché à colocaliser les gènes, à chaque ancêtre, selon les événements qu'ils subiront plus tard (sans introduire de réarrangements chromosomiques supplémentaires) pour que leur perte ou leurs duplications puissent s'expliquer par un seul événement génomique. Par exemple, un certain nombre de gènes devant subir une duplication sur une même branche de l'arbre, seront arrangés de manière contiguë dans les ancêtres qui précèdent cette branche, de façon à ce que la seule duplication du segment de génome qui les contient puisse rendre compte de leurs duplications individuelles.

Cela nécessite de documenter pour chaque famille les branches de l'arbre sur lesquelles elle subit des événements. Ces données sont stockées sous forme d'un vecteur de 0 et de 1, et on définit le score de similarité de deux familles comme le produit scalaire de ces deux vecteurs. Une valeur élevée indique que les familles ont subi des événements sur les mêmes branches. Pour une famille donnée, on peut définir son meilleur «voisin» comme la famille qui maximise le produit scalaire, et sa meilleure position dans le génome comme la position qui maximise la somme des produits scalaires avec les deux familles voisines. Cette dernière définition peut d'ailleurs être étendue à un bloc de gènes pour trouver sa meilleure position, en utilisant le gène à chaque extrémité. Cette maximisation peut être faite du point de vue d'une position du génome : on cherche le bloc qui, inséré à cette position, maximiserait la somme des deux produits scalaires.

Un paramètre global ρ, compris entre 0 et 1, a été introduit et indique, pour chaque ancêtre, quelle proportion des paires de gènes consécutifs est soumise à cette sélection de meilleur voisin / position. Ainsi, on définit un ρ-groupage d'un ensemble des gènes, comme un partitionnement en listes où un gène a ρ chances d'être dans une liste et suivi par son meilleur voisin, et $1-\rho$ chances d'être en fin de liste. Pour garantir que la proportion de gènes colocalisés soit de ρ à tous les ancêtres, le protocole de simulation est adapté à toutes les étapes faisant intervenir un positionnement de gènes. Le génome ancestral initial se définit désormais comme une répartition aléatoire des ρ-groupes des gènes dans des chromosomes. Pour chaque branche, et pour chaque gène qui se duplique, les produits de la duplication sont ρ-groupés. Comme on ne sait pas, a priori, laquelle des deux (ou plus) copies correspond au gène original, on considère que le groupe qui s'insère le mieux (selon la définition de meilleure position du paragraphe précédent) à la position d'origine remplace le gène initial, tandis que les autres sont mis en attente, avec les ρ-groupes des gènes apparus sur la branche. Tous les groupes mis en attente sont finalement insérés, dans ρ cas à leur meilleure position, et dans $1-\rho$ cas, aléatoirement. Ceci garantit qu'à chaque ancêtre, le taux de regroupement des gènes reste ρ.

12.5 Simulation de l'assemblage partiel

Parmi les génomes séquencés, la couverture de séquençage et la continuité de l'assemblage des génomes de vertébrés sont très hétérogènes. Le programme de simulation doit tenir compte de cette information afin d'éviter de donner aux méthodes de reconstruction des génomes aléatoires complets pour des espèces dont le génome réel n'est que partiel, et de biaiser les performances. Par conséquent, les génomes simulés subissent une fragmentation plus ou moins sévère selon trois catégories :

1. espèces séquencées et assemblées entièrement (ou presque), notées *Full* ;

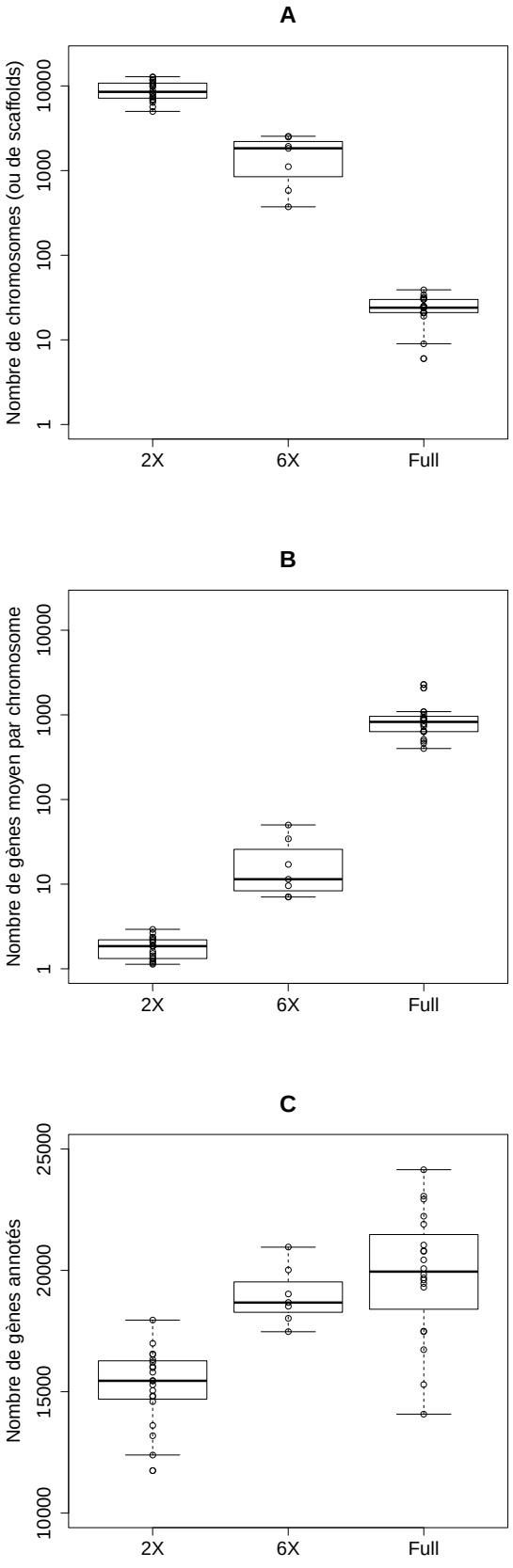


FIGURE 12.4 – Statistiques sur la fragmentation des génomes séquencés. **A :** Distribution du nombre de chromosomes (ou scaffolds) selon les trois catégories de qualité. **B :** Longueur moyenne (en nombre de gènes) des chromosomes (ou scaffolds) selon les trois catégories de qualité. **C :** Nombre de gènes annotés selon les trois catégories de qualité.

2. espèces séquencées presque entièrement, mais partiellement assemblées, notées *6X* en référence à leur taux de couverture moyen ;
3. espèces partiellement séquencées et, de fait, partiellement assemblées, notées *2X*.

D'après la Figure 12.4, les génomes de la catégorie *6X* contiennent en effet à peu près autant de gènes que ceux de la catégorie *Full*. Leur assemblage est cependant encore partiel, et ils sont composés d'environ 1000 scaffolds au lieu d'une vingtaine de chromosomes (nombre moyen chez les vertébrés). Cette situation est simulée en appliquant en toute fin de branche entre 500 et 3000 fissions de chromosomes sur le génome aléatoire simulé.

Les génomes de la catégorie *2X* n'ont de séquencé qu'environ 2/3 du génome réel (estimation en fonction de la couverture comme décrit dans [Lander et Waterman, 1988] et confirmé par la distribution des nombres de gènes), et les plus longs scaffolds contiennent au mieux quelques dizaines de gènes. Le nombre moyen de gènes par scaffolds étant compris entre 1,1 et 2,9, le programme choisit comme objectif une taille moyenne de scaffolds dans cet intervalle, puis fissionne les chromosomes autant que nécessaire. Ensuite, il choisit au hasard entre 55% et 85% des scaffolds pour simuler le manque d'une partie du génome.

Chapitre 13

Paramétrage et validation d'AGORA

Sommaire

Dans ce chapitre nous étudierons comment appliquer les différentes méthodes présentées dans les chapitres 10 et 11 aux génomes de vertébrés. Tout d'abord, nous verrons que dans un cas théorique (aucun événement sur les familles de gènes, uniquement des réarrangements chromosomiques), AGORA se comporte aussi bien que les autres méthodes de reconstruction, tout en permettant de gérer beaucoup plus de marqueurs (gènes), des taux de réarrangements élevés, et la reconstruction simultanée de tous les génomes ancestraux. L'intérêt de la méthode montré, nous décrirons le cheminement chronologique qui a permis d'établir la combinaison d'outils AGORA optimale pour reconstruire les génomes ancestraux de vertébrés. Cela inclut de réaliser des reconstructions sur des jeux de génomes simulés pour paramétrer finement AGORA.

Pour les reconstructions elles-mêmes, nous avons d'abord réalisé une première reconstruction (protocole 1-passe, section 13.2), qui a permis de montrer des limites dans les données, qu'il a fallu dépasser. La section suivante (13.3) décrit le paramétrage d'AGORA (protocole multi-passes) pour obtenir les reconstructions optimales. La Figure 13.7 (page 134) montre le schéma récapitulatif de la procédure que nous utilisons pour AGORA, et que nous allons détailler et expliquer dans ce chapitre.

13.1 Comparaison d'AGORA aux autres méthodes de reconstruction

Le protocole de simulation du chapitre 12 a d'abord servi à valider et montrer les avantages d'AGORA par rapport aux autres méthodes de reconstruction. On connait déjà l'un des résultats : AGORA est capable de traiter un répertoire de gènes qui varie d'un ancêtre à l'autre, contrairement aux autres méthodes, qui demandent que les gènes soient en une

seule copie dans tous les génomes. Ce premier jeu de simulations n'a donc fait intervenir que des réarrangements de chromosomes et a laissé les listes de gènes intactes depuis l'ancêtre *Euteleostomi*. Il n'est donc pas nécessaire de faire intervenir le ρ-groupage des gènes en fonction des événements géniques.

Afin de tester les limites de chaque méthode, nous avons de plus fait varier la taille des génomes simulés, ainsi que le taux de réarrangement de chromosomes. Il en résulte 6 tailles de génomes différentes, définies par un nombre de gènes (et accessoirement, un nombre de chromosomes) : 100 et 500 gènes en 5 chromosomes, 1000, 5000 et 10000 gènes en 10 chromosomes, et 20000 gènes en 20 chromosomes. Bien entendu, le nombre de chromosomes correspond au nombre fixé pour l'ancêtre aléatoire initial *Euteleostomi* et peut varier au cours d'une simulation.

D'autre part, le taux global de réarrangement a été simulé en faisant varier λ_g dans les simulations. Il faut néanmoins remarquer que dans notre définition, le taux de réarrangement est rapporté à une durée (nombre de réarrangements par million d'années). Sur une même période de temps, un génome A 10 fois plus petit qu'un génome B subira autant de réarrangements que le génome B, mais par conséquent avec une fréquence spatiale 10 fois plus élevée. Il faut compenser cela en divisant le taux de réarrangement λ_g par 10. Le Tableau 13.1 donne les valeurs de λ_g équivalentes selon les tailles de génome, par rapport aux taux de réarrangements de base des vertébrés.

	100	500	1000	5000	10000	20000
0,2x	0,0010	0,0050	0,010	0,050	0,10	0,2
0,5x	0,0025	0,0125	0,025	0,125	0,25	0,5
1x	0,0050	0,0250	0,050	0,250	0,50	1,0
2x	0,0100	0,0500	0,100	0,500	1,00	2,0
3x	0,0150	0,0750	0,150	0,750	1,50	3,0

TABLE 13.1 – Valeurs de λ_g utilisées pour la comparaison d'AGORA aux autres méthodes de reconstruction, en fonction de la taille des génomes simulés (nombre de gènes, en colonnes) et du taux de réarrangement voulu (en lignes).

Chaque combinaison de taille et de taux de réarrangement a été testée sur 10 jeux de génomes aléatoires. Ce nombre est suffisant pour donner une tendance des performances et limites de chaque méthode. Les 4 méthodes testées sont MGR [Bourque *et al.*, 2004], MGRA [Alekseyev et Pevzner, 2009], inferCARs [Ma *et al.*, 2006], et bien sûr, AGORA. Pour AGORA, c'est le protocole 1-passe qui a été testé car cette simulation ne fait intervenir aucune modification des listes de gènes alors que le protocole multi-passes demande justement d'établir un sous-jeu de gènes ayant évité les duplications et les délétions.

Les résultats sont donnés dans la Figure 13.1. La figure est composée de 4 tableaux, un par méthode, montrant les performances pour chaque combinaison de taille et de taux de réarrangement. Le fond des cellules est teinté par le nombre d'échecs de la méthode testée (une case est blanche si les dix reconstructions ont donné un résultat, et noire si les dix ont échoué). Cela se produit pour MGRA et MGR sur des génomes gros ou/et réarrangés. MGRA renvoie le message d'erreur *"T-transformation is not complete. Cannot reconstruct genomes"* et s'arrête. D'autre part, alors que toutes les autres reconstructions se terminent au maximum en quelques heures, nous avons du laisser calculer MGR plus de deux semaines sur certains génomes, sans que la reconstruction n'avance beaucoup (le temps total estimé en fonction de l'avancement était de plusieurs mois). Dans ces situations, MGR a été arrêté et nous avons considéré un échec de la méthode.

inferCARs	100	500	1000	5000	10000	20000
0,2x	0 0	0 0	0 0	0 0	0 0	0 0
0,5x	0 0	0 0	0 0	0 0	0 0	0 0
1x	0 0	0 0	0 0	0 0	0 0	4 0
2x	0 0	0 0	0 0	0 0	6 0	24 1
3x	0 0	0 0	0 1	6 0	24 0	40 3

MGR	100	500	1000	5000	10000	20000
0,2x	0 0	0 0	0 0	0 0		
0,5x	0 0	0 0	0 0	1 0		
1x	0 0	7 7	0 0			
2x	0 0	2 2	0 0			
3x	2 2	0 0				

MGRA	100	500	1000	5000	10000	20000
0,2x	0 0	0 0	0 0	0 0	0 0	
0,5x	0 0	0 0	0 0	0 0		
1x	0 0	0 0	0 0	0 0		
2x	0 0	0 0	0 0			
3x	0 0	0 0	0 0			

AGORA	100	500	1000	5000	10000	20000
0,2x	0 0	0 0	0 0	0 0	0 0	0 0
0,5x	0 0	0 0	0 0	1 2	7 2	10 1
1x	0 0	0 0	0 0	7 1	12 3	19 3
2x	0 0	0 0	1 0	12 4	23 13	19 21
3x	0 0	0 0	1 1	10 12	16 23	10 46

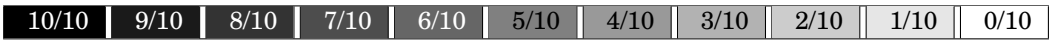


FIGURE 13.1 – Résultats de la comparaison d'AGORA aux autres méthodes de reconstruction (pour *Boreoeutheria*). Chaque tableau est structuré selon la taille des génomes (en colonnes) et le taux de réarrangement (en lignes). Pour chaque case, la couleur du fond est proportionnelle aux nombre d'échecs de la méthode selon la gamme affichée en bas, et les nombres indiquent le nombre d'intervalles prédits et faux (en haut), et non prédits (en bas). Par exemple, sur les 10 simulations de génomes de 20000 gènes à un taux de réarrangement 1x, AGORA a terminé son calcul les 10 fois, a au total prédit 19 intervalles incorrects, et en a raté 3.

Dans chaque case, le premier nombre indique le nombre d'intervalles faux prédits pour *Boreoeutheria* (permet de calculer la spécificité), et le deuxième, le nombre d'intervalles omis pour *Boreoeutheria* (permet de calculer la sensibilité). Ce sont les comptages bruts (sur 10 simulations, donc) qui sont affichés car les pourcentages de spécificité et sensibilité sont compris entre 99,98% et 100%, ce qui est peu pratique pour juger des écarts. Néanmoins, cela signifie que les erreurs de reconstruction sont très peu nombreuses, et que les quatre méthodes permettent assez précisément de reconstruire le génome ancestral pour *Boreoeutheria*.

Il faut, de plus, souligner quelques points négatifs de différentes méthodes.

- MGR reconstruit la phylogénie des espèces dont il a le génome, et cette phylogénie peut, a priori, ne pas correspondre à la phylogénie connue des espèces, ce qui empêcherait d'affecter la reconstruction aux bons ancêtres.
- Les différentes méthodes sont en général prévues et testées sur plusieurs centaines ou milliers de marqueurs. La complexité du programme (selon le nombre de gènes) a alors des conséquences non négligeables lorsqu'on atteint des tailles de 20000 gènes. Ainsi, inferCARS a une complexité (estimée) quadratique en temps d'exécution et en espace mémoire. Une reconstruction sur 20000 gènes met ainsi ≈16 minutes et demande ≈64 Go de mémoire. MGRA et AGORA ont des complexités linéaires et une reconstruction sur 20000 gènes demande (respectivement) ≈100 Mo de mémoire et ≈7 minutes, et ≈1,5 Go de mémoire et ≈23 minutes.

Dans le cas de génomes de vertébrés (≈20000 gènes, et taux de réarrangement 1x), seuls inferCARs et AGORA sont capables de fournir une réponse précise. En rajoutant qu'une reconstruction inferCARs ne donne le génome ancestral que d'un unique génome cible, et qu'il faut lancer d'autres programmes et/ou reconstructions pour avoir les génomes d'autres ancêtres, on se rend compte qu'AGORA (qui reconstruit automatiquement tous les ancêtres) présente des caractéristiques avantageuses pour reconstruire simultanément de nombreux génomes ancestraux, étape indispensable pour une analyse comparative globale de ces génomes. De plus, AGORA est la seule méthode, parmi celles testées ici, à pouvoir gérer des phylogénies de gènes comportant des délétions et des duplications (cf ci-dessous).

13.2 Reconstruction de contigs en une passe

Compte tenu de la validation de la méthode AGORA (1-passe), par rapport aux autres méthodes de reconstructions, nous avons appliqué AGORA sur les données réelles. Puisque ces données incluent des duplications, pertes et gains de gènes, AGORA est ici la seule méthode disponible.

Pour information, les statistiques récoltées après une reconstruction seront :

- le nombre de gènes ;
- le nombre de contigs ;
- la couverture de la reconstruction (nombre de gènes et d'intervalles présents dans les contigs) ;
- les longueurs (en nombre de gènes) des contigs (moyenne, quartiles, et longueurs type N50 [1]).

On gardera en tête que les génomes de vertébrés possèdent (dans le cas des espèces séquencées) entre 20 et 40 chromosomes, et que la taille d'un chromosome est de l'ordre

1. Lorsqu'un génome est fragmenté en contigs, le N50 est la taille de contig telle que l'ensemble des contigs de taille supérieure représente 50% du génome. De la même manière, on définit les N25 et N75.

Ancêtre	Âge (Ma)	Gènes	Contigs	Contigs (> 100 gènes)	Couverture (gènes)		Couverture (intervalles)		25%	50%	75%	N75	N50	N25	Max	Moyenne
Fungi/Metazoa group	1500	2368	12	0	24	1,01%	12	0,51%	2	2	2	2	2	2	2	2,00
Bilateria	580	8833	49	0	105	1,19%	56	0,64%	2	2	2	2	2	2	6	2,14
Chordata	550	10837	139	0	288	2,66%	149	1,38%	2	2	2	2	2	2	6	2,07
Ciona	100	10385	1592	0	4725	45,50%	3133	30,23%	2	2	3	2	3	4	14	2,97
Clupeocephala	320	21931	3729	0	15357	70,02%	11628	53,07%	2	3	5	3	5	8	30	4,12
Percomorpha	190	21428	1512	11	18291	85,36%	16779	78,38%	3	6	13	11	24	52	149	12,10
Tetraodontidae	65	19349	1459	10	16775	86,70%	15316	79,24%	3	5	13	10	24	48	168	11,50
Sauria	267	18340	1190	9	14879	81,13%	13689	74,72%	3	5	13	11	27	56	187	12,50
Neognathae	105	16238	700	24	13546	83,42%	12846	79,21%	3	8	23	21	47	98	240	19,35
Phasianidae	46	15958	676	23	13509	84,65%	12833	80,52%	3	7	22	22	53	111	376	19,98
Euteleostomi	420	20545	3158	0	10641	51,79%	7483	36,46%	2	3	4	2	4	6	24	3,37
Tetrapoda	359	20274	2678	0	13219	65,20%	10541	52,04%	2	3	6	4	6	11	62	4,94
Amniota	326	21455	1379	8	15345	71,52%	13966	65,16%	3	5	12	9	22	52	163	11,13
Mammalia	184	21482	1532	5	15410	71,73%	13878	64,66%	3	5	10	8	20	46	157	10,06
Theria	166	21884	1101	15	16618	75,94%	15517	70,97%	3	7	18	15	33	64	190	15,09
Eutheria	102	28348	1343	14	18979	66,95%	17636	62,26%	2	5	16	16	35	66	224	14,13
Boreoeutheria	95	29479	1117	34	19575	66,40%	18458	62,66%	2	4	20	26	54	93	211	17,52
Euarchontoglires	90	28214	1007	36	19342	68,55%	18335	65,03%	2	4	23	27	57	104	275	19,21
Primates	83	24933	777	40	18399	73,79%	17622	70,73%	2	8	26	28	70	122	315	23,68
Haplorrhini	57	23853	738	42	18280	76,64%	17542	73,60%	3	9	27	28	66	123	391	24,77
Simiiformes	45	24251	720	44	18636	76,85%	17916	73,94%	2	9	29	31	74	129	465	25,88
Catarrhini	31	24357	700	44	18895	77,58%	18195	74,76%	2	7	30	35	80	146	487	26,99
Hominidae	16	23482	661	45	19081	81,26%	18420	78,51%	2	8	34	38	82	138	392	28,87
Homininae	9	23264	713	52	19395	83,37%	18682	80,37%	2	7	31	37	87	148	378	27,20
Homo/Pan group	5	21805	588	56	19251	88,29%	18663	85,67%	2	9	42	48	89	171	314	32,74
Metatheria	148	18257	1879	0	15489	84,84%	13610	74,63%	3	5	11	7	13	22	79	8,24
Atlantogenata	100	22780	1512	0	17742	77,88%	16230	71,31%	3	7	15	10	20	36	85	11,73
Afrotheria	94	21285	1430	2	17589	82,64%	16159	75,99%	3	8	16	11	22	36	118	12,30
Laurasiatheria	88	26300	958	31	19197	72,99%	18239	69,40%	2	6	25	26	54	90	193	20,04
Xenarthra	64	16237	2660	0	7045	43,39%	4385	27,04%	2	2	3	2	3	4	14	2,65
Insectivora	68	16764	2697	0	7076	42,21%	4379	26,15%	2	2	3	2	2	4	14	2,62
Cetartiodactyla	61	22185	1370	3	18438	83,11%	17068	77,00%	3	7	18	13	26	45	108	13,46
Chiroptera	60	18643	3095	0	14675	78,72%	11580	62,18%	2	3	5	3	6	11	61	4,74
Carnivora	56	19665	1414	7	17255	87,74%	15841	80,64%	3	7	16	11	21	36	132	12,20
Lagomorpha	48	19252	1837	2	16527	85,85%	14690	76,38%	3	5	11	8	15	28	114	9,00
Strepsirrhini	69	18276	3169	0	14144	77,39%	10975	60,12%	2	3	5	3	6	9	36	4,46
Glires	81	24605	991	17	18774	76,30%	17783	72,33%	2	9	26	22	42	72	171	18,94
Rodentia	80	23086	946	19	18404	79,72%	17458	75,69%	3	10	25	22	41	74	189	19,45
Sciurognathi	79	22296	1174	9	18184	81,56%	17010	76,36%	3	9	20	15	29	50	161	15,49
Murinae	37	21142	887	28	18834	89,08%	17947	84,97%	3	10	27	23	49	88	236	21,23

TABLE 13.2 – Résultats d'AGORA (données brutes). Statistiques sur les longueurs et nombres de contigs, selon les ancêtres, pour une utilisation d'AGORA (protocole 1-passe) sur les données brutes d'Ensembl.

de la centaine ou du millier de gènes (voir Tableau 13.3). Les ancêtres seront catégorisés en 7 groupes (voir tableau 13.2) : les non-vertébrés, les poissons, les sauriens, la lignée humaine depuis l'ancêtre des vertébrés, les mammifères outgroups de *Boreoeutheria*, les descendants de *Laurasiatheria* et les descendants de *Euarchontoglires* (en dehors de la lignée humaine). L'évolution des performances des reconstructions sur les ancêtres *Boreoeutheria* et *Amniota* au fur et à mesure de l'avancée de la méthode de reconstruction sera représentée dans le Tableau 13.12.

25%	50%	75%	N75	N50	N25	Max	Moyenne
332	649	944	689	952	1367	5054	721,54

TABLE 13.3 – Statistiques sur les longueurs (nombre de gènes) des chromosomes de vertébrés. Les trois premières colonnes indiquent les quartiles de la distribution, les trois suivantes indiquent les valeurs prises par le N75, le N50, et le N25 (voir texte pour la définition), puis les deux dernières la taille du plus grand chromosome et la taille moyenne d'un chromosome de vertébré.

Les résultats de la reconstruction initiale (sur les données brutes d'Ensembl, sans l'insertion des trois espèces supplémentaires, cf sous-section 7.2.3) sont présentés dans le tableau 13.2 et sont plutôt décevants dans l'ensemble. La reconstruction de l'ancêtre *Boreoeutheria* est composée d'environ 1000 contigs de 18 gènes en moyenne. Mais ce qui frappe est l'évolution du nombre de gènes dans les ancêtres qui passe de 20545 chez *Euteleostomi* à 29479 chez *Boreoeutheria* pour revenir à 21805 chez *Homo/Pan group*. Ceci n'a rien de parcimonieux et puisque les génomes de mammifères possèdent en moyenne ≈20000 gènes et que deux mammifères ont en général au moins ≈15000 gènes orthologues, on s'attendait à garder un nombre de gènes relativement constant.

13.2.1 Édition des nœuds de duplication

En remarquant que chaque événement de duplication fait augmenter de 1 le nombre de copies d'un gène, nous nous sommes intéressés à l'inférence des duplications des arbres phylogénétiques dans le pipeline Ensembl Compara.

Le résultat de la création d'un arbre phylogénétique consensus (à partir de 5 méthodes de reconstruction différentes, sous-section 7.1.2) par TreeBest contient quelques singularités sur la position des nœuds de duplication. Il arrive que des nœuds de duplication soient inférés avec un support très faible (voire inexistant, comme dans l'exemple de Figure 13.2) : avec peu d'espèces modernes qui contiennent effectivement le gène en deux copies.

Ce phénomène peut aussi s'observer à travers un autre indicateur : la distribution des tailles des familles des ancêtres. En moyenne, une famille d'un ancêtre A est censée contenir autant de gènes que d'espèces présentes sous A dans la phylogénie des espèces. Si une duplication peu supportée est prédite, alors 2 gènes ancestraux seront définis (au lieu de 1), et un d'entre eux sera relié à peu de gènes modernes. Dans le cas de *Boreoeutheria*, il y a 28 espèces et la Figure 13.3 (à gauche) montre qu'en plus d'un pic à environ 28 gènes, beaucoup de familles ancestrales sont liées à peu de gènes : environ 4000 familles sont composées de 2 gènes ou moins, et 10000 de 8 gènes ou moins. L'ancêtre contient donc énormément de familles peu supportées par les espèces modernes, et dans une quantité suffisante pour expliquer le surplus de gènes ancestraux prédits.

Pour détecter les duplications peu supportées par les espèces modernes, on reprend la notion de score de confiance (introduite par TreeBest). Il s'agit d'une valeur définie

pour chaque nœud de duplication, comprise entre 0 et 1, qui est la proportion d'espèces présentes dans les deux sous-arbres qui suivent la duplication, par rapport aux espèces présentes dans au moins un d'entre eux. Ainsi, une duplication bien supportée (presque toutes les espèces possèdent le gène en deux copies) aura un score proche de 1, et une duplication très peu supportée, un score proche de 0.

La solution a donc été de supprimer les nœuds de duplication ayant un score de confiance faible selon le protocole décrit dans la Figure 13.4 et l'algorithme 13.1. Étant donnée une valeur seuil pour le score de confiance, l'algorithme recherche les nœuds de duplication avec un score inférieur au seuil pour les transformer en nœuds de spécia-

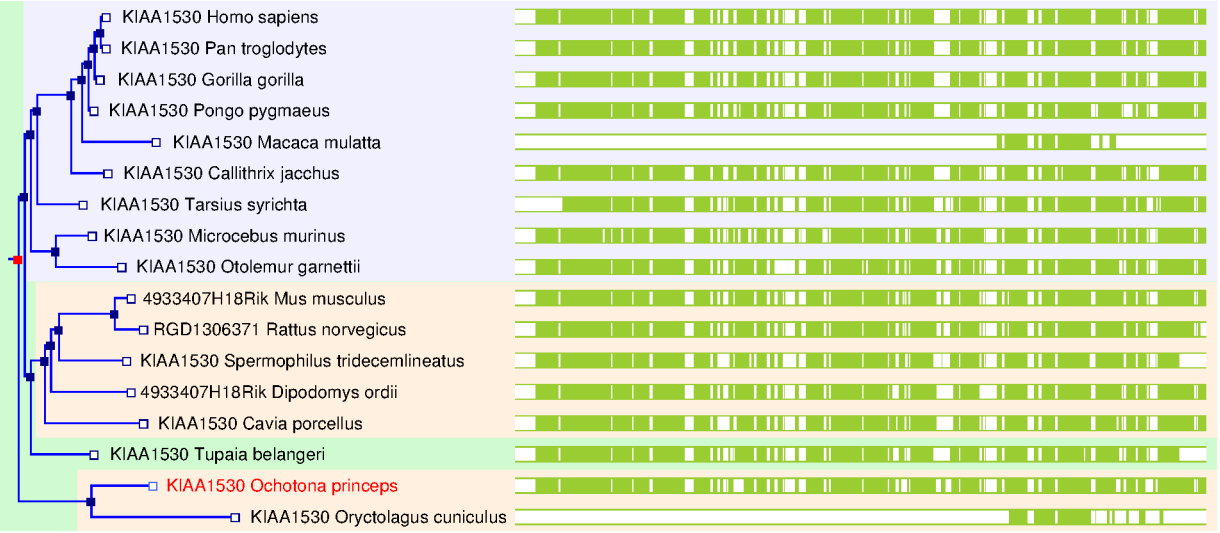


FIGURE 13.2 – Exemple d'arbre phylogénétique avec un nœud de duplication peu supporté. L'arbre phylogénétique est celui du gène KIAA1530, les carrés bleus indiquent des spéciations, les carrés rouges des duplications, et les carrés blancs les espèces modernes. La partie en vert à droite montre l'alignement des protéines correspondantes. Le gène KIAA1530 n'est donc chez aucun mammifère en deux copies et cependant, un nœud de duplication est inféré. Ce nœud est évidemment associé à un score de confiance de 0.

Algorithme 13.1 Édite les nœuds de duplication qui ont un score insuffisant

Variables globales: $\mathscr{A}_e$: l'arbre phylogénétique des espèces.

Entrées: $\mathscr{A}_g$: l'arbre phylogénétique d'un gène, réconcilié avec $\mathscr{A}_e$. s : seuil du score de duplication

1: **si** la racine de $\mathscr{A}_g$ est un nœud de duplication de score < s **alors**
2: Définir Anc, l'ancêtre désigné par la racine de $\mathscr{A}_g$
3: Créer N : un nouveau nœud d'arbre, ayant pour fils les nœuds F_i (les mêmes fils que Anc a dans l'arbre des espèces)
4: Assigner aux F_i les sous-arbres de $\mathscr{A}_g$ commençant au niveau 2
5: Transformer les F_i en nœuds de duplication si nécessaire
6: Appel récursif sur chaque F_i
7: Modifier $\mathscr{A}_g$ pour que sa racine soit N
8: **sinon**
9: Appel récursif sur chaque sous-arbre de $\mathscr{A}_g$ commençant au niveau 1
10: **renvoyer** $\mathscr{A}_g$
11: **fin si**

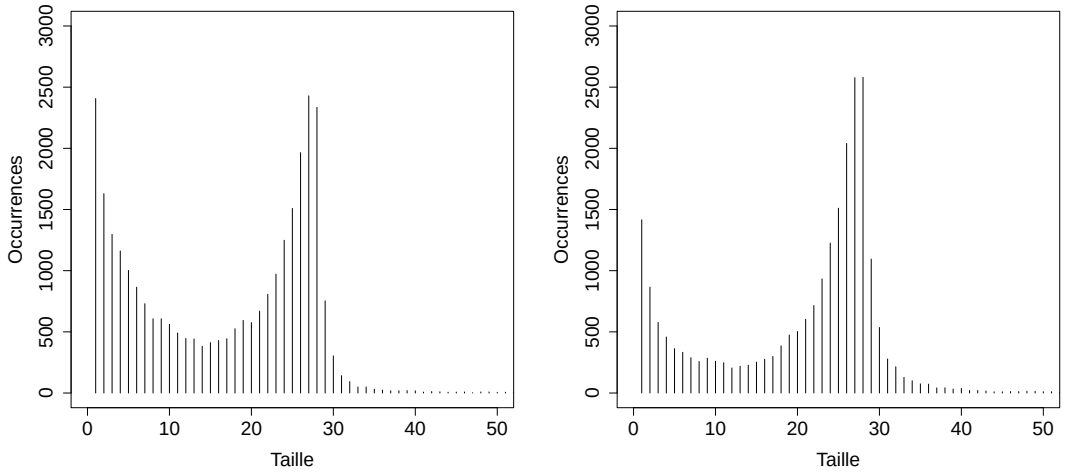


FIGURE 13.3 – Distribution des tailles de familles de *Boreoeutheria* avant (à gauche), et après (à droite) édition des nœuds de duplication.

tion. Les nœuds de duplication ayant un score supérieur au seuil ne sont, quant à eux, pas modifiés et resteront dans l'arbre final. Pour transformer un nœud de duplication en spéciation, il faut réarranger les sous-arbres qui le composent pour les réconcilier avec la phylogénie des espèces. Pour ceci, il peut être nécessaire de créer des nœuds de duplication si on lie deux sous-arbres qui correspondent au même ancêtre (comme Anc3 dans l'exemple de la Figure 13.4) ou à la même espèce (comme Spec1 dans l'exemple de la Figure 13.4). Dans tous les cas, l'algorithme est appelé récursivement sur chaque sous-arbre nouvellement défini pour le réconcilier et vérifier qu'il ne contient pas de nœud de duplication peu supportée.

Les duplications sont de manière générale des événements difficiles à prendre en compte dans les reconstructions et beaucoup de méthodes préfèrent se baser sur des gènes (ou des marqueurs) en 1 copie dans tous les génomes. AGORA utilise tous les gènes référencés par les arbres phylogénétiques, qu'ils soient ou non dupliqués, en utilisant les différentes copies d'un gène s'il en a. Ainsi, si un gène de *Boreoeutheria* est dupliqué (uniquement) chez la souris, AGORA effectuera les comparaisons d'ordre de gènes entre les deux copies de la souris et les copies (uniques) de l'homme, du chien, etc. Pour les reconstructions, AGORA est donc capable, en théorie, de gérer autant de duplications que nécessaire. Cependant, AGORA reste dépendant de la définition des gènes ancestraux, car la présence de nœuds de duplication définit le nombre de copies des gènes, et donc quels gènes peuvent être inclus dans les contigs reconstruits. La Figure 13.5 montre les conséquences sur AGORA de l'édition des nœuds de duplication, en particulier, via le nombre de copies ancestrales prédites.

A.1 Si une duplication est incorrecte (peu supportée), les deux copies qu'elle crée représenteront, en général, le même gène, présent à la même position du génome. Certaines espèces contiendront la première copie (situation notée *(1)*), et d'autres la deuxième copie (situation *(2)*). Les situations *(1)* et *(2)* seront donc vues alternativement dans les génomes modernes.

