HAL
archives-ouvertes.fr

Évolution de l'architecture des génomes : modélisation et reconstruction phylogénétique

Magali Semeria

▶ **To cite this version:**

Magali Semeria. Évolution de l'architecture des génomes : modélisation et reconstruction phylogénétique. Génétique. Université Claude Bernard - Lyon I, 2015. Français. <NNT : 2015LYO10280>. <tel-01298034>

HAL Id: tel-01298034
https://tel.archives-ouvertes.fr/tel-01298034

Submitted on 5 Apr 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

N° 280 - 2015 Année 2015

THÈSE DE L'UNIVERSITÉ DE LYON

Présentée

devant L'UNIVERSITÉ CLAUDE BERNARD LYON 1

pour l'obtention

du DIPLÔME DE DOCTORAT
(arrêté du 7 août 2006)

soutenue publiquement le 9 décembre 2015

par

Magali SEMERIA

Évolution de l'architecture des génomes : modélisation et reconstruction phylogénétique

Directeurs de thèse : Eric TANNIER et Laurent GUÉGUEN

Jury :	Céline BROCHIER-ARMANET	Présidente du Jury
	Vincent BERRY	Rapporteur
	Laurent GUÉGUEN	Directeur de thèse
	Catherine MATIAS	Rapporteuse
	Yann PONTY	Examinateur
	Eric TANNIER	Directeur de thèse

UNIVERSITE CLAUDE BERNARD - LYON 1

Président de l'Université	**M. François-Noël GILLY**
Vice-président du Conseil d'Administration	M. le Professeur Hamda BEN HADID
Vice-président du Conseil des Études et de la Vie Universitaire	M. le Professeur Philippe LALLE
Vice-président du Conseil Scientifique	M. le Professeur Germain GILLET
Directeur Général des Services	M. Alain HELLEU

COMPOSANTES SANTE

Faculté de Médecine Lyon Est – Claude Bernard	Directeur : M. le Professeur J. ETIENNE
Faculté de Médecine et de Maïeutique Lyon Sud – Charles Mérieux	Directeur : Mme la Professeure C. BURILLON
Faculté d'Odontologie	Directeur : M. le Professeur D. BOURGEOIS
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : Mme la Professeure C. VINCIGUERRA
Institut des Sciences et Techniques de la Réadaptation	Directeur : M. le Professeur Y. MATILLON
Département de formation et Centre de Recherche en Biologie Humaine	Directeur : Mme. la Professeure A-M. SCHOTT

COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur : M. le Professeur F. DE MARCHI
Département Biologie	Directeur : M. le Professeur F. FLEURY
Département Chimie Biochimie	Directeur : Mme Caroline FELIX
Département GEP	Directeur : M. Hassan HAMMOURI
Département Informatique	Directeur : M. le Professeur S. AKKOUCHE
Département Mathématiques	Directeur : M. le Professeur Georges TOMANOV
Département Mécanique	Directeur : M. le Professeur H. BEN HADID
Département Physique	Directeur : M. Jean-Claude PLENET
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur : M. Y.VANPOULLE
Observatoire des Sciences de l'Univers de Lyon	Directeur : M. B. GUIDERDONI
Polytech Lyon	Directeur : M. P. FOURNIER
École Supérieure de Chimie Physique Électronique	Directeur : M. G. PIGNAULT
Institut Universitaire de Technologie de Lyon 1	Directeur : M. C. VITON
Institut Universitaire de Formation des Maîtres	Directeur : M. A. MOUGNIOTTE
Institut de Science Financière et d'Assurances	Directeur : M. N. LEBOISNE

Résumé

L'évolution des génomes peut être observée à plusieurs échelles, chaque échelle révélant des processus évolutifs différents. À l'échelle de séquences ADN, il s'agit d'insertions, délétions et substitutions de nucléotides. Si l'on s'intéresse aux gènes composant les génomes, il s'agit de duplications, pertes et transferts horizontaux de gènes. Et à plus large échelle, on observe des réarrangements chromosomiques modifiant l'agencement des gènes sur les chromosomes. Reconstruire l'histoire évolutive des génomes implique donc de comprendre et de modéliser tous les processus à l'œuvre, ce qui reste hors de notre portée. À la place, les efforts de modélisation ont exploré deux directions principales. D'un côté, les méthodes de reconstruction phylogénétique se sont concentrées sur l'évolution des séquences, certaines intégrant l'évolution des familles de gènes. D'un autre côté, les réarrangements chromosomiques ont été très largement étudiés, donnant naissance à de nombreux modèles d'évolution de l'architecture des génomes. Ces deux voies de modélisation se sont rarement rencontrées jusqu'à récemment. Au cours de ma thèse, j'ai développé un modèle d'évolution de l'architecture des génomes prenant en compte l'évolution des gènes et des séquences. Ce modèle rend possible une reconstruction probabiliste de l'histoire évolutive d'adjacences et de l'ordre des gènes de génomes ancestraux en tenant compte à la fois d'événements modifiant le contenu en gènes des génomes (duplications et pertes de gènes), et d'événements modifiant l'architecture des génomes (les réarrangements chromosomiques). Intégrer l'information phylogénétique à la reconstruction d'ordres des gènes permet de reconstruire des histoires évolutives plus complètes. Inversement, la reconstruction d'ordres des gènes ancestraux peut aussi apporter une information complémentaire à la phylogénie et peut être utilisée comme un critère pour évaluer la qualité d'arbres de gènes, ouvrant la voie à un modèle et une reconstruction intégrative.

Abstract

Genomes evolve through processes that modify their content and organization at different scales, ranging from the substitution, insertion or deletion of a single nucleotide to the duplication, loss or transfer of a gene and to large scale chromosomal rearrangements. Extant genomes are the result of a combination of many such processes, which makes it difficult to reconstruct the overall picture of genome evolution. As a result, most models and methods focus on one scale and use only one kind of data, such as gene orders or sequence alignments. Most phylogenetic reconstruction methods focus on the evolution of sequences. Recently, some of these methods have been extended to integrate gene family evolution. Chromosomal rearrangements have also been extensively studied, leading to the development of many models for the evolution of the architecture of genomes. These two ways to model genome evolution have not exchanged much so far, mainly because of computational issues. In this thesis, I present a new model of evolution for the architecture of genomes that accounts for the evolution of gene families. With this model, one can reconstruct the evolutionary history of gene adjacencies and gene order accounting for events that modify the gene content of genomes (duplications and losses of genes) and for events that modify the architecture of genomes (chromosomal rearrangements). Integrating these two types of information in a single model yields more accurate evolutionary histories. Moreover, we show that reconstructing ancestral gene orders can provide feedback on the quality of gene trees thus paving the way for an integrative model and reconstruction method.

Remerciements

Je tiens tout d'abord à remercier Éric et Laurent d'avoir su si bien encadrer cette thèse. Je garde d'excellents souvenirs de nos échanges pendant ces trois années de doctorat : discussions scientifiques stimulantes, promptes relectures et réécritures, gribouillages au tableau, debug-parties, et une petite promenade de santé du côté du Hameau de l'étoile.

Je tiens à remercier Catherine Matias et Vincent Berry pour leur relecture attentive de ce manuscrit. Merci aussi à Céline Brochier-Armanet et Yann Ponty d'avoir accepté de faire partie du jury, ainsi qu'à Sèverine Bérard, Alessandra Carbone, Hugues Roest Crollius et Tristan Lefebure qui ont suivi mon travail.

Pendant ces trois années de doctorat, j'ai eu la chance d'interagir avec de nombreuses personnes au LBBE. Je tiens tout particulièrement à remercier mes collègues au sein de l'équipe BGE et au PRABI pour leur accueil chaleureux et pour leur présence tout au long de ma thèse.

Un grand merci à tous ceux qui ont animé la salle de pause de leurs discussions sérieuses ou, plus souvent, farfelues et de leurs grands éclats de rire. Un merci tout particulier à Marie, Fanny, Héloïse et Aline pour leur présence et leur complicité. Je n'oublie pas non plus Michel et sa réserve de bonbons, Murray et les Bottlenecks, l'incontournable Thomas, Rémi et son pantalon bleu, Dominique et ses grut, Cécile et sa recette de macarons, Guillaume et son sosie, Wandrille et ses sandales, Frédéric, Ghislain, et Jos. Merci à mes grands-frères et sœur de thèse Florent, Yann, Mathieu, Nicolas et Eugénie. Merci à Clément et Christophe pour la motivation du vendredi après-midi et à Laurent J. pour m'avoir fait découvrir Bastien Vivès en pleine rédaction.

Merci à Christine, Philippe, Amandine, Jean-François et Philippe avec qui j'ai eu le plaisir de partager un bureau. Nos discussions, alimentées par les boîtes de biscuits du PRABI vont me manquer.

Merci à Stéphane, Lionel, Simon et Bruno pour leur aide sur le cluster, les conseils en informatique, en culture alternative et en décoration intérieure.

Merci à Nathalie, Odile, Laetitia et Aline pour leur aide sur le plan administratif.

Je dois enfin remercier quelques personnes en dehors du laboratoire qui ont suivi ma thèse de près. Merci donc à Benjamin P. de m'avoir amenée sur les pentes, à Baptiste de m'avoir nourrie pendant que je rédigeais, à Benjamin A. d'avoir vainement tenté de

m’initier au foot, et à Pierre qui, après avoir vu à quoi ressemble la fin d’une thèse, a bravement décidé d’en commencer une.

Un grand merci à ma famille pour leur soutien et pour la confiance aveugle qu’ils placent en moi, et à Nicolas qui prend soin de moi depuis le début.

À Jean-Denis et Bettina Semeria.

Table des matières

1

Introduction : Modéliser l'évolution

1.1 Les arbres

FIGURE 1.1 – Arbre généalogique de Charlemagne. Tiré des Chroniques de Nuremberg de Hartmann Schedel, 1493.

Les arbres, ou plus généralement les structures arborescentes, sont une des manières les plus populaires de représenter et d'organiser l'information. On trouve ainsi

des arbres dans de nombreux domaines, de l'arbre généalogique à l'arborescence des fichiers et répertoires sur un ordinateur. En biologie, l'arbre émerge au dix-neuvième siècle comme le moyen de représenter les relations entres les espèces. Plusieurs biologistes, notamment Lamarck, utilisent ainsi l'arbre pour représenter la classification des espèces. Le travail de Darwin marque un tournant car il utilise l'arbre non pas comme représentation hiérarchique statique mais comme la représentation d'un processus évolutif ayant engendré une multitude d'espèces à partir d'un unique ancêtre commun (Figure 1.2).

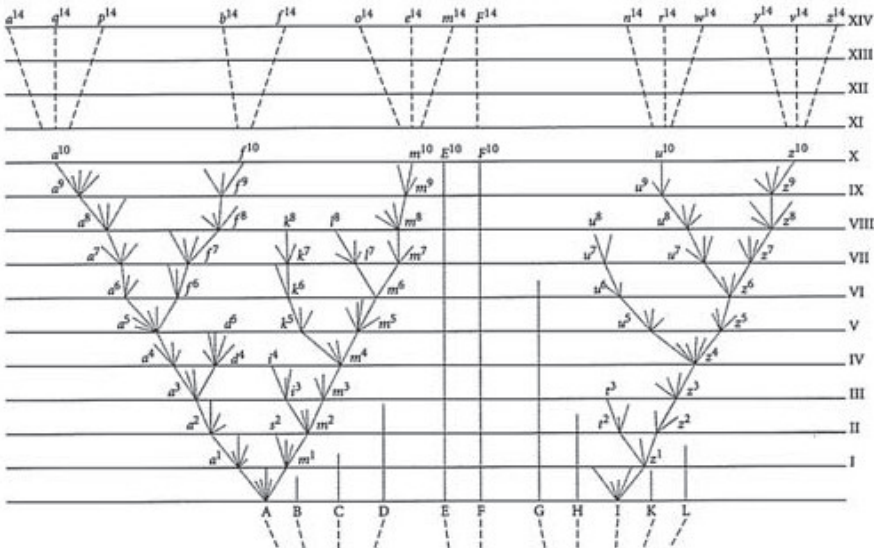


FIGURE 1.2 – Processus de divergence et d'extinction des lignées, tel qu'illustré par Darwin dans *De l'origine des espèces* [36] .

Le processus évolutif décrit par Darwin a depuis été amendé et complété mais sa représentation est toujours utilisée. On appelle **phylogénie** la discipline qui vise à établir les relations de parenté entre différentes entités et **arbres phylogénétiques** les arbres utilisés pour représenter de telles relations, ainsi que pour modéliser le processus évolutif sous-jacent. Pendant longtemps, les arbres phylogénétiques ont été reconstruits presque exclusivement à partir de caractères morphologiques. Cette approche est intuitive mais elle présente plusieurs inconvénients. En premier lieu, les espèces et les caractères observables sont relativement limités. Certaines espèces sont difficiles à observer à cause de leur faible taille, ou de leur habitat. De même, certains caractères

morphologiques sont difficiles à mesurer. Un deuxième obstacle est que certains caractères ont pu apparaître indépendamment chez plusieurs espèces : c'est la convergence évolutive. Un des exemples les plus connus de convergence évolutive est l'acquisition des ailes de manière indépendante pour les oiseaux et pour la chauve-souris. Si on reconstruisait un arbre phylogénétique basé sur ce caractère, on arriverait à la conclusion que la chauve-souris est plus apparentée aux oiseaux qu'aux mammifères.

La mise à disposition de données moléculaires, et notamment de séquences nucléotidiques et protéiques, a permis de s'affranchir de certaines de ces contraintes. Les arbres phylogénétiques basés sur l'information moléculaire font toutefois appel à des méthodes et des modèles plus complexes que ceux basés sur des caractères morphologiques et font face à d'autres types de défis. La phylogénie moléculaire a permis d'étudier les relations de parenté entre les êtres vivants à une résolution infiniment plus fine que la phylogénie traditionnelle. Il est en effet devenu possible de reconstruire des arbres phylogénétiques à partir de séquences correspondant à un seul gène ou une seule protéine. Ce changement de résolution rend nécessaire d'avoir le vocabulaire pour distinguer plusieurs types d'arbres phylogénétiques. On appelle ainsi **arbre d'espèces** un arbre qui décrit les relations de parenté au niveaux des espèces (actuelles ou fossiles) et **arbre de gènes** un arbre qui décrit l'histoire évolutive à l'échelle d'un gène.

Une hypothèse centrale de la phylogénie moléculaire est que l'histoire évolutive des espèces peut se lire dans l'histoire évolutive des séquences constituant les génomes de ces espèces. Cependant, on peut observer des différences importantes entre un arbre d'espèces et un arbre de gènes car l'histoire évolutive des gènes inclut des événements qui ne sont pas forcément observables à l'échelle des espèces, et qui peuvent différer de l'histoire des espèces. Reconnaître ces événements et expliquer les différences entre un arbre de gènes et un arbre d'espèce est donc un enjeu important du domaine.

La phylogénie moléculaire a impliqué des développements méthodologiques considérables depuis les années 1960. Dans les sections suivantes nous aborderons ainsi brièvement les modèles d'évolutions, les algorithmes de reconstruction d'arbres phylogénétiques, et les mesures et tests statistiques permettant d'évaluer la qualité d'un arbre. Une revue complète des méthodes utilisées en phylogénie moléculaire peut être trouvée dans deux livres en particulier : *Inferring Phylogenies* de Joseph Felsenstein [49] et *Concepts et méthodes en phylogénie moléculaire* de Guy Perrière et Céline Brochier-Armanet [109].

L'évolution des séquences par insertions, délétions et substitutions de bases n'est cependant qu'une facette de l'évolution des génomes. Dès 1921, Sturtevant [129] publie une étude démontrant l'existence d'un autre mécanisme évolutif : les réarrangements chromosomiques. En première approximation, imaginons le génome comme une phrase, soumise à un processus évolutif. L'évolution de cette phrase peut être observée à plusieurs résolutions différentes. A l'échelle d'un mot, on observe l'insertion de nouvelles lettres parmi les anciennes, la suppression de certaines lettres, et le remplacement d'autres lettres. Mais si on se place à l'échelle de la phrase toute entière, on constate que l'ordre des mots dans la phrase peut aussi être modifié au cours du temps. La modification de l'ordre des mots affecte aussi la séquence globale des lettres de la phrase, mais pour bien comprendre le processus à l'œuvre, cette échelle d'observation ne suffit pas. Les réarrangements chromosomiques sont des événements évolutifs qui modifient l'agencement des gènes sur les chromosomes. On peut alors parler d'évolution de l'architecture des génomes, ce qui constitue une autre facette de l'évolution des génomes.

Il est possible d'étudier conjointement l'évolution de l'architecture des génomes et l'évolution des gènes et des espèces. Il faut alors articuler l'évolution de l'architecture avec l'évolution des gènes et l'évolution des génomes, et donc trouver pour l'architecture un objet à part sur lequel on observe l'évolution. Après avoir exploré différents aspects des arbres d'espèces et des arbres de gènes, je montrerai ainsi comment formaliser le problème de l'évolution de l'architecture des génomes pour utiliser une représentation sous forme d'arbre. On appellera de tels arbres des **arbres d'adjacences**. Ces trois types d'arbres, et les liens entre les différentes échelles d'évolution qu'il représentent, sont au cœur de ma thèse. Je vais donc détailler leurs caractéristiques, ce qu'ils ont en commun et ce qui les différencie.

1.1.1 Arbre phylogénétique

L'arbre phylogénétique est un objet mathématique. Il est utile d'introduire dès maintenant le vocabulaire associé à cet objet. L'arbre phylogénétique schématisé sur la Figure 1.3 peut ainsi être formellement décrit comme un graphe non orienté (les arêtes n'ont pas de direction), acyclique (il ne comporte pas de cycle) et connexe (il existe un chemin composé d'une suite d'arêtes permettant de relier n'importe quelle paire de

sommets). Les sommets du graphe représentent les entités étudiées qui peuvent être des espèces, des gènes, des adjacences, ou tout autre objet ou concept transmis avec modification au cours des générations. On distingue les *feuilles*, représentant les entités étudiées des autres sommets représentant des entités ancestrales hypothétiques. Les arêtes du graphe, ou *branches* de l'arbre représentent les liens de parenté entre les entités. La longueur d'une branche est un nombre réel associé à l'arête correspondante. Elle représente en général une quantité d'évolution séparant deux sommets. Un arbre peut être raciné par un de ses sommets internes, la *racine* (E sur l'exemple de la figure 1.3). Ceci donne une orientation implicite à toutes les arêtes (depuis la racine jusqu'aux feuilles). Dans ce cas on peut définir une relation d'ordre partiel entre les sommets : un sommet A est plus récent qu'un sommet B si B est sur le chemin entre A et la racine. Dans ce cas A est un descendant de B, et B un ancêtre de A. On peut identifier des ancêtres communs : un sommet D est un ancêtre commun des sommets A et B s'il est sur le chemin entre la racine et les deux sommets A et B. En biologie évolutive tous les arbres sont supposés racinés. Cependant il est fréquent que la position de la racine soit inconnue.