A.2 Avant édition, le graphe d'adjacence (à gauche du signe égal) utilisé pour la reconstruction des contigs de l'ancêtre contient les deux copies. Suivant les poids des

arêtes, une des deux copies va être choisie (laissant l'autre en singleton) ou les deux vont être associées à un des côtés du locus (solutions séparés par des barres obliques). Après édition (à droite de la flèche), le nœud de duplication est supprimé, les deux copies sont fusionnées dans un même gène, et il n'y a plus d'ambiguïté dans le graphe. AGORA reconstruit alors un unique contig. De manière générale, la longueur moyenne des contigs augmente.

B.1 Si une duplication est correcte (bien supportée), les deux copies qu'elle crée représentent des gènes différents, présents à des positions différentes dans les génomes (donc non orthologues). Un génome moderne possèdera donc en général les deux copies à deux positions différentes. Les situations notées *(1)* et *(2)* seront vues simultanément dans les génomes modernes.

B.2 Avant édition, AGORA reconstruit naturellement les deux contigs. Après édition,

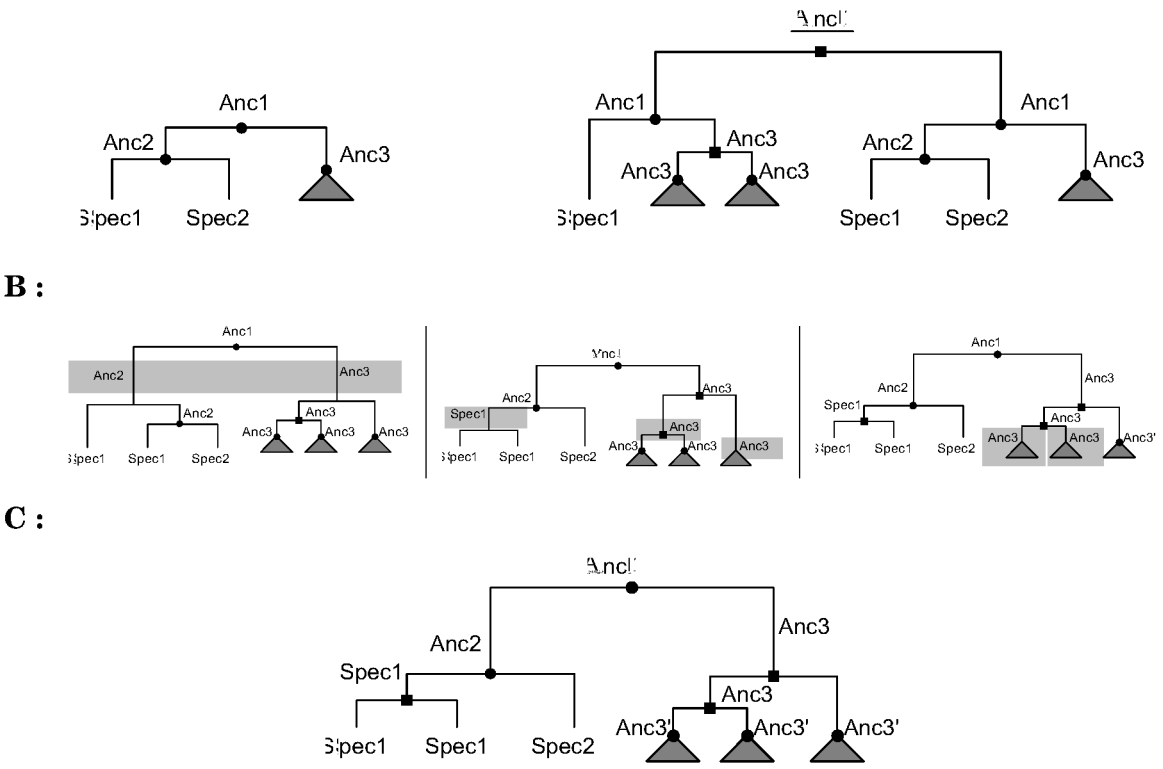
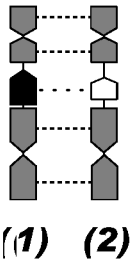
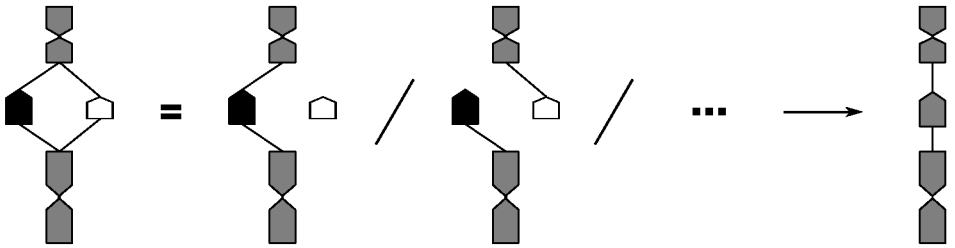


FIGURE 13.4 – Procédure d'édition des nœuds de duplication. **A :** Arbres phylogénétiques des espèces et d'un gène. Le triangle gris indique que de nombreuses espèces (gènes) sont présentes en dessous de Anc3. Les ronds indiquent des spéciations et les carrés des duplications. Un nœud de duplication est souligné s'il possède un score inférieur au seuil et doit être édité. **B :** Le programme crée deux fils Anc2 et Anc3 (comme dans la phylogénie des espèces) et leur attribue les sous-arbres correspondants. Puis, récursivement, il réconcilie la phylogénie de chacun de ces sous-arbres. Ainsi, pour le sous-arbre de Anc2, il faut déplacer le nœud Spec1 et définir un nœud de duplication. Le sous-arbre de Anc3, lui, contient une duplication bien supportée : il faut donc créer un nœud de duplication ancestral. Le programme continue sa marche sur chacun des sous-arbres de Anc3 (triangles gris). **C :** Résultat de la procédure d'édition. Le prime sur Anc3 signifie que ces sous-arbres ont eux-mêmes été modifiés.

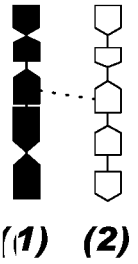
A.1 :



A.2 :



B.1 :



B.2 :

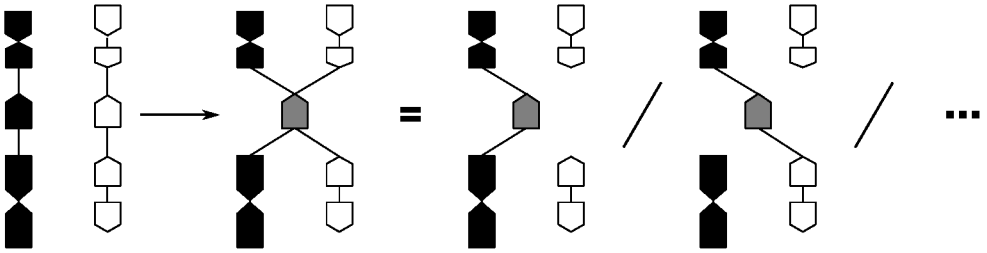


FIGURE 13.5 – Implications sur AGORA de l'édition des nœuds de duplication. Les tirets serrés entre les gènes montrent les liens d'orthologie et les tirets éloignés les liens de paralogie. Le code couleur des gènes montre si le gène est défini en 1 seule copie chez l'ancêtre (gris) ou en 2 copies (noir et blanc). Les *(1)* et *(2)* indiquent des configurations possibles des gènes dans les génomes modernes. Le signe égal, la barre oblique et les points de suspension montrent le graphe d'adjacence construit par AGORA et quelques solutions de contigs qui en résultent. La flèche sépare les situations avant et après édition. **A :** Suppression d'une duplication incorrecte. **B :** Suppression d'une duplication correcte. (voir le texte principal pour les explications)

AGORA ne dispose plus que d'un seul gène ancestral et est obligé de choisir une des deux positions (voire les mélanger) en laissant deux fragments de contigs. De manière générale, la longueur moyenne des contigs diminue.

Nous avons testé toutes les valeurs de seuil pour le score de duplication entre 0 et 0,5, par pas de 0,05 en mesurant la longueur moyenne des contigs reconstruits par AGORA. Le seuil optimal était celui qui permettait d'obtenir les contigs AGORA les plus longs. La Figure 13.6 montre l'évolution du N50 en fonction du seuil défini pour les scores de duplication et a une allure de cloche. Conformément à l'attendu, avec un seuil faible, la plupart des nœuds supprimés sont incorrects et leur délétion entraînent un allongement des contigs. Avec un seuil élevé, de nombreux nœuds corrects sont supprimés, et la taille des contigs qui en résultent diminue. La courbe admet un maximum en 0,30 : ce sera le seuil choisi pour les véritables reconstructions. Cette valeur est amenée à changer selon les versions d'Ensembl et les espèces incorporées. En pratique, depuis la mise en place de ce filtre méthode, la valeur optimale a toujours été de 0,25, 0,30 ou 0,35. Dans Ensembl v57, cela correspond à l'édition de 53038 nœuds de duplication sur un total de 151566 (35.0%). De plus, la Figure 13.3 (à droite) montre que l'édition a bien eu les effets escompté sur la distribution des tailles des familles en diminuant le nombre de «petites» familles.

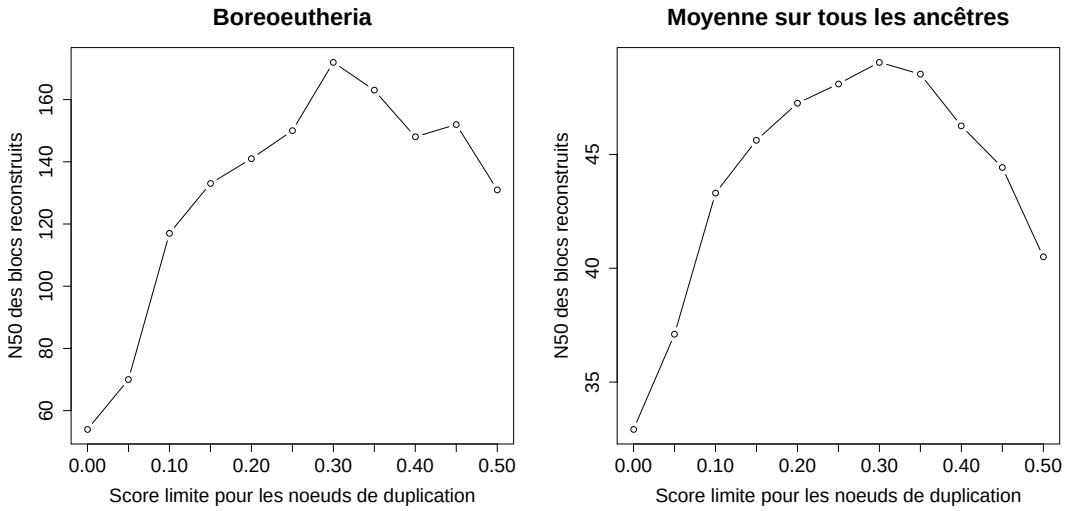


FIGURE 13.6 – Évolution des performances d'AGORA en fonction du seuil d'édition des nœuds de duplication

Après édition, les nouvelles statistiques de reconstruction sont présentées dans le tableau 13.4. Le nombre de gènes est plus constant sur la lignée humaine (maximum à 22137 gènes) et optimal vis à vis de la longueur des contigs. Comme aperçu sur la Figure 13.6, les contigs ont triplé de taille pour *Boreoeutheria*. De manière générale, l'édition des nœuds de duplication peu supportés a été bénéfique pour tous les ancêtres avec une augmentation moyenne de 49,0% du N50. Cette correction des arbres phylogénétiques d'Ensembl Compara doit être intégrée le plus tôt possible dans la mise en place des données, c'est-à-dire avant l'insertion des trois espèces supplémentaires non présentes dans Ensembl (sous-section 7.2.3). Le tableau 13.5 montre les statistiques après correction des données et insertion des nouvelles espèces (les performances de reconstruction sont quasiment identiques).

Ancêtre	Âge (Ma)	Gènes	Contigs	Contigs (> 100 gènes)	Couverture (gènes)		Couverture (intervalles)		25%	50%	75%	N75	N50	N25	Max	Moyenne
Fungi/Metazoa group	1500	2355	12	0	24	1,02%	12	0,51%	2	2	2	2	2	2	2	2,00
Bilateria	580	7949	54	0	120	1,51%	66	0,83%	2	2	2	2	2	2	6	2,22
Chordata	550	9689	147	0	309	3,19%	162	1,68%	2	2	2	2	2	2	7	2,10
Ciona	100	10048	1578	0	4701	46,79%	3123	31,14%	2	2	3	2	3	4	13	2,98
Clupeocephala	320	19660	3510	0	14814	75,35%	11304	57,56%	2	3	5	3	5	9	38	4,22
Percomorpha	190	19998	1728	1	17534	87,68%	15806	79,12%	3	6	13	9	18	33	100	10,15
Tetraodontidae	65	18417	1645	3	16176	87,83%	14531	78,99%	3	5	12	9	18	32	110	9,83
Sauria	267	16653	1105	11	14540	87,31%	13435	80,77%	3	5	14	12	31	65	186	13,16
Neognathae	105	15267	738	18	13260	86,85%	12522	82,13%	4	8	21	18	40	85	213	17,97
Phasianidae	46	15366	755	14	13210	85,97%	12455	81,16%	3	8	21	19	40	71	199	17,50
Euteleostomi	420	17416	3018	0	10615	60,95%	7597	43,67%	2	3	4	2	4	6	28	3,52
Tetrapoda	359	17565	2496	0	12965	73,81%	10469	59,67%	2	4	6	4	7	13	40	5,19
Amniota	326	18294	1212	10	14993	81,96%	13781	75,41%	3	5	13	11	26	62	238	12,37
Mammalia	184	18584	1344	12	15189	81,73%	13845	74,58%	3	5	12	9	25	54	209	11,30
Theria	166	19192	908	22	16361	85,25%	15453	80,60%	3	7	21	19	42	87	300	18,02
Eutheria	102	21382	637	42	17666	82,62%	17029	79,72%	3	8	31	36	78	158	352	27,73
Boreoeutheria	95	22137	411	59	18218	82,30%	17807	80,51%	2	7	42	79	172	276	493	44,33
Euarchontoglires	90	21813	381	58	18225	83,55%	17844	81,88%	2	9	51	80	162	284	580	47,83
Primates	83	20880	491	50	18010	86,25%	17519	83,98%	3	13	38	47	115	197	484	36,68
Haplorrhini	57	20657	501	48	18005	87,16%	17504	84,82%	3	12	39	48	104	194	484	35,94
Simiiformes	45	21454	506	46	18336	85,47%	17830	83,19%	3	11	37	48	109	233	463	36,24
Catarrhini	31	21956	531	51	18600	84,71%	18069	82,37%	2	11	36	49	106	210	468	35,03
Hominidae	16	21816	525	54	18795	86,15%	18270	83,82%	2	10	38	51	112	194	434	35,80
Homininae	9	22029	594	53	19110	86,75%	18516	84,13%	2	9	34	47	98	174	493	32,17
Homo/Pan group	5	21135	520	59	19015	89,97%	18495	87,59%	2	10	43	57	109	196	387	36,57
Metatheria	148	17470	1798	0	15329	87,74%	13531	77,54%	3	5	11	7	14	23	71	8,53
Atlantogenata	100	19512	1186	6	17289	88,61%	16103	82,61%	4	8	17	13	27	49	120	14,58
Afrotheria	94	19172	1202	7	17268	90,07%	16066	83,89%	4	8	18	13	26	47	175	14,37
Laurasiatheria	88	21139	469	64	18260	86,38%	17791	84,24%	3	13	50	54	113	167	298	38,93
Xenarthra	64	15354	2642	0	7060	45,98%	4418	28,81%	2	2	3	2	3	4	14	2,67
Insectivora	68	15694	2673	0	7092	45,19%	4419	28,19%	2	2	3	2	3	4	22	2,65
Cetartiodactyla	61	19790	1050	14	17910	90,50%	16860	85,28%	3	9	22	17	32	61	172	17,06
Chiroptera	60	17666	2979	0	14574	82,50%	11595	65,71%	2	3	5	3	6	13	61	4,89
Carnivora	56	18499	1253	7	17004	91,92%	15751	85,24%	4	8	17	12	23	42	147	13,57
Lagomorpha	48	17971	1653	2	16310	90,76%	14657	81,65%	3	6	12	9	17	31	117	9,87
Strepsirrhini	69	17312	3058	0	14079	81,33%	11021	63,73%	2	3	6	3	6	10	36	4,60
Glires	81	20276	491	59	17950	88,53%	17459	86,19%	3	14	45	45	101	154	474	36,56
Rodentia	80	19844	583	42	17842	89,91%	17259	87,06%	4	13	40	36	69	126	388	30,60
Sciurognathi	79	19676	873	19	17741	90,17%	16868	85,82%	4	10	26	21	42	74	186	20,32
Murinae	37	20193	829	24	18643	92,32%	17814	88,31%	4	11	29	24	50	82	253	22,49

TABLE 13.4 – Résultats d'AGORA (nœuds de duplication corrigés). Statistiques sur les longueurs et nombres de contigs, selon les ancêtres, pour une utilisation d'AGORA (protocole 1-passe) sur les données d'Ensembl avec les nœuds de duplication corrigés.

Ancêtre	Âge (Ma)	Gènes	Contigs	Contigs (> 100 gènes)	Couverture (gènes)		Couverture (intervalles)		25%	50%	75%	N75	N50	N25	Max	Moyenne
Fungi/Metazoa group	1500	2355	13	0	26	1,10%	13	0,56%	2	2	2	2	2	2	2	2,00
Eumetazoa	700	9034	179	0	366	4,05%	187	2,07%	2	2	2	2	2	2	3	2,04
Bilateria	580	12038	224	0	470	3,90%	246	2,05%	2	2	2	2	2	2	6	2,10
Deuterostomia	560	12106	503	0	1075	8,88%	572	4,73%	2	2	2	2	2	2	6	2,14
Chordata	550	12781	538	0	1139	8,91%	601	4,71%	2	2	2	2	2	2	7	2,12
Tunicata	200	9415	121	0	244	2,59%	123	1,31%	2	2	2	2	2	2	4	2,02
Ciona	100	10048	1583	0	4715	46,92%	3132	31,23%	2	2	3	2	3	5	14	2,98
Clupeocephala	320	19753	3522	0	14831	75,08%	11309	57,31%	2	3	5	3	5	9	38	4,21
Percomorpha	190	19998	1743	1	17569	87,85%	15826	79,22%	3	6	13	9	18	33	100	10,08
Tetraodontidae	65	18417	1650	3	16191	87,91%	14541	79,04%	3	5	12	9	18	32	111	9,81
Sauria	267	16792	1111	12	14542	86,60%	13431	80,08%	3	5	14	12	31	64	186	13,09
Neognathae	105	15378	747	18	13281	86,36%	12534	81,61%	4	8	20	17	38	84	213	17,78
Phasianidae	46	15432	773	12	13255	85,89%	12482	80,99%	3	8	21	18	35	69	199	17,15
Euteleostomi	420	18171	3067	0	10787	59,36%	7720	42,53%	2	3	4	2	4	6	28	3,52
Tetrapoda	359	18070	2507	0	13026	72,09%	10519	58,28%	2	4	6	4	7	12	40	5,20
Amniota	326	18637	1226	10	15022	80,60%	13796	74,10%	3	5	13	11	26	59	238	12,25
Mammalia	184	18738	1362	11	15211	81,18%	13849	73,99%	3	5	11	9	24	51	209	11,17
Theria	166	19301	916	21	16387	84,90%	15471	80,24%	3	7	21	19	41	83	301	17,89
Eutheria	102	21445	639	42	17700	82,54%	17061	79,63%	3	8	31	35	77	150	366	27,70
Boreoeutheria	95	22184	406	62	18231	82,18%	17825	80,42%	3	8	47	74	164	251	492	44,90
Euarchontoglires	90	21854	381	61	18226	83,40%	17845	81,73%	2	9	54	81	157	257	580	47,84
Primates	83	20918	495	51	18015	86,12%	17520	83,84%	3	13	37	46	115	193	484	36,39
Haplorrhini	57	20694	503	50	18010	87,03%	17507	84,68%	3	12	38	47	104	192	485	35,81
Simiiformes	45	21484	506	45	18348	85,40%	17842	83,13%	3	11	36	47	118	233	466	36,26
Catarrhini	31	21983	537	52	18616	84,68%	18079	82,32%	2	11	36	46	111	210	466	34,67
Hominidae	16	21835	536	51	18808	86,14%	18272	83,76%	2	11	38	49	100	190	434	35,09
Homininae	9	22044	600	52	19114	86,71%	18514	84,06%	2	9	34	45	97	176	387	31,86
Homo/Pan group	5	21146	527	58	19032	90,00%	18505	87,59%	3	12	41	52	108	182	386	36,11
Metatheria	148	17485	1826	0	15325	87,65%	13499	77,29%	3	5	11	7	13	22	71	8,39
Atlantogenata	100	19512	1185	5	17282	88,57%	16097	82,58%	4	8	18	13	27	48	120	14,58
Afrotheria	94	19172	1207	6	17267	90,06%	16060	83,86%	4	8	18	13	26	45	175	14,31
Laurasiatheria	88	21139	468	63	18264	86,40%	17796	84,27%	3	13	49	52	113	173	299	39,03
Xenarthra	64	15354	2641	0	7059	45,97%	4418	28,81%	2	2	3	2	3	4	14	2,67
Insectivora	68	15694	2674	0	7094	45,20%	4420	28,20%	2	2	3	2	3	4	22	2,65
Cetartiodactyla	61	19790	1050	14	17916	90,53%	16866	85,31%	3	9	22	17	32	60	172	17,06
Chiroptera	60	17666	2980	0	14575	82,50%	11595	65,71%	2	3	5	3	6	13	61	4,89
Carnivora	56	18499	1255	7	17001	91,90%	15746	85,21%	4	8	17	12	23	42	147	13,55
Lagomorpha	48	17971	1656	2	16304	90,72%	14648	81,60%	3	6	12	9	17	31	117	9,85
Strepsirrhini	69	17312	3058	0	14084	81,35%	11026	63,76%	2	3	6	3	6	10	36	4,61
Glires	81	20276	497	59	17956	88,56%	17459	86,19%	3	14	45	45	100	152	336	36,13
Rodentia	80	19844	581	40	17839	89,90%	17258	87,06%	4	13	40	35	68	128	388	30,70
Sciurognathi	79	19676	875	18	17746	90,19%	16871	85,83%	4	10	25	20	42	73	186	20,28
Murinae	37	20193	832	24	18652	92,37%	17820	88,34%	4	11	29	23	50	84	253	22,42

TABLE 13.5 – Résultats d'AGORA (contigs 1-passe). Statistiques sur les longueurs et nombres de contigs, selon les ancêtres, pour une utilisation d'AGORA (protocole 1-passe) sur les données définitives (après correction des nœuds de duplication et insertion des nouvelles espèces).

13.2.2 Nécessité d'une approche en plusieurs passes

Bien que les résultats soient bien meilleurs après l'édition des nœuds de duplication, le N50 pour *Boreoeutheria* n'est toujours que de 164 gènes. Nous avons donc cherché à comprendre pourquoi les contigs étaient aussi courts et s'arrêtaient «aussi tôt». En effet, pour un ancêtre donné, une extrémité de contig est un gène qui n'est jamais suivi du même gène lorsqu'on considère des paires d'espèces informatives pour cet ancêtre. Ce constat est aussi valable pour les singletons en les considérant comme des contigs de taille 1. Le nombre élevé d'extrémités et de singletons (406×2 et 3953 pour *Boreoeutheria*) semble indiquer que des milliers de locus ont subi des réarrangements plusieurs fois indépendamment. Bien que cela puisse être un résultat de choix dans le débat de la

Ancêtre	Taille des familles (nb gènes)			% Singletons dans simulations		
	Intérieur	Extrémités	Singletons	Intérieur	Extrémités	Singletons
Clupeocephala	5,51	5,08	3,90	14,53%	20,86%	45,66%
Percomorpha	4,09	3,59	2,26	2,81%	11,38%	48,71%
Tetraodontidae	2,12	2,00	1,67	3,08%	8,08%	31,84%
Sauria	4,09	3,15	2,26	2,55%	11,61%	46,30%
Neognathae	3,19	2,83	2,11	1,86%	8,60%	38,18%
Phasianidae	2,06	1,86	1,65	2,13%	9,41%	33,19%
Euteleostomi	46,57	44,60	31,29	28,62%	24,02%	53,32%
Tetrapoda	39,36	38,24	25,42	10,46%	14,60%	54,33%
Amniota	38,16	36,24	17,89	2,15%	10,85%	61,61%
Mammalia	34,08	32,84	19,49	2,40%	11,31%	57,22%
Theria	32,73	31,85	15,34	2,27%	16,86%	63,06%
Eutheria	29,90	25,65	9,61	1,34%	19,26%	64,15%
Boreoeutheria	25,55	18,70	6,30	1,11%	25,00%	62,18%
Euarchontoglires	16,00	12,27	4,78	0,97%	22,43%	54,43%
Primates	8,61	7,68	3,59	1,19%	16,97%	52,34%
Haplorrhini	6,90	6,63	3,45	1,52%	16,02%	46,09%
Simiiformes	6,10	5,45	3,10	1,65%	23,69%	52,69%
Catarrhini	5,01	4,42	2,52	1,67%	25,43%	53,51%
Hominidae	3,96	3,51	2,11	1,64%	23,25%	52,50%
Homininae	2,97	2,55	1,67	2,17%	28,24%	55,78%
Homo/Pan group	2,00	1,84	1,36	1,13%	18,69%	46,42%
Metatheria	1,98	2,01	1,80	4,95%	12,90%	39,70%
Atlantogenata	4,30	4,01	2,36	1,33%	7,64%	43,10%
Afrotheria	2,83	2,75	1,78	1,87%	8,59%	40,02%
Xenarthra	1,83	1,78	1,72	32,51%	32,86%	34,34%
Laurasiatheria	9,59	6,69	2,66	0,84%	19,96%	57,76%
Insectivora	1,92	1,84	1,65	32,10%	32,23%	33,27%
Cetartiodactyla	3,57	3,07	1,71	1,36%	8,93%	32,99%
Chiroptera	1,92	1,86	1,72	31,43%	31,75%	33,89%
Carnivora	1,90	1,77	1,38	2,05%	8,76%	35,73%
Lagomorpha	1,99	1,88	1,58	32,91%	34,09%	39,43%
Strepsirrhini	1,90	1,82	1,61	32,18%	32,46%	34,62%
Glires	6,74	6,25	2,62	1,03%	19,32%	53,34%
Rodentia	4,86	5,16	2,38	0,98%	16,46%	49,74%
Sciurognathi	3,86	4,13	2,30	1,32%	10,65%	40,03%
Murinae	2,15	2,46	2,16	2,61%	16,23%	52,13%
Moyenne	*(100%)*	92,20%	60,71%	7,41%	18,32%	47,04%

TABLE 13.6 – Caractéristiques des fins de contigs (extrémités) et des singletons, par rapport aux gènes présents à l'intérieur des contigs. Pour les tailles de familles de gènes, les gènes à l'intérieur des contigs fixent le 100%.

réutilisation des points de cassure (abordé dans les perspectives, section 17.4), nous avons voulu vérifier si le phénomène était réel ou artefactuel.

En particulier, nous avons vérifié, comme précédemment, les tailles des familles. Il en ressort que les gènes présents aux extrémités des contigs correspondent à des familles contenant beaucoup moins de gènes, et les singletons à des familles encore plus petites (Tableau 13.6). De plus, les singletons des reconstructions réelles sont très souvent singletons dans les reconstructions des simulations, alors que les génomes y sont aléatoires. Il en ressort que la caractéristique d'un gène singleton et d'une extrémité de contig n'est pas d'être un lieu de réarrangements intensifs, mais d'être une famille composée peu de gènes. Un contig s'arrête par manque de liens d'orthologie et non à cause de perte d'adjacence conservée (sauf pour les ancêtres éloignés, comme *Euteleostomi*, pour lesquels la différence est moins flagrante). En effet, un gène ayant 2 fois moins de descendants qu'un autre aura environ 4 fois moins de chance d'avoir une adjacence conservée entre deux espèces. Sachant que moins de la moitié des espèces utilisées sont séquencées et assemblées entièrement, on se rend compte que pour certains ancêtres, le risque de ne pas inclure dans les contigs des gènes pour des questions uniquement d'annotation ou de reconstruction phylogénétique sera élevé.

Il est désormais clair que le protocole multi-passes (section 10.1), une fois que l'on aura défini un jeu de gènes propres, permettra d'aller au delà des extrémités actuelles. Il est donc nécessaire d'avoir des gènes dits robustes (sous-section 7.3.3 et sous-section 7.3.2), possédant des représentants dans tous les génomes modernes, et si possible à chaque fois en 1 seule copie. Une autre conclusion est que l'on pourra mettre de côté les singletons des analyses des génomes ancestraux, car ils représentent peu de gènes modernes, et les familles ancestrales correspondantes sont incomplètes (ou fragmentées).

13.3 Optimisation de la reconstruction multi-passes

Nous avons cherché à définir ici la meilleure reconstruction possible pour AGORA, dans le cadre du protocole multi-passes, en réglant les deux paramètres associés.

Le jeu de gènes robustes Deux méthodes (*size* et *events*) ont été définies en sous-section 7.3.3 et sous-section 7.3.2 et dépendent toutes deux de valeurs numériques à régler qui définissent quelles familles sont considérées comme robustes.

La fonction de sélection, f, de la sous-section 9.3.2 Cette fonction choisit quels gènes non-robustes insérer dans les contigs de gènes robustes. Cinq fonctions sont définies : «somme des poids des arêtes maximale», «chemin le plus long», «chemin le plus court», «poids des arêtes maximaux» et «moyenne des poids des arêtes maximale».

Les listes de gènes robustes testées sont les suivantes : $size(T_{\min} = 1,\ T_{\max} = 1)$, $size(T_{\min} = 0,9,\ T_{\max} = 1,1)$, $size(T_{\min} = 0,75,\ T_{\max} = 1,33)$, $events(p_g = 0,20)$, $events(p_g = 0,25)$, $events(p_g = 0,33)$, $events(p_g = 0,50)$, $events(p_g = 0,66)$ et $events(p_g = 0,75)$. En pratique les algorithmes *size* et *events* ont été modifiés pour ne mesurer les pertes et les duplications que sur les espèces séquencées à haute couverture, car on sait que celles séquencées à basse couverture sont encore largement incomplètes. En combinant ces paramètres avec les 5 fonctions de sélection possibles, cela fait $9 \times 5 = 45$ combinaisons de paramètres à tester pour établir le protocole multi-passes optimal, et à comparer au résultat du protocole 1-passe, afin d'être certain d'un gain. Toutes ces combinaisons n'ont pas été testées car, historiquement, la méthode *events* a été mise au point après la mé-

thode *size*, à un moment où la fonction de sélection était déjà choisie. Nous ne disposons donc que de $3 \times 5 + 6 \times 1 = 21$ combinaisons de paramètres.

Nous réutilisons le protocole de simulation du chapitre 12 pour établir des génomes simulés (modernes et ancestraux). La reconstruction sur les génomes simulés modernes fournit des génomes ancestraux que l'on compare à l'attendu : les génomes ancestraux simulés. Les simulations font intervenir exactement les mêmes gènes que les arbres d'Ensembl (après correction des nœuds de duplication) dans 20 chromosomes initiaux, et avec un taux de réarrangement 1x. Nous n'avons pas réussi à définir de valeur optimale de ρ pour le ρ-groupage des gènes. En effet, la valeur de ρ qui permet d'obtenir dans les simulations des longueurs de contigs les plus proches des contigs réels dépend de chaque ancêtre et balaie toute la gamme entre 0 et 1. Nous avons donc conservé les 11 valeurs possibles entre 0 et 1 (par pas de 0,1) et lancé à chaque fois 10 instances de génomes aléatoires. Pour chaque reconstruction, cinq statistiques ont été mesurées.

Tout d'abord, on peut comparer les intervalles des contigs reconstruits aux intervalles des chromosomes attendus, et calculer une sensibilité et une spécificité, au sens classique du terme. En mesurant les cas de vrais positifs V_p (intervalles correctement prédits), de faux positifs F_p (intervalles prédits, et faux), et de faux négatifs F_n (intervalles attendus, et non-prédits), on peut calculer la spécificité $S_n = V_p/(V_p + F_n)$ et la sensibilité $S_p = V_p/(V_p + F_p)$.

Cependant, compte tenu des remarques du paragraphe précédent, les gènes sont singletons plus par essence, que par un taux élevé de réarrangements ou par faiblesse de la reconstruction. Nous avons donc mesuré les mêmes statistiques de sensibilité et de spécificité sur les intervalles des contigs et des chromosomes, en y supprimant d'abord les singletons. Les nouvelles mesures, notées S'_n et S'_p, seront nécessairement plus favorables que S_n et S_p car elles ne compteront pas d'erreur lorsqu'AGORA a laissé des gènes en singletons.

Enfin, nous avons mesuré le nombre d'erreurs de fusions de chromosomes E_c, c'est-à-dire le nombre d'intervalles prédits qui lient des gènes censés être sur des chromosomes différents, et qui produisent un contig qui est la fusion de deux morceaux de chromosomes différents. Ces erreurs, bien qu'intervenant sur un nombre restreint d'intervalles sont critiques pour l'affichage et l'étude du caryotype ancestral. En effet, une seule de ces erreurs peut définir un chromosome ancestral qui est la fusion de deux vrais chromosomes distincts, ce qui peut à son tour rendre visuellement peu crédible la solution et cristalliser des débats [Froenicke *et al.*, 2006] sur la qualité d'une méthode de reconstruction. Alors que dans le même temps, une inversion (si petite soit-elle) à l'intérieur d'un contig est généralement considérée comme moins grave alors qu'elle correspond pourtant à deux intervalles incorrects.

Les résultats sont montrés dans le tableau 13.7 pour *Boreoeutheria* mais sont, bien sûr, disponibles pour tous les autres ancêtres. Parmi les combinaisons avec la méthode *size*, les deux meilleurs paramètres sont f = «chemin le plus long» et *size*($T_{\min} = 1$, $T_{\max} = 1$). Ils permettent respectivement de maximiser la sensibilité, et de minimiser le nombre d'erreurs de fusions de chromosomes.

Ces paramètres doivent être optimisés pour tous les ancêtres, ce qui peut, a priori, donner des valeurs de paramètres optimales différentes. Cependant, nous considérons que la fonction de sélection f est caractéristique de la méthode de reconstruction au sens général (plus précisément du protocole multi-passes) et doit être indépendante des ancêtres. Nous nous attendions d'ailleurs à devoir choisir la fonction «poids des arêtes maximaux» (ou une autre fonction qui maximise une mesure liée aux poids des arêtes) mais les ré-

Jeu de gènes robustes	Fonction de sélection f	S_p	S_n	S'_p	S'_n	E_c
$size(T_{\min}=0,75,\ T_{\max}=1,33)$	somme des poids des arêtes maximale	89,04%	76,51%	98,40%	96,52%	1,67
	chemin le plus long	92,71%	82,93%	98,92%	97,63%	1,67
	chemin le plus court	88,63%	76,00%	98,02%	95,97%	1,68
	poids des arêtes maximaux	88,37%	75,43%	98,20%	96,14%	1,67
	moyenne des poids des arêtes maximale	88,29%	75,30%	98,14%	96,03%	1,67
$size(T_{\min}=0,9,\ T_{\max}=1,1)$	somme des poids des arêtes maximale	89,05%	76,51%	98,39%	96,51%	1,43
	chemin le plus long	92,71%	82,90%	98,92%	97,61%	1,43
	chemin le plus court	88,98%	76,74%	97,84%	95,86%	1,44
	poids des arêtes maximaux	88,37%	75,42%	98,20%	96,12%	1,43
	moyenne des poids des arêtes maximale	88,37%	75,43%	98,13%	96,01%	1,43
$size(T_{\min}=1,\ T_{\max}=1)$	somme des poids des arêtes maximale	89,06%	76,47%	98,40%	96,46%	1,32
	chemin le plus long	92,68%	82,78%	98,92%	97,55%	1,32
	chemin le plus court	89,18%	77,15%	97,70%	95,70%	1,32
	poids des arêtes maximaux	88,40%	75,40%	98,21%	96,08%	1,32
	moyenne des poids des arêtes maximale	88,58%	75,78%	98,11%	95,97%	1,33
$events(p_g=0,20)$	chemin le plus long	92,42%	81,66%	98,31%	95,76%	1,58
$events(p_g=0,25)$	chemin le plus long	92,42%	81,66%	98,31%	95,75%	1,56
$events(p_g=0,33)$	chemin le plus long	92,27%	81,51%	98,09%	95,39%	1,32
$events(p_g=0,50)$	chemin le plus long	92,14%	81,36%	97,89%	95,06%	1,41
$events(p_g=0,66)$	chemin le plus long	92,03%	81,28%	97,76%	94,91%	1,28
1-passe (reconstruction à améliorer)		88,57%	75,14%	98,30%	95,66%	1,58

TABLE 13.7 – Performances, pour *Boreoeutheria*, d'après les simulations, des variantes du protocole multi-passes pour les contigs. S_n et S_p sont la sensibilité et la spécificité (les versions avec un prime sont calculées sans tenir compte des singletons), E_c désigne le nombre d'intervalles moyen, dans des contigs reconstruits, qui lient deux chromosomes différents. Les statistiques du protocole 1-passe sont rajoutées à titre informatif.