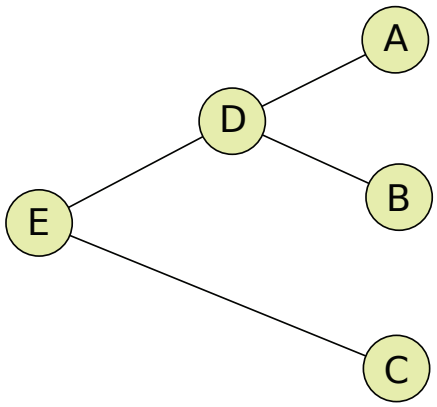


FIGURE 1.3 – Dans cet arbre phylogénétique, A, B et C sont les entités étudiées, D est l'ancêtre commun hypothétique de A et B, et E l'ancêtre commun hypothétique de A, B et C. On dit que A, B et C sont les feuilles de l'arbre et E la racine.

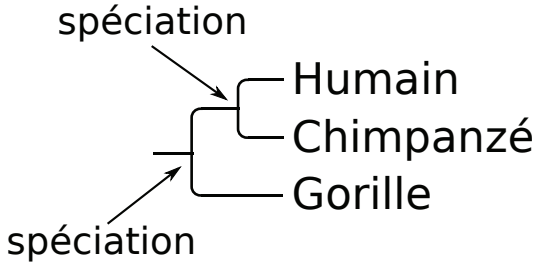


FIGURE 1.4 – Arbre d'espèces représentant les relations de parenté entre l'humain, le chimpanzé et le gorille.

1.1.2 Arbres d'espèces

Les arbres d'espèces représentent l'évolution d'une lignée se divisant pour donner naissance à des lignées différentes jusqu'à obtenir les espèces actuelles (ou fossiles). On appelle spéciation l'événement évolutif décrivant la séparation d'une lignée en deux lignées distinctes (Figure 1.4). Derrière cette description formelle se cache une des questions les plus fascinantes de la biologie évolutive : comment une nouvelle espèce peut-elle apparaître ? Cette question est particulièrement ardue car elle se situe à l'interface de la génétique des populations et de la phylogénie. Dans une récente revue, Sergey Gavrilets utilise la métaphore suivante pour décrire la spéciation [55] : si une population est un nuage de points dans lequel chaque point représente un individu, la spéciation peut être vue comme l'agrégation des points en deux sous-groupes accompagnée de l'apparition d'un isolement reproductif entre les deux sous-groupes. Les mécanismes à l'origine de la séparation et de l'isolement reproductif de deux sous-populations sont loin d'être complètement élucidés. Il est clair que des forces évolutives telles que la sélection naturelle, la dérive génétique et les mutations jouent un grand rôle dans le processus mais il est difficile de quantifier leurs contributions respectives, d'autant qu'il existe de multiples scénarios pouvant aboutir à une spéciation. Le scénario le plus intuitif est peut-être celui où l'isolement géographique d'une partie de la population, associé à une part de dérive génétique et/ou des forces de sélection différentes conduit à l'accumulation de mutations qui rendent à terme impossible la reproduction entre les sous-populations. Il existe cependant des scénarios alternatifs tels que la sélection contre les hétérozygotes ou la différentiation de niche écologique [28].

1.1.3 Arbres de gènes

Un arbre de gènes représente la transmission d'un gène ancestral à travers différentes lignées jusqu'aux espèces étudiées. L'histoire de cette transmission peut être marquée par plusieurs types d'événements évolutifs. Lors d'un événement de spéciation, le gène peut être transmis aux nouvelles lignées, sa séquence pouvant ensuite accumuler des mutations indépendantes dans les lignées ayant divergé. Le gène peut aussi être perdu dans une lignée au cours de son histoire évolutive, ou se dupliquer (Figure 1.5). Dans ce dernier cas le gène existe en deux copies dans la lignée affectée par la duplication. Les duplications jouent un rôle majeur dans la diversification des fonctions codées par

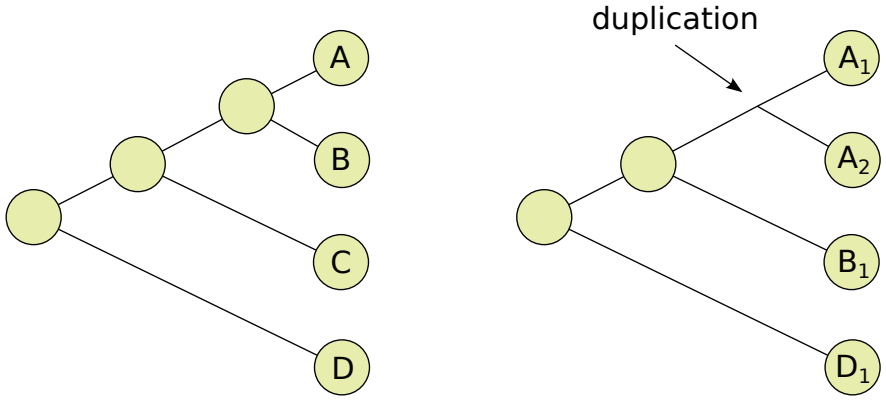


FIGURE 1.5 – Arbre d'espèce (à gauche) et arbre de gènes (à droite). Une duplication conduit à la présence de deux copies du gène : A_1 et A_2. Le gène est perdu dans l'espèce C.

les gènes [105, 68, 14]. Plusieurs scénarios ont été proposés pour expliquer le maintien de gènes dupliqués au cours de l'évolution [65, 69]. Les principaux sont :

- la néo-fonctionnalisation : la duplication du gène entraîne une redondance de la fonction portée par le gène. La pression de sélection est alors relâchée pour une des copies qui va accumuler les mutations. Ces mutations, associées à une sélection positive, peuvent faire émerger une nouvelle fonction [105].

- la sous-fonctionnalisation : les deux copies se spécialisent (avec une part de dérive génétique et une part de sélection). La fonction initiale est divisée entre les deux copies [51, 90].

- l'augmentation de dosage avantageuse [73] : les deux copies conserve une fonction identiques mais leurs niveaux d'expression s'additionnent. Si cette augmentation est bénéfique elle est sélectionnée.

Une des copies du gène dupliqué peut aussi être perdue dans une ou plusieurs lignées après la duplication suite à l'accumulation de mutations délétères. Ce scénario est à l'origine de l'apparition de pseudogènes [142] : des séquences de gènes dégénérées ne codant plus pour des protéines. Un autre scénario possible est la perte d'une partie ou de toute la séquence codant pour une copie du gène suite à une erreur de recopie du chromosome. Il s'agit alors d'un réarrangement chromosomique (voir section 1.1.7.1).

Pendant longtemps on a considéré que la transmission de l'information génétique s'effectuait de manière verticale : chaque génération héritant son matériel génétique de la précédente. Ce mode de transmission justifie d'ailleurs la représentation de l'évolution sous la forme d'un arbre. Mais chez les procaryotes en particulier, l'information génétique peut aussi être transmise horizontalement, c'est à dire entre des individus appartenant à différentes espèces contemporaines. Cet événement évolutif est appelé transfert horizontal de gène, en opposition avec la transmission verticale, héréditaire, des gènes. Les mécanismes cellulaires des transferts horizontaux ont été découverts dans la première moitié du vingtième siècle. On distingue trois mécanismes possibles [136] :

- La transformation [60] : un organisme intègre de l'ADN présent dans l'environnement.
- La transduction [143] : l'échange de matériel génétique entre deux organismes infectés successivement par un bactériophage.
- La conjugaison [78] : l'échange de matériel génétique entre deux organismes via un plasmide.

Les transferts jouent un rôle important dans l'adaptation des organismes à un environnement car ils permettent de "récupérer" des fonctions développées par d'autres espèces. Depuis les années 1950, on sait ainsi que les transferts horizontaux sont fortement impliqués dans l'acquisition de la résistance aux antibiotiques chez les bactéries [39]. On reconnaît aujourd'hui que les transferts horizontaux constituent une force évolutive majeure chez les procaryotes [104].

1.1.4 Étapes de la reconstruction d'un arbre d'espèce

Dans les chapitres 2 et 3, il sera question de reconstruction d'arbres en intégrant diverses informations. Il est donc utile de poser ici les bases d'une reconstruction phylogénétique. La méthode présentée ici est couramment utilisée pour reconstruire un arbre d'espèce à partir de données moléculaires mais il existe beaucoup d'autres méthodes.

La reconstruction d'un arbre d'espèce à partir de séquences nucléotidiques ou protéiques comporte plusieurs étapes, le plus souvent réalisées de manière séquentielle [22, 40] (Figure 1.6).

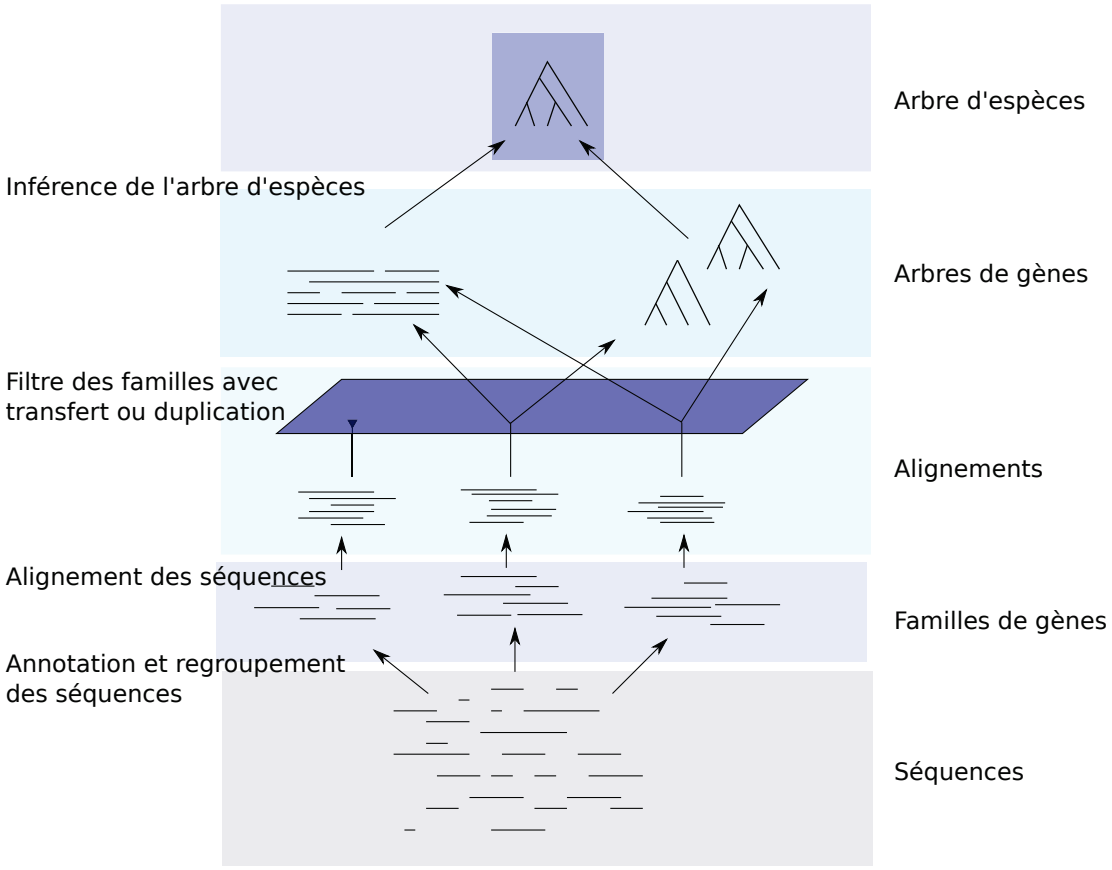


FIGURE 1.6 – Étapes de la reconstruction d'un arbre d'espèces à partir de séquences. Adapté de [22].

1.1.4.1 Choix des séquences et regroupement des gènes en familles d'homologues

La première étape consiste à rechercher des séquences ayant une origine évolutive commune. Ces séquences sont dites homologues. L'homologie est une notion importante en phylogénie moléculaire. Elle se décline en trois sous-types :

- les séquences orthologues sont des séquences dont l'histoire évolutive commune peut être tracée jusqu'à un événement de spéciation.
- les séquences paralogues sont issues d'un événement de duplication.
- les séquences xénologues sont des séquences dont l'histoire évolutive commune remonte à un événement de transfert horizontal.

Il est indispensable que les séquences soient homologues pour reconstruire un arbre. En effet, si les séquences sélectionnées ne partagent pas d'histoire évolutive commune, tenter de reconstruire une histoire commune serait absurde. Mais dire que des séquences sont homologues suppose de déjà connaître leur histoire évolutive, ce qui est l'objectif de la reconstruction. Il faut donc utiliser une autre approche pour estimer si les séquences sont homologues. On fait alors l'hypothèse que les liens de parentés entre les séquences peuvent être estimés à partir du degré de similarité entre les séquences : plus les séquences sont similaires plus leur ancêtre commun est récent. De même, si deux séquences sont plus similaires qu'attendu au hasard, alors elles sont probablement homologues.

La recherche de séquences homologues peut alors être effectuée avec un algorithme de recherche de similarité entre les séquences, tel qu'implémenté dans BLAST [6]. Une autre question liée au choix des séquences est de déterminer, lorsqu'on dispose des deux options, s'il est préférable d'utiliser des séquences nucléotidiques ou des séquences protéiques. La différence porte sur les vitesses d'évolution des séquences. À cause de la redondance du code génétique (plusieurs codons codent pour un même acide aminé), certaines mutations au niveau des nucléotides n'affectent pas la séquence protéique : ce sont des mutations silencieuses. Ces mutations sont soumises à une pression de sélection plus faible que les mutations entraînant un changement d'acide aminé car elles n'ont pas de conséquence sur la protéine. Les séquences protéiques sont donc

plus stables dans le temps. Cette propriété est importante car lorsque des séquences subissent trop de mutations après avoir divergé, elles deviennent trop différentes pour pouvoir être comparées. On appelle ce phénomène la saturation du signal phylogénétique. Le choix des séquences à utiliser dépend alors de la fenêtre temporelle étudiée. Si l'on cherche à reconstruire une histoire évolutive très ancienne, il sera sans doute préférable d'utiliser des séquences protéiques. À l'inverse, pour une histoire évolutive plus récente, les séquences nucléotidiques peuvent être plus informatives car les séquences protéiques n'auront pas forcément eut le temps d'accumuler suffisamment de mutations pour qu'on puisse les comparer. Notons aussi que les vitesses d'évolution des séquences peuvent varier en fonction des gènes. L'ADN ribosomique a ainsi longtemps été le marqueur privilégié pour étudier des relations phylogénétiques anciennes (cf l'arbre du vivant) car il est remarquablement bien conservé à travers les lignées. L'ADN viral, qui évolue lui très rapidement, permet d'étudier des histoires évolutives très récentes. Dans un récent article, Gire et al. reconstituent ainsi la propagation du virus Ebola lors de la dernière épidémie à partir de l'ADN viral prélevé chez des malades [56].

1.1.4.2 Alignement des séquences

Les séquences sélectionnées n'ont pas forcément la même longueur (i.e. le même nombre de nucléotides ou d'acides aminés). Au niveau nucléotidique, les différences de longueurs sont dues aux insertions et délétions de nucléotides. Ces événements peuvent aussi entraîner des pertes et insertions d'acides aminés, visibles dans les séquences protéiques. Soit une séquence ancestrale de longueur n. On appelle *site* la position d'un nucléotide ou d'un acide aminé dans cette séquence. Au cours de l'évolution, des mutations affectent les sites de la séquence ancestrale de manière différente selon les lignées. Lors de substitutions, les sites sont conservés, c'est l'état du caractère associé au site qui change. Lors d'une insertion un nouveau site est crée. Lors d'une délétion, le site est perdu. L'alignement des séquences vise à établir la correspondance entre les sites des séquences sélectionnées et les sites de leur ancêtre commun. Formulé de manière plus rigoureuse, l'alignement vise à identifier les sites homologues, c'est-à-dire les sites descendant d'un même site de la séquence ancestrale. Les résidus qui n'ont pas d'homologues et les résidus perdus sont matérialisés par des emplacement vides, les *gaps*,

dans les séquences concernées. L'objectif des algorithmes d'alignement est d'optimiser le placement des gaps dans les séquences pour expliquer de la meilleure manière possible les différences observées entre les séquences.

1.1.4.3 Reconstruction des arbres de gènes

Une fois les séquences alignées, on peut reconstruire un arbre de gènes correspondant à l'histoire évolutive des séquences. Il existe quatre types d'approches principaux pour reconstruire une topologie :

- Le maximum de parcimonie
- Les méthodes de distance
- Le maximum de vraisemblance
- Les méthodes bayésiennes

Les modèles d'évolution et algorithmes de reconstruction par maximum de vraisemblance sont abordés plus en détail dans la partie 1.2.