Ancêtre	Âge (Ma)	Gènes	Contigs	Contigs (> 100 gènes)	Couverture (gènes)		Couverture (intervalles)		25%	50%	75%	N75	N50	N25	Max	Moyenne
Fungi/Metazoa group	1500	2355	16	0	32	1,36%	16	0,69%	2	2	2	2	2	2	2	2,00
Eumetazoa	700	9034	183	0	374	4,14%	191	2,12%	2	2	2	2	2	2	3	2,04
Bilateria	580	12038	236	0	496	4,12%	260	2,16%	2	2	2	2	2	2	6	2,10
Deuterostomia	560	12106	521	0	1119	9,24%	598	4,95%	2	2	2	2	2	2	6	2,15
Chordata	550	12781	550	0	1171	9,16%	621	4,87%	2	2	2	2	2	2	6	2,13
Tunicata	200	9415	124	0	251	2,67%	127	1,35%	2	2	2	2	2	2	4	2,02
Ciona	100	10048	1440	0	4892	48,69%	3452	34,42%	2	3	4	2	4	6	20	3,40
Clupeocephala	320	19753	3369	0	15006	75,97%	11637	58,97%	2	3	5	3	6	9	31	4,45
Percomorpha	190	19998	767	35	17729	88,65%	16962	84,90%	2	4	11	34	190	441	648	23,11
Tetraodontidae	65	18417	1140	22	16365	88,86%	15225	82,76%	2	5	12	14	45	81	289	14,36
Sauria	267	16792	1055	13	14705	87,57%	13650	81,39%	3	6	16	13	31	64	215	13,94
Neognathae	105	15378	459	33	13561	88,18%	13102	85,31%	3	9	30	36	91	168	546	29,54
Phasianidae	46	15432	683	17	13802	89,44%	13119	85,12%	4	10	25	20	43	77	199	20,21
Euteleostomi	420	18171	2995	0	11007	60,57%	8012	44,14%	2	3	4	3	4	7	29	3,68
Tetrapoda	359	18070	2304	0	13271	73,44%	10967	60,76%	2	4	7	4	8	14	52	5,76
Amniota	326	18637	858	31	15220	81,67%	14362	77,14%	3	5	13	18	68	170	528	17,74
Mammalia	184	18738	1212	18	15419	82,29%	14207	75,90%	3	5	12	11	30	69	317	12,72
Theria	166	19301	709	39	16554	85,77%	15845	82,18%	3	6	22	28	78	169	524	23,35
Eutheria	102	21445	382	48	17985	83,87%	17603	82,16%	2	5	27	104	278	444	1031	47,08
Boreoeutheria	95	22184	168	27	18757	84,55%	18589	83,87%	2	4	13	564	798	1192	1826	111,65
Euarchontoglires	90	21854	214	41	18821	86,12%	18607	85,22%	2	5	55	209	536	842	1765	87,95
Primates	83	20918	1544	0	18565	88,75%	17021	81,45%	5	9	16	10	18	29	94	12,02
Haplorrhini	57	20694	1604	0	18540	89,59%	16936	81,92%	4	9	15	10	17	27	94	11,56
Simiiformes	45	21484	610	47	19038	88,61%	18428	85,86%	4	13	34	34	76	161	378	31,21
Catarrhini	31	21983	273	67	19345	88,00%	19072	86,84%	2	6	97	142	251	419	863	70,86
Hominidae	16	21835	262	58	19712	90,28%	19450	89,16%	3	9	83	135	275	459	877	75,24
Homininae	9	22044	468	50	20073	91,06%	19605	89,02%	4	20	53	50	98	190	472	42,89
Homo/Pan group	5	21146	221	63	19816	93,71%	19595	92,75%	3	15	111	137	312	525	743	89,67
Metatheria	148	17485	2667	0	15199	86,93%	12532	71,75%	3	4	7	4	7	12	33	5,70
Atlantogenata	100	19512	3014	0	17433	89,35%	14419	73,97%	3	5	7	5	7	11	47	5,78
Afrotheria	94	19172	3317	0	17372	90,61%	14055	73,39%	3	4	7	4	6	10	32	5,24
Xenarthra	64	15354	2673	0	7063	46,00%	4390	28,63%	2	2	3	2	3	4	12	2,64
Laurasiatheria	88	21139	182	42	18698	88,45%	18516	87,67%	2	4	59	261	526	757	1152	102,74
Insectivora	68	15694	2740	0	7109	45,30%	4369	27,87%	2	2	3	2	2	3	12	2,59
Cetartiodactyla	61	19790	811	33	18128	91,60%	17317	87,59%	3	10	28	26	51	97	231	22,35
Chiroptera	60	17666	3733	0	14676	83,07%	10943	62,01%	2	3	5	3	4	7	33	3,93
Carnivora	56	18499	2573	0	17226	93,12%	14653	79,30%	3	5	8	5	8	13	35	6,69
Strepsirrhini	69	17312	3588	0	14178	81,90%	10590	61,24%	2	3	5	3	5	7	25	3,95
Lagomorpha	48	17971	2174	0	16443	91,50%	14269	79,49%	3	5	10	6	11	18	97	7,56
Glires	81	20276	441	56	18276	90,14%	17835	88,05%	4	12	43	52	135	238	628	41,44
Rodentia	80	19844	367	58	18068	91,05%	17701	89,29%	3	14	57	71	163	236	476	49,23
Sciurognathi	79	19676	985	13	17920	91,08%	16935	86,16%	4	9	23	17	37	66	197	18,19
Murinae	37	20193	398	57	18788	93,04%	18390	91,16%	3	11	57	69	148	262	475	47,21

TABLE 13.8 – Résultats d'AGORA (contigs multi-passes). Statistiques sur les longueurs et nombres de contigs, selon les ancêtres, pour une utilisation d'AGORA (protocole multi-passes optimal).

sultats montrent que la fonction «chemin le plus long» suffit. Il s'avère que l'architecture des contigs donnée par les gènes robustes est de très bonne qualité et limite l'insertion de gènes inadéquats. La méthode «chemin le plus long» essaie d'insérer le maximum de gènes non-robustes dans chaque intervalle, ce qui, par définition, permet de maximiser la sensibilité. La spécificité, elle, ne faiblit pas car l'algorithme de rajout des gènes non-robustes inclut, dans tous les cas, une procédure de sélection en fonction des poids des arêtes (lorsqu'un gène peut être inclus à plusieurs endroits). L'insertion des gènes non-robustes ne se fait donc pas complètement sans tenir compte des poids des arêtes.

Pour la sélection des gènes robustes, il faut rappeler que les méthodes *size* et *events* ont un esprit différent. *events* sélectionne les meilleurs arbres phylogénétiques en entier. Après application de *events*, les listes de gènes de tous les ancêtres sont cohérentes entre elles, et contiennent tous les gènes ancestraux possibles liés aux mêmes gènes modernes (dans une proportion fixée globalement p_g). Une valeur de p_g est censée définir simultanément les gènes de tous les ancêtres et mélanger les résultats de différentes sélections *events* selon les ancêtres aurait peu de sens. En revanche, la méthode *size* découpe les arbres en sous-arbres jusqu'à ce qu'ils rentrent dans les critères de tailles demandés. Un gène moderne perdra donc (en général) ses homologies distantes après application de *size* et pour des mêmes paramètres $T_{\min}$ et $T_{\max}$, le taux de familles conservées sera plus élevé pour un ancêtre récent que pour un ancêtre vieux (en effet, il est plus facile de trouver des familles sans perte ni duplication en comparant 2 espèces qu'en en comparant 20). Les paramètres $T_{\min}$ et $T_{\max}$ peuvent donc naturellement être relâchés lorsqu'on considère des ancêtres éloignés, voire, pour ceux-ci, revenir au protocole 1-passe.

Ici, les meilleurs résultats sont donnés par $size(T_{\min} = 1,\ T_{\max} = 1)$ et par $events(p_g = 0,33)$ ou $events(p_g = 0,66)$. Ces paramètres (en particulier p_g) varient selon les ancêtres. Pour l'ensemble des ancêtres, la meilleure solution est de combiner plusieurs paramètres de *size* selon les ancêtres et leur âge (plutôt que de choisir une valeur de p_g). Ainsi $size(T_{\min} = 1,\ T_{\max} = 1)$ est utilisé pour les ancêtres *Clupeocephala*, *Theria*, et leurs descendants. $size(T_{\min} = 0,9,\ T_{\max} = 1,1)$ est utilisé pour les ancêtres *Amniota*, et *Sauria* et ses descendants. Enfin, $size(T_{\min} = 0,75,\ T_{\max} = 1,33)$ est utilisé pour les ancêtres *Euteleostomi*, *Tetrapoda* et *Mammalia*. Les résultats optimaux sont donnés sur le tableau 13.8 (résumé pour *Boreoeutheria* et *Amniota* dans le Tableau 13.12). Les contigs de *Boreoeutheria* triplent presque de taille et le N75 passe au-dessus des 500 gènes. C'est-à-dire que trois quarts des gènes de *Boreoeutheria* sont présents dans des contigs de plus de 500 gènes (pour information, seuls 4 chromosomes humains ont une taille inférieure à 500 gènes). Les reconstructions d'*Amniota* augmentent de 50% en moyenne et le N50 est presque triplé.

13.4 Reconstruction en scaffolds

Maintenant qu'AGORA a reconstruit l'ordre des gènes de la manière la plus précise possible (contigs), nous pouvons passer au niveau d'adjacence suivant (l'adjacence des contigs) pour former des scaffolds. Là encore, nous avons le choix entre le protocole 1-passe et le protocole multi-passes. Le protocole multi-passes a été testé avec les blocs robustes $length(l_{\min} = 50)$, $length(l_{\min} = 20)$, $proportion(p = 50)$ et $proportion(p = 70)$. Nous avons réutilisé les mêmes jeux de génomes simulés que dans le paragraphe précédent. En reprenant les mêmes statistiques que précédemment, les résultats du tableau 13.9 montrent que le protocole 1-passe fait beaucoup moins d'erreurs que le protocole multi-passes (quel

	Reconstruction	S_p	S_n	S'_p	S'_n	E_c
Boreoeutheria	(contigs multi-passe)	92,68%	82,78%	98,92%	97,55%	1,32
	scaffolds 1-passe	92,62%	82,87%	98,88%	97,67%	1,79
	scaffolds multi-passes – $length(l_{\min} = 50)$	92,58%	82,87%	98,85%	97,68%	5,72
	scaffolds multi-passes – $length(l_{\min} = 20)$	92,58%	82,87%	98,85%	97,68%	5,49
	scaffolds multi-passes – $proportion(p = 50)$	92,62%	82,87%	98,88%	97,67%	2,16
	scaffolds multi-passes – $proportion(p = 70)$	92,61%	82,87%	98,87%	97,67%	2,96
Amniota	(contigs multi-passe)	90,91%	74,56%	98,8%	94,03%	0,85
	scaffolds 1-passe	90,37%	75,96%	98,65%	96,2%	1,11
	scaffolds multi-passes – $length(l_{\min} = 50)$	90,22%	75,99%	98,50%	96,26%	11,46
	scaffolds multi-passes – $length(l_{\min} = 20)$	90,34%	75,97%	98,61%	96,22%	3,4
	scaffolds multi-passes – $proportion(p = 50)$	90,31%	75,98%	98,59%	96,23%	5,16
	scaffolds multi-passes – $proportion(p = 70)$	90,21%	75,99%	98,49%	96,26%	12,22
Euteleostomi	(contigs multi-passe)	85,94%	48,32%	99,45%	87,2%	0,43
	scaffolds 1-passe	81,97%	49,04%	99,21%	92,7%	0,67
	scaffolds multi-passes – $length(l_{\min} = 50)$	81,82%	49,05%	99,04%	92,72%	6,52
	scaffolds multi-passes – $length(l_{\min} = 20)$	81,96%	49,04%	99,21%	92,70%	0,86
	scaffolds multi-passes – $proportion(p = 50)$	81,01%	49,15%	98,21%	93,07%	51,01
	scaffolds multi-passes – $proportion(p = 70)$	79,93%	49,32%	97,32%	93,76%	116,2

TABLE 13.9 – Performances, pour différents ancêtres, d'après les simulations, des variantes du protocole multi-passes pour les scaffolds. S_n et S_p sont la sensibilité et la spécificité (les versions avec un prime sont calculées sans tenir compte des singletons), E_c désigne le nombre d'intervalles moyen, dans des contigs reconstruits, qui lient deux chromosomes différents. Les statistiques du protocole 1-passe sont rajoutées à titre de comparaison.

Ancêtre	Âge (Ma)	Gènes	Scaffolds	Scaffolds (> 100 gènes)	Couverture (gènes)		Couverture (intervalles)		25%	50%	75%	N75	N50	N25	Max	Moyenne
Fungi/Metazoa group	1500	2355	16	0	32	1,36%	16	0,69%	2	2	2	2	2	2	2	2,00
Eumetazoa	700	9034	181	0	374	4,14%	193	2,14%	2	2	2	2	2	2	4	2,07
Bilateria	580	12038	233	0	496	4,12%	263	2,19%	2	2	2	2	2	2	6	2,13
Chordata	560	12106	513	0	1119	9,24%	606	5,01%	2	2	2	2	2	2	6	2,18
Olfactores	550	12781	542	0	1171	9,16%	629	4,93%	2	2	2	2	2	2	6	2,16
Tunicata	200	9415	120	0	251	2,67%	131	1,39%	2	2	2	2	2	2	4	2,09
Ciona	100	10048	1218	0	4892	48,69%	3674	36,64%	2	3	5	3	5	8	40	4,02
Clupeocephala	320	19753	2838	0	15006	75,97%	12168	61,66%	2	3	6	4	7	14	54	5,29
Percomorpha	190	19998	615	36	17729	88,65%	17114	85,66%	2	4	12	52	354	543	659	28,83
Tetraodontidae	65	18417	677	40	16365	88,86%	15688	85,27%	2	4	11	47	215	335	629	24,17
Sauria	267	16792	549	34	14705	87,57%	14156	84,40%	2	5	16	43	131	373	737	26,79
Neognathae	105	15378	219	33	13561	88,18%	13342	86,87%	3	9	35	144	349	468	747	61,92
Phasianidae	46	15432	192	37	13802	89,44%	13610	88,31%	2	8	64	156	293	390	1252	71,89
Euteleostomi	420	18171	2679	0	11007	60,57%	8328	45,88%	2	3	5	3	5	9	70	4,11
Tetrapoda	359	18070	1707	3	13271	73,44%	11564	64,07%	2	4	9	6	13	27	136	7,77
Amniota	326	18637	609	30	15220	81,67%	14611	78,48%	2	5	14	34	175	440	650	24,99
Mammalia	184	18738	713	30	15419	82,29%	14706	78,57%	2	5	14	27	98	319	673	21,63
Theria	166	19301	362	41	16554	85,77%	16192	83,98%	2	7	30	82	247	434	730	45,73
Eutheria	102	21445	232	39	17985	83,87%	17753	82,86%	2	4	33	264	468	795	1119	77,52
Boreoeutheria	95	22184	133	25	18757	84,55%	18624	84,03%	2	3	29	596	952	1192	1826	141,03
Euarchontoglires	90	21854	128	29	18821	86,12%	18693	85,61%	2	3	53	478	794	1191	1851	147,04
Primates	83	20918	189	43	18565	88,75%	18376	87,93%	2	13	86	198	432	687	1017	98,23
Haplorrhini	57	20694	188	45	18540	89,59%	18352	88,77%	2	16	95	182	377	658	1014	98,62
Simiiformes	45	21484	171	42	19038	88,61%	18867	87,90%	2	6	85	249	546	791	1056	111,33
Catarrhini	31	21983	150	34	19345	88,00%	19195	87,40%	2	2	36	553	724	889	1186	128,97
Hominidae	16	21835	121	32	19712	90,28%	19591	89,81%	2	3	192	533	740	943	1276	162,91
Homininae	9	22044	147	36	20073	91,06%	19926	90,47%	2	3	99	353	627	1098	2005	136,55
Homo/Pan group	5	21146	98	33	19816	93,71%	19718	93,34%	2	3	303	417	626	1056	1984	202,20
Metatheria	148	17485	715	32	15199	86,93%	14484	82,93%	2	6	21	27	67	138	282	21,26
Atlantogenata	100	19512	437	53	17433	89,35%	16996	87,19%	3	11	44	52	125	224	641	39,89
Afrotheria	94	19172	441	51	17372	90,61%	16931	88,40%	3	11	44	51	117	222	450	39,39
Xenarthra	64	15354	2602	0	7063	46,00%	4461	29,09%	2	2	3	2	3	4	16	2,71
Laurasiatheria	88	21139	123	31	18698	88,45%	18575	87,95%	2	3	77	526	735	1069	1636	152,02
Insectivora	68	15694	2644	0	7109	45,30%	4465	28,49%	2	2	3	2	3	4	23	2,69
Cetartiodactyla	61	19790	340	53	18128	91,60%	17788	89,97%	2	5	32	126	297	375	688	53,32
Chiroptera	60	17666	2961	0	14676	83,07%	11715	66,39%	2	3	6	3	6	13	56	4,96
Carnivora	56	18499	264	62	17226	93,12%	16962	91,79%	3	23	86	86	170	261	568	65,25
Lagomorpha	48	17971	731	37	16443	91,50%	15712	87,53%	2	5	16	31	92	239	473	22,49
Strepsirrhini	69	17312	2938	0	14178	81,90%	11240	65,00%	2	3	6	3	6	11	44	4,83
Glires	81	20276	163	35	18276	90,14%	18113	89,42%	2	4	66	332	689	1010	1743	112,12
Rodentia	80	19844	194	46	18068	91,05%	17874	90,16%	2	5	75	227	396	594	926	93,13
Sciurognathi	79	19676	320	48	17920	91,08%	17600	89,54%	3	17	63	73	198	330	782	56,00
Murinae	37	20193	178	42	18788	93,04%	18610	92,25%	2	5	85	269	480	777	1560	105,55

TABLE 13.10 – Résultats d'AGORA (scaffolds 1-passe). Statistiques sur les longueurs et nombres de scaffolds, selon les ancêtres, pour une utilisation d'AGORA (contigs multi-passes optimal, suivi de 1-passe pour établir des scaffolds).

Ancêtre	S_p	S_n	S'_p	S'_n	E_c
Clupeocephala	84,2%	63,57%	96,54%	91,26%	0,68
Percomorpha	91,8%	82,41%	98,66%	97,34%	0,47
Tetraodontidae	94,7%	87,07%	99,33%	97,28%	0,18
Sauria	93,68%	82,24%	98,87%	95,04%	0,59
Neognathae	95,04%	88,13%	99,25%	97,99%	0,33
Phasianidae	95,75%	90,86%	99,47%	98,95%	0,07
Euteleostomi	81,97%	49,04%	99,21%	92,7%	0,67
Tetrapoda	86,81%	63,48%	98,41%	92,21%	0,84
Amniota	90,37%	75,96%	98,65%	96,2%	1,11
Mammalia	90,95%	76,92%	98,83%	95,81%	1,07
Theria	91,3%	79,25%	98,81%	97,2%	1,53
Eutheria	91,04%	79,68%	98,75%	97,36%	1,48
Boreoeutheria	92,62%	82,87%	98,88%	97,67%	1,79
Euarchontoglires	94,25%	86,57%	98,94%	97,73%	1,4
Primates	94,62%	87,2%	99,08%	97,57%	0,68
Haplorrhini	95,4%	88,85%	99,21%	97,7%	0,43
Simiiformes	94,0%	85,88%	98,99%	97,63%	0,34
Catarrhini	93,35%	84,7%	98,88%	97,7%	0,21
Hominidae	94,0%	86,39%	99,13%	98,26%	0,02
Homininae	93,51%	84,36%	99,33%	98,34%	0
Homo/Pan group	96,29%	91,58%	99,49%	99,01%	0
Metatheria	93,97%	80,58%	99,59%	95,38%	1,05
Atlantogenata	95,01%	87,7%	99,16%	97,21%	0,51
Afrotheria	95,94%	88,68%	99,4%	97,16%	0,41
Xenarthra	99,75%	48,61%	100,0%	72,67%	0,15
Laurasiatheria	94,55%	87,24%	99,1%	98,14%	1,17
Insectivora	99,84%	49,2%	99,99%	72,66%	0,15
Cetartiodactyla	95,97%	90,81%	98,99%	97,86%	0,46
Chiroptera	99,92%	50,06%	99,99%	72,83%	0,18
Carnivora	96,11%	89,57%	99,65%	97,84%	0,35
Lagomorpha	99,8%	48,15%	99,99%	72,07%	0,15
Strepsirrhini	99,89%	49,34%	100,0%	72,73%	0,15
Glires	94,93%	88,08%	99,2%	97,95%	0,93
Rodentia	95,6%	89,45%	99,31%	98,14%	0,94
Sciurognathi	96,06%	89,89%	99,41%	97,8%	0,71
Murinae	95,39%	89,67%	99,49%	98,91%	0,35

TABLE 13.11 – Performances, pour différents ancêtres, d'après les simulations, des reconstructions AGORA en scaffolds. S_n et S_p sont la sensibilité et la spécificité (les versions avec un prime sont calculées sans tenir compte des singletons), E_c désigne le nombre d'intervalles moyen, dans des contigs reconstruits, qui lient deux chromosomes différents.

que soit le jeu de contigs robustes). Nous avons donc effectué les reconstructions en scaffolds avec le protocole 1-passe (statistiques de longueurs de blocs sur le tableau 13.10).

13.5 Performances réelles d'AGORA

La procédure finale pour reconstruire les ordres de gènes dans les ancêtres est schématisée sur la Figure 13.7. La Tableau 13.12 montre le chemin parcouru dans ce chapitre

	Méthode AGORA	Gènes	Contigs	Couverture (gènes)		Couverture (intervalles)	
Boreoeutheria	*sans édition des nœuds de duplication*						
	contigs / 1-passe *	29479	1117	19575	66,40%	18458	62,66%
	avec édition des nœuds de duplication						
	contigs / 1-passe *	22137	411	18218	82,30%	17807	80,51%
	contigs / 1-passe	22184	406	18231	82,18%	17825	80,42%
	contigs / multi-passes	22184	168	18757	84,55%	18589	83,87%
	scaffolds / 1-passe	22184	133	18757	84,55%	18624	84,03%
Amniota	*sans édition des nœuds de duplication*						
	contigs / 1-passe *	21455	1379	15345	71,52%	13966	65,16%
	avec édition des nœuds de duplication						
	contigs / 1-passe *	18294	1212	14993	81,96%	13781	75,41%
	contigs / 1-passe	18637	1226	15022	80,60%	13796	74,10%
	contigs / multi-passes	18637	858	15220	81,67%	14362	77,14%
	scaffolds / 1-passe	18637	609	15220	81,67%	14611	78,48%

	Méthode AGORA	25%	50%	75%	N75	N50	N25	Max	Moyenne
Boreoeutheria	*sans édition des nœuds de duplication*								
	contigs / 1-passe *	2	4	20	26	54	93	211	17,52
	avec édition des nœuds de duplication								
	contigs / 1-passe *	2	7	42	79	172	276	493	44,33
	contigs / 1-passe	3	8	47	74	164	251	492	44,90
	contigs / multi-passes	2	4	13	564	798	1192	1826	111,65
	scaffolds / 1-passe	2	3	29	596	952	1192	1826	141,03
Amniota	*sans édition des nœuds de duplication*								
	contigs / 1-passe *	3	5	12	9	22	52	163	11,13
	avec édition des nœuds de duplication								
	contigs / 1-passe *	3	5	13	11	26	62	238	12,37
	contigs / 1-passe	3	5	13	11	26	59	238	12,25
	contigs / multi-passes	3	5	13	18	68	170	528	17,74
	scaffolds / 1-passe	2	5	14	34	175	440	650	24,99
Moyenne des vertébrés		332	649	944	689	952	1367	5054	721,54

TABLE 13.12 – Évolution des statistiques de reconstruction de *Boreoeutheria* et *Amniota* au fur et à mesure de l'amélioration de la méthode. L'astérisque indique les reconstructions avant insertion des 3 espèces supplémentaires. Les moyennes sur l'ensemble des génomes des vertébrés sont rappelées à titre indicatif.

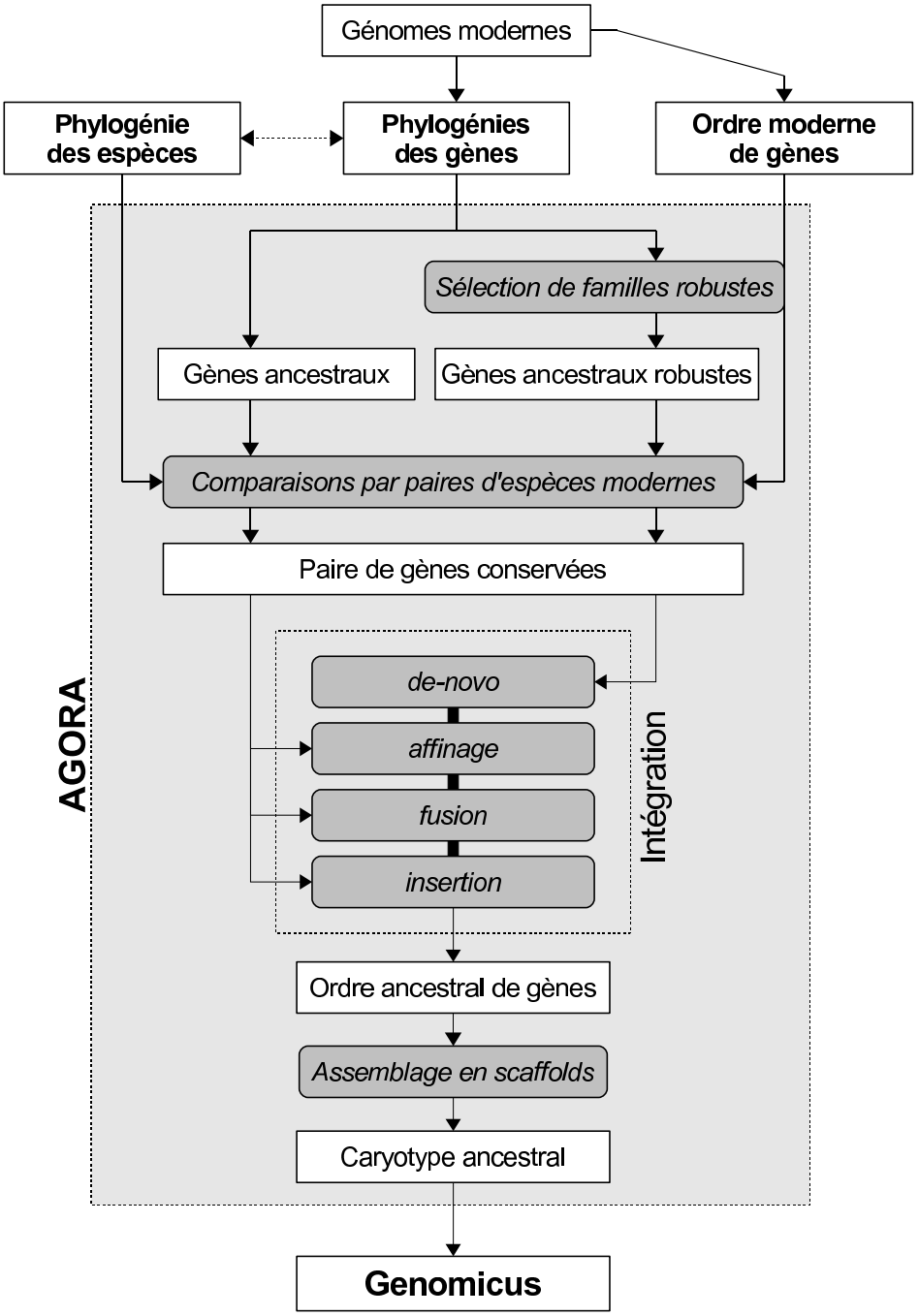


FIGURE 13.7 – Schéma récapitulatif de la procédure de reconstruction de l'ordre des gènes dans AGORA. Les données utilisées en entrée d'AGORA sont les annotations disponibles dans la base de données Ensembl (annotations des gènes et arbres phylogénétiques). Dans ce chapitre, nous avons paramétré AGORA pour extraire des sous-jeux de gènes, dits robustes, pour produire l'ossature de la reconstruction de l'ordre de gènes, avant de réinsérer les autres gènes. Enfin, nous appliquons une reconstruction en scaffolds (qui permet d'atteindre le caryotype pour de nombreux ancêtres, voir chapitre 14). Les données sont, en fin de processus, mises à disposition de la communauté, via le serveur Genomicus (chapitre 15).

(sur *Boreoeutheria* et *Amniota*), pour la configuration et l'optimisation des reconstructions AGORA. Les performances finales (longueur des scaffolds : tableau 13.10, performances d'après simulations : Tableau 13.11) montrent que les mammifères sont très bien reconstruits (excepté pour les ancêtres situés à la convergence d'espèces séquencées à faible couverture). Au-delà, le signal d'adjacence est trop faible pour pouvoir être aussi bien exploité et les scaffolds restent de taille assez limitée. Il sera donc nécessaire d'utiliser les reconstructions basées sur *walktrap* et *concorde*. Nous n'avons pas encore paramétré ces approches, mais une tentative de *walktrap* sur l'ancêtre *Euteleostomi* (section suivante) augure de bons résultats.

13.6 Duplications de génomes

Pour les reconstructions basées sur l'exploitation de duplications complètes de génomes, il n'y a pas de paramètres à régler dans les méthodes AGORA. Il a donc uniquement fallu appliquer les méthodes vues au chapitre 11 pour les ancêtres *Teleostei* et *Chordata*. Enfin, pour *Euteleostomi* (le plus vieil ancêtre des vertébrés), pour lequel les méthodes AGORA basées sur l'ordre des gènes éprouvent des difficultés, nous avons testé la méthode *walktrap*, afin de vérifier que le résultat était proche de la reconstruction de *Chordata* post-duplication. Les statistiques sont données sur le Tableau 13.13 et confirment l'intérêt des méthodes basées sur les duplications complètes de génomes : les ancêtres sont reconstruits avec une couverture (nombre de gènes inclus dans les chromosomes) et un nombre de chromosomes inégalés par les méthodes AGORA classiques (quelque dizaines de chromosomes, avec un N50 de plus de 500 gènes). De plus, on remarque que la reconstruction avec *walktrap* de l'ancêtre *Euteleostomi* est équivalente à un caryotype complet, malgré son âge.

Ancêtre	Gènes	Chromosomes	Chromosomes (> 100 gènes)	Singletons
Teleostei (pré-dup)	14712	40	14	43
Chordata (pré-dup)	5169	20	11	0
Chordata (post-dup)	15265	44	38	546
Euteleostomi (walktrap)	18171	97	28	3195

Ancêtre	25%	50%	75%	N75	N50	N25	Max	Moyenne
Teleostei (pré-dup)	4	26	651	1035	1274	1406	1432	366,73
Chordata (pré-dup)	2	148	487	487	635	705	761	258,45
Chordata (post-dup)	149	279	497	294	502	635	908	331,52
Euteleostomi (walktrap)	2	2	205	385	693	913	1160	154,39
Moyenne des vertébrés	332	649	944	689	952	1367	5054	721,54

TABLE 13.13 – Statistiques de longueurs de blocs pour les reconstructions liées aux duplications complètes de génome.

Chapitre 14

Comparaison des résultats d'AGORA aux références

Sommaire

Le chapitre précédent a montré la faisabilité des approches AGORA pour la reconstruction de génomes ancestraux chez les vertébrés. La méthode a été paramétrée de manière optimale pour obtenir les génomes les plus proches des caryotypes ancestraux. Dans ce chapitre, nous allons comparer les résultats d'AGORA (y compris les reconstructions pas encore validées par des simulations) aux consensus existant pour les trois espèces ancestrales présentées dans l'introduction (chapitre 3).

14.1 *Boreoeutheria*

Les reconstructions chez les vertébrés ont été systématiquement menées avec la méthode basée sur les adjacences (reconstruction en contigs, puis en scaffolds). La reconstruction AGORA de l'ancêtre *Boreoeutheria* est composée de 25 scaffolds de plus de 150 gènes, et qui incluent 80,36% des gènes ancestraux (95,04% lorsqu'on ne tient pas compte des singletons). Ces 25 scaffolds sont en correspondance 1-1 avec les 24 chromosomes ancestraux définis dans l'introduction, à la différence près de la fusion des chromosomes humains 7 et 16 qui n'est pas vue.

Nous avons comparé l'ordre des gènes dans les scaffolds AGORA à l'ordre prédit dans les 29 CARs de Ma *et al.* [2006]. Leur reconstruction prédit 1309 adjacences ancestrales parmi 1338 segments conservés entre les mammifères. Nous avons transcrit les coordonnées de ces segments sur les nouveaux assemblages des génomes, et sélectionné les

	$l \geqslant 150$	$100 < l < 50$	$50 < l \geqslant < 10$	$10 > l$	Singletons
Nombre de scaffolds	25	5	15	88	3427
Nombre de gènes inclus	17827	379	301	250	3427

TABLE 14.1 – Statistiques sur les longueurs de scaffolds de *Boreoeutheria*.

925 segments conservés qui contenaient au moins un gène ancestral non-singleton dans AGORA, et qui pouvaient donc être comparés. Parmi les 896 adjacences qu'ils représentent, 850 sont également prédites par AGORA, 3 lient des extrémités de scaffolds AGORA, et 43 adjacences totalement incompatibles. Dans l'autre sens, ils prédisent 2 adjacences fausses (qu'AGORA a évité) et en ratent 4 qu'AGORA prédit. Une analyse plus poussée des différences est en cours, afin de préciser quelle configuration est correcte.

14.2 *Teleostei* pré-duplication

La reconstruction de l'ancêtre *Teleostei* pré-duplication a été effectuée par la méthode de la synténie dédoublée (avec comme paramètre pour le critère de proximité des gènes dans les espèces dupliquées : $d = 25$ gènes) AGORA calcule le même caryotype ancestral que [Kasahara *et al.*, 2007], mais en fusionnant leurs chromosomes f et g. Cette fusion, jamais prédite par les autres études semble artefactuelle et causée par l'utilisation simultanée de toutes les espèces de poissons téléostéens séquencés. Le Tableau 14.2 montre une concordance malencontreuse entre les chromosomes du poisson zèbre et de l'épinoche qui contiennent les fragments des chromosomes ancestraux f et g. En particulier, la paire de chromosomes 14/7 du poisson zèbre correspond exactement à ces deux chromosomes ancestraux. Notre programme de reconstruction utilise simultanément tous les génomes et reprend, au moment de la comparaison des blocs de synténie dédoublée, le système d'interpolation défini par l'Équation 6.1 pour calculer un score qui indiquera si deux blocs partagent la même alternance sur les chromosomes dupliqués. Dans notre étude, les blocs issus des chromosomes f et g ont donc eu un score de similarité élevé pour le poisson zèbre et l'épinoche, et cette valeur a été propagée à la valeur ancestrale moyenne, utilisée pour le clustering, ce qui a favorisé leur fusion.