1.1.4.4 Ignorer les incongruences

Depuis Zuckerkandl et Pauling [144], l'hypothèse de base de la phylogénie moléculaire est que l'histoire évolutive des espèces peut être lue dans les séquences. Cependant, l'histoire des gènes est marquée par des événements évolutifs, tels que les transferts horizontaux, les duplications, les pertes de gènes, et le tri de lignées incomplet qui ne reflètent pas des relations de parenté entre les espèces. À cause de ces événements, la topologie d'un arbre de gène ne correspond pas toujours à la phylogénie des espèces [94]. On dit alors que l'arbre de gène est incongruent avec l'arbre des espèces. Lorsqu'on reconstruit un arbre d'espèces à partir de différentes familles de gènes, le problème est alors de déterminer comment reconstruire une topologie unique à partir d'arbres de gènes discordants. L'approche la plus courante consiste à simplement éliminer de la reconstruction les familles suspectées de contenir des gènes dupliqués, ou transférés horizontalement [22]. On peut alors reconstruire l'arbre d'espèce en amalgamant les arbres de gènes (c'est l'approche *supertree* [16, 117]) ou les alignements (l'approche

supermatrix [41]) relativement consensuels. Cette approche est contestée car elle implique d'ignorer une partie du signal phylogénétique. Dans un article de 2010, Boussau et Daubin passent en revue des approches alternatives permettant de reconstruire la topologie en utilisant toute l'information disponible [22]. Ces approches impliquent de prendre en compte les événements évolutifs spécifiques à chaque famille de gènes pour expliquer les incongruences de topologies.

1.1.5 Incongruences entre les arbres de gènes et l'arbre des espèces

Les topologies incongruentes montrent qu'il existe plusieurs histoires évolutives différentes selon l'échelle d'observation qu'on choisit. À l'échelle des espèces, on observe la transmission et la divergence des lignées donnant naissance à de nouvelles espèces : c'est l'histoire des spéciations. À l'échelle du gène, on observe l'histoire des spéciations, des duplications, des pertes, et des transferts horizontaux. Et à l'échelle de la population, on observe la transmission de plusieurs allèles correspondant à un même gène. La transmission des allèles est parfois visible à l'échelle du gène, c'est le *tri de lignées incomplet*. Ces histoires différentes sont imbriquées les unes dans les autres. Dans un article de 1997, Maddison propose ainsi de représenter les arbres de gènes à l'intérieur des arbres d'espèces [94], comme illustré sur la figure 1.7 . D'après Maddison, l'arbre des espèces est en fait composé d'une distribution d'arbres de gènes, au sein de laquelle on peut distinguer un signal phylogénétique correspondant à la phylogénie des espèces ; la variance de cette distribution traduisant la diversité des histoires évolutives des familles de gènes.

Les incongruences entre les arbres d'espèces et les arbres de gènes montrent les défis liés à l'intégration de plusieurs échelles dans la reconstruction de l'histoire évolutive des génomes. Confronter des topologies d'arbres de gènes et d'arbre des espèces et chercher à expliquer les incongruences entre ces topologies permet d'un autre côté de mieux comprendre les histoires évolutives représentées. Dans la partie 1.1.8 et le chapitre 2, nous verrons comment articuler l'histoire évolutive des gènes avec l'histoire évolutive de l'architecture des génomes. Dans le chapitre 3, nous verrons que l'intégration de ces deux échelles d'évolution apporte un regard critique sur la reconstruction des arbres de gènes.

Dans cette partie, nous allons voir comment des duplications, pertes et transferts

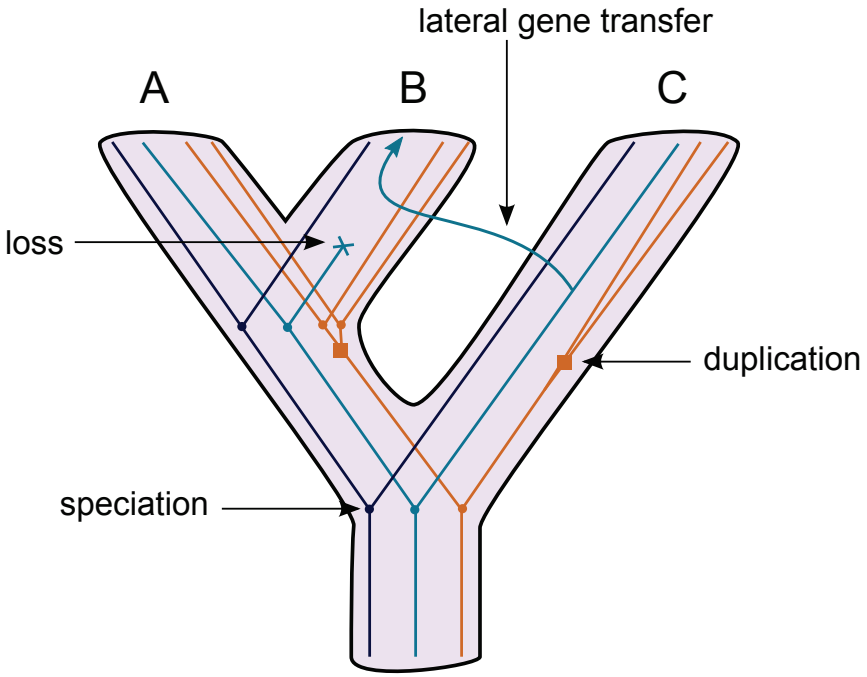


FIGURE 1.7 – Trois arbres de gènes (en trait fin, bleu foncé, bleu claire et orange) sont représentés à l'intérieur d'un arbre d'espèces. Chaque famille de gènes subit des événements qui lui sont propres : duplications pour la famille orange et transfert et perte pour la famille bleu clair. La famille bleu foncé suit l'arbre des espèces. Ces événements peuvent conduire à des différences de topologies entre les arbres de gènes et l'arbre des espèces.

horizontaux peuvent être à l'origine d'incongruences. Cependant, les incongruences ne sont pas forcément liées à des phénomènes biologiques. Dans certains cas, elles reflètent plutôt des artefacts de reconstruction des arbres. Expliquer une incongruence par une duplication, un transfert ou du tri de lignées incomplet suppose alors que les arbres de gènes décrivent bien l'histoire des gènes, ce qui est loin d'être toujours acquis (voir section 1.1.4 et chapitre 3).

1.1.5.1 Les duplications et les pertes de gènes

Après une duplication ancestrale, certaines lignées peuvent perdre une des deux copies du gène. Si aucune lignée ne conserve les deux copies, il est nécessaire de disposer de l'arbre d'espèces pour deviner l'existence de la duplication. On appelle ce phénomène

la *paralogie cachée*. Son effet possible sur la topologie d'un arbre de gène est illustré sur la figure 1.8.

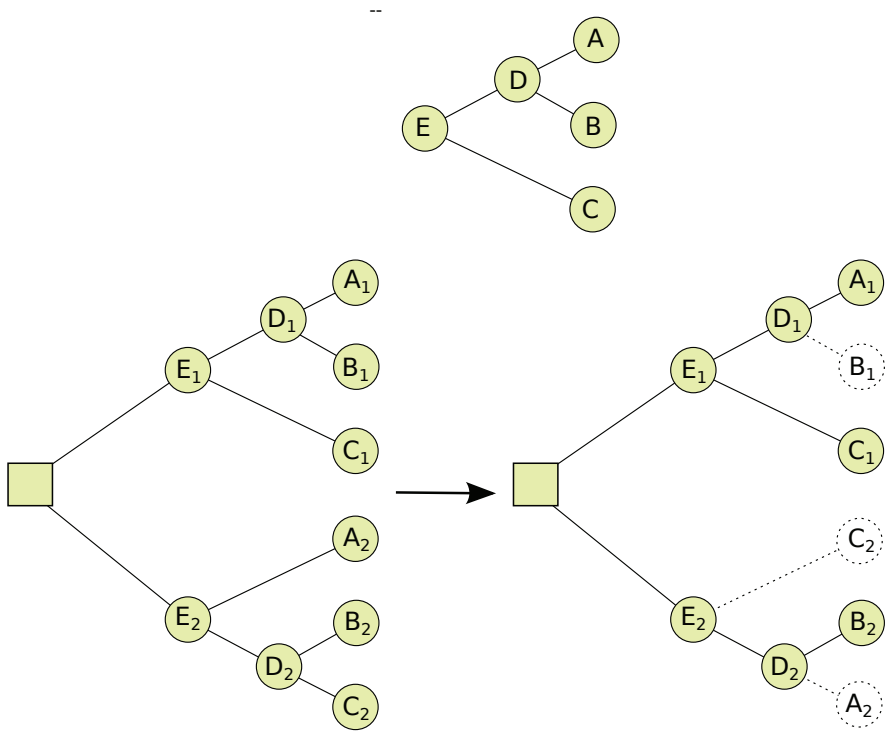


FIGURE 1.8 – En haut, l'arbre d'espèces. En bas à gauche, l'arbre de gènes avant les pertes. En bas à droite l'arbre de gènes après les pertes. Après la duplication dans l'espèce E, une copie du gène est perdue dans la lignée conduisant à l'espèce B, tandis que l'autre copie est perdue dans les lignées conduisant à A et C. La topologie de l'arbre de gène ne correspond alors pas à celle de l'arbre d'espèces.

1.1.5.2 Les transferts horizontaux

Les transferts horizontaux augmentent le degré de similarité des séquences entre les lignées concernées par le transfert. La similarité des séquences ne signifie alors pas que les deux espèces ont un ancêtre commun particulièrement récent, mais simplement qu'il y a eu un échange de matériel génétique plus récent que leur spéciation. Dans un arbre de gène, les gènes impliqués dans un transfert seront ainsi plus proches qu'attendu étant donné une phylogénie des espèces (Figure 1.9).

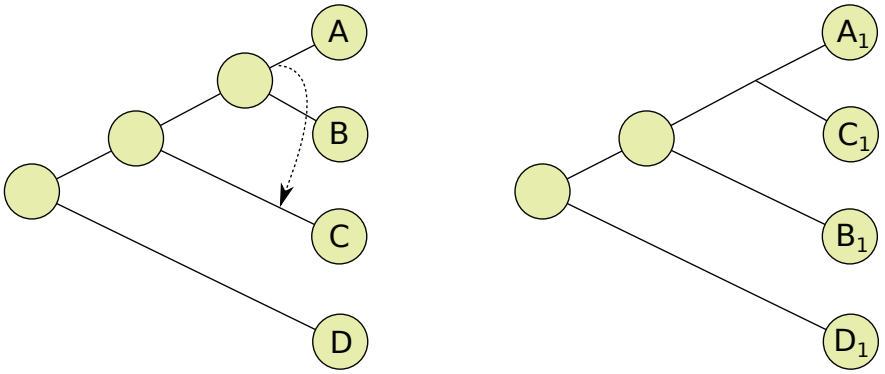


FIGURE 1.9 – Arbre d'espèce (à gauche) et arbre de gènes (à droite). Effet d'un transfert horizontal sur la topologie d'un arbre de gène. Le transfert horizontal de la lignée conduisant à l'espèce A vers la lignée conduisant à l'espèce C. La séquence de A_1 est alors plus proche de celle de C_1 que de celle de B_1, ce qui conduit à la topologie $((A_1, C_1), B_1)$ pour l'arbre de gènes.

1.1.5.3 Le tri de lignées incomplet

Un dernier phénomène à l'origine de topologies incongruentes est le tri de lignées incomplet. Pour comprendre ce phénomène, il faut rappeler que l'évolution des espèces implique des processus évolutifs œuvrant au niveau des populations. Dans le cas du tri de lignées incomplet, un polymorphisme ancestral est transmis aux lignées descendantes puis éliminé de manière indépendante dans chaque lignée. Des lignées proches peuvent ainsi éliminer un allèle différend, tandis que des lignées plus éloignées conserveront le même allèle. En reconstruisant l'arbre de ces séquences, on regroupera donc les lignées ayant conservé le même allèle, ce qui ne correspond pas forcément à la phylogénie des espèces (Figure 1.10). Le tri de lignées incomplet affecte surtout les lignées caractérisées par un nombre de générations limité et des tailles de populations importantes [94]. Une récente étude estime ainsi que 30% des gènes humains sont plus proches des gènes du gorille que de ceux du chimpanzé à cause de ce phénomène [120].

1.1.5.4 Réconciliation

À cause des duplications, des pertes, des transferts horizontaux et du tri de lignées incomplet, la topologie des arbres de gènes ne correspond pas toujours à la phylogénie

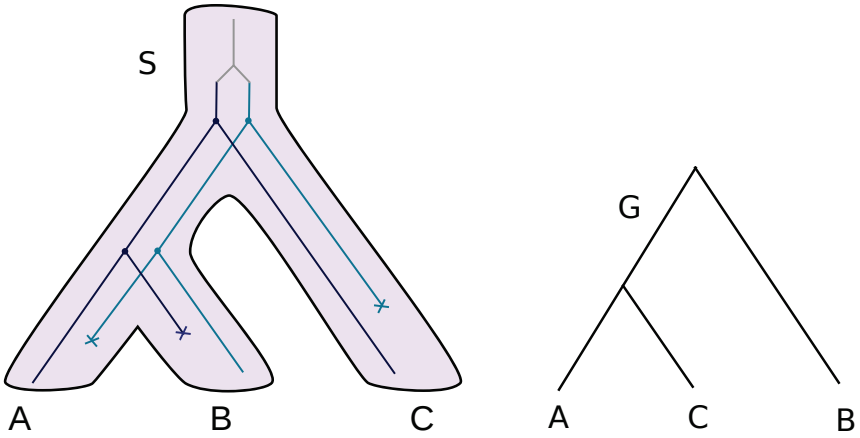


FIGURE 1.10 – Effet du tri de lignées incomplet sur la topologie d'un arbre de gènes. À gauche, l'arbre des espèces à l'intérieur duquel sont représentées en bleu clair et en bleu foncé les histoires évolutives de deux allèles. Les deux allèles co-existent dans la population de l'espèce ancestrale, puis sont fixés de manière différentielle dans les populations des espèces A, B et C. L'allèle bleu clair est ainsi perdu dans A et C et l'allèle bleu foncé dans B. Ce phénomène, appelé tri des lignées incomplet, peut conduire à une topologie de l'arbre de gènes G qui ne correspond pas à l'arbre d'espèces. Ici, les espèces A et C ont ainsi hérité de l'allèle bleu foncé alors que l'espèce C a hérité de l'allèle bleu clair. Lorsqu'on reconstruit l'arbre de gènes (à droite), on ne distingue pas les différents allèles du gène. Ici, la séquence du gène chez A est plus proche de la séquence du gène chez C que chez B. L'arbre de gènes présente donc la topologie ((A,C),B).

des espèces. Si la recherche d'un arbre d'espèce s'en trouve compliquée [40], étudier les différences entre des topologies incongruentes permet de mieux comprendre l'histoire évolutive de chaque famille de gènes. Le principe de la réconciliation est de déterminer les événements ayant conduits aux différences de topologies observées entre un arbre d'espèces et des arbres de gènes. On doit à Goodman et collaborateurs le terme *réconciliation* et le premier algorithme permettant de réconcilier un arbre de gène avec un arbre des espèces [58]. L'algorithme de Goodman est basé sur le principe de maximum de parcimonie (voir 1.2). Il effectue la réconciliation en recherchant le scénario avec un nombre de duplications et de pertes minimal qui permet d'expliquer les différences de topologies entre les deux arbres. Dans le même esprit, Doyon et al. et Nguyen et al. proposent une méthode de réconciliation parcimonieuse en termes de duplications, pertes et transferts [43] [101].

1.1.5.5 Au delà de la réconciliation : des modèles d'évolution complexes

Les méthodes de réconciliation ont inspiré le développement de modèles d'évolution de familles de gènes prenant en compte les duplications, pertes, transferts horizontaux, et le tri de lignées incomplet. En utilisant ces modèles, on intègre la réconciliation dans le processus d'inférence des arbres de gènes au lieu de réconcilier les arbres a posteriori. Un des premiers modèles développés fut celui d'Arvestad et collaborateurs en 2003. Leur modèle prend en compte les duplications et le pertes dans sa première version [9], et ajoute l'évolution des séquences [8] dans une deuxième version. En 2009, Akerborg et collaborateurs proposent un modèle similaire autorisant aussi des taux de substitutions différents selon les lignées [2]. Parallèlement au développement de ces modèles, plusieurs méthodes permettant d'inférer des phylogénies en présence de tri de lignées incomplet sont publiées [95, 86]. En 2012, Rasmussen et Kellis publient une méthode permettant de prendre en compte les duplications, les pertes et le tri de lignées incomplet [115]. L'année suivante, Szöllősi et collaborateurs proposent une méthode d'inférence avec duplications, pertes et transferts [133]. À ce jour, il n'existe pas de méthodes prenant en compte les duplications, les pertes, les transferts et le tri de lignées incomplet [132]. L'approche probabiliste est particulièrement appropriée pour l'inférence d'arbres de gènes, mais aussi des arbres d'espèces en présence de tous ces événements [132]. Avec Phyldog [24], Boussau et collaborateurs développent une méthode pour inférer conjointement les arbres de gènes et l'arbre des espèces, en modélisant les duplications, les pertes, et l'évolution des séquences.

1.1.6 Un arbre du vivant ?

FIGURE 1.11 – Représentation schématique de l'arbre du vivant avec les trois grands domaines. Tiré de [134]. Cette figure a été omise pour des raisons de droit d'auteur.

L'idée de représenter les relations de parenté entre toutes les espèces dans un unique arbre était déjà bien présente au dix-neuvième siècle (1). L'essor de la phylogénie moléculaire à partir des années 1960 relance les efforts pour reconstruire un *arbre du vivant* à partir de marqueurs moléculaires. Dans les années 1970, Woese et Fox utilisent ainsi l'ARN ribosomique 16s pour reconstruire une phylogénie des procaryotes. Ils

établissent la structure de l'arbre du vivant en trois grand domaines : les bactéries, les archées, et les eucaryotes [139] (Figure 1.11). Avec l'augmentation du volume de données issues du séquençage, on dispose de plus en plus de marqueurs, et de plus en plus d'espèces à inclure dans l'arbre du vivant. Mais, comme nous l'avons vu dans la partie précédente, les phylogénies de gènes peuvent être contradictoires. Avoir plus de séquences disponibles ne facilite donc pas forcément l'inférence de la phylogénie des organismes [40].