Toutes les études automatiques précédentes étaient basées sur le génome de tétraodon, ou de medaka, et n'étaient donc pas impactées par la colocalisation des chromosomes

Chromosome ancestral	Chromosomes modernes							
	Tétraodon		Épinoche		Medaka		Poisson zèbre	
a	10	14	XV	XVIII	22	24	17 et 13	20
b	8 et 10	21	XX	X	16	11	16	19
c	2	3 et 5	XVI	I	21	2	9	6 1
d	2 et 18	17	V et IX	VI	19 et 1	15	12 et 1	13
e	2 et 18	3	V et IX	XI	19 et 1	8	12 et 1	3
f	18	20	IX	VII	1	18	1 et 14	7
g	1	7	IV	VII	10	14	14	21 7
h	7	10 et 16	VII	I	14	13	21, 5, et 10	15 et 18
i	4	12	XIV	XIII	12	9	10 et 21	5 et 8
j	5	13	II	XIX	3	6	7	25 et 18
k	13	19	XIX	IV	6	23	25 et 18	4
l	9	11	XII	XVII	7	5	23 et 8	11 et 6
m	1, 6, et 16	15	XXI et VIII	III	4 et 20	17	24 et 7	2

TABLE 14.2 – Répartition des chromosomes ancestraux *Teleostei* pré-duplication sur les chromosomes des espèces modernes, organisée selon les paires issues de la duplication complète. La nomenclature des chromosomes ancestraux provient de Kasahara *et al.* [2007].

f et g chez le poisson zèbre et l'épinoche. Cependant, en ajoutant que dans la reconstruction avec medaka [Kasahara *et al.*, 2007], les auteurs avaient remarqué une translocation du chromosome 18 sur le chromosome 10 d'un fragment du chromosome ancestral f, on peut se poser la question de la validité du modèle où f et g sont deux chromosomes ancestraux différents.

Il faudrait pour cela étudier la localisation des chromosomes f et g dans les espèces non-dupliquées. Si ces chromosomes sont colocalisés ou mélangés, on pourra, par parcimonie, conclure que les f et g représentent le même chromosome ancestral pré-duplication.

14.3 *Chordata* pré/post-duplication

La reconstruction des génomes pré-2R et post-2R a été menée avec les méthodes basées sur les ohnologues uniquement, en se servant du génome *Amniota* comme génome dupliqué, et avec comme paramètre $p_{\min} = 10^{-5}$ pour découper les scaffolds d'*Amniota* en segments non-réarrangés. Le caryotype pré-2R est composé de 11 chromosomes (et donc celui post-2R de 44 chromosomes). Nous n'avons pas encore comparé ce caryotype au résultat de [Nakatani *et al.*, 2007] et Putnam *et al.* [2008], et ne pouvons donc juger de la qualité de la reconstruction.

14.4 Évolution du caryotype chez les vertébrés

Dans les trois figures suivantes, les caryotypes dessinés comprennent les scaffolds de pus de 100 gènes (s'il y en a au moins 20, sinon aucune taille minimale n'est imposée), mais, dans tous les cas, jamais plus de 50 scaffolds ne seront dessinés. Les données chiffrées sur les comparaisons des ancêtres entre eux et la concordance des reconstructions ne sont pas encore disponibles, nous nous contenterons pour l'instant d'une description visuelle. Seuls les génomes reconstruits avec les scaffolds disposent d'un ordre de gènes. Pour les autres, les gènes sont dessinés selon l'ordre des chromosomes de l'espèce référence.

La première figure (Figure 14.1) montre une partie des caryotypes des espèces ancestrales reconstruits chez les vertébrés (les nœuds de concours des espèces séquencées à haute couverture) avec la méthode «classique» (ordre de gènes ancestral, dans les scaffolds). On retrouve visuellement la conclusion du tableau 13.10 : les ancêtres éloignés (en particulier *Euteleostomi*) sont très mal reconstruits (du point de vue de la continuité), ce qui confirme l'intérêt des méthodes de clustering ou utilisant les duplications complètes, pour pouvoir cibler des ancêtres très anciens.

La figure suivante (Figure 14.2) montre les caryotypes chez les poissons, en prenant comme référence le génome pré-duplication reconstruit par la méthode de synténie dédoublée. Malgré des méthodes différentes, il semble y avoir concordance entre le génome pré-duplication et les génomes reconstruits en scaffolds (*Percomorpha* et *Tetraodontidae*).

Enfin, nous avons testé le clustering en chromosomes (avec *walktrap*) sur l'ancêtre *Euteleostomi* (car c'était l'ancêtre le moins bien reconstruit après la phase scaffolds). La Figure 14.3 compare 4 génomes reconstruits avec des méthodes différentes et indépendantes : *Amniota* (scaffolds), *Teleostei* (synténie dédoublée), *Euteleostomi* (*walktrap*) et *Chordata* pré- et post-2R (appariement des ohnologues). La comparaison des génomes n'est pas encore chiffrée, mais visuellement, les génomes concordent nettement, ce qui validerait simultanément les 4 approches.

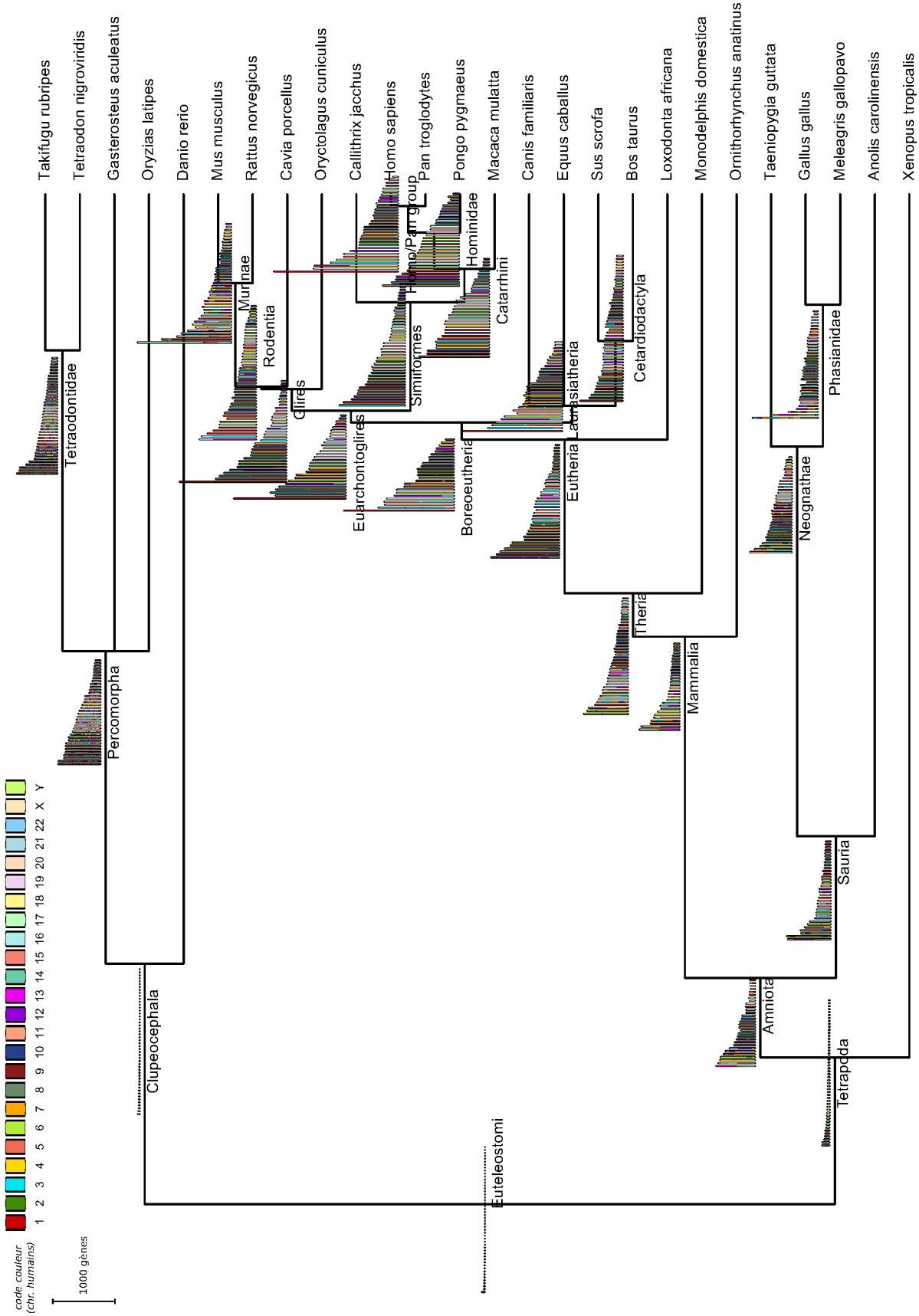


FIGURE 14.1 – Ensemble des reconstructions AGORA (scaffolds) chez les vertébrés (avec un code couleur selon les chromosomes humains.

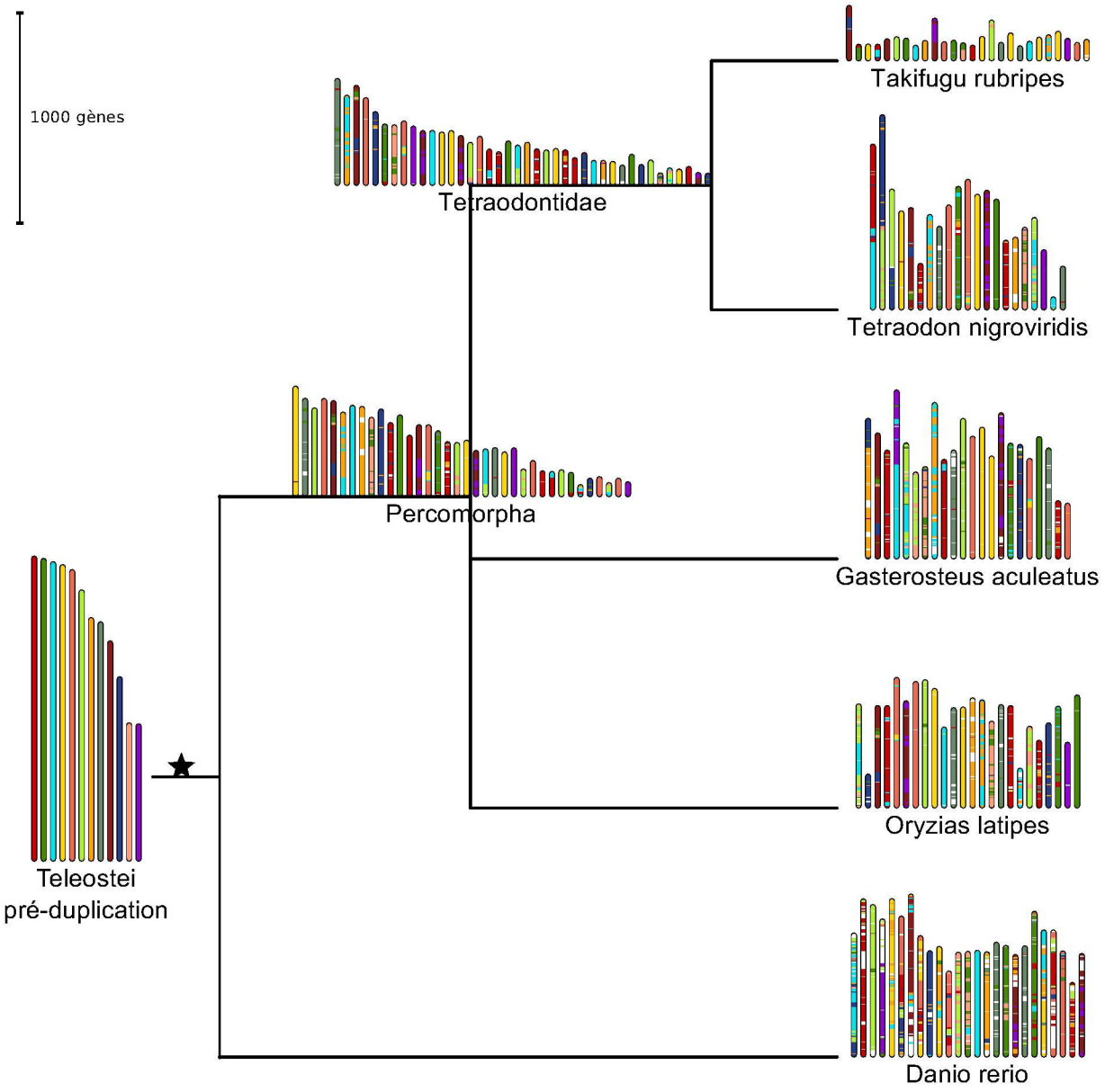


FIGURE 14.2 – Évolution du caryotype chez les poissons. L'ancêtre pré-duplication est reconstruit par la synténie dédoublée, les ancêtres *Percomorpha* et *Tetraodontidae* sont reconstruits par la méthode classique (scaffolds).

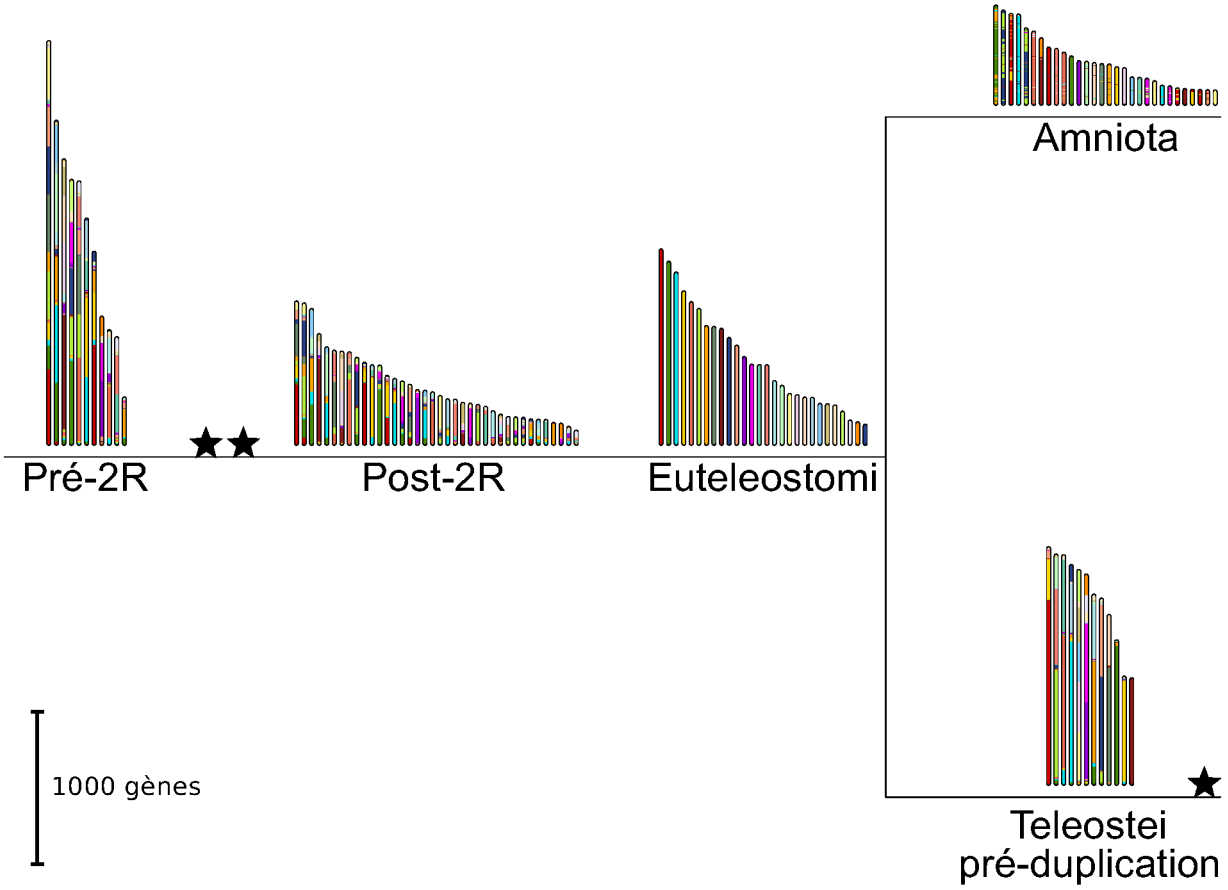


FIGURE 14.3 – Ensemble des reconstructions AGORA proche de l'origine des vertébrés. Les ancêtres pré-2R et post-2R sont reconstruits par clustering des ohnologues, l'ancêtre *Teleostei* pré-duplication est reconstruit par blocs de synténie dédoublée, l'ancêtre *Amniota* est reconstruit par la méthode classique (scaffolds), l'ancêtre *Euteleostomi* est reconstruit en clusterisant les scaffolds avec *walktrap*.

Chapitre 15

Navigateur de génomes : *Genomicus*

Sommaire

Nous avons défini dans les chapitres précédents une série d'outils pour comparer les génomes modernes et reconstruire des génomes ancestraux. Afin de pouvoir analyser ces résultats et les rendre accessibles à la communauté, il était nécessaire d'avoir un outil qui permette de visualiser ces données. Des outils existent pour comparer des génomes, mais ils sont généralement limités à la comparaison simultanée de deux ou trois espèces [Derrien *et al.*, 2007, Sinha et Meller, 2007, Courcelle *et al.*, 2008, Lyons *et al.*, 2008], ou proposent un choix restreint d'espèces [Byrne et Wolfe, 2005, Dong *et al.*, 2009], ce qui limite rapidement la profondeur de l'analyse et sa facilité. De plus, il nous était nécessaire d'avoir un outil capable de comparer l'état ancestral aux états modernes (ce qui n'était pas le cas de [Jensen *et al.*, 2009, Pan *et al.*, 2005]), et ainsi s'affranchir des traditionnelles comparaisons d'espèces modernes entre elles.

Le manque d'outils s'explique par le fait que la génomique comparative est encore relativement jeune, car elle n'est possible que depuis l'avènement du séquençage systématique des génomes des espèces qui nous entourent. Nous avons développé et mis en place un navigateur de génomes, *Genomicus*, entièrement fondé sur les données d'AGORA, et développé dans la seule optique de comparaison de génomes, qui permet rapidement d'analyser l'évolution d'un génome ou d'une de ses régions, sur de grandes distances phylogénétiques.

15.1 Présentation

Genomicus est accessible en ligne depuis début 2009 sur http://www.dyogen.ens.fr/genomicus/ [1]. Les données proviennent d'Ensembl et des reconstructions AGORA et sont mises à jour simultanément avec Ensembl (approximativement tous les deux mois) tandis que les versions précédentes sont archivées et restent disponibles. Depuis sa mise en ligne, Genomicus est continûment utilisé par des dizaines d'équipes dans le monde (voir Tableau 15.1). Genomicus est également accessible depuis les pages d'Ensembl (via

1. Les données complètes sont téléchargeables au format texte sur le serveur FTP ftp://ftp.biologie.ens.fr/pub/dyogen/genomicus/.

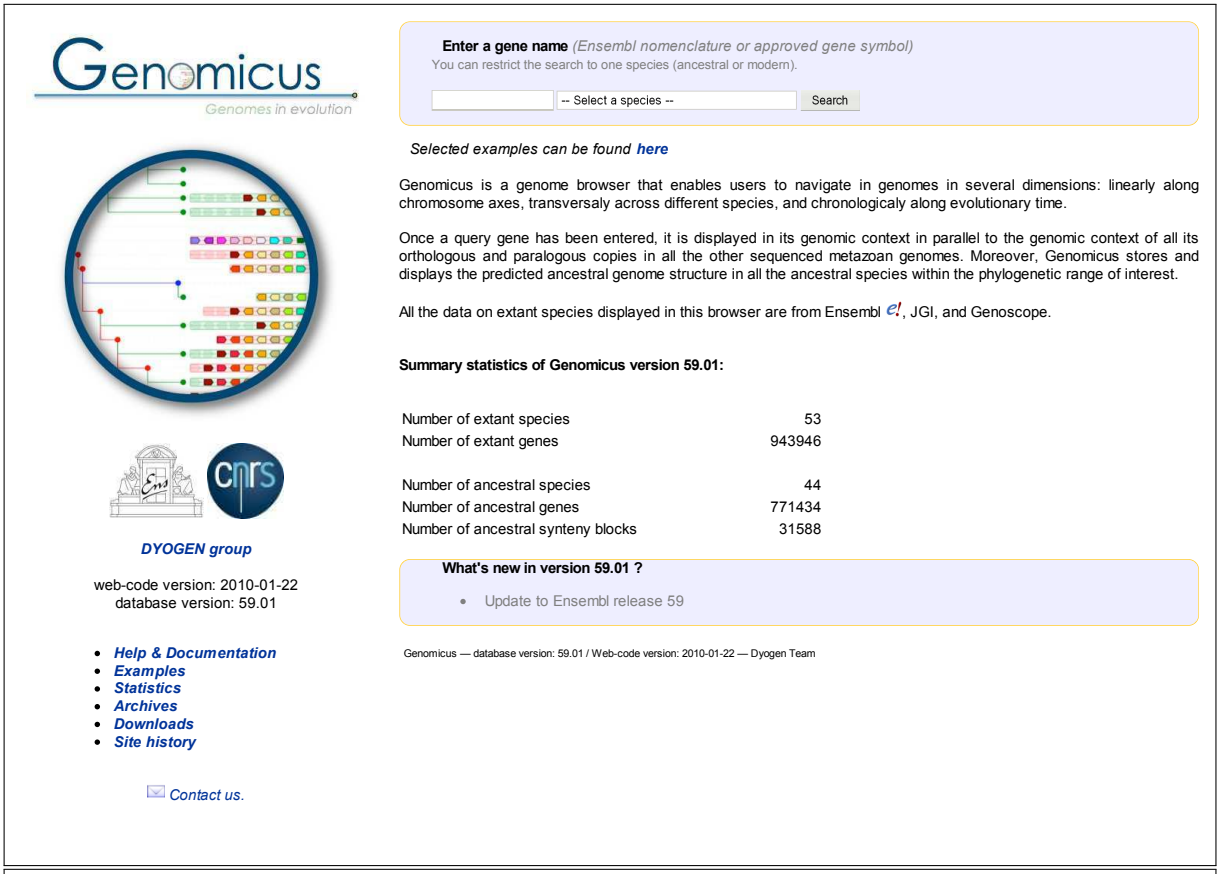


FIGURE 15.1 – Captures d'écran de Genomicus. **Haut :** Page d'accueil. **Bas :** Vue PhyloView sur le gène *fgf8a* (*fibroblast growth factor 8*) du poisson zèbre.

le lien *View synteny in Genomicus* des arbres de protéines) et le moteur de recherche *Bioinformatic Harvester* [2] [Liebel *et al.*, 2005] dans l'onglet «Évolution».

A :

Mois	Visiteurs différents	Visites	Hits	Bande passante
Oct 2009	734	1299	17334	1,92 Go
Nov 2009	803	1320	12525	1,55 Go
Déc 2009	629	974	7916	1,17 Go
Jan 2010	750	1186	8633	3,29 Go
Fév 2010	794	1207	20601	4,76 Go
Mar 2010	1421	2577	27383	13,45 Go
Avr 2010	967	1657	22767	8,02 Go
Mai 2010	773	1314	20931	4,22 Go
Juin 2010	654	1180	15638	4,18 Go
Juil 2010	587	1134	15959	3,68 Go
Aoû 2010	573	1405	14381	3,24 Go
Sep 2010	592	1241	16225	5,52 Go

B :

Pays	Nombre de pages	Bande passante
Europe (pays non identifié)	67336	4,25 Go
États-Unis	53103	4,59 Go
France	33698	4,87 Go
Norvège	10036	2,16 Go
Japon	6896	2,43 Go
Grande Bretagne	4345	1,10 Go
Singapour	4081	0,67 Go
Canada	3808	1,04 Go
Chine	3034	0,98 Go
Allemagne	2484	1,00 Go
Portugal	2327	0,78 Go

TABLE 15.1 – Statistiques d'utilisation du site Genomicus. **A :** Nombre de visites mensuelles, entre octobre 2009 et septembre 2010. Le pic d'utilisation en mars correspond à la parution de l'article Muffato *et al.* [2010]. **B :** Origine des connexions au site durant cette même période.

Techniquement, Genomicus est composé de scripts Perl, exécutés sur un serveur web Apache2 (grâce à mod_perl) et interfacés à une base de données MySQL. Les dessins sont constitués de commandes XHTML-SVG-CSS et rendus par le navigateur web de l'utilisateur. Du JavaScript (AJAX) permet d'augmenter l'interaction avec l'utilisateur.

Les vues actuellement disponibles sont accédées via un gène dit «de référence» et une espèce ancestrale, qui indique la profondeur phylogénétique de l'étude.

PhyloView montre le voisinage du gène de référence, ainsi que le voisinage de ses orthologues et paralogues dans les autres espèces. Les espèces sont ordonnées selon l'arbre phylogénétique du gène de référence. Ainsi, une espèce peut être en double si elle contient deux gènes paralogues, tout deux homologues du gène de référence.

2. http://harvester.fzk.de/

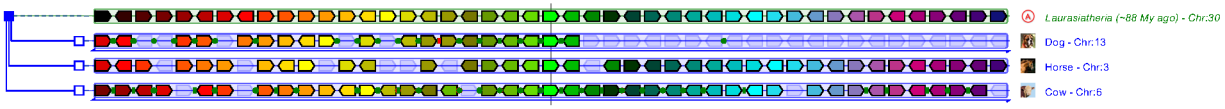
AlignView montre l'alignement du voisinage du gène de référence sur les autres espèces. Ici, une espèce peut être répartie sur plusieurs lignes si la région de référence y est éclatée sur plusieurs chromosomes.

Dans ces deux vues, le gène de référence est initialement au centre avec 15 gènes de chaque côté. Les homologues (orthologues et paralogues dans un code couleur différent) des gènes du voisinage du gène de référence sont coloriés dans les autres génomes. Les arbres peuvent être édités (réduction / développement, affichage de certains nœuds) pour clarifier la vue. Des liens permettent de se déplacer facilement dans la vue, de changer de gène de référence, et de revenir sur un navigateur classique de génomes (Ensembl, UCSC ou NCBI) pour accéder aux informations spécifiques d'un gène / d'une espèce.

Des éléments conservés non-codants on été identifiés et sont affichés dans Genomicus en plus de l'ordre des gènes. Ils ont été extraits de l'alignement multiple de 46 génomes de vertébrés (généré et disponible via l'UCSC). Trois niveaux de conservation sont distingués (affichés dans différentes couleurs) : les mammifères (homme + chien + vache + souris), les amniotes (mammifères + poulet) et les vertébrés (amniotes + poisson-zèbre). Les éléments conservés sont définis à partir d'une ancre de 20 bases, ayant un minimum d'identité (entre 60% et 80%, selon l'éloignement des espèces comparées) et étendue tant qu'il n'y a pas plus de 2 bases consécutives non strictement identiques entre les espèces comparées. Les zones annotées comme exons sont au préalable exclues de l'alignement.

15.2 Exemple d'utilisation

A :



B :

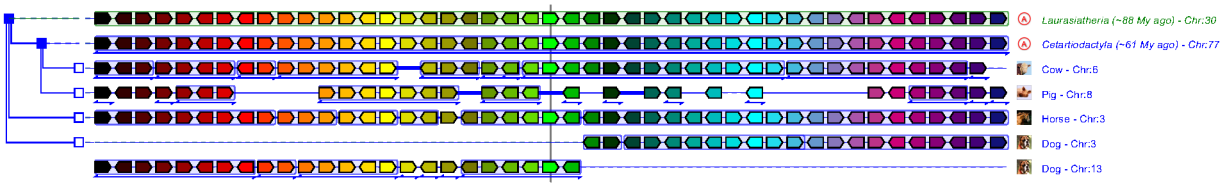


FIGURE 15.2 – Étude d'un réarrangement grâce à Genomicus. **A :** PhyloView sur le gène *PHOX2B* (*paired-like homeobox 2b*) du cheval. **B :** AlignView sur le même gène.

Sur la Figure 15.2.A (PhyloView), La région autour de PHOX2B est relativement bien conservée entre la vache et le cheval (ce qui permet de définir la configuration ancestrale). En revanche, seule la partie gauche est également synténique avec le chien. Grâce à AlignView, on voit que les gènes de la partie droite sont en effet sur un autre chromosome du chien. On peut donc conclure que la région ancestrale est restée quasiment identique chez la vache et le cheval, mais a subi une cassure (ou une transposition) chez le chien. D'autre part, PhyloView ne montre pas la région homologue du cochon car PHOX2B n'y est pas annoté, et cette vue fonctionne sur l'arbre phylogénétique du gène de référence. À l'inverse, AlignView aligne les régions dans leur globalité, ce qui permet d'inclure ici le

cochon. À ce propos, bien que la synténie sur le locus soit globalement conservée avec le cochon, l'ordre local des gènes l'est très peu. À ce stade, on ne peut pas savoir si cela est effectivement dû à un nombre important de réarrangements, ou si cela est dû aux données disponibles (erreurs dans l'assemblage ou dans l'annotation). Les éléments non-codants conservés sont indiqués par des cercles de couleurs.

15.3 Futures améliorations

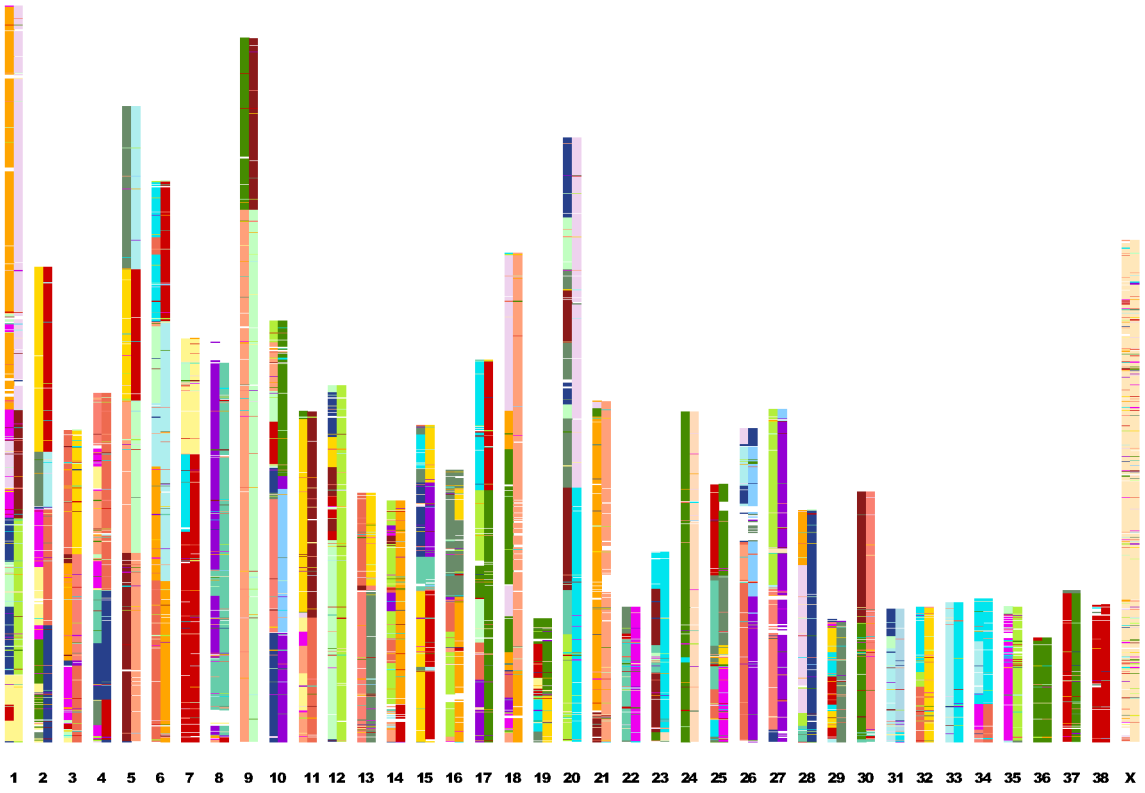


FIGURE 15.3 – Caryotype du génome du chien, avec un code couleur selon les chromosomes de la souris et de l'homme. Les ruptures communes de couleur indiquent (comme sur le quart supérieur du chromosome 9) des réarrangements spécifiques du chien (par rapport à *Euarchontoglires*), tandis que les ruptures de couleurs dans une seule bande (comme sur le tiers supérieur du chromosome 10) montrent des réarrangements chez la souris (ou l'homme, selon la bande).

Les voies de développement de Genomicus sont deux nouvelles vues, actuellement disponibles, en interne, en ligne de commande :

KaryoView (exemple Figure 15.3) Le caryotype d'une espèce est dessiné, et chacun de ses gènes est colorié selon le chromosome de son orthologue dans l'autre espèce. Cette vue permet de repérer rapidement les points de réarrangements inter-chromosomiques par les changements de couleur. En revanche, au sein d'une bande de même couleur, rien ne garantit que l'ordre des gènes soit conservé, et qu'il n'y ait eu aucun réarrangement.

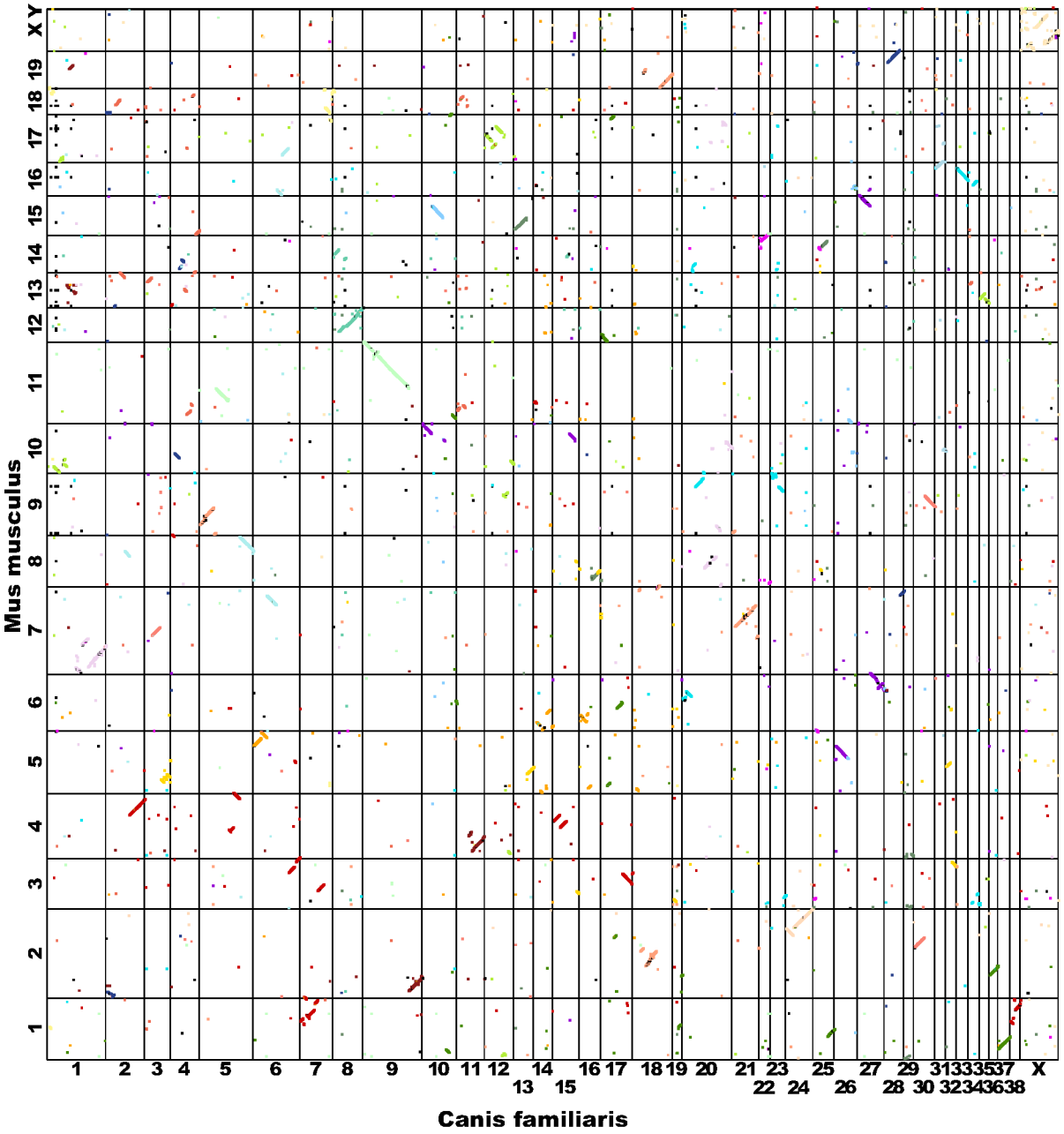


FIGURE 15.4 – Répartition des orthologues entre le génome du chien et de la souris, avec un code couleur selon les chromosomes humains. On observe que le réarrangement du chromosome 9 du chien est spécifique du chien (le point de cassure est partagé entre l'homme et la souris) et que les réarrangements du chromosome 10 (sur le côté droit) sont spécifiques de la souris (le chromosome du chien est éclaté sur plusieurs chromosomes de la souris, mais avec la même couleur de chromosome humain).