La notion même de transfert horizontal remet en cause la pertinence de représenter l'évolution sous forme d'arbre. Dans un article de 1999, Doolittle fait état de la controverse sur la place à accorder aux transferts horizontaux dans la classification du vivant [42]. Pour sa part, Doolittle soutien que l'abondance des transferts horizontaux dans l'histoire évolutive des êtres vivants rend obsolète toute tentative de représentation de l'arbre du vivant (Figure 1.12). Ce point de vue n'est pas adopté par toute la communauté. Peu après, Philippe et al. proposent une représentation alternative de l'arbre du vivant pour les procaryotes (Figure 1.13). D'après cette représentation, l'histoire évolutive verticale (l'histoire des spéciations) peut être reconstituée à partir d'un nombre limité de gènes qui ne sont jamais transférés au cours de l'évolution. Cette histoire est aussi portée, dans une moindre mesure par des gènes rarement transférés. Les gènes restants, qui ont été beaucoup transférés, ne portent pas de signal phylogénétique et représentent la composante horizontale de l'évolution. En d'autres termes, l'arbre du vivant est bruité par les transferts. Cette représentation de l'arbre du vivant est appuyée par deux études montrant la présence d'un signal vertical dans des jeux de données avec transferts [25, 38].

FIGURE 1.12 – Représentation schématique de l'histoire évolutive du vivant. D'après Doolittle, une représentation en réseau est plus appropriée qu'une représentation sous forme d'arbre. Tiré de Doolittle et al. 1999 [42]. Cette figure a été omise pour des raisons de droit d'auteur.

Pour pouvoir trancher entre ces deux représentations (arbre et réseau), il faudrait savoir quelle a été la fréquence des transferts au cours de l'évolution, et le nombre de gènes transférés dans chaque lignée. Ces deux questions sont étudiées dans de très nombreux articles tentant d'apporter une réponse à la controverse. Par exemple, d'après

FIGURE 1.13 – Représentation schématique de l'histoire évolutive des procaryotes. Les flèches vertes représentent les transferts horizontaux. L'arbre tracé en trait fin noir représente des gènes qui ne sont jamais transférés. La zone bleue représente les gènes qui ne sont presque jamais transférés, et avec lesquels on peut tout de même reconstruire une phylogénie. La zone grise texturée représente les gènes très transférés au cours de leur histoire évolutive. La phylogénie ne permet pas de représenter leur histoire évolutive, qui est essentiellement horizontale. Tiré de Philippe et al. 2003 [113]. Cette figure a été omise pour des raisons de droit d'auteur.

Dagan et Martin [35], les gènes permettant de reconstruire une phylogénie "verticale" représentent en moyenne 1% des génomes procaryotes. D'un autre côté, Galtier et Daubin montrent qu'on peut trouver un signal phylogénétique moyen, malgré la présence de transferts horizontaux [54]. La notion de signal moyen est contestée par Bapteste et collaborateurs [11], qui mettent aussi en garde contre les biais des méthodes de reconstruction et le fait de vouloir reconstruire un arbre à tout prix. Ces articles montrent qu'il n'est pas évident de décider de la place à accorder aux transferts horizontaux en phylogénie. Dans son rapport sur l'article de Bapteste et al. (publié à la fin de l'article), Galtier pose le problème de cette manière : l'évolution a une composante verticale et une composante horizontale. Que faire en cas d'incongruences ? Et à partir de quelle quantité de transferts doit-on abandonner la représentation verticale ? [11]. Dans un article très récent, Daskalakis et Roche montrent qu'il est toujours possible de reconstruire la phylogénie des espèces si chaque gène est associé à un taux de transfert constant [37]. En 2012, Abby et collaborateurs proposent une approche différente et montrent qu'il est possible de reconstruire un arbre phylogénétique en tenant compte des transferts horizontaux, et non malgré eux [1].

1.1.7 Évolution de l'architecture des génomes

Depuis les années 2000, on observe une forte augmentation du volume de données moléculaires disponibles. Cette augmentation concerne aussi bien le nombre d'espèces que le nombre de gènes par espèce. Alors qu'on reconstruisait auparavant les histoires évolutives des espèces à partir des quelques marqueurs moléculaires disponibles, il devient possible d'étudier les relations entre les espèces à partir de génomes de plus en plus

complets (on parle alors parfois de *phylogénomique)* [40]). Ce changement d'échelle n'est cependant pas trivial du point de vue méthodologique [40].

Une difficulté majeure dans l'application des méthodes de reconstruction phylogénétique à l'échelle des génomes est que la plupart de ces méthodes sont basées sur l'hypothèse que les gènes d'un génome évoluent indépendamment les uns des autres. Or, les gènes sont organisés en structures qui contraignent l'évolution de ces derniers. Un niveau d'organisation basique est l'agencement, ou ordre, des gènes sur le chromosome. L'évolution de l'ordre des gènes est l'objet principal de la modélisation réalisée au cours de ma thèse. Au cours de l'évolution des génomes, l'ordre des gènes est modifié par des événements appelés réarrangements chromosomiques. L'effet des réarrangements est visible sur les chromosomes des espèces actuelles. On retrouve en effet des séquences similaires entre des chromosomes appartenant à deux espèces différentes, mais ces séquences peuvent occuper des positions différentes sur les chromosomes (Figure 1.14), voire être réparties sur plusieurs chromosomes dans une des espèces (Figure1.15).

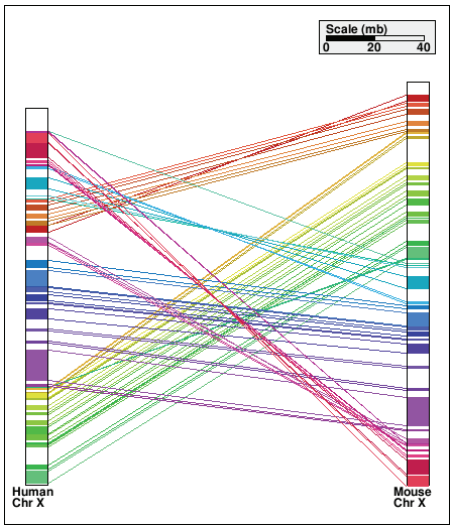


FIGURE 1.14 – Blocs de synténie conservés entre le chromosome X humain et le chromosome X murin. Tiré du serveur Cynteny [126]

L'histoire des réarrangements fait partie de l'histoire évolutive des génomes. Dans ce manuscrit, j'ai choisi de parler d'évolution de l'architecture des génomes pour décrire cet aspect de l'évolution des génomes. Dans la littérature liée à l'étude et à la modélisation de l'évolution de l'architecture des génomes, on utilise couramment deux

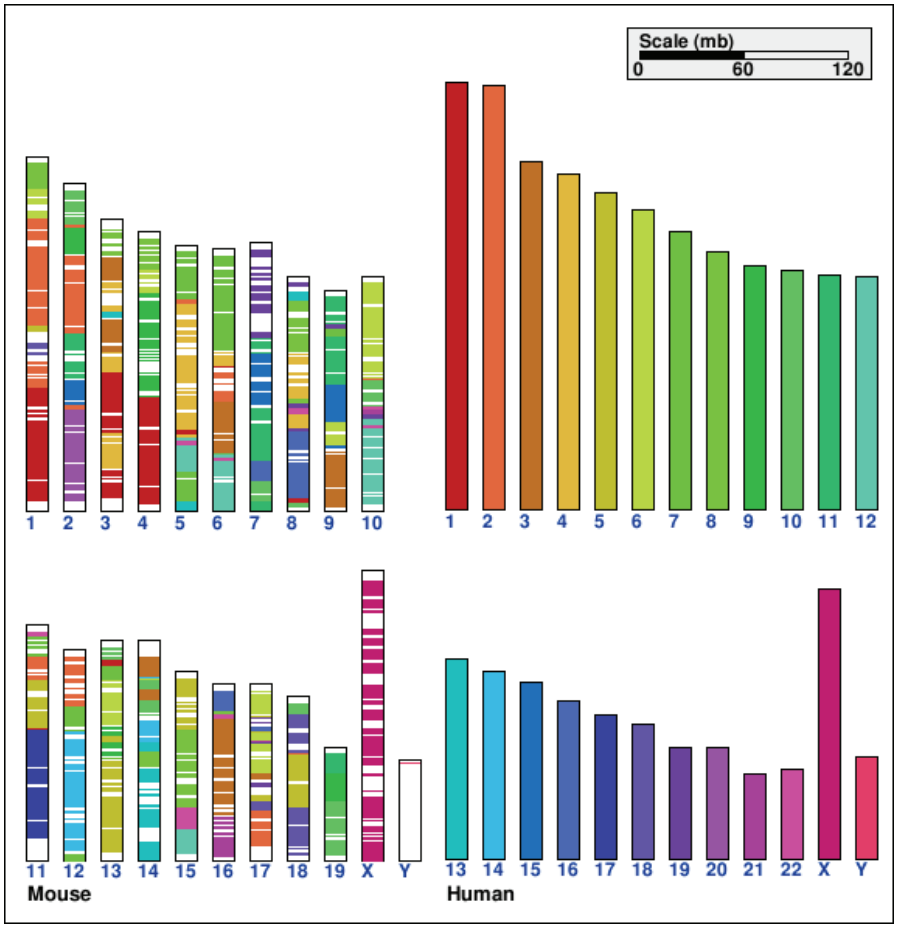


FIGURE 1.15 – Comparaison des génomes humain et murin. On peut faire correspondre des séquences appartenant au chromosome 1 humain à des séquences appartenant aux chromosomes 1, 3, 4, 5, 8, 11 et 13 de la souris. Tiré du serveur Cynteny [126].

autres termes, qu'il est utile d'introduire ici.

Ordre des gènes : On peut représenter un chromosome comme une séquence de gènes. Les positions respectives des gènes sur les chromosomes définissent l'ordre des gènes sur le chromosome, établi à partir des coordonnées chromosomiques des gènes. Dans la base de données Ensembl, ces coordonnées incluent le numéro de chromosome, le brin et deux positions sur le chromosome. Ces deux positions correspondent au début du transcrit du gène le plus en amont sur le chromosome et à la fin du trans-

crit du gène le plus en aval sur le chromosome. Si les coordonnées des gènes étudiés ne se chevauchent pas, il est alors facile d'ordonner les gènes. Dans le cas contraire, il faut décider d'un critère pour ordonner les gènes qui se chevauchent, ce qui introduit cependant une incertitude dans l'ordre des gènes utilisé. Un critère possible est de ne se baser que sur le début du gène [12].

Synténie : La synténie décrit la relation entre plusieurs gènes *voisins* dans un génome. La notion de voisinage peut se comprendre à plusieurs niveaux, ce qui donne une certaine ambiguïté à la définition de synténie. Ainsi, la synténie peut décrire :

- des gènes présents sur le même chromosome. C'est la définition originale de la synténie [106].
- des gènes voisins sur un chromosome. Il faut alors définir une distance, un seuil à partir duquel on considère que deux gènes sont voisins. Cette définition est plus particulièrement utilisée lorsqu'on parle de *blocs de synténie* [126].
- l'ordre des gènes sur un chromosome. Cette définition est plus particulièrement utilisée lorsqu'on parle de *reconstruction de synténie ancestrale* [31].

Par ailleurs, le mot synténie est utilisé pour décrire la conservation de l'organisation et/ou du voisinage des gènes entre plusieurs génomes. On parle alors plutôt de *blocs de synténie conservés* entre les chromosomes de différentes espèces (Figure 1.14). Dans ce manuscrit, on parlera de *synténie* pour décrire l'ordre des gènes sur un chromosome.

Même si au cours de ma thèse je me suis concentrée sur l'évolution de l'architecture des génomes, gardons à l'esprit qu'il existe de nombreux autres niveaux d'organisation au sein des des génomes dont l'évolution peut être modélisée : la structure tridimensionnelle des génomes, les voies métaboliques, les complexes protéiques...

1.1.7.1 Les réarrangements chromosomiques

Le terme réarrangement regroupe plusieurs processus distincts. Ainsi, on parle de réarrangement lors de la duplication, de la perte, de la translocation, de l'inversion d'une portion de chromosome, ou lors d'une fusion ou d'une fission de deux chromosomes (Figure 1.16). Au cours d'une duplication, une portion du chromosome est copiée et

insérée à une autre position sur le chromosome. Les duplications peuvent être observées à très large échelle, de plusieurs gènes, jusqu'à des chromosomes ou même des génomes entiers. Avec les fusions et les fissions de chromosomes, les duplications et les pertes de chromosomes entiers peuvent modifier le nombre de chromosomes des génomes. La fusion et la fission sont deux réarrangements inverses qui consistent respectivement à fusionner deux chromosomes et à diviser un chromosome en deux entités. L'inversion est un réarrangement observable à une résolution plus fine. Il conduit à un ré-agencement des gènes sur le chromosome. La translocation est l'échange de deux fragments appartenant à deux chromosomes différents.

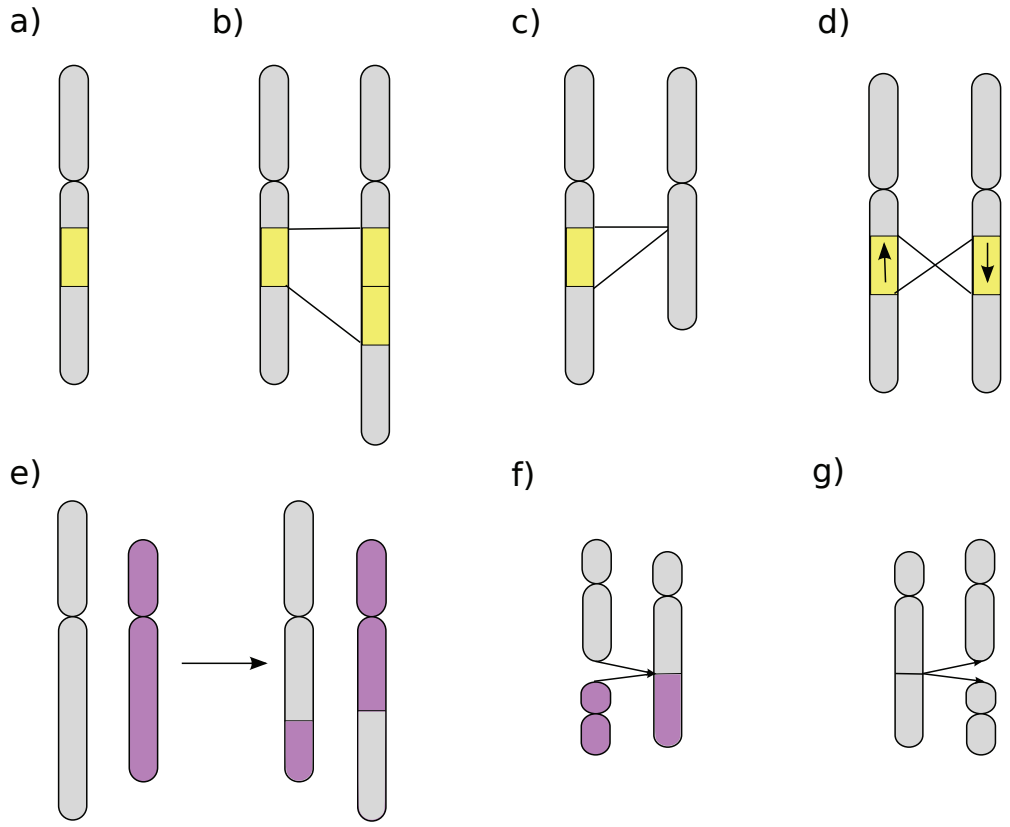


FIGURE 1.16 – a) chromosome de référence b) duplication d'une portion du chromosome c) perte d'une portion du chromosome d) inversion d'une portion du chromosome e) translocation f) fusion de deux chromosomes g) fission d'un chromosome.

La duplication et la perte de portions de chromosome sont directement liées à la duplication et à la perte de gènes. En revanche, les autres types de réarrangements n'af-

fectent pas directement le contenu en gènes des génomes. La figure 1.17 illustre la modification de l'ordre des gènes suite à une inversion. Les réarrangements tels que les inversions ont été associés à l'apparition et à la perte de gènes [52].

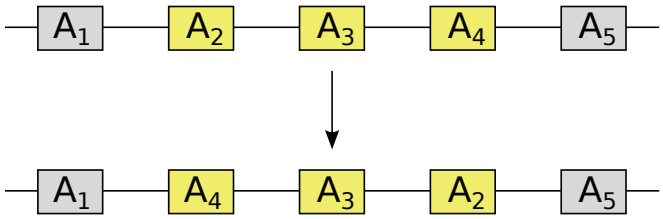


FIGURE 1.17 – Inversion de la portion du chromosome contenant les gènes A_2, A_3, et A_4. L'ordre des gènes est $A_1A_2A_3A_4A_5$ avant l'inversion et $A_1A_4A_3A_2A_5$ après l'inversion.

Depuis la découverte en 1921 d'une inversion sur un chromosome de la drosophile [129], les réarrangements ont été largement étudiés et leurs origines et conséquences partiellement élucidées. Au niveau moléculaire, les réarrangements surviennent lors de la réparation d'une cassure double brin de l'ADN [112]. Face à une cassure simple brin, le processus cellulaire de réparation consiste en effet à reconstituer le fragment perdu en utilisant l'information portée par l'autre brin. En revanche, dans le cas d'une cassure double brin, les deux brins d'ADN sont coupés à des positions proches et la réparation peut être assimilée à un recollage des fragments d'ADN. Ce type de cassure peut être causé par un agent mutagène extérieur, par un accident au cours de la réplication de l'ADN, ou plus simplement lors d'un crossing-over au cours de la méiose. Chez l'humain, les réarrangements chromosomiques ont fait l'objet de nombreuses études, notamment parce que nombre de maladies génétiques sont causées par des réarrangements. La base de données Chromosomal Variation in Man [19] recense par exemple 24000 "anormalités chromosomiques" associées à des pathologies. A l'échelle d'une population, les réarrangements participent à la variabilité génétique entre les individus. Certains réarrangements ont pu être associés à des phénotypes particuliers, chez les eucaryotes [32] et chez les procaryotes [108]. Il a par exemple été montré que deux inversions étaient responsables d'un polymorphisme de taille chez *Keyacris scurra* (Figure 1.18). Chez les procaryotes, des réarrangements seraient impliqués dans la divergence et la virulence de différentes souches de *Yersinia pestis* [82].

A l'échelle de l'évolution des espèces, les inversions semblent intervenir dans la

FIGURE 1.18 – Inversions de fragments chromosomiques et phénotypes chez *Keyacris scurra*. Tiré de [32]. Cette figure a été omise pour des raisons de droit d'auteur.

différenciation des chromosomes sexuels en provoquant l'arrêt progressif de la recombinaison [10, 79].

Dès les années 1930, Sturtevant et Dobzhansky s'intéressent à l'aspect évolutif des réarrangements et proposent une phylogénie de *Drosophila pseudoobscura* basée sur des inversions [130]. Les inversions semblent jouer un rôle dans le processus de spéciation chez les drosophiles [102]. Il est cependant difficile de quantifier la contribution des réarrangements aux spéciations à l'échelle de l'arbre du vivant [28]. Chez les mammifères, on observe une corrélation entre les taux de spéciations et de réarrangements [27] mais aucun lien de causalité n'a pu être établi [28]. La variation des taux de réarrangements entre des lignées différentes est une question centrale dans l'étude des réarrangements. Chez les vertébrés, Burt et al. suggèrent que les taux de réarrangements ont pu varier entre 0.2 et 2 réarrangements par million d'années au cours de l'évolution [26], des ordres de grandeurs retrouvés plus tard par Murphy et al. [98] (Figure 1.19).

FIGURE 1.19 – Taux de réarrangement au cours de l'évolution des mammifères, tels qu'estimés par Murphy et al. 2005. Tiré de [98]. Cette figure a été omise pour des raisons de droit d'auteur.

On trouve des taux de réarrangements sensiblement plus élevés chez les invertébrés [114, 32]. Un exemple de taux de réarrangement extrêmement élevé peut être trouvé chez les nématodes : de l'ordre de 40 à 100 réarrangements par million d'années d'après Coghlan et al. [33]. Il faut cependant considérer ces chiffres avec précaution. D'une part, les tailles de génomes peuvent varier considérablement entre les lignées. Si on divise les taux de réarrangements par la taille des génomes, on obtient une vision moins biaisée des variations du taux de réarrangement. Coghlan et al. [32] rapportent avec cette correction des taux de réarrangements deux fois plus élevés chez la drosophile que chez les rongeurs, et cent fois plus élevés chez les nématodes que chez la drosophile. D'autre part, l'estimation de taux de réarrangements est sensible à la qualité des données et à la méthode utilisée [32]. En ce qui concerne les données, la qualité de l'as-

semblage des génomes influe fortement sur le nombre de réarrangements inférés dans les lignées concernées, et donc sur l'estimation des taux de réarrangement. En effet, toute erreur d'assemblage peut conduire à inférer un réarrangement fictif. En ce qui concerne les méthodes, l'estimation des taux de réarrangements dépend de la fiabilité de l'estimation des temps de divergence des lignées, de la méthode de détection des blocs de synténie conservés entre les espèces, et de la résolution à laquelle on étudie les réarrangements (i.e. le nombre de paires de bases impliquées dans le réarrangement).

Une autre question centrale dans l'étude des réarrangements est de savoir si les taux de réarrangements varient selon les régions des génomes. En 1984, Nadeau et Taylor proposent une théorie selon laquelle les réarrangements peuvent survenir aléatoirement à n'importe quelle position du génome [99]. Le modèle de Nadeau et Taylor est appelé modèle de cassure aléatoire (*random breakage model*). À faible résolution, des études réalisées chez les vertébrés semblent soutenir ce modèle [46]. En 2003, avec la mise à disposition de génomes complètement séquencés et annotés, il devient possible d'étudier les réarrangements à une résolution beaucoup plus fine [110]. Pevzner et Tesler montrent alors que les petits réarrangements (i.e. les blocs de synténie < 1 million de paires de bases (Mbp) [89]), jusque-là invisibles, ne soutiennent pas le modèle de cassure aléatoire. Ils proposent alors une théorie alternative selon laquelle il existe des régions plus susceptibles que d'autres aux réarrangements : c'est le modèle des régions fragiles (*fragile breakage model*) [111]. Ces deux théories contradictoires ont suscité une controverse, résumée entre autres dans un article de Alekseyev et Pevzner en 2007 [4].

1.1.7.2 L'architecture des génomes en génomique comparative et en phylogénie

L'étude de la synténie conservée entre les espèces a plusieurs champs d'applications intéressants. En premier lieu, la synténie peut être utilisée pour annoter des gènes. En effet, un ordre des gènes conservé au cours de l'évolution peut indiquer une interaction fonctionnelle entre des gènes. Le cas le plus connu est celui des opérons chez les bactéries, où des gènes voisins participent à une même fonction biologique. Même chez les eucaryotes, la proximité spatiale des gènes peut être un indice d'une contrainte fonctionnelle. Dans un récent article, Naville et al. utilisent ainsi la conservation de l'ordre des gènes pour prédire une interaction de régulation entre deux régions du chromosome X humain [100]. La synténie conservée peut aussi être utilisée comme un com-

plément aux méthodes de recherches de similarité entre les séquences pour trouver des gènes orthologues [59, 5, 70]. L'information contenue dans des ordres de gènes peut être utilisée pour résoudre des relations phylogénétiques. Il existe ainsi de nombreux exemples de phylogénies entièrement reconstruites à partir de l'ordre des gènes [119]. Il est également possible d'utiliser la synténie en complément d'une analyse phylogénétique basée sur les séquences. Dans un article de 2006, Duret et al. utilisent cette approche pour élucider l'origine évolutive du gène Xist, responsable de l'inactivation d'un exemplaire du chromosome X humain [44].

1.1.7.3 Modéliser l'évolution de l'architecture des génomes

A la fin des années 1990 et au début des années 2000 ont été publiées plusieurs méthodes permettant de reconstruire l'ordre des gènes dans des génomes ancestraux. Parmi ces méthodes émergent BP-Analysis [17], MGR [20, 21] et Grappa [96, 97], qui servent encore souvent de référence aujourd'hui. Ces trois méthodes ont en commun une représentation des génomes sous forme de permutation de caractères. Au sein d'un génome, chaque caractère représente un gène. On utilise ici une définition libérale du gène : ce n'est pas forcément une séquence codante mais n'importe quel locus conservé. Les gènes homologues sont représentés par le même caractère dans les différents génomes étudiés. On distingue les gènes encodés sur le brin anti-sens des gènes encodés sur le brin sens avec l'aide du signe "-" précédant le caractère associé au gène anti-sens. Les permutations de caractères ne contiennent pas tous les gènes des génomes, seulement ceux pour lesquels on a identifié des orthologues dans les autres génomes. BP-Analysis, MGR et Grappa reconstruisent l'ordre des gènes chez l'ancêtre des génomes étudiés en recherchant la permutation de caractères qui permet de retrouver tous les génomes étudiés en effectuant un minimum d'actions.

FIGURE 1.20 – Génomes mitochondriaux de l'humain, la drosophile et l'oursin encodés sous forme de permutations de caractères. Un génome ancestral (A) est reconstruit avec MGR. Tiré de [20]. Cette figure a été omise pour des raisons de droit d'auteur.

Si elles permettent de reconstruire un ordre des gènes ancestral de manière élégante, ces méthodes deviennent très coûteuses en temps de calcul, voire inutilisables, lorsque

le nombre de génomes étudiés et le nombre de gènes par génome augmentent. Un autre inconvénient de ces méthodes est leur difficulté à prendre en compte toute la complexité des génomes : gènes dupliqués, duplications complètes de génomes ...

Une manière de résoudre ces difficultés est de représenter les génomes non pas comme une permutation de caractères mais comme un ensemble d'adjacences. Pour comprendre ces deux représentations différentes, imaginons un génome comme un graphe dans lequel les gènes sont les sommets et les adjacences sont les arêtes qui relient les gènes consécutifs. Pour communiquer l'ordre des gènes dans ce génome, il n'est pas nécessaire de re-dessiner tout le graphe. On peut en effet se contenter de donner l'information nécessaire en donnant un ensemble ordonné des sommets : c'est la représentation par une permutation de caractères. Une autre manière de procéder est de donner l'ensemble des arêtes, en précisant quel sommet doit être rattaché à chaque extrémité des arrêtes : c'est la représentation par un ensemble d'adjacences (Figure 1.21).

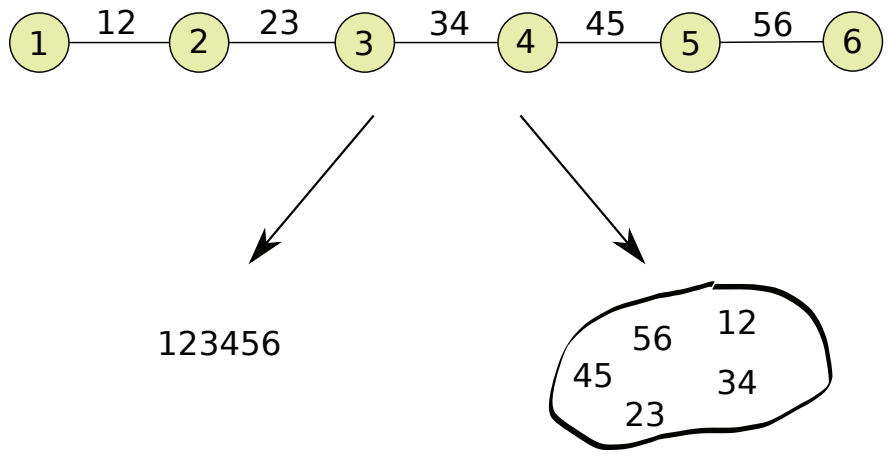


FIGURE 1.21 – Soit un génome constitué des gènes 1, 2, 3, 4, 5, 6. Les gènes consécutifs dans l'ordre des gènes sont des sommets reliés par une arête nommée d'après les sommets qu'elle relie. Les gènes 1 et 2 sont ainsi reliés par l'arête 12. On peut encoder ce génome de deux manières différentes : comme une permutation de caractères et comme un ensemble d'adjacences.

Les adjacences permettent donc de décrire localement l'ordre des gènes sur un chromosome. Comme les permutations de caractères, les adjacences décrivent un sous-ensemble des gènes des génomes. Une définition plus rigoureuse de l'adjacence est donc la suivante :

Adjacence : Soit A l'ensemble des gènes d'un génome, et $B \subset A$ un sous-ensemble de gènes que l'on va étudier. On considère que deux gènes appartenant à B sont adjacents si les deux gènes sont sur le même chromosome et qu'il n'existe pas d'autre gène appartenant à B localisé entre ces deux gènes sur le chromosome.

Les méthodes qui utilisent cette représentation font l'hypothèse que les adjacences d'un même génome évoluent indépendamment les unes des autres. Ce n'est en réalité pas le cas : les inversions ont par exemple à leurs extrémités deux adjacences qui évoluent ensemble.

L'hypothèse reste cependant intéressante du point de vue de la modélisation. D'une part, elle permet de réduire le problème de reconstruction de l'ordre des gènes dans un génome ancestral à celui de la reconstruction d'adjacences ancestrales indépendantes. D'autre part, la phylogénie moléculaire fait l'hypothèse qu'au sein d'un génome, chaque gène évolue indépendamment des autres. Les adjacences représentent une dépendance entre deux gènes. Faire l'hypothèse que les adjacences sont indépendantes constitue donc un pas vers la modélisation des dépendances réelles entre les gènes.

Une fois définies les adjacences des génomes étudiés, il devient possible d'encoder l'ordre des gènes en binaire. Les deux génomes représentés sur la figure 1.22 se différencient par l'inversion de l'ordre des gènes 2, 3 et 4. Cette inversion conduit à des adjacences différentes entre les deux génomes. Dans le génome G_1, on trouve ainsi les adjacences 12, 23, 34, 45 et 56 tandis que le génome G_2 contient les adjacences 14, 43, 32, 25, 56. Si on considère que *l'orientation* des adjacences est importante (i.e. l'adjacence 23 n'est pas équivalente à l'adjacence 32), alors G_1 et G_2 n'ont qu'une seule adjacence commune. Le motif binaire correspondant à cette situation est représenté dans la table 1.1. Si au contraire on considère que les adjacences ne sont pas orientées, G_1 et G_2 ont en commun trois adjacences, et le motif binaire correspondant est celui de la table 1.2.

TABLEAU 1.1 – Présence et absence des adjacences orientées dans les génomes G_1 et G_2

	12	23	34	45	56	14	43	32	25
G_1	1	1	1	1	1	0	0	0	0
G_2	0	0	0	0	1	1	1	1	1

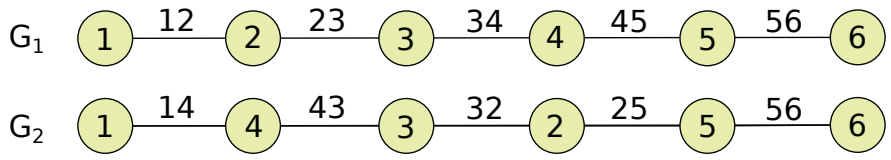


FIGURE 1.22 – Ordre des gènes dans les génomes G_1 et G_2. Les génomes sont représentés par des graphes dans lesquels les sommets sont les gènes et les arêtes sont les adjacences entre les gènes. G_1 et G_2 se distinguent par une inversion de l'ordre des gènes 2, 3 et 4. Cette inversion conduit à des adjacences différentes entre les deux génomes.

TABLEAU 1.2 – Présence et absence des adjacences non-orientées dans les génomes G_1 et G_2

	12	23	34	45	56	14	25
G_1	1	1	1	1	1	0	0
G_2	0	1	1	0	1	1	1

Ces deux représentations, adjacences orientées et non-orientées, correspondent à deux précisions différentes. Par ailleurs, on peut parler d'adjacence entre différents objets : entre deux gènes, deux locus, ou deux blocs de séquences orthologues.

Le motif binaire permet d'encoder la présence et l'absence des adjacences observées dans les génomes étudiés. L'adjacence est donc un caractère à deux états, transmis par spéciation et dont l'état est modifié par des réarrangements chromosomiques, ainsi que par les duplications et les pertes de gènes (Figure 1.23). Lorsqu'on considère que la phylogénie des espèces est connue, on peut alors utiliser les états observés dans les espèces étudiées pour inférer l'histoire évolutive des adjacences (Figure 1.24) [119]. On peut ensuite reconstruire l'ordre des gènes dans des génomes ancestraux à partir des adjacences inférées comme présentes.

À ma connaissance, la première méthode de reconstruction d'ordre des gènes ancestraux à partir d'adjacences évoluant indépendamment à être publiée fut celle de Ma et collaborateurs en 2006 [93]. La première étape de leur méthode consiste à définir des blocs de séquences orthologues dans les génomes étudiés (l'humain, la souris, rat et le chien). Les adjacences sont ensuite placées entre les blocs consécutifs dans les génomes. Dans un deuxième temps, on reconstruit les adjacences ancestrales avec un algorithme

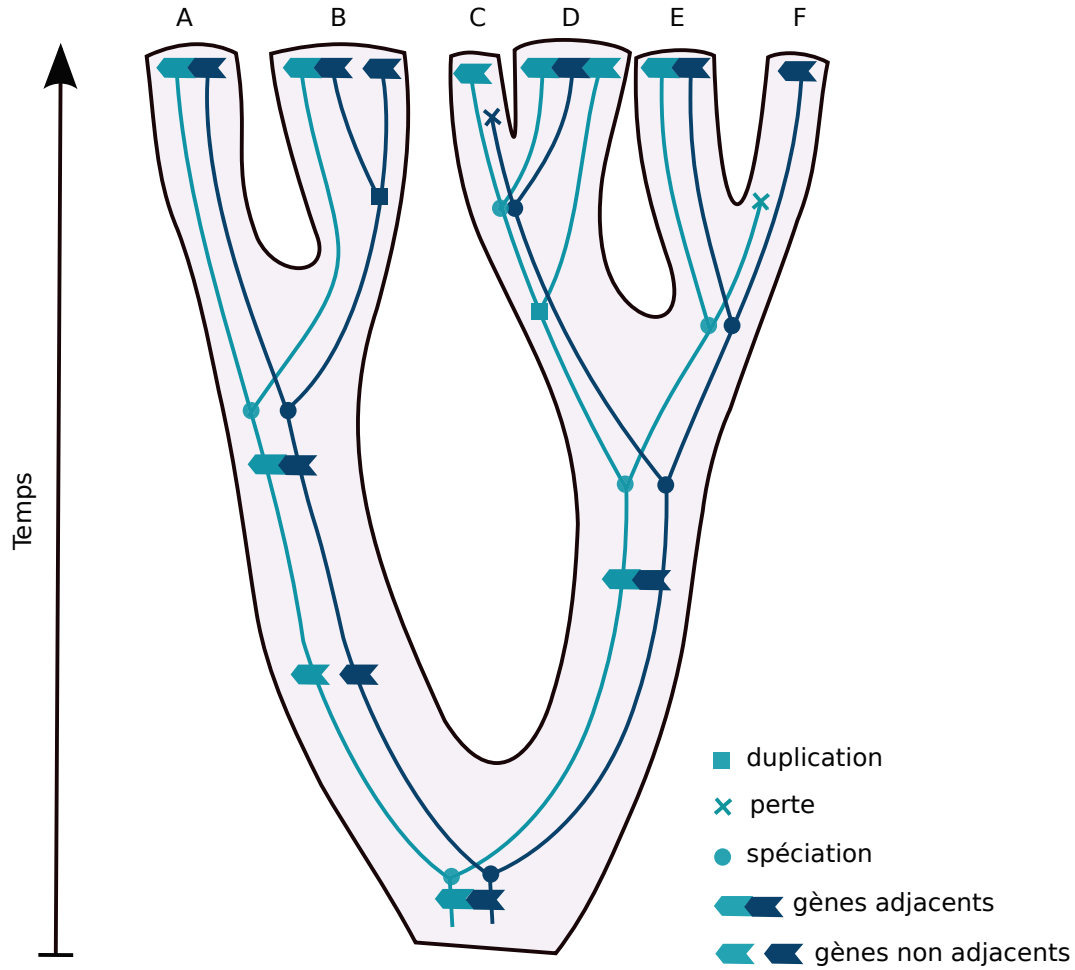


FIGURE 1.23 – Transmission de l'adjacence entre un gène bleu clair et un gène bleu foncé au cours de leur histoire évolutive. L'adjacence est présente à la racine de l'arbre. Après la première divergence des lignées, elle est conservée dans la lignée de droite, mais perdue puis regagnée dans la lignée de gauche suite à deux réarrangements successifs. Dans la lignée conduisant à l'espèce B le gène bleu foncé est dupliqué et l'adjacence est transmise à une copie du gène, et perdue pour l'autre. Dans la lignée conduisant à D elle est transmise aux deux copies. Elle est perdue suite à la perte d'un des deux gènes dans les lignées conduisant à C et F.