MatrixView (exemple Figure 15.4) Cette vue[3], également appelée *dot-plot*, représente en deux dimensions (1 pour chaque génome, toutes les positions génomiques mises bout à bout) les couples de gènes orthologues. Une région synténique apparaît comme un nuage de points. Si de plus, l'ordre est conservé, les points forment une diagonale

Ces deux vues seront étendues au paradigme de comparaison simultanée de multiples génomes. Par exemple en dessinant tous les caryotypes des espèces (modernes ou ancestrales) de l'arbre en fonction d'une espèce de référence, ou en coloriant un caryotype (éventuellement un chromosome) avec tous les autres génomes. Pour MatrixView, on peut comparer n génomes en organisant sous forme de table les comparaisons de chaque paire de génomes. Dans tous les cas, ces deux nouvelles vues demanderont nécessairement une refonte de l'interface utilisateur, en particulier en définissant les passages d'une vue à l'autre grâce à des scénarios d'utilisation. En particulier, Genomicus offrira trois points d'entrée (trois types d'analyses possibles) : l'étude d'un locus, d'un gène, ou la comparaison générale de deux espèces (ancestrales et / ou modernes).

Enfin, Genomicus a été développé à un moment où seul le protocole 1-passe d'AGORA existait. Il manque en particulier une signalétique pour faire ressortir les différentes étapes de la reconstruction dans le cas de l'utilisation du protocole multi-passes. Genomicus ne dispose pas non plus de vue dédiée à la duplication de génomes. Pour l'instant, ces génomes ancestraux peuvent être représentés comme des génomes ancestraux sans duplication de génome. L'utilisateur ne peut donc pas distinguer les informations de paralogie aussi facilement que dans une vue centrée sur la duplication complète du génome.

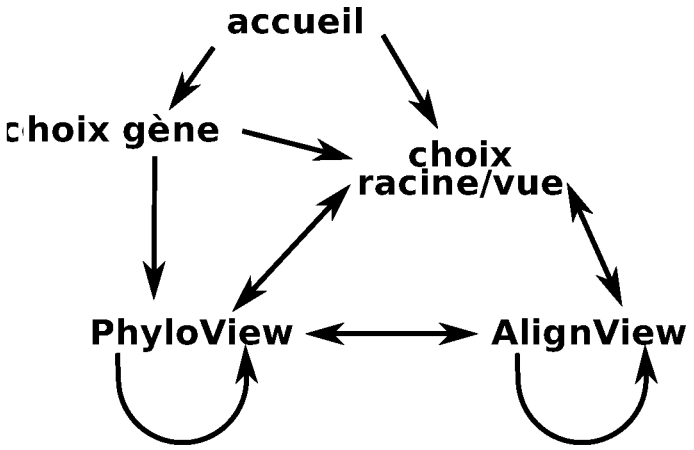


FIGURE 15.5 – Schéma simplifié de navigation dans Genomicus. Ne sont représentés que les liens entre les fonctions de recherche et les vues AlignView et PhyloView.

3. Cet outil a été utilisé pour visualiser la conservation de synténie entre *Oikopleura dioica* et les autres chordés, dans le cadre du papier relatant l'analyse de son génome [Denoeud *et al.*, 2010]. Les résultats ne sont pas décrits dans ce manuscrit car ils relèvent de la comparaison classique, *pairwise*, de génomes, sans faire intervenir de reconstructions ancestrales.

Cinquième partie

Discussion & Perspectives

Chapitre 16

Discussion

Sommaire

La méthode AGORA est désormais établie et validée, les données générées sont disponibles à la communauté sur le serveur Genomicus, et les efforts sont désormais concentrés sur l'analyse des génomes ancestraux (chapitre suivant) pour permettre une compréhension globale de l'évolution des gènes et des génomes des vertébrés. Cependant, certaines des étapes d'AGORA ne sont pas encore optimales et pourraient être améliorées (précision, efficacité) ou étendues pour gérer davantage de scénarios évolutifs. Dans ce chapitre, nous reprenons certaines étapes clés du processus de reconstruction et énumérons les limites et améliorations possibles.

16.1 Avantages, limites, et risques de l'approche locale

Nous avons privilégié, dans les algorithmes de reconstruction AGORA, les méthodes fondées sur des parcours de graphes (dont les arêtes représentent des adjacences de gènes ou de contigs). La parcours de graphe implique qu'une adjacence ancestrale ne peut être prédite que si elle est vue dans des comparaisons d'espèces modernes, utiles pour l'ancêtre en question.

L'avantage de cette approche est la qualité des reconstructions qui en découlent. En effet, une adjacence n'est sélectionnée chez un ancêtre que si elle est vue dans au moins une comparaison d'espèces, c'est-à-dire dans deux génomes modernes. Dans la plupart des cas, les algorithmes qui emploient cette approche prennent en considération le nombre de comparaisons qui soutiennent chaque adjacence. Une adjacence ne peut être choisie que si aucune autre adjacence, vue dans plus de comparaisons, n'était possible. Une reconstruction est donc composée des meilleures adjacences possibles, et ceci se retrouve dans les performances d'AGORA (plus de 98% de spécificité, sans tenir compte des singletons). Cependant, ce processus appelle deux commentaires.

Le premier concerne la capacité d'AGORA à détecter et prédire les adjacences (sensibilité). En effet, une adjacence qui n'est jamais vue dans des génomes modernes ne peut pas (dans le cas du protocole 1-passe) être prédite chez des ancêtres. Ainsi, la moindre rupture

de conservation de l'ordre des gènes entraîne des coupures dans les contigs reconstruits. C'est par exemple le constat qui a été fait lors de l'analyse des résultats du protocole 1-passe (les contigs butaient sur des gènes non-robustes, sous-section 13.2.2) et qui a mené à l'utilisation du protocole multi-passes. Grâce à l'utilisation d'un jeu de gènes robustes et du protocole multi-passes, les reconstructions peuvent inclure des adjacences ancestrales qui ne sont jamais vues dans les génomes modernes (Figure 10.2), ce qui améliore nettement les performances d'AGORA (Tableau 13.12). La sélection de sous-ensembles de marqueurs «robustes», alliée à l'utilisation du protocole multi-passes permet d'adapter AGORA aux biais présents dans les données que le protocole 1-passe n'est pas capable de dépasser. Il s'agit du premier axe de récursivité (Figure 16.1) : AGORA permet de séparer les marqueurs en niveaux de robustesse et de reconstruire le génome ancestral en traitant récursivement des groupes de marqueurs.

Une fois les contigs obtenus à partir des adjacences de gènes conservés, AGORA offre un niveau supplémentaire de reconstruction (les scaffolds) en utilisant les adjacences de contigs. AGORA dispose donc d'un deuxième axe de récursivité (Figure 16.1), en permettant de faire une reconstruction sur un ensemble de marqueurs correspondant au résultat d'une autre reconstruction (en laissant toujours le choix des protocoles). En remarquant que les scaffolds sont des contigs de contigs, on peut donc imaginer reconstruire des contigs de scaffolds, et ainsi de suite. Les variations d'AGORA sont donc immenses et permettent de gérer de nombreuses situations, mais toujours en utilisant de la conservation d'adjacences.

Avec la reconstruction en scaffolds, les performances dans les reconstructions chez les vertébrés se sont améliorées. Cependant, la reconstruction d'ancêtres éloignés (*Amniota*, par exemple) n'atteint pas encore le niveau d'un caryotype complet. En effet, passé une certaine distance évolutive, l'adjacence devient un critère trop stringent pour comparer les génomes. Il serait intéressant d'inclure la notion intermédiaire d'intervalle commun [Chauve et Tannier, 2008], qui permet d'identifier des ensembles de marqueurs groupés dans plusieurs génomes, mais sans conservation d'ordre, et qui est une forme d'adjacence encore plus relâchée. Le résultat sur les génomes ancestraux aurait un degré de précision intermédiaire : certaines régions seraient définies par un contenu non-ordonné (que l'on pourrait éventuellement résoudre avec le voyageur de commerce) mais cela permettrait d'avoir une meilleure continuité dans les génomes ancestraux reconstruits.

L'autre point est la prédiction d'intervalles incorrects (spécificité) et le risque de fusion de chromosomes qui en découle (particulièrement visibles quand les reconstructions atteignent la taille de chromosomes, comme au niveau de reconstruction «scaffolds»). Bien que nous n'en ayons pas observé dans les reconstructions réelles de *Boreoeutheria*, de telles erreurs peuvent se produire d'après les simulations. Par exemple, dans le cas du protocole 1-passe, toujours pour *Boreoeutheria*, sur les 27374 adjacences conservées de gènes issues des comparaisons d'espèces, seules 17797 (65,0%) se retrouvent dans la reconstruction finale. Cela signifie que les 9577 autres adjacences sont autant de pièges dressés devant AGORA, éventuellement capables de créer des fausses fusions de chromosomes dans les reconstructions. La situation a, là encore, été améliorée avec le protocole multi-passes. Pour la première étape (*de-novo*), sur les 6226 adjacences possibles, 5703 (91,6%) sont effectivement choisies par AGORA. Cela confirme l'intérêt de faire une reconstruction sur des gènes dits robustes (c'est-à-dire bien annotés, et non sujets à la délétion ou la duplication) : le risque de fausses fusions a significativement diminué, et reste relativement maitrisé dans les reconstructions d'AGORA. Toutefois, par précaution, il est envisageable de rajouter une étape de curation manuelle (vérification des réarrange-

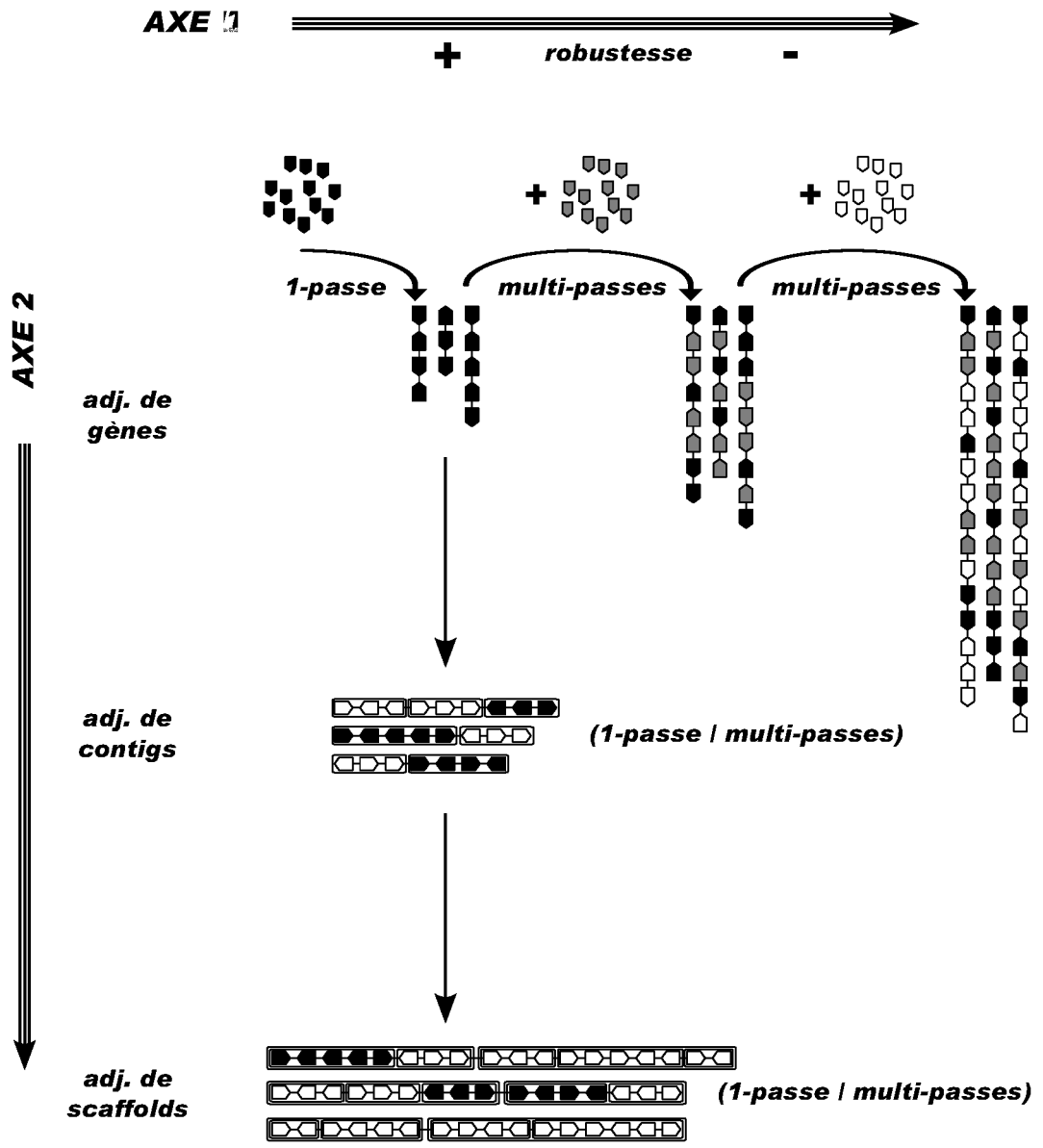


FIGURE 16.1 – Axes de récursivité d'AGORA. AGORA sait traiter en entrée la partition de l'ensemble des marqueurs selon des degrés de robustesse. AGORA échafaude une première reconstruction avec les marqueurs les plus robustes (protocole 1-passe, étape nommée *de-novo*), puis inclut successivement les autres ensembles de marqueurs par degré décroissant de robustesse (protocoles multi-passes, en utilisant la reconstruction précédente à la place de *de-novo*). D'autre part, AGORA permet d'utiliser à la place des gènes les contigs eux-mêmes comme marqueurs, et ainsi de reconstruire des contigs de contigs, en extrayant les adjacences des contigs, puis éventuellement des contigs de contigs de contigs, et ainsi de suite. Là encore, il est possible d'utiliser les protocoles 1-passe ou multi-passes, en définissant des niveaux de robustesse parmi les contigs. Le choix des protocoles et des catégories de robustesse est fortement dépendant des données et doit systématiquement être adapté en fonction.

ments plus grand qu'une certaine taille, à fixer en nombre de gènes) des reconstructions pour assurer une bonne qualité des génomes ancestraux, avant de les rendre disponibles sur Genomicus.

Chauve *et al.* [2010] souligne l'importance du jeu de données pour les reconstructions de génomes ancestraux. En particulier, les auteurs proposent de mesurer la stabilité des reconstructions en faisant varier les ensembles de marqueurs, et d'associer aux adjacences ancestrales une mesure de conservation. Cette mesure est déjà disponible dans AGORA : chaque adjacence ancestrale correspond à une arête d'un graphe, et dispose donc d'un poids, entier, indiquant le nombre de comparaisons d'espèces qui la supporte. Nous avons pour projet de mettre à disposition ces données sur Genomicus, pour pointer les adjacences «faibles». Quant à la stabilité des reconstructions par rapport au jeu de données, on a pu estimer dans la sous-section 13.2.2 les différences entre le protocole 1-passe et le protocole multi-passes avec des gènes «robustes» : on peut conclure que la fragmentation des chromosomes ancestraux en contigs n'est absolument pas stable (pour *Boreoeutheria*, l'écart entre les nombres de contigs est un facteur proche de 3), mais les simulations (tableau 13.7) montraient que la spécificité était équivalente dans les deux protocoles. Autrement dit, le contenu et l'ordre des gènes dans les contigs étaient stables.

16.2 Avantages, limites, et risques de l'approche globale

Nous avons privilégié dans AGORA des méthodes basées sur des parcours de graphe, avec en particulier des méthodes de résolution qui choisissent séparément les «bonnes» adjacences, sans utiliser d'optimisation globale. Les méthodes d'optimisation (typiquement des recherches de flots de coûts minimaux) ont été volontairement mises à l'écart, en raison des déboires de *MGR*. Toutefois, il n'y a aucune preuve que ces algorithmes, intégrés à nos données de gènes ancestraux et à nos protocoles de reconstruction, donnent de mauvais résultats, il pourrait être intéressant de les tester pour s'en assurer.

En dehors de la reconstruction d'adjacences ancestrales, deux autres techniques sont disponibles dans AGORA : le clustering (avec *walktrap*) et le voyageur de commerce (avec *concorde*). Ces deux méthodes fournissent un résultat global (des groupes de contigs ou un ordre de contigs) sans que l'on puisse justifier chaque décision (appartenance de deux contigs à un même groupe, adjacence de deux contigs) car les décisions ne sont pas prises à un niveau local.

Plus précisément, la méthode de clustering *walktrap* fonctionne sur le principe de marches aléatoires dans un graphe de similarité pour identifier des sous-graphes denses, fortement connectés. La décision de mettre deux contigs dans le même groupe est donc prise en étudiant leur voisinage dans le graphe, et non en comparant directement les deux contigs. Nous avons observé une limite (empirique) sur le graphe que *walktrap* traite : les objets à clusteriser doivent former de tels sous-graphes de quelques dizaines de sommets au minimum. Par exemple, lorsque *walktrap* est utilisé pour définir les chromosomes ancestraux (lorsqu'on a épuisé toutes les informations d'adjacence), le résultat est de bonne qualité lorsque les contigs à clusteriser sont assez nombreux (plusieurs centaines, ce qui est équivalent à avoir une dizaine de gènes en moyenne par contig), ce qui est le cas pour *Amniota* ou *Euteleostomi*, mais pas pour des ancêtres récents.

Les limites de *concorde* sont, elles, davantage d'ordre technique (distances représentées par des valeurs entières, contrainte sur les distances pour utiliser des sommets orien-

tés), sachant que la résolution d'un problème de voyageur de commerce devient, dans tous les cas, difficile lorsqu'on dépasse le millier de nœuds. L'utilisateur peut guider le résultat en définissant une distance particulière qui favorise la juxtaposition de deux sommets donnés, mais ne peut avoir la garantie que l'ordre final les fera apparaître en voisins car le résultat de *concorde* reste une optimisation globale. Il pourrait être intéressant de mettre en place une méthode hybride, qui implique des décisions prises sur des raisonnements de parcimonie (grâce à des comparaisons avec des espèces clés), et qui se rabat sur une optimisation via un voyageur de commerce sur le reste.

Pour finir, revenons sur une optimisation possible de la méthodologie *walktrap* pour clusteriser des contigs en chromosomes ancestraux. Pour l'instant, la définition de la mesure de similarité se fait de la manière suivante : pour chaque paire de contigs, vérifier s'ils sont vus sur le même chromosome dans chaque espèce moderne, définir ainsi des 0 et 1 qui seront interpolés par l'Équation 6.1 pour définir une probabilité de synténie ancestrale. Ce modèle ne fait pas intervenir le risque d'observer par coïncidence deux contigs sur le même chromosome. Ainsi, pour une espèce de n chromosomes contenant autant de marqueurs sur chaque chromosome, deux marqueurs pris au hasard ont approximativement 1 chance sur n d'appartenir au même chromosome. Ce risque est d'autant plus élevé sur une espèce qui contient peu de chromosomes, or le nombre de chromosomes peut varier d'un facteur 4 d'une espèce à l'autre (le génome de l'opossum possède 9 paires chromosomes et celui du chien 39). De plus, au fur et à mesure des réarrangements, un génome se mélange, jusqu'à apparaître comme un arrangement aléatoire des gènes, par rapport au génome de départ. Pour la reconstruction d'un ancêtre donné, la confiance qu'on peut accorder en la synténie qu'on y observe décroit donc avec la distance évolutive qui les sépare. Tous ces éléments ne sont pas encore pris en compte, et il serait donc sage d'utiliser un modèle plus fin de prédiction de synténie ancestrale, basé sur le nombre et la taille des chromosomes de chaque espèce, et les longueurs des branches.

16.3 Arbres phylogénétiques des gènes

Comme vu dans la sous-section 13.2.1, les arbres phylogénétiques sont peu fiables quant à l'inférence des nœuds de duplication. Nous avons développé une solution ad-hoc pour gérer le problème en supprimant les nœuds douteux et éditant les arbres phylogénétiques correspondants pour qu'ils restent réconciliés avec la phylogénie des espèces. La solution ultime est évidemment de disposer d'une méthode de reconstruction d'arbre phylogénétique infaillible. C'est une tâche ardue, et il est dans tous les cas opportun d'inclure dans AGORA des filtres pour vérifier la cohérence des données (la valeur optimale du seuil est d'ailleurs recalculée à chaque version d'Ensembl). La méthode d'édition actuelle pourrait être améliorée selon les points suivants :

Gestion des espèces séquencées à faible couverture. Tout d'abord, le score de confiance utilisé est celui fourni par Ensembl Compara. Pour rappel, il est égal au rapport entre le nombre d'espèces présentes dans les deux sous-arbres qui suivent la duplication, par rapport aux nombre d'espèces présentes dans au moins un d'entre eux. Partant du constat qu'un génome séquencé à faible couverture ne contient qu'environ 2/3 du génome total, on s'attend à ce que les gènes des espèces à faible couverture soient souvent absents des arbres. Il serait donc naturel de mesurer le score de confiance uniquement sur les espèces séquencées entièrement.

Degré de modification de l'arbre phylogénétique. Sur un autre plan, la modification d'un arbre se fait en déplaçant les nœuds de duplication peu supportés vers les feuilles de l'arbre (dans le but de les transformer en duplications terminales indépendantes). Une solution intermédiaire (et moins invasive) aurait été de pousser les nœuds de duplication ancêtre par ancêtre vers les feuilles, jusqu'au nœud qui permette d'avoir un score de confiance suffisant.

Seuils dépendants selon les ancêtres. Tous les nœuds de duplication de tous les ancêtres sont actuellement confrontés au même seuil, calculé sur la moyenne des N50 des reconstructions de tous les ancêtres. Cela cache le fait que la valeur optimale globale ne l'est pas nécessairement pour chaque ancêtre. Il est donc possible d'adapter le seuil à chaque ancêtre afin d'optimiser plus finement les arbres.

Utilisation de la synténie. La solution la plus reconnue pour avoir des arbres phylogénétiques de bonne qualité (en dehors d'augmenter le nombre de génomes séquencés) est l'inclusion de mesures de synténie de gènes. Le principe sous-jacent est que des gènes orthologues ont, en général, des gènes voisins orthologues entre eux et que les gènes paralogues sont, en général, à des positions différentes dans les génomes. Une telle mesure a été utilisée dans Ensembl (jusqu'à la version 41) avant qu'ils ne basculent sur une approche via les arbres phylogénétiques. La synténie est régulièrement employée pour mesurer la justesse d'un jeu de gènes orthologues [Vilella *et al.*, 2009] et des méthodes s'en servent pour définir des jeux de paires (ou clusters) de gènes orthologues [Burgetz *et al.*, 2006, Swidan *et al.*, 2006], mais seul *SYNERGY* [Wapinski *et al.*, 2007] est capable d'utiliser la synténie pour guider la reconstruction d'arbres phylogénétiques.

Dans notre cas (édition des arbres Ensembl), il faut définir pour chaque gène une mesure de synténie de ses voisins et utiliser cette mesure pour les nœuds de duplication à la place du score de confiance. La difficulté réside dans la circularité de la définition, car l'orthologie des gènes voisins est elle-même soumise à l'édition de nœuds de duplication.

Nous avons rapidement testé chacun de ces points séparément. Les résultats sont encourageants car les nouveaux arbres édités permettent des reconstructions de qualité presque équivalente à la méthode d'édition actuelle. Cependant, nous n'avons pas testé la combinaison de ces différentes approches. Il serait sûrement possible de définir un test bien plus discriminant de détection des nœuds de duplication peu supportés. On rajoutera que la valeur de bootstrap calculée dans les arbres d'Ensembl serait un autre facteur à inclure dans le test.

16.4 Cohérence des ancêtres entre eux

AGORA est organisé de telle manière que les génomes ancestraux sont reconstruits indépendamment (du point de vue de la reconstruction en elle-même) bien qu'utilisant exactement les mêmes données de départ (position des gènes et arbres phylogénétiques). Cela provient du fait que les paires de gènes conservées sont identifiées pour chaque comparaison d'espèces modernes, puis proposées à chaque ancêtre qui (indépendamment) les valide en fonction de son propre contenu en gènes (Figure 8.1). Chaque ancêtre possède donc son propre jeu de gènes ancestraux, sa propre liste de paires conservées entre espèces modernes, et suit son propre processus de sélection (paramètres de sélection des

familles robustes différents). L'avantage de cette indépendance est que les erreurs présentes à un ancêtre ne sont pas systématiquement propagées aux ancêtres l'entourant. La conséquence malheureuse de cette approche est que des divergences peuvent exister entre les reconstructions. Par exemple, il est possible que trois gènes a, b et c soient dans l'ordre abc aux ancêtres $n-1$ et $n+1$, mais dans l'ordre acb à l'ancêtre n, ce qui est théoriquement possible, mais peu parcimonieux. Deux solutions sont envisageables.

La première serait de comparer les ancêtres entre eux pour identifier de telles inconsistances, et de les résoudre (selon un raisonnement par parcimonie). Les corrections apportées aux ancêtres permettraient de définir un jeu de génomes ancestraux cohérents entre eux.

La seconde serait de reconstruire simultanément les ancêtres, en imposant toute décision prise à un ancêtre à tous les autres. Cela impliquerait soit de traiter simultanément les parcours de graphes, soit de construire un unique multi-graphe coloré qui contiendrait toute l'information de chaque ancêtre et les liens entre les ancêtres. Dans les deux cas, la difficulté tient au contenu différent en gènes (pertes, gains, et duplications), ce qui empêche de propager directement les intervalles d'un ancêtre à l'autre. En effet, une paire de gènes d'un ancêtre peut être séparée dans un autre ancêtre par un gène qui est apparu depuis. Il serait donc nécessaire de traiter une notion d'ordre relâché qui transcrirait le fait que deux gènes peuvent être voisins, à des insertions de gènes près, ce qui dépassait les notions d'adjacences (strictes) utilisées dans nos méthodes [1].

16.5 Duplications complètes de génomes

Les duplications complètes de génomes peuvent être traitées dans AGORA grâce à des méthodes spéciales. À ce stade, que ce soit pour les reconstructions pré- ou post-duplication, les chromosomes reconstruits ne sont que des ensembles non-ordonnés de blocs non-ordonnés. Autrement dit, les chromosomes sont uniquement des ensembles de gènes. Pour ordonner chaque chromosome, deux méthodes s'offrent à nous. La première solution est d'appliquer la méthode de la section 10.4 (voyageur de commerce) pour ordonner ces gènes, Cette méthode fonctionne en général difficilement car un chromosome peut contenir plusieurs milliers de gènes, ce qui est proche de la limite de résolution en temps raisonnable. L'autre solution est de réutiliser des reconstructions d'AGORA basées sur l'adjacence des gènes.

Dans cette optique, il faut prendre les contigs ou scaffolds reconstruits pour l'ancêtre le plus proche de la duplication, les assigner «en entier» au cluster pré- (ou post-) duplication avec lequel il partage le plus de gènes (tous les gènes d'un même contig doivent être assignés au même cluster). Les chromosomes sont alors comme des ensembles de contigs, et non plus de gènes. On dispose donc exactement du même type de données qu'après utilisation de *walktrap* (reconstruction de chromosomes ancestraux, section 10.3) et on peut là encore revenir à une application du voyageur de commerce (qui est la seule technique disponible ici lorsque l'on ne dispose pas d'information d'adjacence). Comme le nombre d'objets à ordonner a diminué (on est passé de quelques milliers de gènes à quelques dizaines/centaines de contigs), le voyageur de commerce a toutes les chances de trouver une solution. On aura donc un ordre ancestral reconstruit avant et après la duplication, ordre ancestral qui utilise au maximum les données d'adjacence (AGORA) avant de basculer sur de l'optimisation (voyageur de commerce).

1. Dans le paradigme des DCJ (*Double Cut and join*), Yancopoulos et Friedberg [2009] introduisent la notion de *ghost adjacency* pour tenir compte de ce phénomène

Cependant, cette approche utilise une grave approximation : le génome ancestral reconstruit grâce à une duplication complète se trouve juste avant ou juste après cette duplication, mais dans tous les cas strictement entre deux nœuds de spéciations de la phylogénie des espèces (les nœuds auxquels une reconstruction AGORA d'ordre de gènes est disponible). Les listes de gènes et les reconstructions de contigs ou scaffolds sont, elles, définies uniquement aux nœuds ancestraux de spéciation. Des réarrangements ont donc pu se produire entre ces différents points temporels. L'approximation est d'autant plus valable que les points sont rapprochés, mais reste une approximation.

Il est néanmoins possible de disposer de l'ordre des gènes dans un génome pré-duplication, en particulier si l'ordre des gènes est bien conservé entre les espèces dupliquées, comme c'est le cas chez les levures [Gordon *et al.*, 2009]. De plus, de la même manière que les réarrangements «classiques» des génomes ont été abordés d'un point de vue combinatoire, il est possible de traiter les duplications de génomes théoriquement. Les techniques de résolutions se nomment *Guided genome halving* [Sankoff *et al.*, 2007, Gavranović et Tannier, 2010] car il s'agit de fusionner le contenu des chromosomes post-duplication. Elles ne permettent pas encore de traiter automatiquement les génomes comme AGORA (génomes fragmentés ou partiellement assemblés), d'autant plus que les difficultés d'assignation correcte des orthologues, paralogues (pollution par les duplications segmentales) compliquent la tâche.

Enfin, il est de toute importance de développer un outil de mesure de la qualité des reconstructions AGORA qui utilisent les duplications complètes. Pour l'instant, le protocole de simulation ne permet pas de modéliser de tels événements. Les reconstructions ont pu toutefois être validées par comparaison avec d'autres méthodes, mais cela ne permet pas de quantifier précisément le taux d'erreur.

Chapitre 17

Perspectives

Sommaire

Ce chapitre décrit un certain nombre d'analyses en cours de réalisation qui illustrent les applications possibles d'AGORA dans le but de répondre à des questions sur l'évolution des génomes. Les reconstructions d'AGORA incluent la liste des gènes de chaque ancêtre, ainsi que leur ordre dans des chromosomes (ou fragments de chromosomes), ce qui permet une vision similaire à celle d'un génome moderne à la fin du processus de séquençage / assemblage / annotation des gènes. Cependant, elles gagneront une valeur ajoutée quand certaines propriétés biologiques seront annotées (comme pour les génomes modernes), comme la séquence des protéines ou des annotations fonctionnelles (circuits de régulation, voies métaboliques, ontologies, domaines protéiques). D'autre part, la génomique comparative se voit offrir un nouveau champ d'investigation en la comparaison d'ancêtres entre eux (dans le sens de l'évolution), ce qui permet par exemple l'étude des réarrangements chromosomiques, ou des occurrences de sélection positive.

17.1 Ajout d'autres marqueurs dans les reconstructions (ncRNAs, CNEs)

Les reconstructions AGORA sont pour l'instant limitées aux gènes codant pour des protéines, alors que de nombreuses autres annotations fonctionnelles pourraient être intégrées dans les génomes ancestraux reconstruits (à condition qu'un signal d'orthologie soit disponible et puisse être retrouvé entre les génomes modernes). On pourrait inclure les éléments non-codants conservés (CNEs), qui sont potentiellement des sites de fixation de facteurs de transcription, les séquences répétées ancestrales, et les ARNs non-codants

(ncRNAs). La présence et la position des CNEs, en particulier, semble être une des forces majeures qui contraignent l'ordre des gènes et leur configuration (*Genomic Regulatory Blocks*, Kikuta *et al.* [2007]). Intégrer ces éléments dans les génomes ancestraux reconstruits fournirait une ressource apte à mesurer de manière quantitative ce phénomène.

Cependant, la position des CNEs dans les génomes modernes, et les distances avec leurs cibles semblent peu contraintes, ce qui rendra difficile le travail de positionnement ancestral. Une solution (en réutilisant les méthodes existantes) serait d'utiliser le protocole multi-passes en assimilant les gènes codant pour des protéines aux marqueurs robustes, et les CNEs aux marqueurs non-robustes. La reconstruction de voies métaboliques fait l'objet d'un projet européen, NeuroXSys, pour lequel une post-doctorante a été recrutée dans le laboratoire. Un des objectifs sera de faire évoluer les reconstructions AGORA, pour inclure ces annotations.

17.2 Extension à d'autres familles d'organismes

Depuis la création d'EnsemblGenomes, les possibilités de reconstruction d'AGORA se sont multipliées. En effet, les mêmes données qu'Ensembl (au même format) sont maintenant disponibles pour (les plantes et les champignons). Les données sont en cours de test et montrent des résultats encourageants. C'est pourquoi des collaborations sont en cours pour faire fructifier cette base de connaissances qui est en train d'émerger.

17.2.1 Plantes

Nous avons tout d'abord testé les reconstructions sur les plantes en utilisant les données directement fournies par EnsemblGenomes (arbre phylogénétique des espèces sur la Figure 17.1.A). Les données étant déjà disponibles dans les mêmes formats que pour les vertébrés, le processus n'a pas demandé de formatage particulier. Les paramètres d'AGORA n'ont pas encore été optimisés, et sont pour l'instant :

1. seuil d'édition des nœuds de duplication à 0,00 ;
2. reconstruction de contigs multi-passes avec les gènes robustes fournis par la méthode $size(T_{\min} = 1, T_{\max} = 1)$;
3. reconstruction de scaffolds multi-passes avec les contigs robustes $length(l_{\min} = 20)$.