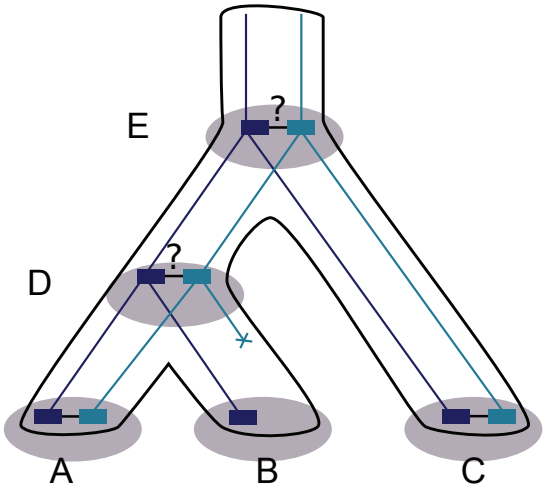


FIGURE 1.24 – Lorsque la phylogénie des espèces est connue, on infère l'état (i.e. la présence ou l'absence) de l'adjacence dans les génomes ancestraux D et E étant donnés les états dans les génomes étudiés A, B et C.

de parcimonie (voir 1.2) pour obtenir l'agencement des blocs chez l'ancêtre des génomes étudiés.

Peu après, Bhutkar et collaborateurs [15] publient une méthode similaire et reconstruisent l'ordre des gènes chez l'ancêtre de huit génomes de drosophiles. En 2008, Chauve et Tannier proposent une généralisation de la méthode de Ma et al. [31]. Dans leur version, la définition des blocs d'orthologues ancestraux est plus souple et la reconstruction des adjacences ancestrales prend en compte plusieurs scénarios possibles, chaque scénario étant associé à un poids. Une des limites de ces trois méthodes est qu'elles ne peuvent pas prendre en compte les gènes dupliqués ni les pertes de gènes, ce qui limite le nombre de marqueurs utilisables. Bhutkar et al. choisissent ainsi le "meilleur homologue" [15] en présence de gènes dupliqués. En 2008, Ma et collaborateurs publient une nouvelle version de leur méthode prenant en compte les duplications [92]. En 2012, Gagnon et collaborateurs proposent une généralisation de la première version de la méthode de Ma et al. qui prend en compte les duplications complètes de génomes [53]. En 2010, Ma et al. publient une troisième version de leur méthode et l'insèrent pour la première fois dans un cadre probabiliste [91]. La transition d'une méthode de reconstruction parcimonieuse à une méthode probabiliste est intéressante car elle permet d'explorer d'avantage de solutions, et d'associer une pro-

babilité a posteriori à chaque adjacence ancestrale inférée. On obtient ainsi une vision plus nuancée de l'ordre des gènes ancestraux : certaines adjacences sont plus fiables que d'autres. L'inconvénient de passer à une méthode probabiliste se trouve bien souvent dans l'augmentation des temps de calcul. Pour limiter cet effet, deux méthodes récentes utilisent RAxML, un programme de reconstruction phylogénétique probabiliste particulièrement rapide : MLWD [83] et PMAG (et son extension PMAG+) [67]. Avec l'utilisation de RAxML, il devient possible de reconstruire des ordres de gènes ancestraux à partir de 100 espèces et 100000 gènes par espèce [83], des ordres de grandeurs difficiles à atteindre autrement. PMAG et MLWD tolèrent la présence de gènes dupliqués mais ne proposent pas de modèle spécifique pour l'effet des duplications et les pertes de gènes sur l'ordre des gènes. Ces deux méthodes construisent un alignement de caractères binaires à partir des adjacences présentes et absentes dans les génomes étudiés et utilisent RAxML pour reconstruire la topologie en fonction de l'alignement. Une autre limite de ces deux méthodes est que l'ordre des gènes dans une espèce ancestrale ne peut être reconstruit qu'à partir du sous arbre dont la racine correspond à l'espèce ancestrale dont on recherche l'ordre des gènes. Par exemple, en prenant l'arbre de la figure 1.24, il faudrait couper l'arbre au niveau de l'espèce D pour reconstruire l'ordre des gènes de D à partir des adjacences chez A et B. Cette approche pose problème car on n'utilise pas l'information contenue dans le reste de l'arbre.

En 2012, Bérard et collaborateurs proposent avec DeCo [12] une méthode de reconstruction d'adjacences ancestrales basée sur un algorithme de parcimonie, des arbres de gènes réconciliés pour les duplications et les pertes de gènes, et un système de coûts de présence et d'absence d'une adjacence. L'inférence d'une adjacence ancestrale est basée sur le calcul du coût de présence et d'absence de l'adjacence. Les modalités du calcul différent selon les événements évolutifs associés aux gènes impliqués dans l'adjacence. Ceci permet de modéliser explicitement l'impact des duplications et des pertes de gènes sur l'évolution des adjacences. En 2013, Patterson et collaborateurs étendent DeCo pour inclure les transferts horizontaux [107].

1.1.8 Les arbres d'adjacences

Un autre point intéressant de l'article de Bérard et al. est que les auteurs introduisent la notion d'*arbre d'adjacences* pour représenter l'histoire évolutive d'une adjacence. L'exis-

tence d'un arbre d'adjacences était implicite dans la plupart des méthodes citées plus haut. Il existait d'ailleurs déjà des phylogénies reconstruites à partir d'ordres de gènes [89, 119]. Mais les arbres d'adjacences sont rarement représentés. On peut alternativement représenter l'évolution de l'adjacence entre deux arbres de gènes, comme illustré sur les figures 1.23 et 1.24 [141, 132]. L'avantage de cette dernière représentation est qu'elle permet de visualiser l'évolution des gènes, des adjacences, et des espèces en même temps. L'inconvénient est que l'évolution de l'adjacence n'est pas représentée directement comme l'évolution des gènes et des espèces, par les branches d'un arbre phylogénétique. Le patron de descendance est déduit des autres objets. Alors qu'une adjacence peut être considérée elle-même comme un objet évolutif dont l'état est modifié en fonction des réarrangements et des événement évolutifs affectant les gènes à ses extrémités (duplications, pertes, transferts horizontaux). L'histoire de la transmission et des modifications de l'adjacence peut donc être représentée par un arbre, similaire aux arbres de gènes. Les figures 1.25, 1.26 et 1.27 montrent comment représenter un arbre d'adjacences étant donné la présence et l'absence de l'adjacence dans les génomes étudiés et ancestraux, et les arbres des gènes impliqués dans l'adjacence. La perte d'une adjacence suite à la perte d'un gène et suite à un réarrangement sont représentées ici de la même manière (Figure 1.26). Dans DeCo et dans Harpi, le modèle présenté dans le chapitre 2, ces deux événements sont modélisés de manière distincte.

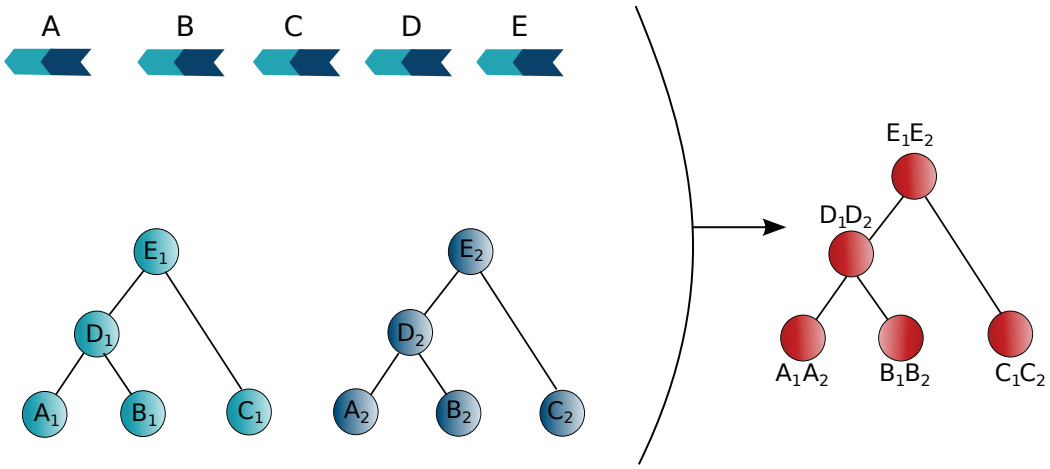


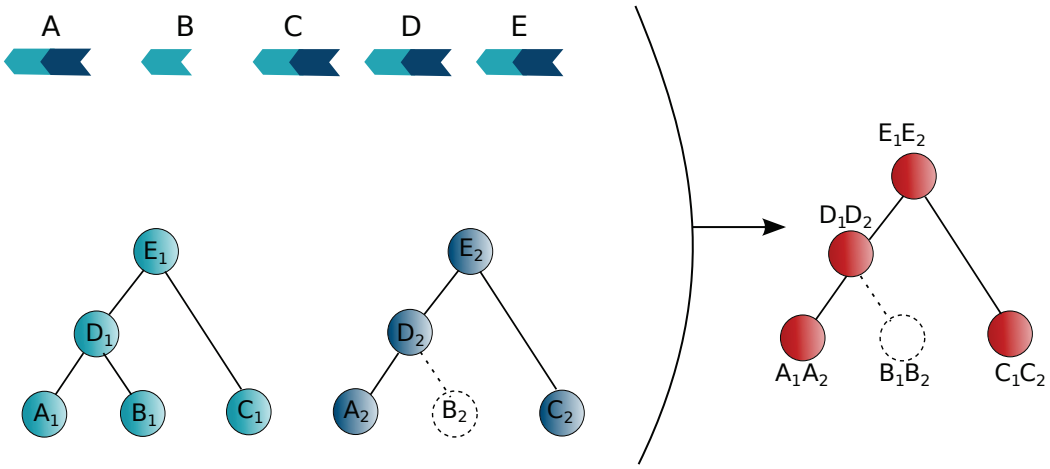
FIGURE 1.25 – Arbre d'adjacences sans duplication ni perte de gène. Les histoires évolutives des gènes bleu clair et bleu foncé sont caractérisées uniquement par des spéciations. Dans ce cas, tous les génomes (A, B,C,D,E) contiennent les deux gènes. Aucun réarrangement n'est intervenu dans ces génomes, l'adjacence est donc conservée (en haut). L'arbre d'adjacences correspondant à ce scénario suit donc la succession des spéciations (en rouge). L'adjacence, présente entre les deux gènes chez l'espèce E, est transmise à tous les descendants.

1.2 Modèles et méthodes

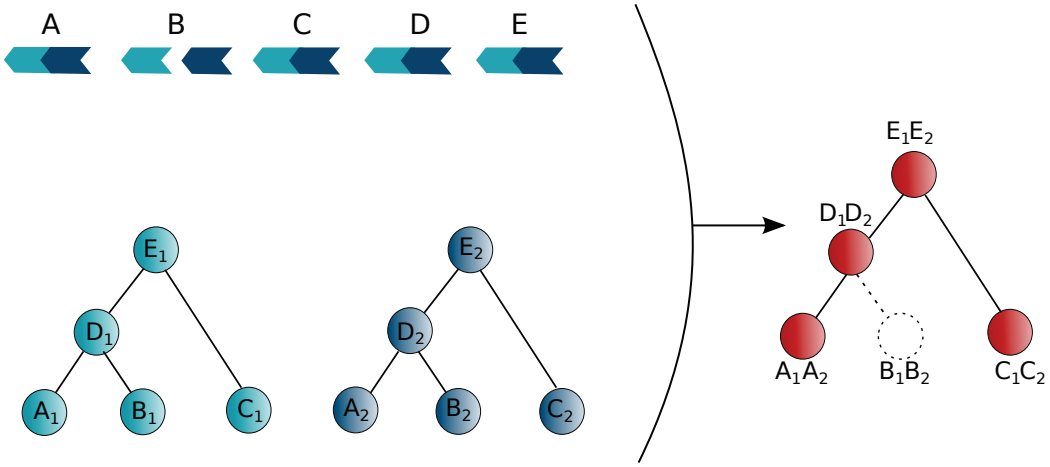
L'évolution des génomes est la conséquence de processus observables à différentes échelles : substitution, insertion et délétions de nucléotides ou d'acides aminés, perte et duplication de gènes, transmission d'allèles différente selon les lignées, et réarrangements chromosomiques. Modéliser l'évolution des génomes implique donc de prendre en compte les modifications apportées par tous ces processus pour reconstruire les changements d'états des génomes au cours de l'évolution. Le problème est que la taille des génomes conduit à un très grand nombre d'états possibles, ce qui rend difficile la modélisation de l'évolution des génomes avec un unique modèle.

Pour pallier cette difficulté, une approche pour modéliser de l'évolution des génomes consiste à décomposer le problème de manière à réduire le nombre d'états à prendre en compte. Le développement de modèles pour l'évolution des génomes a alors emprunté deux directions principales : la modélisation de l'évolution du *contenu*

FIGURE 1.26 – Arbres d'adjacences avec perte.



(a) Le gène bleu foncé est perdu dans l'espèce B, conduisant à une perte dans la lignée correspondante dans l'arbre du gène bleu foncé, et à une perte de l'adjacence dans l'arbre d'adjacences.



(b) Aucune duplication ni perte ne marque les histoires évolutives des gènes, mais un réarrangement conduit à la perte de l'adjacence dans B. Dans l'arbre d'adjacences, cela se traduit par une perte, comme dans le cas d'une perte de gène.

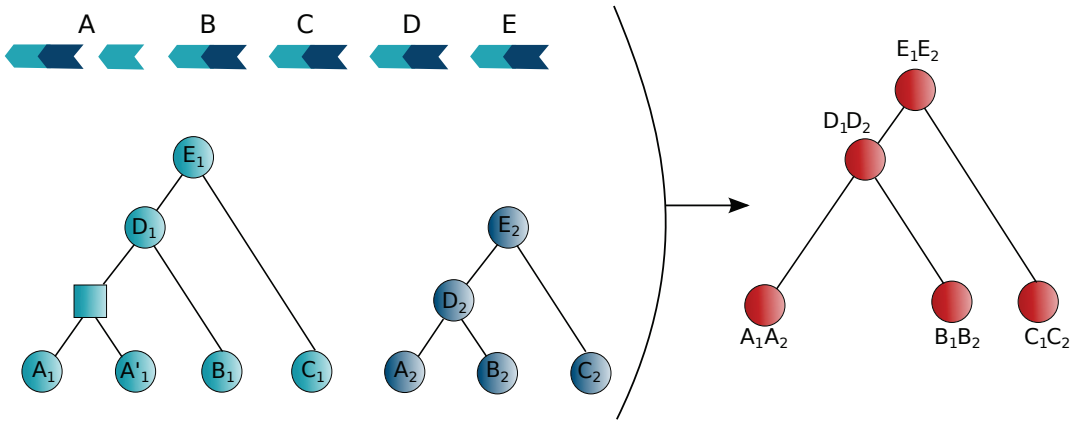


FIGURE 1.27 – Arbre d'adjacences avec duplication de gène. Le gène bleu clair est dupliqué dans l'espèce A. Ceci se traduit par une duplication visible dans l'arbre du gène bleu clair (carré bleu clair). L'adjacence (non orientée) est transmise à une seule des deux copies du gène bleu clair. Il n'y a donc pas de traces de la duplication dans l'arbre d'adjacences.

des génomes et la modélisation de l'évolution de *l'architecture* des génomes. La modélisation de l'évolution du *contenu* des génomes est passée par plusieurs changements d'échelles. Dans un premier temps, on a représenté l'évolution de familles de gènes, et non plus d'espèces, au moyen d'arbres de gènes. Dans un deuxième temps, on a réduit le problème à l'échelle du résidu et développé des modèles d'évolution à l'échelle du nucléotide, du codon, et de l'acide aminé. Cette échelle permet de travailler avec des nombres d'états possibles très restreints : quatre états pour les nucléotides, vingt-deux pour les acides aminés et soixante-quatre pour les codons. La modélisation de l'évolution de l'architecture des génomes a elle aussi nécessité un changement d'échelle. La reconstruction d'ordres des gènes ancestraux avec des modèles basés sur des permutations de caractères est un problème NP-complet [1] lorsqu'on augmente la taille des

[1] Un problème NP-complet a deux caractéristiques [34, 71] :

-
- Il est possible de **vérifier** une solution à ce problème en temps polynomial. i.e. il appartient à la classe des problèmes NP.
-
- Ce problème est au moins aussi difficile que tous les autres problèmes de la classe NP.

génomes ou le nombre de génomes étudiés (voir 1.1.7.3). Il est par contre possible de réduire le nombre d'états en ramenant le problème de l'évolution de l'ordre des gènes à celui de l'évolution indépendante d'adjacences. Avec cette approche, on réduit le nombre d'états à deux : la présence et l'absence de l'adjacence.

Nous nous intéressons ici à des modèles d'évolution *probabilistes*. Ces modèles sont des modèles de Markov décrivant l'évolution d'un objet (un site ou une adjacence par exemple). Ils sont basés sur un nombre d'états qui dépend de l'objet étudié et sur des taux de transitions entre ces états représentant les mutations. Les modèles d'évolution permettent de quantifier les mutations modifiant l'objet étudié en un temps donné. Définir un tel modèle d'évolution consiste à établir la matrice des probabilités de transition entre les différents états en fonction du temps $P(t)$.

1.2.1 Modèle d'évolution : le modèle binaire

Il existe de nombreux modèles d'évolution en phylogénie. Dans ce manuscrit, je ne détaillerai que le modèle binaire car il est particulièrement adapté à la modélisation de l'évolution indépendante d'adjacences et a inspiré le modèle d'évolution développé au cours de ma thèse (voir chapitre 2). Le modèle binaire décrit l'évolution d'un objet à deux états. Appliqué à l'évolution d'une adjacence, les deux états sont l'absence et la présence d'une adjacence (Figure 1.28).