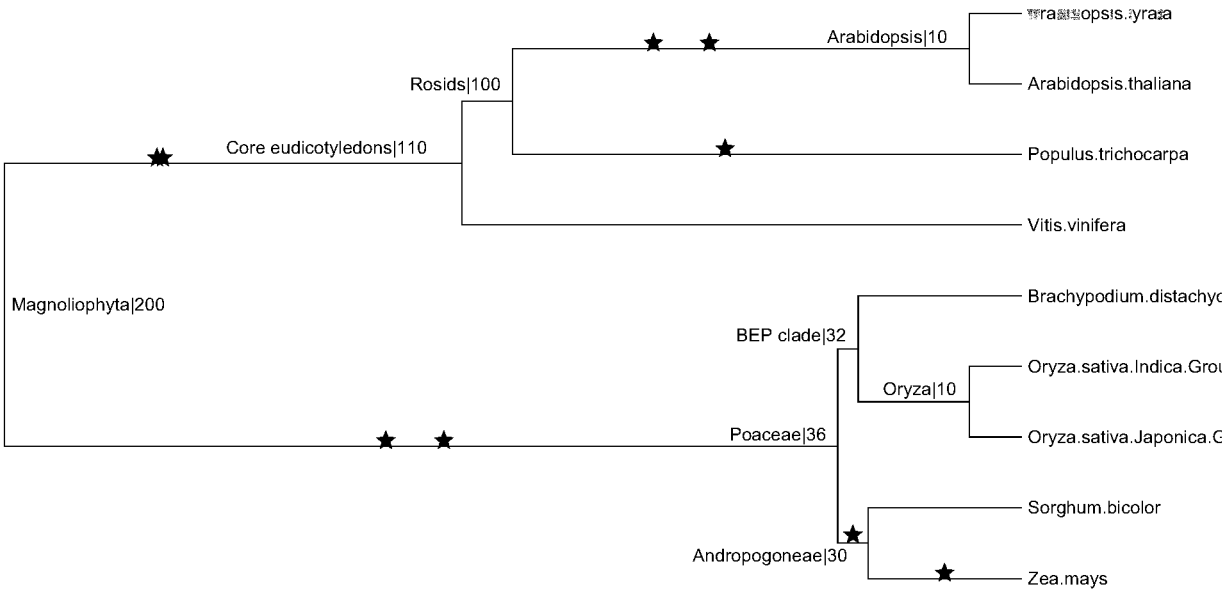
Les résultats sont représentés sur le tableau 2.1 et sont très bons pour certains ancêtres (en particulier l'ancêtre commun des deux espèces d'arabidopsis – *Arabidopsis* –, celui des deux espèces de riz – *Oryza* –, et pour *BEP clade*) avec des N50 de plus de 1000 gènes. Cela confirme le fort potentiel d'AGORA et la généricité de ses méthodes.

D'autre part, comme de nombreuses polyploïdisations ont eu lieu au cours de l'évolution des génomes des plantes [Otto et Whitton, 2000, de Peer *et al.*, 2009], il sera de toute importance de combiner les reconstructions d'ordre de gènes à celles basées sur les duplications de génomes.

17.2.2 Levures

Les reconstructions ancestrales chez les levures (arbre phylogénétique des espèces sur la Figure 17.1.B) ont demandé un travail supplémentaire de traitement des données. En effet, celles-ci étaient dispersées sur trois base de données : EnsembGenomes, le Broad Institute, et YGOB [Byrne et Wolfe, 2005]. Pour incorporer toutes les espèces de levure,

A :



B :

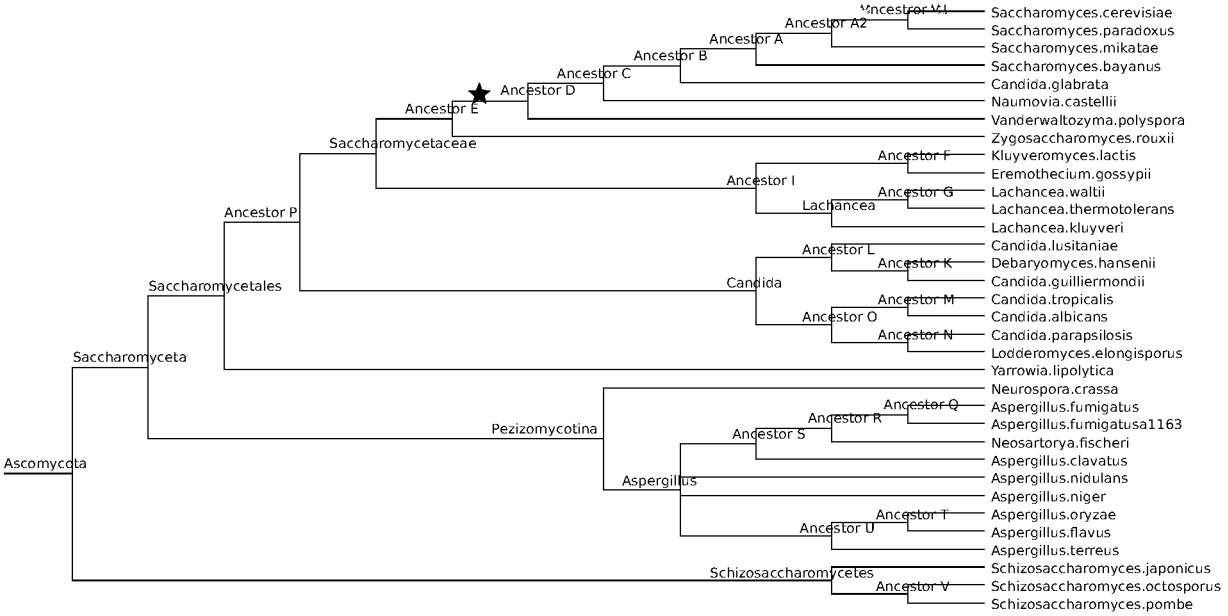


FIGURE 17.1 – Arbre phylogénétique des espèces de plantes (**A**) et de levures (**B**). Les étoiles indiquent les événements de tetraploïdisations documentés, la double étoile indique une hexaploïdisation (triplement du contenu en chromosomes).

Ancêtre	Âge (Ma)	Gènes	Contigs	Contigs (> 100 gènes)	Couverture (gènes)		Couverture (intervalles)		25%	50%	75%	N75	N50	N25	Max	Moyenne
Magnoliophyta	200	13742	1163	0	3114	22,66%	1951	14,22%	2	2	3	2	2	4	52	2,68
Core eudicotyledons	110	19449	1837	0	11241	57,80%	9404	48,40%	2	3	7	5	10	20	90	6,12
Rosids	100	19655	1923	1	12268	62,42%	10345	52,69%	2	3	7	5	11	22	123	6,38
Arabidopsis	10	25690	196	29	22296	86,79%	22100	86,09%	2	5	30	462	1579	2169	2246	113,76
Poaceae	36	26692	702	21	17637	66,08%	16935	63,49%	2	2	5	191	625	1354	1880	25,12
BEP clade	32	26618	487	19	19157	71,97%	18670	70,19%	2	2	4	634	1023	2205	2684	39,34
Oryza	10	31135	328	14	25887	83,14%	25559	82,14%	2	2	6	1106	2009	3324	3570	78,92
Andropogoneae	30	27323	984	30	19468	71,25%	18484	67,70%	2	3	11	33	112	634	1156	19,78
Ascomycota		3753	170	0	368	9,81%	198	5,30%	2	2	2	2	2	2	6	2,16
Saccharomyceta		6841	260	0	569	8,32%	309	4,53%	2	2	2	2	2	2	6	2,19
Saccharomycetales		6760	525	0	1233	18,24%	708	10,50%	2	2	3	2	2	3	7	2,35
Ancestor P		7012	733	0	1981	28,25%	1248	17,85%	2	2	3	2	3	4	10	2,70
Saccharomycetaceae		7735	90	12	4797	62,02%	4707	61,01%	2	2	38	114	340	614	658	53,30
Ancestor E		6456	74	11	4903	75,94%	4829	75,03%	2	2	36	165	467	1714	1714	66,26
Ancestor D		6520	189	17	4896	75,09%	4707	72,42%	2	3	14	56	132	238	314	25,90
Ancestor C		6547	250	17	5067	77,39%	4817	73,80%	2	5	16	22	109	164	245	20,27
Ancestor B		6213	211	11	5060	81,44%	4849	78,30%	3	8	21	28	80	170	266	23,98
Ancestor A		6083	76	19	5320	87,46%	5244	86,49%	3	23	96	103	180	378	557	70,00
Ancestor A2		6244	65	18	5407	86,60%	5342	85,83%	2	17	130	149	208	380	433	83,18
Ancestor A1		6084	52	21	5451	89,60%	5399	89,03%	2	44	174	173	240	340	706	104,83
Ancestor I		7174	45	6	4960	69,14%	4915	68,70%	2	2	39	772	1652	1658	1658	110,22
Ancestor F		5137	94	14	4794	93,32%	4700	91,85%	5	20	69	65	118	229	280	51,00
Lachancea		6998	31	7	4993	71,35%	4962	71,11%	2	2	7	674	936	1347	1347	161,06
Ancestor G		6757	24	7	4881	72,24%	4857	72,09%	2	3	308	667	726	1349	1349	203,38
Candida		6544	642	0	4861	74,28%	4219	64,67%	2	5	9	6	12	23	92	7,57
Ancestor L		6290	213	14	5163	82,08%	4950	78,95%	3	11	26	25	60	124	245	24,24
Ancestor K		6103	52	12	5227	85,65%	5175	85,07%	2	3	56	249	522	656	917	100,52
Ancestor O		5959	333	7	5217	87,55%	4884	82,24%	3	7	15	13	34	75	208	15,67
Ancestor M		5878	216	14	5376	91,46%	5160	88,08%	4	8	18	27	92	197	315	24,89
Ancestor N		5453	175	16	5068	92,94%	4893	90,06%	4	10	22	27	109	201	378	28,96
Pezizomycotina		9220	822	0	1889	20,49%	1067	11,60%	2	2	2	2	2	3	9	2,30
Aspergillus		16104	556	13	8326	51,70%	7770	48,31%	2	2	3	257	545	983	1411	14,97
Ancestor S		11228	308	12	8368	74,53%	8060	71,91%	2	2	4	317	710	1377	1565	27,17
Ancestor R		10680	129	7	9096	85,17%	8967	84,12%	2	2	4	1219	1429	2693	2693	70,51
Ancestor Q		9689	40	9	9354	96,54%	9314	96,33%	2	4	7	844	1293	1506	1581	233,85
Ancestor U		13675	475	15	8098	59,22%	7623	55,83%	2	2	3	202	380	532	919	17,05
Ancestor T		12939	90	9	11028	85,23%	10938	84,67%	2	3	5	873	1328	1880	2128	122,53
Schizosaccharomycetes		4499	525	0	2997	66,61%	2472	55,19%	2	4	8	4	8	13	40	5,71
Ancestor V		4475	94	15	3476	77,68%	3382	75,91%	2	9	42	61	126	179	265	36,98

TABLE 17.1 – Statistiques sur les longueurs et nombres de scaffolds, selon les ancêtres, pour une utilisation d'AGORA sur des données de génomes de plantes et de levures.

il a donc été nécessaire d'exécuter le pipeline d'Ensembl Compara sur l'ensemble des protéines des trois bases pour reconstruire les arbres phylogénétiques, avant la reconstruction de génomes proprement dite. Comme pour les plantes, la reconstruction AGORA n'est pas encore optimisée et paramétrée spécifiquement pour les plantes. Les longueurs du tableau 2.1 montrent que certains ancêtres sont extrêmement bien reconstruits (en particulier les ancêtres I, Q, R,et T), ce qui permet aussi d'espérer des caryotypes ancestraux complets.

Chez les levures, la reconstruction de référence [Gordon *et al.*, 2009] est celle de l'ancêtre qui précède juste la duplication complète du génome propre aux *Saccharomycetaceae* (plus précisément entre les nœuds D et E). Nous l'avons comparée à une version préliminaire du génome reconstruit par AGORA (à l'ancêtre E). La reconstruction de référence est composée de 4703 gènes répartis en 8 chromosomes. La reconstruction AGORA (contig établis par le protocole 1-passe, avec 11 espèces, en réutilisant les données d'orthologie de YGOB) est composée de 4713 gènes répartis en 103 contigs. Les différences sont (une fois tenu compte de la différence de liste de gènes) :

- 95 paires de gènes non-vues par AGORA (ce qui explique la différence de nombre de contigs (103 au lieu de 8) ;
- 3 inversions, qui peuvent s'expliquer par la différence de position dans l'arbre phylogénétique des deux reconstructions (Figure 17.2) ;
- 75 adjacences AGORA qui auraient dû inclure 1 gène (dans 72 cas) ou 2 gènes (dans 3 cas).

La comparaison montre que les reconstructions concordent (surtout en remarquant qu'il s'agissait d'une reconstruction préliminaire d'AGORA). Cette comparaison sur une version préliminaire des reconstructions AGORA sont très encourageantes

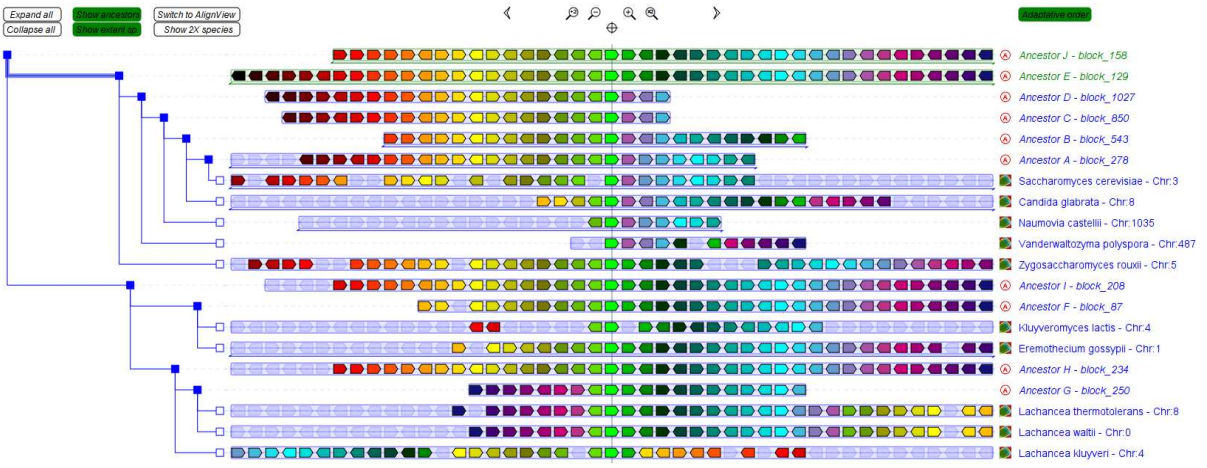


FIGURE 17.2 – Exemple de différence explicable entre AGORA et Gordon *et al.* [2009]. Dans cet exemple, deux configurations principales coexistent chez les levures étudiées, selon si la partie de droite (coloriée du vert au bleu au violet) est inversée ou non. AGORA a correctement (du point de vue de la parcimonie) reconstruit les ordres ancestraux, ce qui permet de proposer qu'une inversion a eu lieu entre les ancêtres E et D. La duplication complète (et donc l'ancêtre reconstruit par Gordon *et al.* [2009]) est située entre les ancêtres E et D, et peut contenir l'inversion si cela se justifie par l'étude du chromosome paralogue. La différence entre les deux reconstructions serait alors justifiée, et non une erreur de l'une ou l'autre.

17.2.3 Procaryotes

L'extension d'AGORA au monde des procaryotes soulève en revanche des nouveaux problèmes, en particulier à cause du nombre élevé de transferts horizontaux de gènes, chose qui n'est pour l'instant pas gérée par le pipeline de reconstructions d'arbres phylogénétiques TreeBest. Il faudrait rajouter au pipeline une nouvelle classe de nœuds (en plus des nœuds de spéciation et de duplication), pour tenir compte de ces événements en gardant une cohérence et une précision dans les arbres, ou choisir un autre pipeline [Penel *et al.*, 2009]. De plus, la circularité des chromosomes entraine une reformulation des problèmes : on ne cherche plus des chemins dans les graphes, mais un unique cycle, et le clustering en chromosomes ancestraux n'a plus de sens. Enfin, cela impliquerait une modification de l'interface utilisateur Genomicus et éventuellement de la base de données.

17.2.4 Extensions de Genomicus

L'utilisation d'AGORA sur d'autres familles d'organismes permettra de créer des serveurs Genomicus pour chacune d'entre elles (vertébrés, plantes, insectes, levures, etc). Ces différents Genomicus peuvent être interconnectés dans un réseau qui offrirait une interface unifiée de recherche, et d'étude de l'évolution des génomes.

La multiplication des ressources affichées dans Genomicus pourrait nécessiter une nouvelle organisation des données dans la base, en particulier une mise sous forme normale des tables. Le maintien de Genomicus dans le laboratoire, la mise à jour régulière des données, et l'extension aux autres phylum seront assurés par une ingénieur de recherche de l'équipe, et par des collaborations en cours.

17.3 Séquence ancestrale

Les reconstructions d'AGORA sont pour l'instant focalisées sur la présence et l'ordre des gènes chez les ancêtres, ce qui fournit une vue globale sur l'architecture des génomes. Cependant, afin de passer à un niveau supérieur d'analyse (identification d'éléments fonctionnels, de signatures de sélection – positive ou négative –), il faudrait disposer des séquences, nucléiques ou protéiques, ancestrales. Dans AGORA, compte tenu de la position centrale des gènes et des protéines, via les arbres phylogénétiques, une démarche naturelle serait d'exploiter les alignements multiples des gènes réalisés par TreeBest. Plusieurs outils permettent déjà, gène par gène, de reconstruire la séquence ancestrale, comme certains modules spécifiques de PAML [Yang, 2007].

À plus grande échelle, il faut fournir au programme d'inférence de séquence des alignements multiples de génomes entiers. Il faut d'abord noter que la donnée de l'ordre des gènes dans des contigs AGORA peut servir de guide pour aligner les séquences intergéniques des gènes, ce qui peut ostensiblement améliorer la couverture des alignements multiples. Les jalons de la reconstruction nucléotidique de l'ancêtre *Boreoeutheria* ont été posés dans l'article Blanchette *et al.* [2004a]. Les auteurs y décrivent le processus de prédiction de plus d'un mégabase d'ADN ancestral englobant le locus *CFTR*. La base de leur analyse est un alignement multiple (au niveau nucléique) entre les séquences orthologues de 19 espèces de mammifères placentaires, effectué avec TBA [Blanchette *et al.*, 2004b]. L'inférence de chaque nucléotide ancestral a été effectué par une approche de maximum de vraisemblance [Guindon et Gascuel, 2003, Hasegawa *et al.*, 1985]. Le taux d'erreurs estimé (par des simulations) décroît avec le nombres d'espèces de mammifères utilisées

dans l'alignement (6% pour 5 génomes, 1% pour 20 génomes) et les auteurs montrent que *Boreoeutheria* est l'ancêtre qui permet d'atteindre la meilleure précision. En pratique, le taux d'erreur global était de 4% (en incluant des régions répétées), et de moins de 1% pour la protéine CFTR en elle-même (sur les séquences nucléiques et protéiques). La technique a depuis été mise à jour [Paten *et al.*, 2008], pour mêler dans un même programme (*Ortheus*), la construction de l'alignement multiple, la prédiction des insertions et des délétions, et l'inférence de la séquence ancestrale. Cela confirme qu'avec suffisamment de génomes séquencés, il sera possible d'avoir la séquence ancestrale, du moins pour *Boreoeutheria*, en incluant même des éléments répétés / transposables ancestraux.

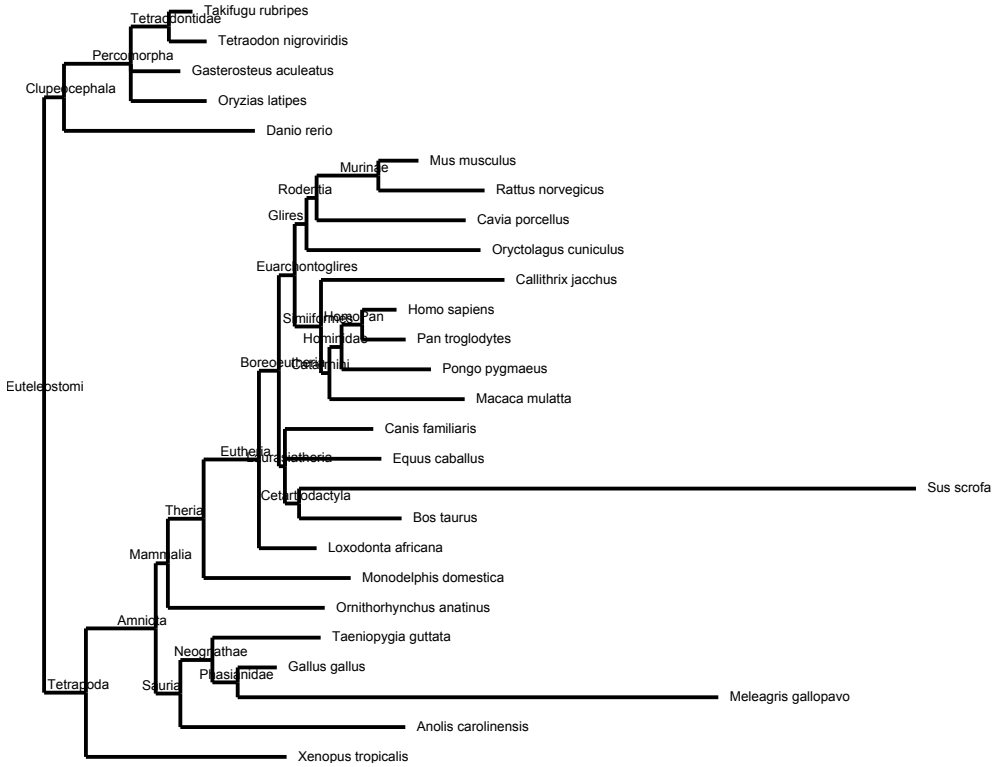
17.4 Estimation du nombre de réarrangements

Posséder les génomes des ancêtres des vertébrés permet d'établir des points de passage de l'évolution au cours du temps. La comparaison de deux génomes ancestraux successifs renvoie la liste des événements qui ont modifié le génome dans un intervalle de temps donné. Bien que l'analyse des points de réarrangements des génomes chez les vertébrés dépasse le cadre de cette thèse (il s'agit du projet de thèse d'une doctorante de l'équipe), nous pouvons cependant poser les bases d'un tel travail.

La technique la plus simple est la comparaison directe de toutes les paires d'ancêtres successifs. En particulier, l'algorithme décrit en section 8.2 permet d'aligner deux génomes, et en particulier ceux de deux ancêtres (en utilisant les paramètres $R = R_{\text{inters}}$ et $n_{\text{insert}} = 0$). Toutes les bordures de régions alignées sont des points où des réarrangements ont interrompu la colinéarité des génomes comparés. L'arbre phylogénétique de la Figure 17.3.A utilise comme longueurs de branches le nombre d'intervalles différents entre deux ancêtres. Une étude manuelle a de plus montré que tous les intervalles différents ne représentaient pas tous des «vrais» réarrangements, ce qui explique les branches particulièrement longues du cochon domestique (*Sus scrofa*) et du dindon sauvage (*Meleagris gallopavo*). Il en est sorti un ensemble de filtres (non expliqués ici) qui ont permis d'établir une nouvelle liste de réarrangements Figure 17.3.B. L'analyse de ces données est en cours.

L'étude des réarrangements identifiés sur les reconstructions d'AGORA permettra tout particulièrement d'établir un modèle de leur distribution, et de participer au débat existant dans la communauté. Il s'agit de déterminer si les réarrangements se produisent uniformément dans le génome (*Random Breakage Model*, Nadeau et Taylor [1984], Ma *et al.* [2006]), s'ils sont concentrés dans certaines régions (*Fragile Breakage Model*, Pevzner et Tesler [2003a]), si certaines régions les excluent (*Genomic Regulatory Blocks*, Kikuta *et al.* [2007]), ou s'il s'agit d'un mélange de ces modèles (Becker et Lenhard [2007]). La réponse pourrait avoir des conséquences importantes sur la validité des méthodes de reconstruction de génomes ancestraux. En effet, toutes les méthodes font plus ou moins appel à la parcimonie, et la possibilité que les génomes puissent réutiliser des points de cassure plus qu'attendu au hasard (qu'il existe des régions fragiles) mettrait à mal les raisonnements sous-jacents. Pour les reconstructions en elles-mêmes, le risque majeur est de ne pas considérer certains réarrangements (certaines configurations) sous prétexte qu'elles se produiraient sur le même locus qu'un réarrangement déjà établi.

A :



B :

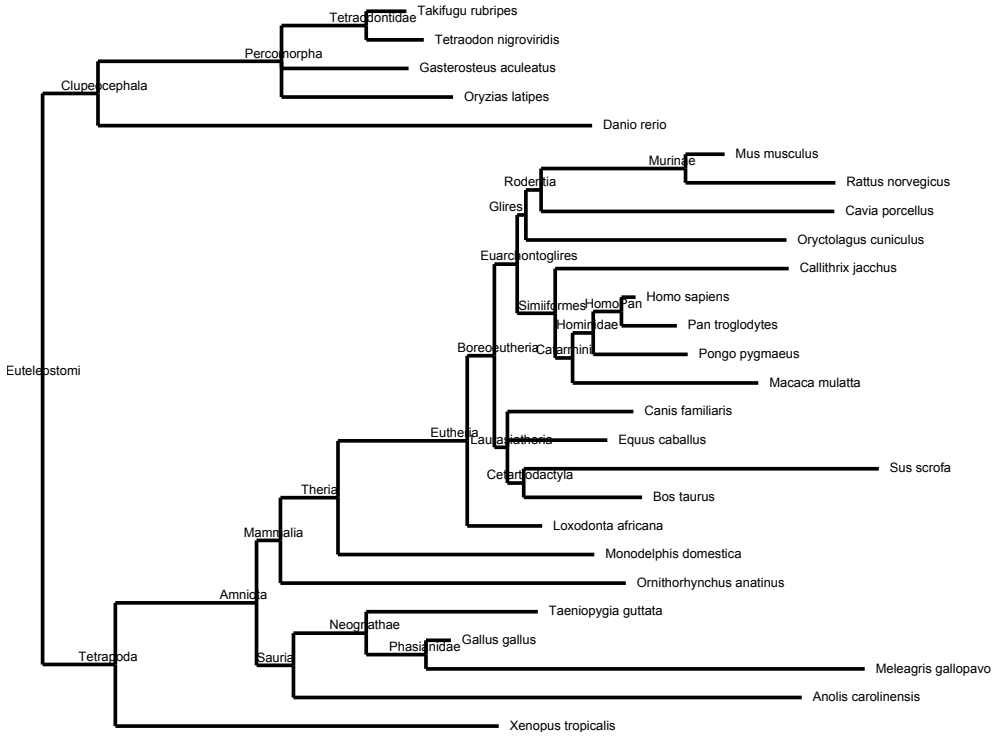


FIGURE 17.3 – Taux de réarrangements chez les vertébrés, tirés des reconstructions d'AGORA.

17.5 Fonction des gènes et sélection positive

La plupart des analyses fonctionnelles utilisent des ontologies de gènes (en particulier, *Gene Ontology*, encore appelé GO, Consortium [2008]) pour tester un enrichissement ou une déplétion d'un ensemble de gènes, selon certaines propriétés ou fonctions biologiques, En théorie, la même approche s'applique dans le cadre des génomes ancestraux, pour estimer les fonctions biologiques qui ont évolué entre génomes successifs. Dans le but d'accomplir de tels tests, il sera impératif de transférer les annotations disponibles des génomes modernes sur les génomes ancestraux, en tenant compte de leurs catalogues de gènes, et des connaissances sur la présence de certaines voies métaboliques. Le projet *GO-paint* permet de faire ceci et est disponible sur le site de la base de données *PantherDB* [Mi *et al.*, 2010].

D'autre part, la sélection positive est le moteur principal de l'émergence de nouvelles fonctions au cours de l'évolution des espèces. Les génomes ancestraux que nous avons reconstruit définissent un cadre idéal de lecture de l'évolution et des événements de sélection positive (anciens ou récents), en les replaçant dans un contexte temporel (la phylogénie des espèces). L'intérêt est encore plus fort depuis la découverte de l'existence de points chauds de sélection positive chez les vertébrés [Enard *et al.*, 2010].

Nous avons commencé un travail d'analyse des fonctions à partir des reconstructions AGORA. Le but est d'étudier l'évolution des processus biologiques et leur établissement au cours de l'évolution En première approximation, toutes les annotations GO des gènes humains sont reportées sur les gènes ancestraux dont ils sont issus. Puis, avec le génome humain en référence, nous avons sélectionné :

1. les cas de duplication en tandem conservée (un gène ancestral s'est dupliqué et les duplicats sont toujours voisins dans le génome humain), groupés selon l'ancêtre à partir duquel ils sont dupliqués ;
2. les cas d'adjacences conservées (deux gènes voisins dans un génome ancestral sont toujours voisins dans le génome humain), groupés selon le plus vieil ancêtre qui les contient.

La Figure 17.4 montre la répartition de tous les cas de duplication en tandem conservée dans le génome humain, tandis que la Figure 17.5 montre le détail par ancêtre pour les chromosomes 6 et X/Y humains. On peut ainsi repérer des cas de colocalisation de duplication en tandem conservées, au cours du temps, en comparant les ancêtres entre eux. Dans le cas de la paire de chromosomes X/Y, certains cas de colocalisation correspondent à des gènes impliqués dans les fonctions de reproduction sexuelle. Dans le cas du chromosome 6, on retrouve les gènes du CMH (complexe majeur d'histocompatibilité) et des clusters de récepteurs olfactifs, tous deux connus pour être soumis de fortes pressions de sélection.

Il est donc possible d'étudier pour chaque ancêtre l'enrichissement ou la déplétion de chaque terme GO dans les deux sous-ensembles de gènes définis. Les deux figures 17.7 et 17.6 montrent l'ensemble des termes GO associés à un enrichissement, en rouge, (ou une déplétion, en vert) significative (avec un seuil de probabilité à 10^{-4}), triés selon l'ancêtre auquel le facteur d'enrichissement est le plus élevé. Sur chaque ligne, l'intensité de rouge ou de vert est proportionnelle au facteur d'enrichissement (ou de déplétion) et normalisée par la valeur maximale.

L'analyse des termes et des gènes identifiés est en cours, en particulier pour étudier les liens entre les gènes liés au développement embryonnaire et des nœuds anciens (précédent *Euteleostomi*), ou les gènes liés au système immunitaire ou à la mise en place des

chromosomes sexuels chez les mammifères et des ancêtres plus récents. De manière générale, ces études seront à relier à celles destinées à identifier des éléments fonctionnels et des signatures de sélection (positive ou négative).

En effet, plusieurs méthodes existent pour identifier sur des branches précises de l'arbre des espèces les cas de sélection positive (par exemple en se basant sur le ratio de mutations non-synonymes vs synonymes, *dn/ds*, comme *Codeml* du package *PAML*). Ces calculs peuvent être effectués systématiquement sur les gènes utilisés par AGORA, en reprenant les alignements multiples fournis par le pipeline TreeBest (en particulier, des ratios *dn/ds* y sont déjà disponibles), et pourront, à terme, enrichir les reconstructions et Genomicus.

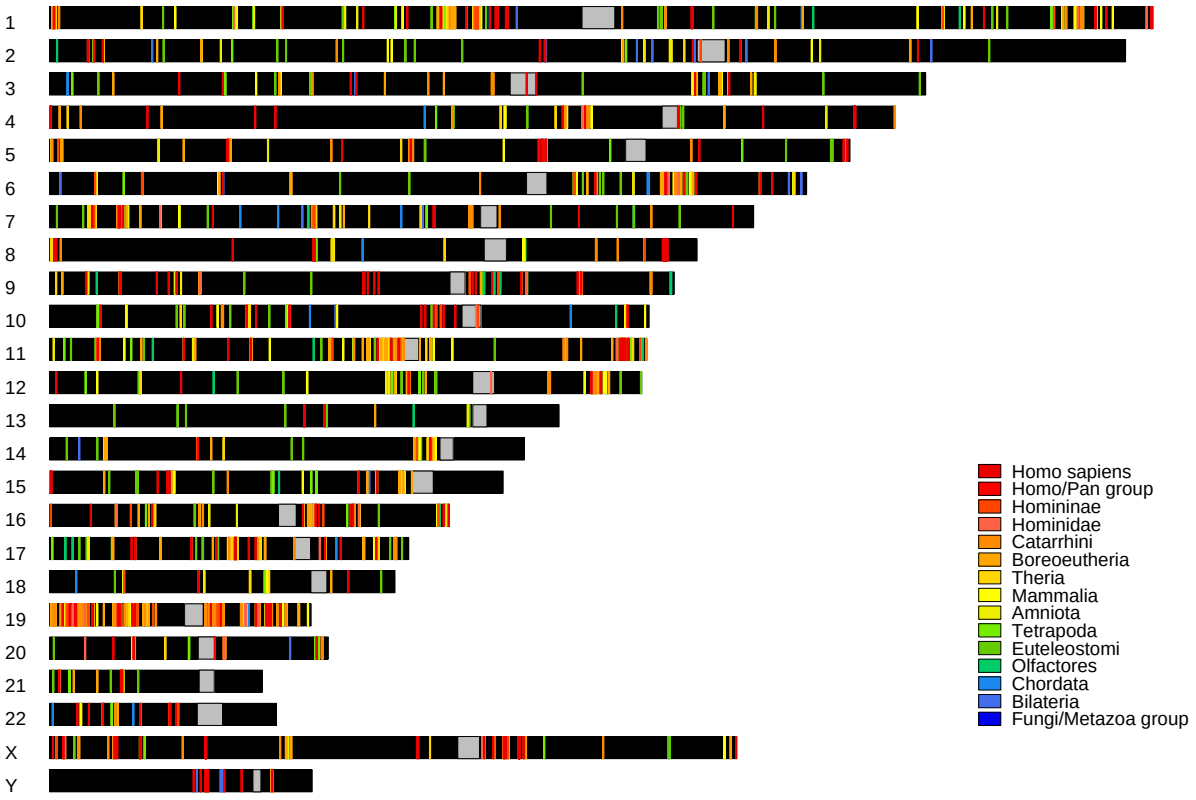


FIGURE 17.4 – Localisations des duplications en tandem dans le génome humain, avec un code couleur indiquant l'âge du plus vieil ancêtre chez qui la duplication est diagnostiquée.

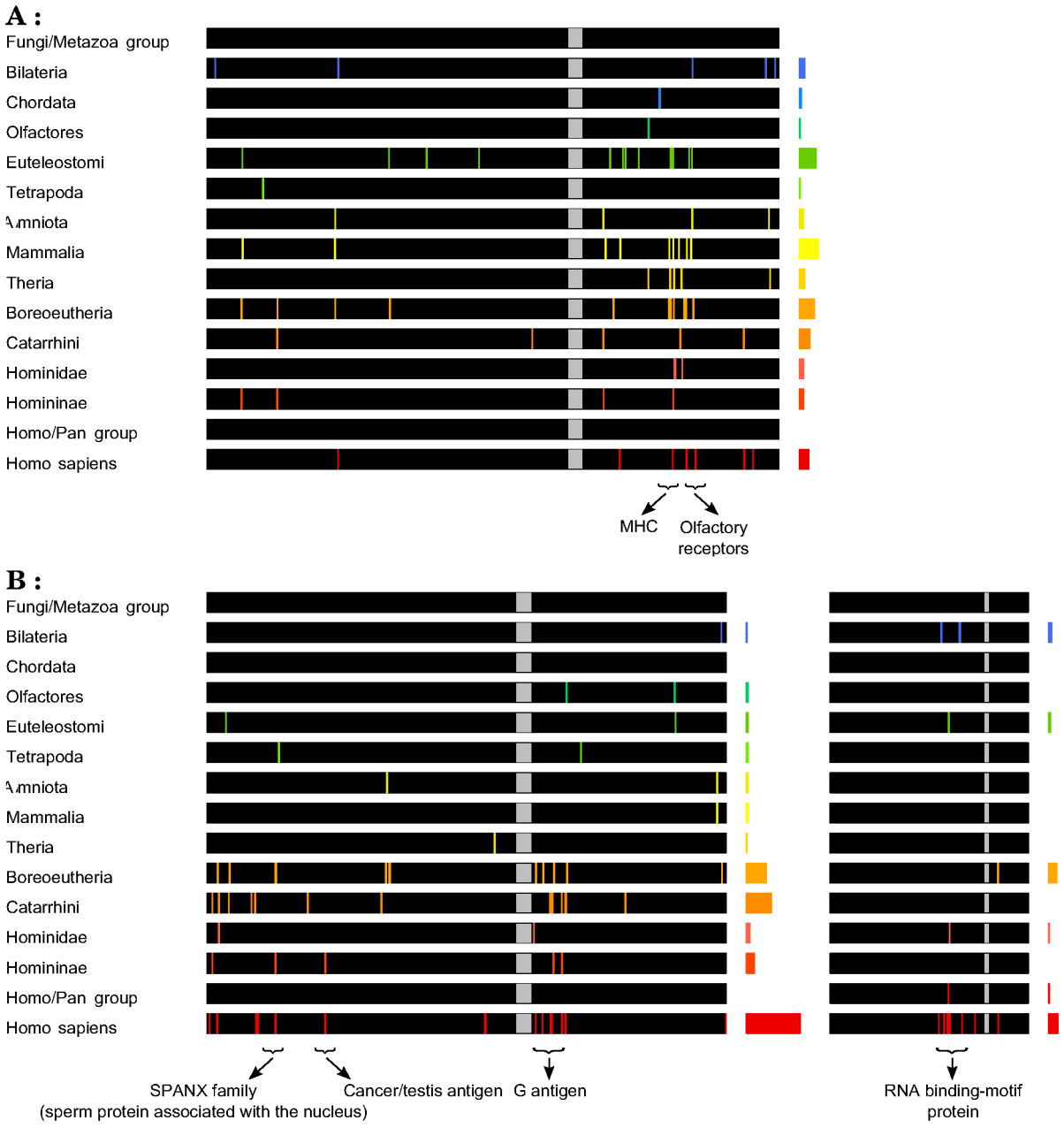


FIGURE 17.5 – Exemple de fonctions associées à des gènes dupliqués en tandem et colocalisés (chromosomes humains 6, X, et Y).

FIGURE 17.6 – Analyse *Gene Ontology* des duplications en tandem conservées.

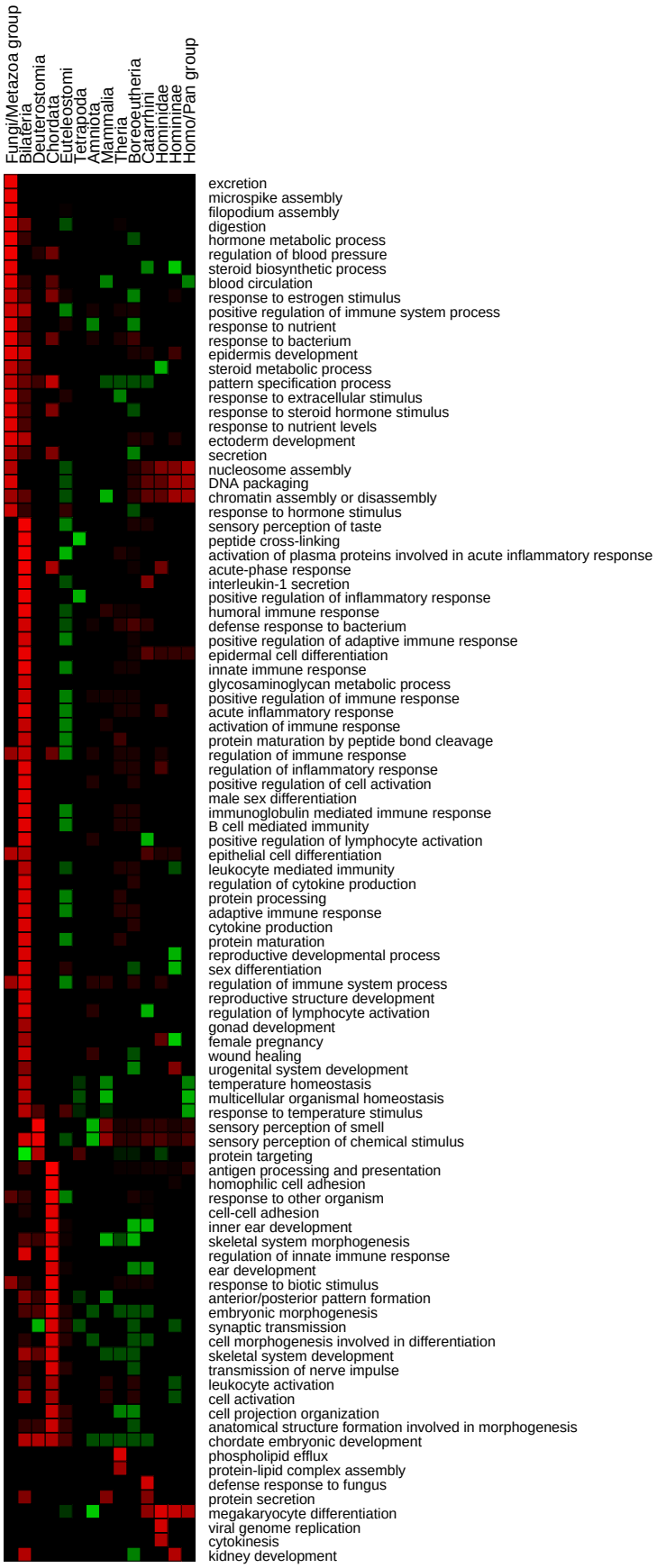


FIGURE 17.7 – Analyse *Gene Ontology* des adjacences de gènes conservées.

Sixième partie

Annexes

Annexe A

AGORA

A.1 Règles générales de cohérence des fichiers

Afin d'offrir un regard cohérent sur la structure des génomes existants et reconstruits, certaines règles doivent être établies et respectées tout au long du processus de reconstruction. Elles permettent d'alléger les programmes et méthodes de vérifications redondantes, et également de permettre la création du navigateur de génomes Genomicus, présenté au chapitre suivant.

Un arbre phylogénétique définit la liste des espèces d'intérêt et leurs relations. Toute espèce existante référencée est définie comme la liste de ses gènes (nom et coordonnées). Il ne doit y avoir aucun doublon dans les noms de gènes au sein d'une espèce, et à travers plusieurs espèces. Les arbres phylogénétiques des gènes lient un sous-ensemble des gènes des espèces séquencées, et les arbres doivent être réconciliés avec la phylogénie des espèces. Les nœuds n-aires de la phylogénie des espèces doivent également être n-aires dans les arbres des gènes. La liste des gènes de chaque ancêtre doit être cohérente avec les arbres de gènes, et toute version filtrée de ces gènes (comme pour l'extraction des familles robustes) doit en désigner un sous-ensemble. Chaque résultat d'une étape de reconstruction doit présenter la position (évidemment unique) de tous les gènes de tous les ancêtres, en incluant donc les singletons.

A.2 Gestionnaire de reconstruction AGORA

La suite AGORA vient avec un gestionnaire qui permet de lancer plusieurs reconstructions en :

- tirant partie de la disponibilité de plusieurs cœurs sur une machine ;
- assurant une cohérence dans le placement et le nommage des différents fichiers (intermédiaires et finaux) ;
- remplissant automatiquement les paramètres de chaque programme.

Le gestionnaire lit un fichier de configuration qui contient la liste des tâches à effectuer et leurs dépendances (au sein d'un protocole multi-passes, par exemple), les organise dans un graphe orienté (supposé acyclique) dont le parcours de chaque composante connexe donne la suite des programmes à exécuter. Si l'organisation des fichiers est paramétrée, le gestionnaire permet en fait de gérer les deux situations opposées :

- une étape isolée du processus de reconstruction en spécifiant uniquement les paramètres de la méthode ayant des valeurs différentes de celles par défaut (le gestionnaire se chargeant de remplir les autres) ;

– une série de reconstructions avec toutes les combinaisons de valeurs de plusieurs paramètres.

Le gestionnaire a en particulier été utilisé pour tester sur des génomes aléatoires de nombreuses combinaisons de paramètres possibles (175 combinaisons sur 110 jeux de génomes aléatoires) et trouver la combinaison optimale. Il est également utilisé pour tester les paramètres sur de nouvelles familles d'organismes (plantes et levures).

Ci-dessous est inséré le fichier de configuration du gestionnaire AGORA qui permet les reconstructions de contigs multi-passes, en sélectionnant des jeux de gènes robustes spécifiques pour chaque ancêtre. Le format permet de facilement définir de nouveaux niveaux de robustesse, ainsi que l'enchaînement des algorithmes pour la reconstruction, et la sélection de la version finale des contigs.

```
#######################################
# General configuration file for AGORA #
#######################################


# Section 1: folder structure
##############################
> Files

speciesTree = /users/ldog/muffato/workspace/data57/phylTree.full.conf
genes = /users/ldog/muffato/workspace/data57/genes/genesST.%(name)s.list.bz2

ancGenesData = ancGenes/%(filt)s/ancGenes.%(name)s.list.bz2
ancGenesOutput = ancGenes/%(filt)s.txt.bz2
ancGenesLog = ancGenes/%(filt)s.log

pairwiseOutput = diags/pairwise/pairs-%(filt)s.list.bz2
pairwiseLog = diags/pairwise/pairs-%(filt)s.log

integrBlocks = diags/integr/%(method)s/anc/contigs.%(name)s.list.bz2
integrOutput = diags/integr/%(method)s/graph.txt.bz2
integrLog = diags/integr/%(method)s/log


# Section 2: ancestral gene sets
#################################
> AncGenes

1 = size 1 1
2 = size 0.9 1.1
3 = size 0.75 1.33


# Section 3: pairwise comparisons
#################################
> Pairwise

all = conservedPairs
1 = conservedPairs
```

```
2 = conservedPairs
3 = conservedPairs


# Section 4: integration
########################
> Integration

denovo (all)

denovo (1) ! Theria,Clupeocephala
[/denovo-size-custom/] copy ! Theria,Clupeocephala

denovo (2) ! =Amniota,Sauria
[/denovo-size-custom/] copy ! =Amniota,Sauria

denovo (3) ! =Mammalia,=Tetrapoda,=Euteleostomi,/Euteleostomi
[/denovo-size-custom/] copy ! =Mammalia,=Tetrapoda,=Euteleostomi,/Euteleostomi

refine -func=0,32|100,40t|10000 -timeout=120 (all) ! Fungi/Metazoa^group
extend +onlySingletons (all)
halfinsert (all)

[/final] copy
```

A.3 Composantes d'AGORA

Module	Nb. lignes		Description
ancsequence	210	169	Application de l'algorithme d'interpolation à la reconstruction de séquence
breakpoints	823	593	Extraction des points de réarrangements
concorde	508	358	Voyageur de commerce
data	2247	1518	Formatage des données
dcs	1404	887	Duplication de génomes (synténie dédoublée)
simulations	1524	990	Simulations
synteny	2996	2126	Reconstructions AGORA (ordre de gènes)
walktrap	391	236	Clustering des blocs en chromosomes
misc	609	447	Outils pour la comparaison de génomes
total	10712	7324	

TABLE A.1 – Taille (nombre de lignes avec et sans commentaires) des différents modules d'AGORA.

A.4 Utilisation d'AGORA

La figure suivante (A.1 montrent la procédure suivie à chaque mise à jour de la base de données Ensembl, pour mettre à jour les données de notre côté.

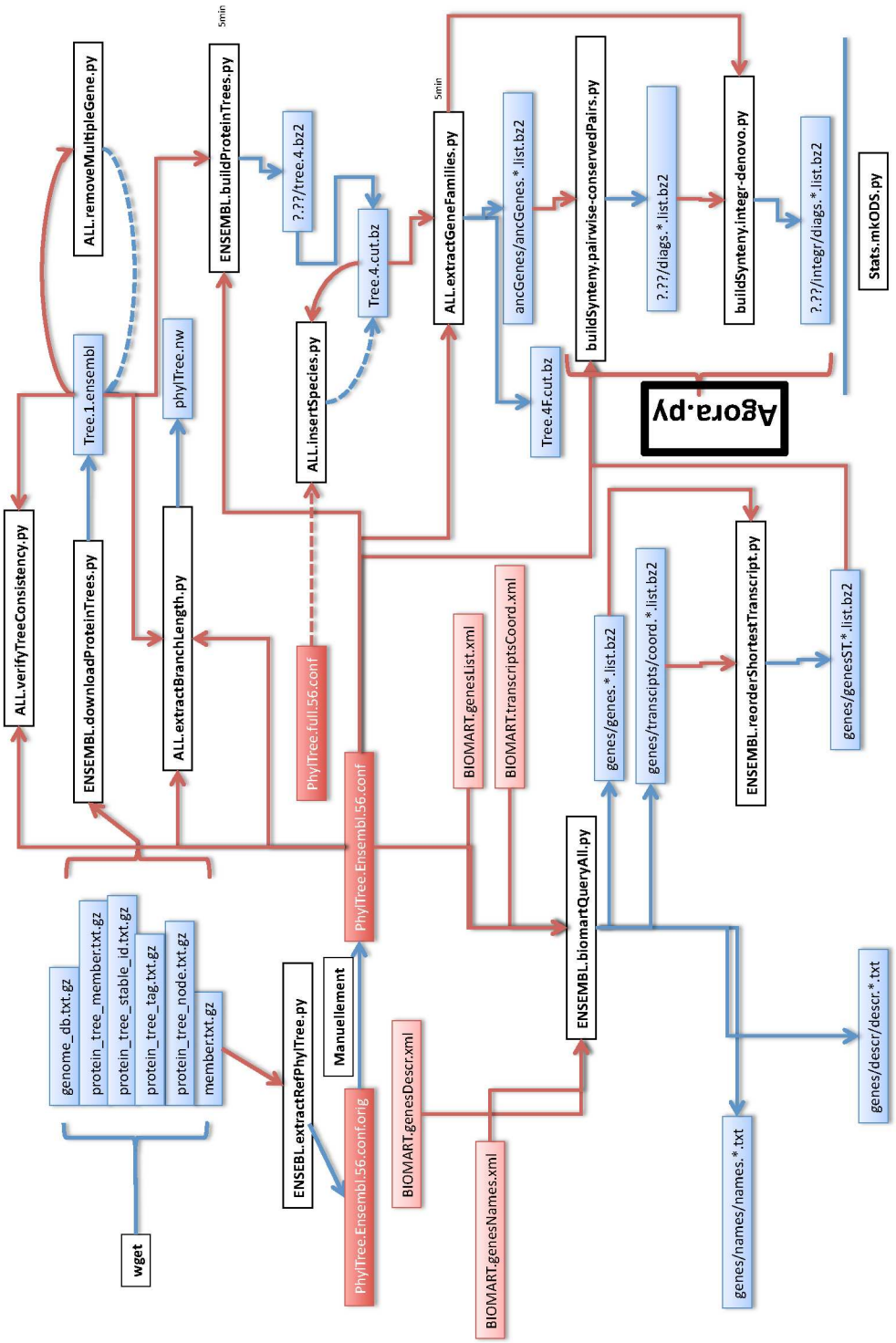


FIGURE A.1 – Procédure de mise à jour des données (reconstructions AGORA).

Annexe B

Genomicus

B.1 Schéma de la base de données

La Figure B.1 montre le schéma des tables dans la base de données de Genomicus. En rose sont indiquées les tables qui permettent les deux vues.

- PhyloView est organisé autour des phylogénies des gènes (tables `Gene` et `Tree`).
- AlignView est organisé autour des phylogénies des espèces (tables `Species` et `SpeciesTree`).

La table `Orthologs` fait le lien entre les espèces en stockant toutes les paires de gènes orthologues (par groupes), elle est primordiale pour les deux vues. Les tables en jaune représentent les données utiles pour l'interface utilisateur, comme la table `Search` qui permet d'interroger la base avec des références d'autres bases de données (noms des gènes, identifiants RefSeq, etc). Les tables en vert représentent les données annexes, les «pistes» d'information que l'on peut rajouter sur les génomes ancestraux, comme les éléments non-codants conservés (tables `CNE` et `CNE_items`).

B.2 Composantes de Genomicus

Module	Nb. lignes		Description
`database`	938	655	Création des tables
`cgi-bin`	2703	1807	Scripts Perl web
`perl-lib`	1762	1166	Librairies Perl
`css`	472	382	Feuilles de style
`js`	69	60	Code JavaScript
total	5944	3980	

TABLE B.1 – Taille (nombre de lignes avec et sans commentaires) des différents modules de Genomicus.

B.3 Mise à jour de la base de données

Par analogie avec la A.1, la B.2 montre la procédure suivie à chaque mise à jour de la base de données Ensembl, pour mettre à jour Genomicus.

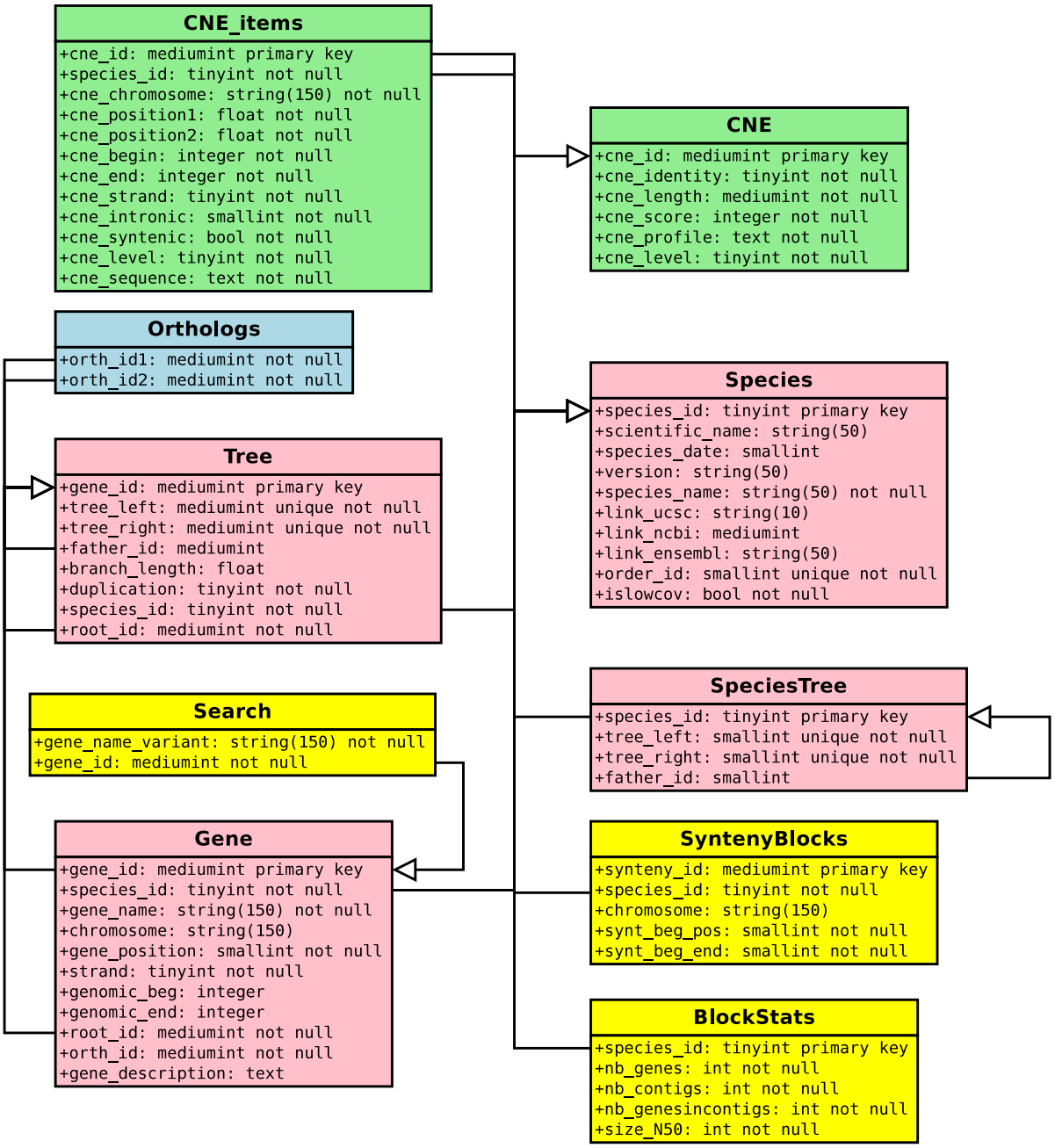


FIGURE B.1 – Schéma des tables de la base de données de Genomicus.

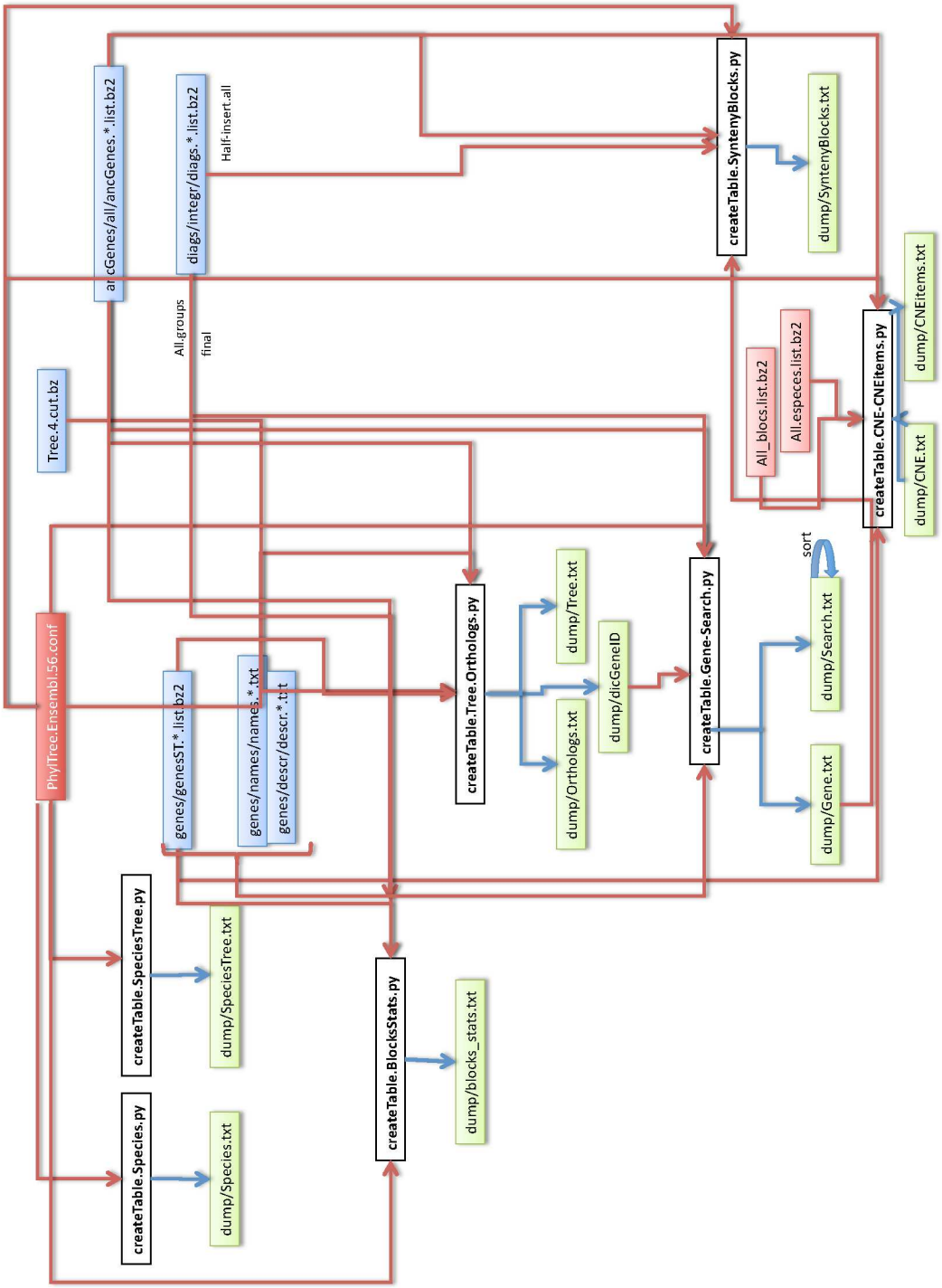


FIGURE B.2 – Procédure de mise à jour des données (base de données Genomicus).

Annexe C

concorde

Malgré ses performances, *concorde* montre quelques limitations sur les données qu'il a à traiter. Il faut ainsi donner les distances sous forme de nombres entiers, et veiller aux dépassements de capacité lors des calculs internes (l'exécutable disponible est prévu pour une architecture de 32 bits, et les nombres entiers y sont bornés à 2^{32}). Dès lors, si on désire ordonner n nœuds, il faut s'assurer (condition forte) que le poids de chaque arête ne dépasse pas un certain $d_{\max}$, à fixer, inférieur à $2^{32}/(n-1)$. On utilise la fonction $x \mapsto d_{\max} \exp(-x/d_{\max})$ pour borner à $d_{\max}$ une distance x calculée. Puisqu'il faut en plus prendre la partie entière des distances avant de les utiliser dans le graphe, on multipliera d'abord les distances par 10^d, pour conserver une précision de d décimales. Par commodité, on utilise 10^9 à la place de 2^{32} pour conserver des puissances de 10. La transformation finale (représentée Figure C.1) est donc :

$$x \mapsto \left\lfloor 10^d d_{\max} \exp(-x/d_{\max}) \right\rfloor \text{ où typiquement } d_{\max} = 10^{\lfloor 9-d-\log_{10}(n) \rfloor} \tag{C.1}$$

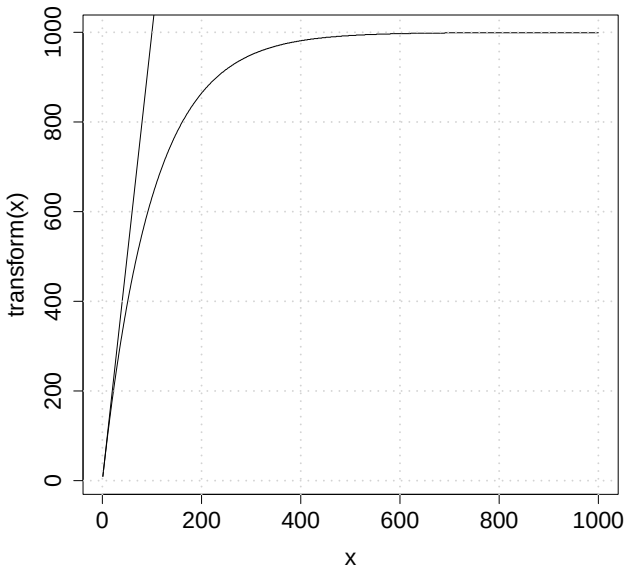


FIGURE C.1 – Transformation d'une distance pour utilisation par *concorde*. La courbe correspond au résultat de la transformation de l'Équation C.1 sur une distance x (en abscisse entre 0 et 1000). Ici, $d_{\max} = 100$ et $d = 1$. La droite correspond à l'équation $y = 10x$, la transformation de base pour ajouter 1 décimale.

Table des figures

Liste des tableaux

Liste des algorithmes

Bibliographie

Michael Abrouk, Florent Murat, Caroline Pont, Joachim Messing, Scott Jackson, Thomas Faraut, Eric Tannier, Christophe Plomion, Richard Cooke, Catherine Feuillet, et Jérôme Salse. Palaeogenomics of plants : synteny-based modelling of extinct ancestors. *Trends Plant Sci*, 15(9) :479–487, Sep 2010. doi: 10.1016/j.tplants.2010.06.001.
Cité page 20.

Max A Alekseyev et Pavel A Pevzner. Breakpoint graphs and ancestral genome reconstructions. *Genome Res*, 19(5) :943–957, May 2009. doi: 10.1101/gr.082784.108.
Cité 3 fois pages 15, 103 et 112.

David L. Applegate, Robert E. Bixby, Vasek Chvátal, et William J. Cook. *The Traveling Salesman Problem : A Computational Study*. Princeton University Press, 2006. URL http://www.tsp.gatech.edu/book/index.html.
Cité page 33.

Jean-Marc Aury, Olivier Jaillon, Laurent Duret, Benjamin Noel, Claire Jubin, Betina M Porcel, Béatrice Ségurens, Vincent Daubin, Véronique Anthouard, Nathalie Aiach, Olivier Arnaiz, Alain Billaut, Janine Beisson, Isabelle Blanc, Khaled Bouhouche, Francisco Câmara, Sandra Duharcourt, Roderic Guigo, Delphine Gogendeau, Michael Katinka, Anne-Marie Keller, Roland Kissmehl, Catherine Klotz, France Koll, Anne Le Mouël, Gersende Lepère, Sophie Malinsky, Mariusz Nowacki, Jacek K Nowak, Helmut Plattner, Julie Poulain, Françoise Ruiz, Vincent Serrano, Marek Zagulski, Philippe Dessen, Mireille Bétermier, Jean Weissenbach, Claude Scarpelli, Vincent Schächter, Linda Sperling, Eric Meyer, Jean Cohen, et Patrick Wincker. Global trends of whole-genome duplications revealed by the ciliate paramecium tetraurelia. *Nature*, 444(7116) :171–178, Nov 2006. doi: 10.1038/nature05230.
Cité 2 fois pages 18 et 20.

D. A. Bader, B. M. Moret, et M. Yan. A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. *J Comput Biol*, 8 (5) :483–491, 2001. doi: 10.1089/106652701753216503.
Cité page 12.

Thomas S Becker et Boris Lenhard. The random versus fragile breakage models of chromosome evolution : a matter of resolution. *Mol Genet Genomics*, 278(5) :487–491, Nov 2007. doi: 10.1007/s00438-007-0287-0.
Cité page 167.

Richard Bellman. Dynamic programming treatment of the travelling salesman problem. *J. ACM*, 9(1) :61–63, 1962. ISSN 0004-5411. doi: 10.1145/321105.321111.
Cité page 33.

Mathieu Blanchette, Eric D Green, Webb Miller, et David Haussler. Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Res*, 14(12) :2412–2423, Dec 2004a. doi: 10.1101/gr.2800104.
Cité 2 fois pages 23 et 166.

Mathieu Blanchette, W. James Kent, Cathy Riemer, Laura Elnitski, Arian F A Smit, Krishna M Roskin, Robert Baertsch, Kate Rosenbloom, Hiram Clawson, Eric D Green, David Haussler, et Webb Miller. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res*, 14(4) :708–715, Apr 2004b. doi: 10.1101/gr.1933104.
Cité page 166.

Guillaume Bourque et Pavel A Pevzner. Genome-scale evolution : reconstructing gene orders in the ancestral species. *Genome Res*, 12(1) :26–36, Jan 2002. doi: 10.1101/gr.12.1.26. URL http://www.genome.org/cgi/content/full/12/1/26.
Cité 3 fois pages 13 et 14.

Guillaume Bourque, Pavel A Pevzner, et Glenn Tesler. Reconstructing the genomic architecture of ancestral mammals : lessons from human, mouse, and rat genomes. *Genome Res*, 14(4) :507–516, Apr 2004. doi: 10.1101/gr.1975204.
Cité 2 fois pages 14 et 112.

Guillaume Bourque, Evgeny M Zdobnov, Peer Bork, Pavel A Pevzner, et Glenn Tesler. Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. *Genome Res*, 15(1) :98–110, Jan 2005. doi: 10.1101/gr.3002305.
Cité 2 fois pages 14 et 104.

Guillaume Bourque, Glenn Tesler, et Pavel A Pevzner. The convergence of cytogenetics and rearrangement-based models for ancestral genome reconstruction. *Genome Res*, 16 (3) :311–313, Mar 2006. doi: 10.1101/gr.4631806.
Cité page 23.

Ingrid J Burgetz, Salimah Shariff, Andy Pang, et Elisabeth R M Tillier. Positional homology in bacterial genomes. *Evol Bioinform Online*, 2 :77–90, 2006.
Cité page 158.

Kevin P Byrne et Kenneth H Wolfe. The yeast gene order browser : combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res*, 15 (10) :1456–1461, Oct 2005. doi: 10.1101/gr.3672305.
Cité 2 fois pages 143 et 162.

Alberto Caprara. Formulations and hardness of multiple sorting by reversals. In *RECOMB '99 : Proceedings of the third annual international conference on Computational molecular biology*, pages 84–93, New York, NY, USA, 1999. ACM. ISBN 1-58113-069-4. doi: http ://doi.acm.org/10.1145/299432.299461.
Cité page 13.

E. A. Carver et L. Stubbs. Zooming in on the human-mouse comparative map : genome conservation re-examined on a high-resolution scale. *Genome Res*, 7(12) :1123–1137, Dec 1997. doi: 10.1101/gr.7.12.1123.
Cité page 13.

Cedric Chauve et Eric Tannier. A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes. *PLoS Comput Biol*, 4(11) :e1000234, Nov 2008. doi: 10.1371/journal.pcbi.1000234.
Cité 4 fois pages 16, 23 et 154.

Cedric Chauve, Haris Gavranovic, Aida Ouangraoua, et Eric Tannier. Yeast ancestral genome reconstructions : the possibilities of computational methods ii. *J Comput Biol*, 17(9) :1097–1112, Sep 2010. doi: 10.1089/cmb.2010.0092.
Cité page 156.

B. P. Chowdhary, T. Raudsepp, L. Frönicke, et H. Scherthan. Emerging patterns of comparative genome organization in some mammalian species as revealed by zoo-fish. *Genome Res*, 8(6) :577–589, Jun 1998. doi: 10.1101/gr.8.6.577.
Cité page 9.

Gene Ontology Consortium. The gene ontology project in 2008. *Nucleic Acids Res*, 36 (Database issue) :D440–D444, Jan 2008. doi: 10.1093/nar/gkm883.
Cité page 169.

International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, 432(7018) :695–716, Dec 2004. doi: 10.1038/nature03154.
Cité page 14.

Emmanuel Courcelle, Yoann Beausse, Sébastien Letort, Olivier Stahl, Romain Fremez, Catherine Ngom-Bru, Jérôme Gouzy, et Thomas Faraut. Narcisse : a mirror view of conserved syntenies. *Nucleic Acids Res*, 36(Database issue) :D485–D490, Jan 2008. doi: 10.1093/nar/gkm805.
Cité page 143.

Yves Van de Peer, Steven Maere, et Axel Meyer. The evolutionary significance of ancient genome duplications. *Nat Rev Genet*, 10(10) :725–732, Oct 2009. doi: 10.1038/nrg2600.
Cité page 162.

Yves Van de Peer, Steven Maere, et Axel Meyer. 2r or not 2r is not the question anymore. *Nat Rev Genet*, 11(2) :166, Feb 2010. doi: 10.1038/nrg2600-c2.
Cité page 20.

Paramvir Dehal et Jeffrey L Boore. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol*, 3(10) :e314, Oct 2005. doi: 10.1371/journal.pbio.0030314.
Cité page 20.

France Denoeud, Simon Henriet, Sutada Mungpakdee, Jean-Marc Aury, Corinne Da Silva, Henner Brinkmann, Jana Mikhaleva, Lisbeth Charlotte Olsen, Claire Jubin, Cristian Cañestro, Jean-Marie Bouquet, Gemma Danks, Julie Poulain, Coen Campsteijn, Marcin Adamski, Ismael Cross, Fekadu Yadetie, Matthieu Muffato, Alexandra Louis, Stephen Butcher, Georgia Tsagkogeorga, Anke Konrad, Sarabdeep Singh, Marit Flo Jensen, Evelyne Huynh Cong, Helen Eikeseth-Otteraa, Benjamin Noel, Véronique Anthouard, Betina M Porcel, Rym Kachouri-Lafond, Atsuo Nishino, Matteo Ugolini, Pascal Chourrout, Hiroki Nishida, Rein Aasland, Snehalata Huzurbazar, Eric Westhof, Frédéric Delsuc, Hans Lehrach, Richard Reinhardt, Jean Weissenbach, Scott W

Roy, François Artiguenave, John H Postlethwait, J. Robert Manak, Eric M Thompson, Olivier Jaillon, Louis Du Pasquier, Pierre Boudinot, David A Liberles, Jean-Nicolas Volff, Hervé Philippe, Boris Lenhard, Hugues Roest Crollius, Patrick Wincker, et Daniel Chourrout. Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science*, Nov 2010. doi: 10.1126/science.1194167.
Cité 3 fois pages 6, 48 et 149.