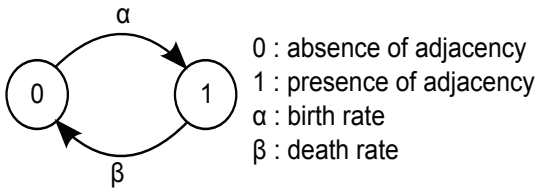


FIGURE 1.28 – Modèle de Markov à deux états. α est le taux de transition de l'état 0 à l'état 1. β est le taux de transition de l'état 1 à l'état 0.

Ce modèle est formalisé par la matrice des taux instantanés Q, aussi appelé générateur :

$$Q = \begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix} \tag{1.1}$$

Cette matrice comporte deux paramètres :

- α : le taux de transition instantané de l'état 0 vers l'état 1.
- β : le taux de transition instantané de l'état 1 vers l'état 0.

À partir de cette matrice, on définit $P(t) = e^{Qt}$ la matrice des probabilités de transition entre les états pendant un temps t :

$$P(t) = \begin{pmatrix} \frac{\beta+\alpha e^{-\lambda t}}{\alpha+\beta} & \frac{\alpha-\alpha e^{-\lambda t}}{\alpha+\beta} \\ \frac{\beta-\beta e^{-\lambda t}}{\alpha+\beta} & \frac{\alpha+\beta e^{-\lambda t}}{\alpha+\beta} \end{pmatrix} \tag{1.2}$$

Où $\lambda = \alpha + \beta$ est le taux d'évolution moyen du modèle.

Cette matrice s'interprète comme suit :

- la probabilité de rester dans l'état 0 pendant un temps t est $\frac{\beta+\alpha e^{-\lambda t}}{\alpha+\beta} = P_{00}(t)$.
- la probabilité de passer de l'état 0 à l'état 1 en un temps t est $\frac{\alpha-\alpha e^{-\lambda t}}{\alpha+\beta} = P_{01}(t)$.
- la probabilité de passer de l'état 1 à l'état 0 en un temps t est $\frac{\beta-\beta e^{-\lambda t}}{\alpha+\beta} = P_{10}(t)$.
- la probabilité de rester dans l'état 1 pendant un temps t est $\frac{\alpha+\beta e^{-\lambda t}}{\alpha+\beta} = P_{11}(t)$.

En phylogénie, on utilise la matrice $P(t)$ pour calculer les probabilités de transition entre les états le long d'une branche. Le temps t correspond alors à la longueur de la branche. La difficulté est qu'avec $P(t)$ telle que nous l'avons définie, t est indissociable de λ. Autrement dit, la longueur de la branche est indissociable du taux d'évolution du processus. Pour résoudre cette difficulté, on normalise Q de façon à ce que l'unité de temps corresponde à un changement d'état par site sur la distribution stationnaire.

La distribution stationnaire correspondant au modèle binaire est la suivante :

$$\begin{cases} \pi_0 = \frac{\beta}{\alpha+\beta} \\ \pi_1 = \frac{\alpha}{\alpha+\beta} \end{cases}$$

La condition de normalisation est $\alpha\pi_0 + \beta\pi_1 = 1$, ce qui équivaut à $\frac{2\alpha\beta}{\alpha+\beta} = 1$. On normalise donc Q pour obtenir Q' :

$$Q' = \frac{\alpha+\beta}{2\alpha\beta}\begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix} \tag{1.3}$$

ce qui donne $P'(t)$:

$$P'(t) = \begin{pmatrix} \frac{\beta+\alpha e^{-\lambda t}}{\alpha+\beta} & \frac{\alpha-\alpha e^{-\lambda t}}{\alpha+\beta} \\ \frac{\beta-\beta e^{-\lambda t}}{\alpha+\beta} & \frac{\alpha+\beta e^{-\lambda t}}{\alpha+\beta} \end{pmatrix} \tag{1.4}$$

Où $\lambda = \frac{(\alpha+\beta)^2}{2\alpha\beta}$.

Les paramètres α et β étant dépendants du fait de la normalisation, on peut se ramener à un seul paramètre $\varkappa = \frac{\alpha}{\beta}$. Dans ce cas, $P(t)$ s'écrit :

$$P(t) = \begin{pmatrix} \frac{1+\varkappa e^{-\lambda t}}{\varkappa+1} & \frac{\varkappa-\varkappa e^{-\lambda t}}{\varkappa+1} \\ \frac{1-e^{-\lambda t}}{\varkappa+1} & \frac{\varkappa+e^{-\lambda t}}{\varkappa+1} \end{pmatrix} \tag{1.5}$$

Où $\lambda = \frac{(\varkappa+1)^2}{2\varkappa}$.

C'est cette paramétrisation qui est utilisée au chapitre 2.

1.2.2 Inférence

Plusieurs méthodes de reconstruction on été évoquées plus haut. Ces méthodes peuvent être appliquées à la reconstruction d'arbre de gènes, d'espèces, d'adjacences, ou de tout autre objet évolutif. C'est pour cette raison que j'ai choisi de les présenter dans une partie séparée. De même, il existe des modèles d'évolution pour tout type d'objet évolutif. Reconstruire la topologie d'un arbre consiste à hiérarchiser les relations de parentés entre les objets étudiés, ces objets pouvant être des séquences, des espèces, des adjacences, ou n'importe quel objet évolutif. Il existe différentes méthodes pour cela. Celle abordée dans ce manuscrit est le maximum de vraisemblance. Les modèles d'évolution probabilistes sont utilisés pour reconstruire des arbres avec une méthode de maximum de vraisemblance.

1.2.2.1 Le maximum de vraisemblance

La méthode du maximum de vraisemblance a été développée par Fisher dans les années 1920 [3]. Son application à la reconstruction phylogénétique est plus récente, le premier algorithme permettant de calculer une vraisemblance de manière efficace à partir de séquences nucléotidique ayant été développé par Felsenstein en 1981 [47]. Le principe de la reconstruction par maximum de vraisemblance est de trouver les paramètres qui maximisent la probabilité d'obtenir les états de caractères observés. Les paramètres incluent la topologie, les longueurs de branches de l'arbre, et les paramètres du modèle d'évolution utilisé. La reconstruction par maximum de vraisemblance implique (1) de savoir calculer la vraisemblance pour une topologie, des longueurs de branches et un modèle d'évolution à paramètres donnés, et (2) de trouver la topologie, les longueurs de branches et les paramètres du modèle qui maximisent la vraisemblance.

Dans un premier temps, nous allons voir comment calculer la vraisemblance d'un arbre étant donné un modèle, des longueurs de branches et une topologie. On prendra l'exemple d'un alignement de séquences nucléotidiques comprenant m sites. Soit D les états de caractères observés (les données) et T les longueurs de branches. La vraisemblance, notée L d'après l'anglais *likelihood*, est la probabilité d'obtenir les données étant données la topologie de l'arbre et les longueurs de branches :

$$L = Prob(D|T) \tag{1.6}$$

Si on fait l'hypothèse que les sites de l'alignement sont indépendants, la vraisemblance s'écrit alors :

$$L = \prod_{i=1}^{m} Prob(D^i|T) \tag{1.7}$$

Avec D^i les états du caractère observé au site i. Il suffit alors de savoir calculer la vraisemblance à un site L^i.

$$L^i = Prob(D^i|T) \tag{1.8}$$

La figure 1.29 représente une topologie avec les états observés à un site. Pour calculer la vraisemblance de l'arbre à ce site, étant données les longueurs de branches, il faut prendre en compte toutes les combinaisons d'états possibles pour les nœuds internes

de l'arbre :

$$L^i = \sum_x \sum_y \sum_z \sum_w Prob(A, C, C, C, G, x, y, z, w|T) \tag{1.9}$$

Où x, y, z et $w \in [A, C, G, T]$.

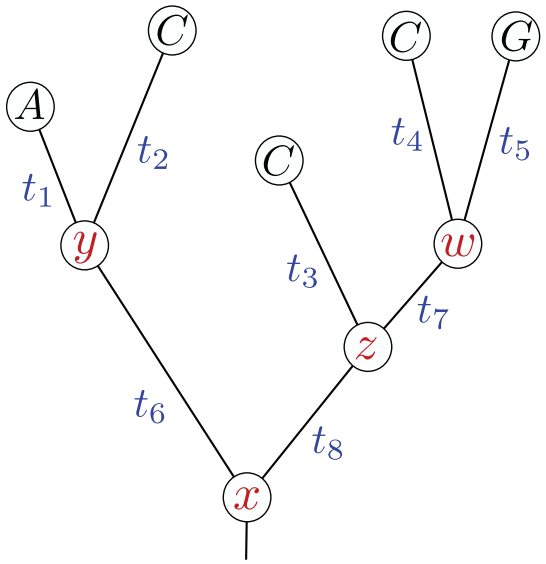


FIGURE 1.29 – Calcul de la vraisemblance. Les états de caractères observés sont en noir (A,C,C,C,G). Ceux inconnus sont en rouge (x,y,z,w). Les longueurs de branches sont en bleu (t_1...t_8). Adapté de [49].

Si on fait l'hypothèse que les branches sont indépendantes, on peut décomposer l'équation 1.9 :

$$L^i = \sum_x \sum_y \sum_z \sum_w Prob(x) Prob(y|x, t_6) Prob(A|y, t_1) Prob(C|y, t_2) Prob(z|x, t_8)$$
$$Prob(C|z, t_3) Prob(w|z, t_7) Prob(C|w, t_4) Prob(G|w, t_5) \tag{1.10}$$

La probabilité de x à la racine, $Prob(x)$, peut être obtenue de deux manières. Soit on considère que le modèle est stationnaire et que $Prob(x)$ est la probabilité de x à l'équilibre du modèle. Soit on estime $Prob(x)$ au cours de la reconstruction par maximum de vraisemblance. On sait calculer toutes les autres composantes de cette équation

avec la matrice des probabilités de transition du modèle d'évolution car $Prob(j|i,t) = P_{ij}(t)$. Cependant, même si le calcul de chaque composante est simple, le temps nécessaire au calcul d'une vraisemblance augmente considérablement lorsqu'on augmente le nombre de nœuds dans l'arbre, ou le nombre d'états possibles. Pour cette raison, la formule de la vraisemblance (1.10) n'est pas directement utilisée dans les implémentations du maximum de vraisemblance. À la place, Felsenstein a proposé l'algorithme d'élagage (*pruning*), basé sur le principe de la programmation dynamique [47]. L'idée de l'algorithme d'élagage est que certaines opérations sont redondantes et qu'il suffit de les effectuer une fois puis de réutiliser leur résultat de multiples fois. Si on écrit la formule de la vraisemblance en décalant les opérateurs de somme vers la droite, on obtient :

$$L^i = \sum_x Prob(x)(\sum_y Prob(y|x,t_6)Prob(A|y,t_1)Prob(C|y,t_2)) \\ (\sum_z Prob(z|x,t_8)Prob(C|z,t_3) \\ (\sum_w Prob(w|z,t_7)Prob(C|w,t_4)Prob(G|w,t_5))) \quad (1.11)$$

La structure de cette équation reproduit la topologie de l'arbre : (A,C)(C,(C,G)). L'algorithme d'élagage tire parti de cette structure en calculant la vraisemblance de manière récursive. En partant des feuilles, on remonte dans l'arbre en calculant à chaque nœud les *vraisemblances partielles* du nœud. Une vraisemblance partielle correspond à la probabilité d'obtenir les états de caractère observés chez les descendants du nœud, sachant l'état du nœud en question. Comme on doit considérer tous les états possibles pour le nœud, on calcule plusieurs vraisemblances partielles pour chaque nœud. En parcourant l'arbre de cette façon, le calcul est récursif. Sur l'arbre illustré sur la figure 1.30 , la vraisemblance partielle du nœud k dans l'état s et au site i s'écrit ainsi :

$$L_k^i(s) = (\sum_x Prob(x|s,t_l)L_l^i(x))(\sum_y Prob(y|s,t_m)L_m^i(y)) \quad (1.12)$$

Où l et m sont les nœuds descendants de k , x est l'état de l, y est l'état de m, t_l et t_m sont les longueurs des branches partant de k et menant respectivement à l et m, $L_l^i(x)$

est la vraisemblance partielle de l à l'état x et $L^i_m(y)$ est la vraisemblance partielle de m à l'état y.

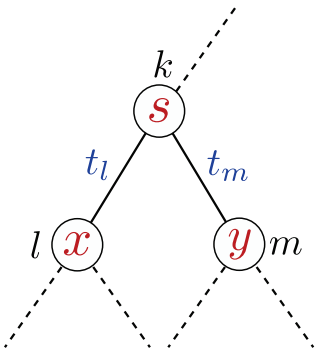


FIGURE 1.30 – En noir : nom des nœuds, en rouge : état des nœuds, en bleu : longueurs des branches.

Savoir calculer une vraisemblance étant données des valeurs de paramètres, comme nous venons de le faire est la première étape d'une reconstruction phylogénétique par maximum de vraisemblance. La deuxième étape consiste à rechercher les valeurs des paramètres qui maximisent la vraisemblance.

La reconstruction par maximum de vraisemblance était très peu utilisée avant les années 1990, à cause de son coût en temps de calcul [109]. Depuis, avec l'amélioration des performances des ordinateurs, elle est devenue la méthode standard en phylogénie. Parmi les implémentations les plus connues, on trouve Phyml [63] et RAxML [128].

L'algorithme d'élagage permet de calculer la vraisemblance d'un arbre de manière récursive, en parcourant l'arbre des feuilles vers la racine. Il est cependant parfois utile de pouvoir calculer des vraisemblances conditionnelles aux nœuds intermédiaires. Dans ce cas, si on se contente de parcourir le sous-arbre sous un nœud n des feuilles vers n, on n'exploite pas l'information contenue dans le reste de l'arbre pour calculer la vraisemblance à n. Pour calculer cette vraisemblance en prenant en compte le reste de l'arbre il faut utiliser une double récursion : dans un premier temps on calcule la vraisemblance à la racine avec l'algorithme d'élagage, et dans un deuxième temps on calcule les vraisemblances conditionnelles aux nœuds intermédiaires en utilisant un algorithme de calcul de vraisemblance basé sur le parcours de l'arbre de la racine vers les feuilles. Pour la deuxième étape, on peut utiliser l'algorithme développé par Boussau et Gouy

[23] pour les modèles non-réversibles. Dans le chapitre 2, c'est cette procédure qui est utilisée pour inférer l'état d'adjacences ancestrales. On dit alors qu'on infère des *probabilités a posteriori* pour les adjacences ancestrales.

1.2.2.2 Maximiser la vraisemblance avec la recherche locale

Dans l'exemple présenté ci-dessus, nous n'avons testé qu'une seule topologie et un seul ensemble de paramètres pour le modèle et les longueurs de branches. Pour reconstruire un arbre par maximum de vraisemblance, il faudrait en théorie explorer tout l'espace des topologies et toutes les valeurs des paramètres. Cavalli-Sforza et Edwards ont montré qu'il existaient pour n feuilles $\frac{(2n-3)!}{2^{n-2}(n-2)!}$ arbres racinés possibles. Avec cinq feuilles, on devrait donc tester 105 topologies. Avec dix feuilles, il faudrait en tester 34459425 et avec vingt environ $8.2 * 10^{21}$. Lorsque le nombre de feuilles augmente, il devient rapidement impossible de tester toutes les topologies. On utilise alors une heuristique pour explorer l'espace des topologies et trouver une solution optimale, au moins localement.

Le principe de des méthodes d'exploration de l'espace des topologies par recherche locale est le suivant : en partant d'une topologie initiale, on effectue des petites modifications aléatoires. À chaque modification, on évalue si la topologie modifiée est meilleure (plus vraisemblable) qu'avant la modification. Si c'est le cas on modifie l'arbre est on tente une autre modification pour améliorer encore la topologie. Si ça n'est pas le cas, on annule la modification et on en essaie une autre. Ce principe est en fait celui de la recherche locale *hill climbing* : on explore l'espace proche du point de départ et on essaie de se diriger vers un optimum en recherchant à chaque itération une solution un peu meilleure que la précédente. Après chaque mouvement, on regarde si la vraisemblance augmente ou pas et, éventuellement on recalcule les paramètres qui maximisent la vraisemblance. La procédure exacte dépend de l'implémentation.

Avec ce type d'heuristique, le risque est de parvenir à un optimum local, différent de l'optimum global. Mais l'optimum local n'est pas forcément une bonne solution, il peut-être très éloigné de l'optimum global. Et l'optimum global n'est pas forcément la vraie solution, il est seulement une solution optimale étant donné le système de coût utilisé pour évaluer les solutions. Appliqué à la reconstruction de la topologie, cela veut dire que l'on peut parvenir à une "bonne" topologie selon le modèle utilisé, mais cette topologie n'est pas forcément optimale, et ne correspond pas forcément à l'histoire évo-

lutive des objets. Il existe différentes manières de modifier localement la topologie. Les mouvements les plus couramment utilisés sont le *nearest neighbour interchange* (NNI) et le *subtree pruning and regrafting* (SPR) [109, 49].

La reconstruction par maximum de vraisemblance n'implique pas forcément d'optimiser tous les paramètres. On peut par exemple rechercher les valeurs des paramètres du modèle d'évolution et les longueurs de branches qui maximisent la vraisemblance, avec une topologie fixée à l'avance. C'est cette dernière application du maximum de vraisemblance qui est présentée dans le chapitre 2.

1.2.3 Évaluer et comparer des arbres

Comme on l'a vu avec les méthodes d'exploration de l'espace des topologies, la topologie optimale retenue est une estimation de la vraie histoire évolutive des objets étudiés. Il est donc utile d'avoir une mesure du support statistique de cette estimation. Il existe plusieurs méthodes pour obtenir un tel support. La plus utilisée en phylogénie est la technique du *bootstrap* [48]. Soit un alignement de séquences comportant n sites. La procédure du bootstrap se déroule comme suit :

1. On reconstruit l'arbre.

2. On tire aléatoirement et avec remise n sites de l'alignement. On répète cette étape x fois pour obtenir x alignements de n sites.

3. Pour chaque alignement ainsi ré-échantillonné, on reconstruit l'arbre. On obtient ainsi x arbres.

4. Pour chaque branche interne de l'arbre obtenu à l'étape 1, on compte le nombre d'arbres parmi x présentant cette même branche. On reporte le pourcentage obtenu sur la branche de l'arbre de l'étape 1.