Thomas Derrien, Catherine André, Francis Galibert, et Christophe Hitte. Autograph : an interactive web server for automating and visualizing comparative genome maps. *Bioinformatics*, 23(4) :498–499, Feb 2007. doi: 10.1093/bioinformatics/btl618.
Cité page 143.

Fred S Dietrich, Sylvia Voegeli, Sophie Brachat, Anita Lerch, Krista Gates, Sabine Steiner, Christine Mohr, Rainer Pöhlmann, Philippe Luedi, Sangdun Choi, Rod A Wing, Albert Flavier, Thomas D Gaffney, et Peter Philippsen. The ashbya gossypii genome as a tool for mapping the ancient saccharomyces cerevisiae genome. *Science*, 304(5668) : 304–307, Apr 2004. doi: 10.1126/science.1095781.
Cité page 18.

Gauthier Dobigny, Jean-François Ducroz, Terence J Robinson, et Vitaly Volobouev. Cytogenetics and cladistics. *Syst Biol*, 53(3) :470–484, Jun 2004. doi: 10.1080/10635150490445698.
Cité page 9.

Xianjun Dong, David Fredman, et Boris Lenhard. Synorth : exploring the evolution of synteny and long-range regulatory interactions in vertebrate genomes. *Genome Biol*, 10(8) :R86, 2009. doi: 10.1186/gb-2009-10-8-r86.
Cité page 143.

David Enard, Frantz Depaulis, et Hugues Roest Crollius. Human and non-human primate genomes share hotspots of positive selection. *PLoS Genet*, 6(2) :e1000840, 2010. doi: 10.1371/journal.pgen.1000840.
Cité page 169.

Malcolm A Ferguson-Smith et Vladimir Trifonov. Mammalian karyotype evolution. *Nat Rev Genet*, 8(12) :950–962, Dec 2007. doi: 10.1038/nrg2199.
Cité 3 fois pages 9, 10 et 23.

Walter M. Fitch. Toward defining the course of evolution : Minimum change for a specific tree topology. *Systematic Zoology*, 20(4) :406–416, 1971. ISSN 00397989. URL `http://www.jstor.org/stable/2412116`.
Cité page 15.

Paul Flicek, Bronwen L Aken, Benoit Ballester, Kathryn Beal, Eugene Bragin, Simon Brent, Yuan Chen, Peter Clapham, Guy Coates, Susan Fairley, Stephen Fitzgerald, Julio Fernandez-Banet, Leo Gordon, Stefan Gräf, Syed Haider, Martin Hammond, Kerstin Howe, Andrew Jenkinson, Nathan Johnson, Andreas Kähäri, Damian Keefe, Stephen Keenan, Rhoda Kinsella, Felix Kokocinski, Gautier Koscielny, Eugene Kulesha, Daniel Lawson, Ian Longden, Tim Massingham, William McLaren, Karine Megy, Bert Overduin, Bethan Pritchard, Daniel Rios, Magali Ruffier, Michael Schuster, Guy Slater, Damian Smedley, Giulietta Spudich, Y. Amy Tang, Stephen Trevanion, Albert Vilella, Jan

Vogel, Simon White, Steven P Wilder, Amonida Zadissa, Ewan Birney, Fiona Cunningham, Ian Dunham, Richard Durbin, Xosé M Fernández-Suarez, Javier Herrero, Tim J P Hubbard, Anne Parker, Glenn Proctor, James Smith, et Stephen M J Searle. Ensembl's 10th year. *Nucleic Acids Res*, 38(Database issue) :D557–D562, Jan 2010. doi: 10.1093/nar/gkp972.
Cité page 45.

L. Froenicke. Origins of primate chromosomes - as delineated by zoo-fish and alignments of human and mouse draft genome sequences. *Cytogenet Genome Res*, 108(1-3) :122–138, 2005. doi: 10.1159/000080810.
Cité 2 fois pages 9 et 23.

Lutz Froenicke, Montserrat Garcia Caldés, Alexander Graphodatsky, Stefan Müller, Leslie A Lyons, Terence J Robinson, Marianne Volleth, Fengtang Yang, et Johannes Wienberg. Are molecular cytogenetics and bioinformatics suggesting diverging models of ancestral mammalian genomes ? *Genome Res*, 16(3) :306–310, Mar 2006. doi: 10.1101/gr.3955206.
Cité 2 fois pages 23 et 126.

Haris Gavranović et Eric Tannier. Guided genome halving : provably optimal solutions provide good insights into the preduplication ancestral genome of saccharomyces cerevisiae. *Pac Symp Biocomput*, pages 21–30, 2010.
Cité page 160.

R. Glas, J. A. Marshall Graves, R. Toder, M. Ferguson-Smith, et P. C. O'Brien. Cross-species chromosome painting between human and marsupial directly demonstrates the ancient region of the mammalian x. *Mamm Genome*, 10(11) :1115–1116, Nov 1999. doi: 10.1007/s003359901174.
Cité page 9.

Jonathan L Gordon, Kevin P Byrne, et Kenneth H Wolfe. Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern saccharomyces cerevisiae genome. *PLoS Genet*, 5(5) :e1000485, May 2009. doi: 10.1371/journal.pgen.1000485.
Cité 7 fois pages 18, 19, 160, 165 et 188.

Stéphane Guindon et Olivier Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*, 52(5) :696–704, Oct 2003. doi: 10.1080/10635150390235520.
Cité page 166.

J. A. Hartigan. Minimum mutation fits to a given tree. *Biometrics*, 29(1) :pp. 53–65, 1973. ISSN 0006341X. URL http://www.jstor.org/stable/2529676.
Cité page 15.

M. Hasegawa, Y. Iida, T. Yano, F. Takaiwa, et M. Iwabuchi. Phylogenetic relationships among eukaryotic kingdoms inferred from ribosomal rna sequences. *J Mol Evol*, 22(1) : 32–38, 1985. doi: 10.1007/BF02105802.
Cité page 166.

S. Blair Hedges, Joel Dudley, et Sudhir Kumar. Timetree : a public knowledge-base of divergence times among organisms. *Bioinformatics*, 22(23) :2971–2972, Dec 2006. doi: 10.1093/bioinformatics/btl505.
Cité page 22.

A. L. Hughes, J. da Silva, et R. Friedman. Ancient genome duplications did not structure the human hox-bearing chromosomes. *Genome Res*, 11(5) :771–780, May 2001. doi: 10.1101/gr.160001.
Cité page 20.

Olivier Jaillon, Jean-Marc Aury, Frédéric Brunet, Jean-Louis Petit, Nicole Stange-Thomann, Evan Mauceli, Laurence Bouneau, Cécile Fischer, Catherine Ozouf-Costaz, Alain Bernot, Sophie Nicaud, David Jaffe, Sheila Fisher, Georges Lutfalla, Carole Dossat, Béatrice Segurens, Corinne Dasilva, Marcel Salanoubat, Michael Levy, Nathalie Boudet, Sergi Castellano, Véronique Anthouard, Claire Jubin, Vanina Castelli, Michael Katinka, Benoît Vacherie, Christian Biémont, Zineb Skalli, Laurence Cattolico, Julie Poulain, Véronique De Berardinis, Corinne Cruaud, Simone Duprat, Philippe Brottier, Jean-Pierre Coutanceau, Jérôme Gouzy, Genis Parra, Guillaume Lardier, Charles Chapple, Kevin J McKernan, Paul McEwan, Stephanie Bosak, Manolis Kellis, Jean-Nicolas Volff, Roderic Guigó, Michael C Zody, Jill Mesirov, Kerstin Lindblad-Toh, Bruce Birren, Chad Nusbaum, Daniel Kahn, Marc Robinson-Rechavi, Vincent Laudet, Vincent Schachter, Francis Quétier, William Saurin, Claude Scarpelli, Patrick Wincker, Eric S Lander, Jean Weissenbach, et Hugues Roest Crollius. Genome duplication in the teleost fish tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*, 431(7011) :946–957, Oct 2004. doi: 10.1038/nature03025.
Cité 5 fois pages 17, 18, 19 et 24.

Lars J Jensen, Michael Kuhn, Manuel Stark, Samuel Chaffron, Chris Creevey, Jean Muller, Tobias Doerks, Philippe Julien, Alexander Roth, Milan Simonovic, Peer Bork, et Christian von Mering. String 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res*, 37(Database issue) :D412–D416, Jan 2009. doi: 10.1093/nar/gkn760.
Cité page 143.

Masahiro Kasahara, Kiyoshi Naruse, Shin Sasaki, Yoichiro Nakatani, Wei Qu, Budrul Ahsan, Tomoyuki Yamada, Yukinobu Nagayasu, Koichiro Doi, Yasuhiro Kasai, Tomoko Jindo, Daisuke Kobayashi, Atsuko Shimada, Atsushi Toyoda, Yoko Kuroki, Asao Fujiyama, Takashi Sasaki, Atsushi Shimizu, Shuichi Asakawa, Nobuyoshi Shimizu, Shin-Ichi Hashimoto, Jun Yang, Yongjun Lee, Kouji Matsushima, Sumio Sugano, Mitsuru Sakaizumi, Takanori Narita, Kazuko Ohishi, Shinobu Haga, Fumiko Ohta, Hisayo Nomoto, Keiko Nogata, Tomomi Morishita, Tomoko Endo, Tadasu Shin-I, Hiroyuki Takeda, Shinichi Morishita, et Yuji Kohara. The medaka draft genome and insights into vertebrate genome evolution. *Nature*, 447(7145) :714–719, Jun 2007. doi: 10.1038/nature05846.
Cité 8 fois pages 18, 19, 24, 25, 138 et 139.

Manolis Kellis, Bruce W Birren, et Eric S Lander. Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisiae. *Nature*, 428(6983) : 617–624, Apr 2004. doi: 10.1038/nature02424.
Cité 2 fois pages 18 et 19.

Claus Kemkemer, Matthias Kohn, Hildegard Kehrer-Sawatzki, Peter Minich, Josef Högel, Lutz Froenicke, et Horst Hameister. Reconstruction of the ancestral ferungulate karyotype by electronic chromosome painting (e-painting). *Chromosome Res*, 14(8) : 899–907, 2006. doi: 10.1007/s10577-006-1097-7.
Cité page 9.

Claus Kemkemer, Matthias Kohn, David N Cooper, Lutz Froenicke, Josef Högel, Horst Hameister, et Hildegard Kehrer-Sawatzki. Gene synteny comparisons between different vertebrates provide new insights into breakage and fusion events during mammalian karyotype evolution. *BMC Evol Biol*, 9 :84, 2009. doi: 10.1186/1471-2148-9-84.
Cité 5 fois pages 9, 11, 23 et 24.

P. J. Kersey, D. Lawson, E. Birney, P. S. Derwent, M. Haimel, J. Herrero, S. Keenan, A. Kerhornou, G. Koscielny, A. Kähäri, R. J. Kinsella, E. Kulesha, U. Maheswari, K. Megy, M. Nuhn, G. Proctor, D. Staines, F. Valentin, A. J. Vilella, et A. Yates. Ensembl genomes : extending ensembl across the taxonomic space. *Nucleic Acids Res*, 38(Database issue) :D563–D569, Jan 2010. doi: 10.1093/nar/gkp871.
Cité page 46.

Hiroshi Kikuta, Mary Laplante, Pavla Navratilova, Anna Z Komisarczuk, Pär G Engström, David Fredman, Altuna Akalin, Mario Caccamo, Ian Sealy, Kerstin Howe, Julien Ghislain, Guillaume Pezeron, Philippe Mourrain, Staale Ellingsen, Andrew C Oates, Christine Thisse, Bernard Thisse, Isabelle Foucher, Birgit Adolf, Andrea Geling, Boris Lenhard, et Thomas S Becker. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res*, 17(5) :545–555, May 2007. doi: 10.1101/gr.6086307.
Cité 2 fois pages 162 et 167.

David G Knowles et Aoife McLysaght. Recent de novo origin of human protein-coding genes. *Genome Res*, 19(10) :1752–1759, Oct 2009. doi: 10.1101/gr.095026.109.
Cité page 103.

E. S. Lander et M. S. Waterman. Genomic mapping by fingerprinting random clones : a mathematical analysis. *Genomics*, 2(3) :231–239, Apr 1988. doi: 0888-7453/88.
Cité page 109.

Dan Larhammar, Lars-Gustav Lundin, et Finn Hallböök. The human hox-bearing chromosome regions did arise by block or chromosome (or even genome) duplications. *Genome Res*, 12(12) :1910–1920, Dec 2002. doi: 10.1101/gr.445702.
Cité page 20.

Urban Liebel, Bjoern Kindler, et Rainer Pepperkok. Bioinformatic "harvester" : a search engine for genome-wide human, mouse, and rat protein resources. *Methods Enzymol*, 404 :19–26, 2005. doi: 10.1016/S0076-6879(05)04003-6.
Cité page 145.

Eric Lyons, Brent Pedersen, Josh Kane, Maqsudul Alam, Ray Ming, Haibao Tang, Xiyin Wang, John Bowers, Andrew Paterson, Damon Lisch, et Michael Freeling. Finding and comparing syntenic regions among arabidopsis and the outgroups papaya, poplar, and grape : Coge with rosids. *Plant Physiol*, 148(4) :1772–1781, Dec 2008. doi: 10.1104/pp.108.124867.
Cité page 143.

Jian Ma, Louxin Zhang, Bernard B Suh, Brian J Raney, Richard C Burhans, W. James Kent, Mathieu Blanchette, David Haussler, et Webb Miller. Reconstructing contiguous regions of an ancestral genome. *Genome Res*, 16(12) :1557–1565, Dec 2006. doi: 10.1101/gr.5383506.
Cité 9 fois pages 15, 16, 17, 23, 105, 112, 137 et 167.

Aoife McLysaght, Karsten Hokamp, et Kenneth H Wolfe. Extensive genomic duplication during early chordate evolution. *Nat Genet*, 31(2) :200–204, Jun 2002. doi: 10.1038/ng884.
Cité page 20.

Huaiyu Mi, Qing Dong, Anushya Muruganujan, Pascale Gaudet, Suzanna Lewis, et Paul D Thomas. Panther version 7 : improved phylogenetic trees, orthologs and collaboration with the gene ontology consortium. *Nucleic Acids Res*, 38(Database issue) : D204–D210, Jan 2010. doi: 10.1093/nar/gkp1019.
Cité page 169.

Webb Miller, Kate Rosenbloom, Ross C Hardison, Minmei Hou, James Taylor, Brian Raney, Richard Burhans, David C King, Robert Baertsch, Daniel Blankenberg, Sergei L Kosakovsky Pond, Anton Nekrutenko, Belinda Giardine, Robert S Harris, Svitlana Tyekucheva, Mark Diekhans, Thomas H Pringle, William J Murphy, Arthur Lesk, George M Weinstock, Kerstin Lindblad-Toh, Richard A Gibbs, Eric S Lander, Adam Siepel, David Haussler, et W. James Kent. 28-way vertebrate alignment and conservation track in the ucsc genome browser. *Genome Res*, 17(12) :1809–1822, Dec 2007. doi: 10.1101/gr.6761107.
Cité page 20.

Matthieu Muffato et Hugues Roest Crollius. Paleogenomics in vertebrates, or the recovery of lost genomes from the mist of time. *BioEssays*, 30(2) :122–134, Feb 2008. doi: 10.1002/bies.20707.
Cité 2 fois pages 5 et 28.

Matthieu Muffato, Alexandra Louis, Charles-Edouard Poisnel, et Hugues Roest Crollius. Genomicus : a database and a browser to study gene synteny in modern and ancestral genomes. *Bioinformatics*, 26(8) :1119–1121, Apr 2010. doi: 10.1093/bioinformatics/btq079.
Cité 3 fois pages 4, 5 et 145.

Florent Murat, Jian-Hong Xu, Eric Tannier, Michael Abrouk, Nicolas Guilhot, Caroline Pont, Joachim Messing, et Jérôme Salse. Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. *Genome Res*, Sep 2010. doi: 10.1101/gr.109744.110.
Cité page 20.

William J Murphy, Denis M Larkin, Annelie Everts van der Wind, Guillaume Bourque, Glenn Tesler, Loretta Auvil, Jonathan E Beever, Bhanu P Chowdhary, Francis Galibert, Lisa Gatzke, Christophe Hitte, Stacey N Meyers, Denis Milan, Elaine A Ostrander, Greg Pape, Heidi G Parker, Terje Raudsepp, Margarita B Rogatcheva, Lawrence B Schook, Loren C Skow, Michael Welge, James E Womack, Stephen J O'brien, Pavel A

Pevzner, et Harris A Lewin. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science*, 309(5734) :613–617, Jul 2005. doi: 10.1126/science.1111387.
Cité 2 fois page 14.

William J Murphy, Thomas H Pringle, Tess A Crider, Mark S Springer, et Webb Miller. Using genomic data to unravel the root of the placental mammal phylogeny. *Genome Res*, 17(4) :413–421, Apr 2007. doi: 10.1101/gr.5918807.
Cité page 48.

Stefan Müller, Melanie Hollatz, et Johannes Wienberg. Chromosomal phylogeny and evolution of gibbons (hylobatidae). *Hum Genet*, 113(6) :493–501, Nov 2003. doi: 10.1007/s00439-003-0997-2.
Cité page 9.

J. H. Nadeau et B. A. Taylor. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc Natl Acad Sci U S A*, 81(3) :814–818, Feb 1984. URL http://www.pnas.org/cgi/content/full/81/3/814.
Cité page 167.

Sridevi Nagarajan, Willem Rens, James Stalker, Tony Cox, et Malcolm A Ferguson-Smith. Chromhome : a rich internet application for accessing comparative chromosome homology maps. *BMC Bioinformatics*, 9 :168, 2008. doi: 10.1186/1471-2105-9-168.
Cité page 9.

Yoichiro Nakatani, Hiroyuki Takeda, Yuji Kohara, et Shinichi Morishita. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res*, 17(9) :1254–1265, Sep 2007. doi: 10.1101/gr.6316407.
Cité 10 fois pages 18, 20, 25, 26, 86, 87, 104, 139 et 187.

Kiyoshi Naruse, Minoru Tanaka, Kazuei Mita, Akihiro Shima, John Postlethwait, et Hiroshi Mitani. A medaka gene map : the trace of ancestral vertebrate proto-chromosomes revealed by comparative gene mapping. *Genome Res*, 14(5) :820–828, May 2004. doi: 10.1101/gr.2004004.
Cité page 19.

S. Ohno, U. Wolf, et N. B. Atkin. Evolution from fish to mammals by gene duplication. *Hereditas*, 59(1) :169–187, 1968.
Cité page 20.

S. P. Otto et J. Whitton. Polyploid incidence and evolution. *Annu Rev Genet*, 34 :401–437, 2000. doi: 10.1146/annurev.genet.34.1.401.
Cité 3 fois pages 17 et 162.

Michal Ozery-Flato et Ron Shamir. An o(n$^{3/2}$sqrt(log n)) algorithm for sorting by reciprocal translocations. In Moshe Lewenstein et Gabriel Valiente, editors, *CPM*, volume 4009 of *Lecture Notes in Computer Science*, pages 258–269. Springer, 2006. ISBN 3-540-35455-7.
Cité page 12.

Xiaokang Pan, Lincoln Stein, et Volker Brendel. Synbrowse : a synteny browser for comparative sequence analysis. *Bioinformatics*, 21(17) :3461–3468, Sep 2005. doi: 10.1093/bioinformatics/bti555.
Cité page 143.

Georgia Panopoulou et Albert J Poustka. Timing and mechanism of ancient vertebrate genome duplications – the adventure of a hypothesis. *Trends Genet*, 21(10) :559–567, Oct 2005. doi: 10.1016/j.tig.2005.08.004.
Cité page 20.

Benedict Paten, Javier Herrero, Stephen Fitzgerald, Kathryn Beal, Paul Flicek, Ian Holmes, et Ewan Birney. Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res*, 18(11) :1829–1843, Nov 2008. doi: 10.1101/gr.076521.108.
Cité page 167.

Simon Penel, Anne-Muriel Arigon, Jean-François Dufayard, Anne-Sophie Sertier, Vincent Daubin, Laurent Duret, Manolo Gouy, et Guy Perrière. Databases of homologous gene families for comparative genomics. *BMC Bioinformatics*, 10 Suppl 6 :S3, 2009. doi: 10.1186/1471-2105-10-S6-S3.
Cité page 166.

Pavel Pevzner et Glenn Tesler. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc Natl Acad Sci U S A*, 100(13) :7672–7677, Jun 2003a. doi: 10.1073/pnas.1330369100.
Cité page 167.

Pavel Pevzner et Glenn Tesler. Genome rearrangements in mammalian evolution : lessons from human and mouse genomes. *Genome Res*, 13(1) :37–45, Jan 2003b. doi: 10.1101/gr.757503.
Cité page 13.

Pascal Pons et Matthieu Latapy. Computing communities in large networks using random walks (long version). *Journal of Graph Algorithms and Applications*, 10(2) :191–218, Dec 2005. URL http://arxiv.org/abs/physics/0512106.
Cité page 35.

Arjun B Prasad, Marc W Allard, N. I. S. C. Comparative Sequencing Program, et Eric D Green. Confirming the phylogeny of mammals by use of large comparative sequence data sets. *Mol Biol Evol*, 25(9) :1795–1808, Sep 2008. doi: 10.1093/molbev/msn104.
Cité page 48.

Nicholas H Putnam, Mansi Srivastava, Uffe Hellsten, Bill Dirks, Jarrod Chapman, Asaf Salamov, Astrid Terry, Harris Shapiro, Erika Lindquist, Vladimir V Kapitonov, Jerzy Jurka, Grigory Genikhovich, Igor V Grigoriev, Susan M Lucas, Robert E Steele, John R Finnerty, Ulrich Technau, Mark Q Martindale, et Daniel S Rokhsar. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science*, 317(5834) :86–94, Jul 2007. doi: 10.1126/science.1139158.
Cité 2 fois pages 27 et 48.

Nicholas H Putnam, Thomas Butts, David E K Ferrier, Rebecca F Furlong, Uffe Hellsten, Takeshi Kawashima, Marc Robinson-Rechavi, Eiichi Shoguchi, Astrid Terry, Jr-Kai Yu, E. Lia Benito-Gutiérrez, Inna Dubchak, Jordi Garcia-Fernàndez, Jeremy J

Gibson-Brown, Igor V Grigoriev, Amy C Horton, Pieter J de Jong, Jerzy Jurka, Vladimir V Kapitonov, Yuji Kohara, Yoko Kuroki, Erika Lindquist, Susan Lucas, Kazutoyo Osoegawa, Len A Pennacchio, Asaf A Salamov, Yutaka Satou, Tatjana Sauka-Spengler, Jeremy Schmutz, Tadasu Shin-I, Atsushi Toyoda, Marianne Bronner-Fraser, Asao Fujiyama, Linda Z Holland, Peter W H Holland, Nori Satoh, et Daniel S Rokhsar. The amphioxus genome and the evolution of the chordate karyotype. *Nature*, 453(7198) : 1064–1071, Jun 2008. doi: 10.1038/nature06967.
Cité 7 fois pages 20, 23, 27, 48, 139 et 187.

R. Puttagunta, L. A. Gordon, G. E. Meyer, D. Kapfhamer, J. E. Lamerdin, P. Kantheti, K. M. Portman, W. K. Chung, D. E. Jenne, A. S. Olsen, et M. Burmeister. Comparative maps of human 19p13.3 and mouse chromosome 10 allow identification of sequences at evolutionary breakpoints. *Genome Res*, 10(9) :1369–1380, Sep 2000. doi: 10.1101/gr.145200.
Cité page 13.

F. Richard, M. Lombard, et B. Dutrillaux. Reconstruction of the ancestral karyotype of eutherian mammals. *Chromosome Res*, 11(6) :605–618, 2003. doi: 10.1023/A :1024957002755.
Cité page 9.

Terence J Robinson, Aurora Ruiz-Herrera, et Lutz Froenicke. Dissecting the mammalian genome–new insights into chromosomal evolution. *Trends Genet*, 22(6) :297–301, Jun 2006. doi: 10.1016/j.tig.2006.04.002.
Cité page 23.

Y. Rumpler et B. Dutrillaux. Chromosomal evolution in malagasy lemurs. i. chromosome banding studies in the genuses lemur and micrcebus. *Cytogenet Cell Genet*, 17(5) : 268–281, 1976.
Cité page 8.

Jérôme Salse, Michael Abrouk, Stéphanie Bolot, Nicolas Guilhot, Emmanuel Courcelle, Thomas Faraut, Robbie Waugh, Timothy J Close, Joachim Messing, et Catherine Feuillet. Reconstruction of monocotelydoneous proto-chromosomes reveals faster evolution in plants than in animals. *Proc Natl Acad Sci U S A*, 106(35) :14908–14913, Sep 2009. doi: 10.1073/pnas.0902350106.
Cité page 18.

David Sankoff, Chunfang Zheng, et Qian Zhu. Polyploids, genome halving and phylogeny. *Bioinformatics*, 23(13) :i433–i439, Jul 2007. doi: 10.1093/bioinformatics/btm169.
Cité page 160.

H. Scherthan, T. Cremer, U. Arnason, H. U. Weier, A. Lima de Faria, et L. Frönicke. Comparative chromosome painting discloses homologous segments in distantly related mammals. *Nat Genet*, 6(4) :342–347, Apr 1994. doi: 10.1038/ng0494-342.
Cité page 8.

Marielle C Schneider, Adilson A Zacaro, Ricardo Pinto-Da-Rocha, Denise M Candido, et Doralice M Cella. A comparative cytogenetic analysis of 2 bothriuridae species and overview of the chromosome data of scorpiones. *J Hered*, 100(5) :545–555, 2009. doi: 10.1093/jhered/esp023.
Cité page 9.

Amit U Sinha et Jaroslaw Meller. Cinteny : flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. *BMC Bioinformatics*, 8 :82, 2007. doi: 10.1186/1471-2105-8-82.
Cité page 143.

Mansi Srivastava, Emina Begovic, Jarrod Chapman, Nicholas H Putnam, Uffe Hellsten, Takeshi Kawashima, Alan Kuo, Therese Mitros, Asaf Salamov, Meredith L Carpenter, Ana Y Signorovitch, Maria A Moreno, Kai Kamm, Jane Grimwood, Jeremy Schmutz, Harris Shapiro, Igor V Grigoriev, Leo W Buss, Bernd Schierwater, Stephen L Dellaporta, et Daniel S Rokhsar. The trichoplax genome and the nature of placozoans. *Nature*, 454(7207) :955–960, Aug 2008. doi: 10.1038/nature07191.
Cité page 27.

R. Stanyon, M. Rocchi, O. Capozzi, R. Roberto, D. Misceo, M. Ventura, M. F. Cardone, F. Bigoni, et N. Archidiacono. Primate chromosome evolution : ancestral karyotypes, marker order and neocentromeres. *Chromosome Res*, 16(1) :17–39, 2008. doi: 10.1007/s10577-007-1209-z.
Cité page 9.

Marta Svartman, Gary Stone, et Roscoe Stanyon. The ancestral eutherian karyotype is present in xenarthra. *PLoS Genet*, 2(7) :e109, Jul 2006. doi: 10.1371/journal.pgen.0020109.
Cité page 8.

Firas Swidan, Eduardo P C Rocha, Michael Shmoish, et Ron Y Pinter. An integrative method for accurate comparative genome mapping. *PLoS Comput Biol*, 2(8) :e75, Aug 2006. doi: 10.1371/journal.pcbi.0020075.
Cité page 158.

Marie Sémon et Kenneth H Wolfe. Rearrangement rate following the whole-genome duplication in teleosts. *Mol Biol Evol*, 24(3) :860–867, Mar 2007. doi: 10.1093/molbev/msm003.
Cité 2 fois pages 18 et 101.

Eric Tannier, Anne Bergeron, et Marie-France Sagot. Advances on sorting by reversals. *Discrete Appl. Math.*, 155(6-7) :881–888, 2007. ISSN 0166-218X. doi: 10.1016/j.dam.2005.02.033.
Cité page 12.

Glenn Tesler. Grimm : genome rearrangements web server. *Bioinformatics*, 18(3) :492–493, Mar 2002. doi: 10.1093/bioinformatics/18.3.492.
Cité page 13.

Albert J Vilella, Jessica Severin, Abel Ureta-Vidal, Li Heng, Richard Durbin, et Ewan Birney. Ensemblcompara genetrees : Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res*, 19(2) :327–335, Feb 2009. doi: 10.1101/gr.073585.107.
Cité 2 fois pages 46 et 158.

Iain M Wallace, Orla O'Sullivan, Desmond G Higgins, et Cedric Notredame. M-coffee : combining multiple sequence alignment methods with t-coffee. *Nucleic Acids Res*, 34 (6) :1692–1699, 2006. doi: 10.1093/nar/gkl091.
Cité page 46.

Ilan Wapinski, Avi Pfeffer, Nir Friedman, et Aviv Regev. Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics*, 23(13) :i549–i558, Jul 2007. doi: 10.1093/bioinformatics/btm193.
Cité page 158.

Michael Westerman, Robert W Meredith, et Mark S Springer. Cytogenetics meets phylogenetics : A review of karyotype evolution in diprotodontian marsupials. *J Hered*, Jun 2010. doi: 10.1093/jhered/esq076.
Cité page 9.

J. Wienberg, A. Jauch, R. Stanyon, et T. Cremer. Molecular cytotaxonomy of primates by chromosomal in situ suppression hybridization. *Genomics*, 8(2) :347–350, Oct 1990. doi: 10.1016/0888-7543(90)90292-3.
Cité page 8.

Johannes Wienberg. The evolution of eutherian chromosomes. *Curr Opin Genet Dev*, 14 (6) :657–666, Dec 2004. doi: 10.1016/j.gde.2004.10.001.
Cité page 9.

Derek E Wildman, Monica Uddin, Juan C Opazo, Guozhen Liu, Vincent Lefort, Stephane Guindon, Olivier Gascuel, Lawrence I Grossman, Roberto Romero, et Morris Goodman. Genomics, biogeography, and the diversification of placental mammals. *Proc Natl Acad Sci U S A*, 104(36) :14395–14400, Sep 2007. doi: 10.1073/pnas.0704342104.
Cité page 48.

L. G. Wilming, J. G R Gilbert, K. Howe, S. Trevanion, T. Hubbard, et J. L. Harrow. The vertebrate genome annotation (vega) database. *Nucleic Acids Res*, 36(Database issue) : D753–D760, Jan 2008. doi: 10.1093/nar/gkm987.
Cité page 45.

K. H. Wolfe et D. C. Shields. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387(6634) :708–713, Jun 1997. doi: 10.1038/42711.
Cité page 17.

Ian G Woods, Catherine Wilson, Brian Friedlander, Patricia Chang, Daengnoy K Reyes, Rebecca Nix, Peter D Kelly, Felicia Chu, John H Postlethwait, et William S Talbot. The zebrafish gene map defines ancestral vertebrate chromosomes. *Genome Res*, 15(9) : 1307–1314, Sep 2005. doi: 10.1101/gr.4134305.
Cité page 19.

Sophia Yancopoulos et Richard Friedberg. Dcj path formulation for genome transformations which include insertions, deletions, and duplications. *J Comput Biol*, 16(10) : 1311–1338, Oct 2009. doi: 10.1089/cmb.2009.0092.
Cité 2 fois pages 13 et 159.

Sophia Yancopoulos, Oliver Attie, et Richard Friedberg. Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics*, 21(16) : 3340–3346, Aug 2005. doi: 10.1093/bioinformatics/bti535.
Cité page 12.

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 chromosome painting among human, aardvark, and elephant (superorder afrotheria) reveals the likely eutherian ancestral karyotype. *Proc Natl Acad Sci U S A*, 100(3) :1062–1066, Feb 2003. doi: 10.1073/pnas.0335540100.
Cité page 9.

Fengtang Yang, Alexander S Graphodatsky, Tangliang Li, Beiyuan Fu, Gauthier Dobigny, Jinghuan Wang, Polina L Perelman, Natalya A Serdukova, Weiting Su, Patricia Cm O'Brien, Yingxiang Wang, Malcolm A Ferguson-Smith, Vitaly Volobouev, et Wenhui Nie. Comparative genome maps of the pangolin, hedgehog, sloth, anteater and human revealed by cross-species chromosome painting : further insight into the ancestral karyotype and genome evolution of eutherian mammals. *Chromosome Res*, 14(3) :283–296, 2006. doi: 10.1007/s10577-006-1045-6.
Cité page 9.

Ziheng Yang. Paml 4 : phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, 24 (8) :1586–1591, Aug 2007. doi: 10.1093/molbev/msm088.
Cité page 166.

J. J. Yunis et O. Prakash. The origin of man : a chromosomal pictorial legacy. *Science*, 215 (4539) :1525–1530, Mar 1982. doi: 10.1126/science.7063861.
Cité page 8.

Hao Zhao et Guillaume Bourque. Recovering genome rearrangements in the mammalian phylogeny. *Genome Res*, 19(5) :934–942, May 2009. doi: 10.1101/gr.086009.108.
Cité 2 fois pages 15 et 104.

Résumé

Les études biologiques se limitent rarement à l'analyse des données d'une unique espèce, et couvrent en général une période de l'évolution (par comparaison de plusieurs espèces), ce qui permet de replacer les caractéristiques des génomes dans un cadre évolutif. Cela revient, en général implicitement, à décrire le génome d'une (ou plusieurs) espèces ancestrales. Malheureusement, les mêmes raisonnements, avec des données similaires, sont souvent répétés à cause de l'absence d'une base de données qui regrouperait de manière intégrée les connaissances dont on dispose sur les génomes ancestraux.

Ce travail de thèse décrit une nouvelle méthode, appelée AGORA (*Algorithms for Gene Order Reconstruction in Ancestors*), pour reconstruire de manière automatique et systématique l'ordre des gènes et les caryotypes de toutes les espèces ancestrales dans une phylogénie donnée. AGORA est capable de gérer les duplications de gènes, les délétions, et les gains, et interprète de manière réaliste des phylogénies complexes de gènes. Nous avons appliqué la méthode chez 46 espèces de vertébrés séquencées et annotées (en utilisant 8 espèces outgroups supplémentaires) pour reconstruire des ordres de gènes ancestraux dans 43 génomes ancestraux sur près de 600 millions d'années d'évolution. Les performances d'AGORA ont été mesurées par des simulations de génomes de vertébrés, et par confrontation à des génomes ancestraux déjà connus. Les données, présentées graphiquement dans un serveur web nommé Genomicus fournissent une nouvelle ressource pour des études comme l'évolution des gènes, ou les réarrangements dans les génomes.

Mots clés : Génomes ancestraux, génomique comparative, évolution des génomes, algorithmique des graphes

---

Abstract

Biological studies rarely limit to the single-genome-analysis, and often include several species, thus encompassing an entire window of genome evolution (by the comparison of several species), and adding time and evolution as a new dimension to the study. Generally, this includes defining characters of ancestral genomes. With the lack of a wide ancestral genomes database, studies are often performed several times.

Here we describe a new method, named AGORA (*Algorithms for Gene Order Reconstruction in Ancestors*) to automatically and systematically reconstruct gene order and karyotypes in all the ancestral species of a given phylogeny. AGORA can handle different gene content between species (duplications, gains, and loss) by using accurate gene phylogenies as input. We applied AGORA on 46 sequenced & annotated vertebrate genomes (using 8 outgroups genomes) to reconstruct ancestral gene order in 43 ancestral genomes on a 600 million years time-frame. AGORA performances were estimated using simulated datasets, and comparison with other studies. The results can be freely browsed and downloaded from a new web server, Genomicus, dedicated to the study of genome evolution, helping areas such as gene evolution, or genome rearrangements.

Keywords : Ancestral genomes, comparative genomics, genome evolution, graph algorithms