Le principe derrière le bootstrap est de perturber le signal phylogénétique contenu dans les données (l'alignement) et de vérifier à quel point la topologie reconstruite est robuste à ces perturbations. Un branchement soutenu par un fort bootstrap est normalement interprété comme fiable. Cependant, dans certains cas, un fort bootstrap peut indiquer un biais systématique dans les données. Avec des données nucléotidiques, il

peut s'agir d'un biais de composition, de l'attraction des longues branches, et d'une hétérogénéité dans le temps des taux de mutation [89]. Dans un article de 2005, Delsuc et collaborateurs montrent ainsi l'exemple d'une phylogénie pour laquelle il est possible de reconstruire deux topologies différentes [40]. Les deux topologies sont contradictoires mais les branchements au cœur du conflit sont soutenus par de forts bootstraps dans les deux topologies. Les auteurs montrent que l'attraction d'une longue branche est responsable de cette inconsistance.

L'évaluation de la vraisemblance d'un arbre permet d'utiliser des méthodes d'évaluation alternatives au bootstrap : les tests de vraisemblance. Le principe général de ces tests est de comparer des topologies par une comparaison de leur vraisemblance afin de déterminer si elles sont statistiquement équivalentes ou si une topologie est significativement meilleure que la ou les autres. Dans le chapitre 3, nous verrons que les tests de vraisemblance sont utiles pour comparer des topologies reconstruites à partir d'informations différentes (voir 3.3).

Un des tests les plus ancien est le test de Kishino et Hasegawa (KH) [72]. Le KH est utilisé uniquement pour comparer deux topologies. Les deux hypothèses envisagées par le KH sont (1) l'hypothèse nulle que que les deux topologies peuvent générer les données de façon sensiblement équiprobable et (2) l'hypothèse alternative que l'une des deux topologies permet de générer les données avec une plus grande probabilité. Sous l'hypothèse alternative, une topologie est alors significativement plus vraisemblable que l'autre. Une restriction importante dans l'usage du KH est que les deux topologies comparées doivent être définies indépendamment des données. Il ne faut ainsi pas utiliser le KH lorsqu'une des deux topologies à comparer est la topologie reconstruite en maximum de vraisemblance. Utilisé dans ce contexte, le KH aura tendance à rejeter des topologies alternatives pourtant valables. On appelle cela un biais de sélection ; car sélectionner les topologies en fonction d'un critère basé sur les données (le maximum de vraisemblance par exemple) va biaiser le résultat du test.

Le test de Shimodaira et Hasegawa (SH) [124] est une alternative au KH qui permet de corriger le biais de sélection. Il permet la comparaison de topologies multiples, avec une étape de ré-échantillonnage basée sur le bootstrap. Il présente cependant deux contraintes : les topologies doivent être définies indépendamment des données et la topologie la plus vraisemblable pour chaque réplicat de bootstrap doit faire partie des topologies comparées. Ces deux contraintes semblent paradoxales. Mais si on consi-

dère le SH dans le cadre d'une reconstruction phylogénétique, le paradoxe est résolu en incluant tout l'espace des topologies possibles dans la comparaison. Il n'y a alors pas de biais de sélection et on est sûr que la topologie la plus vraisemblable pour chaque réplicat du bootstrap se trouve dans la comparaison. Le principe du SH est, pour chaque réplicat, d'estimer la topologie la plus vraisemblable à partir de l'alignement, et de comparer simultanément toutes les autres topologies à la topologie la plus vraisemblable. La principale limite du SH est qu'il faudrait en théorie inclure toutes les topologies possibles dans l'analyse. En pratique, cela n'est pas toujours possible quand le nombre de topologies est trop grand. On peut alors être tenté de sélectionner les topologies testées, ce qui entraîne des biais [109]. Il a par ailleurs été montré que les performances du SH se dégradent quand le nombre d'arbres testés est grand [109]. Dans ce cas de figure, le SH est souvent jugé trop conservateur.

Le test approximativement non biaisé (AU, pour *Approximately Unbiased*) [123] a été développé comme alternative au KH et au SH. Il est, comme le SH, basé sur une procédure de ré-échantillonnage de type bootstrap, la distinction étant que le nombre de sites échantillonnés est variable. Le AU teste autant de jeux d'hypothèses qu'il y a de topologies dans la comparaison. Par exemple, s'il y a x topologies testées, on teste pour chaque topologie i :

- H_0 : la topologie i est la meilleure topologie parmi les x topologies testées et H_1 : la topologie i n'est pas la meilleure topologie parmi les x topologies testées.

À la fin de la procédure, on obtient une probabilité seuil pour chaque topologie testée. Si la probabilité est inférieure au risque de première espèce choisi, alors la probabilité s'interprète comme le rejet de la topologie correspondante.

Les tests KH, SH et AU sont implémentés, entre autres, dans le logiciel CONSEL [125]. Les biais associés à chaque test sont discutés dans plusieurs articles, notamment celui de Goldman et collaborateurs [57].

1.3 Travail accompli dans cette thèse

Les deux facettes évoquées dans ce chapitre, évolution des séquences et évolution de l'architecture des génomes, ont très peu échangé au cours des décennies qui ont vu leur

développement. Ma thèse se situe cependant à leur interface. Dans le chapitre 2, je présenterai ainsi au moyen d'un article un modèle d'évolution concernant l'architecture des génomes. Ce modèle innove car il réunit trois caractéristiques principales :

- il utilise et s'inspire des arbres phylogénétiques pour reconstruire l'évolution d'adjacences.
- il s'agit d'un modèle probabiliste.
- il modélise explicitement des événements évolutifs affectant les gènes (duplications et pertes de gènes).

Dans le chapitre 3, je montrerai que modéliser l'évolution de l'architecture des génomes peut apporter une information sur la qualité d'arbres phylogénétiques. Les chapitres 2 et 3 présentent deux façons différentes d'utiliser la synténie, l'une en modélisant l'évolution d'adjacences et l'autre en analysant la linéarité des génomes ancestraux. La première approche peut être intégrée dans un modèle multi-échelle de l'évolution des génomes. La deuxième est plus difficile à intégrer dans un cadre probabiliste. Il est cependant possible d'utiliser cette information pour proposer des corrections sur des arbres de gènes. Dans un deuxième temps, on peut alors évaluer ces corrections avec des modèles d'évolution.

2

Évolution d'adjacences dans des arbres de gènes réconciliés

2.1 DeCo et Harpi : de la parcimonie à la vraisemblance

L'article présenté dans ce chapitre introduit le modèle Harpi (Histories of Adjacencies, Reconciliation and Probabilistic Inference). La méthode développée s'inscrit dans la lignée des méthodes reconstruisant des adjacences ancestrales présentées aux chapitre précédent. Elle emprunte un formalisme proche de DeCo, développé par Bérard et collaborateurs. Afin de bien appréhender le fonctionnement de Harpi, il est donc utile de présenter brièvement DeCo. DeCo reconstruit des adjacences ancestrales en prenant en compte les événements évolutifs affectant les contenus en gènes des génomes : duplications et pertes de gènes. Il requiert trois types d'information, donnés en entrée :

- des adjacences dans les génomes étudiés
- des arbres de gènes
- un arbre des espèces

La reconstruction d'adjacences ancestrales passe ensuite par ces étapes, réalisées séquentiellement :

- les arbres de gènes sont réconciliés avec l'arbre des espèces : chaque nœud est annoté avec une espèce actuelle ou ancestrale et un événement évolutif (spéciation,

duplication). Des nœuds supplémentaires, représentant les pertes de gènes, sont ajoutés. La méthode de réconciliation implémentée dans DeCo est une réconciliation parcimonieuse en termes de duplications et de pertes de gènes.

- les adjacences sont regroupées en *classes d'équivalence*. L'objectif de cette étape est de regrouper les adjacences qui peuvent être homologues. En première approximation, deux adjacences peuvent être homologues si les gènes aux extrémités sont homologues. Plus formellement, deux adjacences A-B et C-D font partie de la même classe d'équivalence si elles respectent les conditions suivantes :

  - Les gènes A et C sont dans un même arbre de gènes noté G_1, et les gènes B et D sont dans une même arbre de gènes noté G_2.
  - Si G_1 et G_2 sont deux arbres différents, alors il existe un gène E dans G_1 et un gène F dans G_2 tels que E et F appartiennent au même génome ancestral, E étant un ancêtre de A et C, et F étant un ancêtre de B et D.
  - Si G_1 et G_2 correspondent à un unique arbre, il existe deux gènes distincts E et F descendant du dernier ancêtre commun à A, B, C et D, tels que E est un ancêtre de A et C, et F est un ancêtre de B et D.

  Pour chaque classe, on définit l'adjacence potentielle entre E et F comme la racine.

- Pour chaque classe d'équivalence, on parcourt les deux arbres de gènes réconciliés des feuilles à la racine de la classe. Pour chaque paire de nœuds dans ce parcours, on calcule le coût de mettre et de ne pas mettre une adjacence ancestrale entre ces nœuds. Ce calcul prend en compte la nature des événements évolutifs associés aux nœuds. Une fois arrivés à la racine de la classe, on revient sur nos pas pour déterminer les adjacences ancestrales qui minimisent le coût total de la classe. Le résultat obtenu est un ensemble d'arbre d'adjacences, reconstruits selon le principe de maximum de parcimonie.

Harpi reste fidèle à DeCo jusqu'à l'étape de réconciliation. La construction des classes d'adjacence reste similaire, la seule différence étant le choix de la racine pour la classe. Le point de divergence entre les deux méthodes est la reconstruction des adjacences ancestrales. DeCo adapte les algorithmes de parcimonie sur des caractères bi-

naires de Fitch [50] et Sankoff [118]. Avec Harpi, nous avons adapté l'algorithme de Felsenstein (maximum de vraisemblance sur des caractères binaires) à la reconstruction d'adjacences ancestrales avec des arbres de gènes réconciliés. Ceci a nécessité le développement de nouvelles fonctionnalités, dont les principales sont un modèle binaire d'évolution avec DL, un algorithme de calcul de vraisemblance avec DL, et un arbre des adjacences possibles sur lequel le modèle peut être appliqué. Le tableau 2.1 résume les différences entre DeCo et Harpi.

TABLEAU 2.1 – Apports de Harpi par rapport à DeCo

	DeCo	Harpi
Principe	Maximum de parcimonie	Maximum de vraisemblance
Modèle	Système de coût avec DL	Modèle d'évolution avec DL
Paramètres	Coût d'un gain d'adjacence	$\varkappa$ = taux naissance / mort d'une adjacence
	Coût d'une perte d'adjacence	Fréquences à la racine
		Longueurs de branche de l'arbre des espèces
Objet	Arbres d'adjacences parcimonieux	Arbre des adjacences possibles
Sortie	Adjacences ancestrales	Probabilités a posteriori d'adjacences ancestrales
		Longueurs de branches optimisées
		$\varkappa$ optimisé
		Fréquences à la racine

La transition vers un cadre probabiliste a un coût : la méthode devient plus complexe et l'optimisation des paramètres prend du temps. Mais les avantages gagnés sont considérables. Harpi permet ainsi de reconstruire une histoire évolutive des adjacences ancestrales plus nuancée en calculant des probabilités a posteriori pour chaque adjacence ancestrale possible, et en estimant au passage des valeurs pour tous les paramètres de la méthode. Un autre avantage est l'intégration sur tous les scénarios possibles, alors que la parcimonie impose de choisir un scénario parmi de nombreux scénarios équivalents. À plus long terme, le passage à une méthode probabiliste ouvre le champ à une possible intégration dans des modèles à d'autres échelles.

L'article décrit Harpi plus en détail, montre un application à la phylogénie de 12 espèces de drosophiles et compare les adjacences ancestrales reconstruites à celles trou-

vées par DeCo sur le même jeu de données.

2.2 Harpi : un modèle probabiliste d'évolution d'adjacences dans des arbres de gènes réconciliés

Semeria et al. BMC Bioinformatics 2015, 16(Suppl 14):S5
http://www.biomedcentral.com/1471-2105/16/S14/S5

BMC Bioinformatics

RESEARCH Open Access

Probabilistic modeling of the evolution of gene synteny within reconciled phylogenies

Magali Semeria[1], Eric Tannier[1,2], Laurent Guéguen[1*]

From 13th Annual Research in Computational Molecular Biology (RECOMB) Satellite Workshop on Comparative Genomics
Frankfurt, Germany. 4-7 October 2015

Abstract

Background: Most models of genome evolution concern either genetic sequences, gene content or gene order. They sometimes integrate two of the three levels, but rarely the three of them. Probabilistic models of gene order evolution usually have to assume constant gene content or adopt a presence/absence coding of gene neighborhoods which is blind to complex events modifying gene content.

Results: We propose a probabilistic evolutionary model for gene neighborhoods, allowing genes to be inserted, duplicated or lost. It uses reconciled phylogenies, which integrate sequence and gene content evolution. We are then able to optimize parameters such as phylogeny branch lengths, or probabilistic laws depicting the diversity of susceptibility of syntenic regions to rearrangements. We reconstruct a structure for ancestral genomes by optimizing a likelihood, keeping track of all evolutionary events at the level of gene content and gene synteny. Ancestral syntenies are associated with a probability of presence.
We implemented the model with the restriction that at most one gene duplication separates two gene speciations in reconciled gene trees. We reconstruct ancestral syntenies on a set of 12 *drosophila* genomes, and compare the evolutionary rates along the branches and along the sites. We compare with a parsimony method and find a significant number of results not supported by the posterior probability. The model is implemented in the Bio++ library. It thus benefits from and enriches the classical models and methods for molecular evolution.

Background

Genomes evolve through processes that modify their content and organization at different scales, ranging from substitutions, insertions or deletions of single nucleotides to large scale chromosomal rearrangements. Extant genomes are the result of a combination of many such processes, which makes it difficult to reconstruct the big picture of genome evolution. Instead, most models and methods focus on one scale and use only one kind of data, such as gene orders or sequence alignments.

Models based on sequence alignments were first developed in the 1960's and underwent steady development until reaching a high level of complexity [1]. In a recent development, they have been extended to include gene content, modeling duplications, losses and transfers of genes with reconciliation methods [2,3]. Reconciled gene trees account for evolutionary events at both the sequence level and the gene family level. They thus yield a better representation of genome evolution and pave the way for approaches integrating other levels of information [4,5].

In parallel extant gene orders have long been used to infer evolutionary relationships between organisms and to reconstruct ancestral genomes [6-8]. Although the early stages of their development were computational challenges, methods based on gene orders gradually overcame theoretical and computational constraints so that they can now handle unequal gene content, multichromosomal genomes, whole genome duplications and dozens of genomes with large amounts of genes [9-11],

* Correspondence: laurent.gueguen@univ-lyon1.fr
[1]Laboratoire de Biométrie et Biologie Évolutive UMR CNRS 5558, Université Claude Bernard Lyon 1, 43 boulevard du 11 novembre 1918, 69622 Villeurbanne, France
Full list of author information is available at the end of the article

BioMed Central

© 2015 Semeria et al. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

and can be inserted into probabilistic frameworks [12-17].

All ingredients are present to integrate gene order and sequence evolution models, yet this leap has not been taken, mostly because of computational issues. Reconstructing gene order histories is often hard [18]. A computational solution to reconstruct gene orders and scale up with the size of datasets is to see a genome as a set of independently evolving adjacencies, *i.e.* the links between consecutive genes [19]. One can reconstruct ancestral gene orders following three main steps:

- Group potentially homologous adjacencies (they connect homologous pairs of genes)
- For each group, reconstruct the common history of adjacencies, by recovering ancestral ones
- Assemble the ancestral adjacencies in each ancestral species to obtain ancestral chromosomes

The assumption that adjacencies evolve independently allows quick computations at the second step: the size of the data can be an order of magnitude larger than without the assumption. But an optimization assembly step is required because of possible conflicts between adjacencies wrongly assumed independent [20].

Another difficulty is the integration of gene content dynamics. Often probabilistic solutions are limited to invariable gene content [12-14]. A solution is to encode altogether the presence and absence of genes and adjacencies as binary characters and use a binary sequence evolution model [15,16], but it lacks an evolutionary model of gene content and order dynamics. Gene profiles [21] or reconciled gene trees [22,10] are more promising for integration with sequence evolution models. They were mainly used with parsimony methods to reconstruct ancestral adjacencies, which makes it difficult to combine with a model at a different scale.

We propose a probabilistic model of adjacency evolution accounting for gene duplications and losses, using extant gene orders and reconciled gene trees. We base on the parsimony algorithm of DeCo [10] that we adapt to Felsenstein's maximum likelihood algorithm [1] with a birth/death process that models the evolution of adjacencies. We compute the most likely adjacencies in ancestral genomes and the quantity of gains and losses of adjacencies in all the branches of a species trees, thus providing an insight into the dynamics of rearrangements in these lineages. The model is implemented in Bio++ [23], the present form allowing at most one duplication node between two speciation nodes in gene trees. We compute the likelihood of gene orders in a set of 12 drosophila whose genomes are annotated in the Ensembl Metazoa [24] database. We optimize branch lengths in a species phylogeny and construct ancestral genomes. We compare the results with a parsimony approach, showing that while most adjacencies inferred by parsimony have a good probability, a non negligible proportion (> 11%) are not supported (posterior probability < 0.5).

Methods

Input

Species tree

A rooted species tree is a binary tree that describes the evolutionary relationships between organisms. The leaves of the tree are available species, internal nodes are ancestral species. The species tree has branch lengths indicating the quantity of expected evolution. Branch lengths can be also estimated as an output.

Adjacencies

An ordered set of genes is represented by a set of *adjacencies*, which are pairs of consecutive genes. For example, a genome A containing the sequence of genes a_1 – a_2 – a_3 – a_4 contains adjacencies a_1a_2, a_2a_3, and a_3a_4. Adjacencies are not oriented, meaning that a_1a_2 is equivalent to a_2a_1.

Gene trees

Genes are grouped into homologous families across genomes. The evolutionary history of each family is represented by a rooted gene tree. Gene trees are reconciled with the species tree (see precomputation below).

Principle

The principle is illustrated on Figure 1. It consists in reconstructing hypothetical ancestral adjacencies, modeling the evolution of adjacencies, computing a maximum likelihood of the model given the data, and computing the *a posteriori* probability of presence for each ancestral adjacency.

In this section, we give an overview of the main steps in our method. All these steps are detailed in the following sections, except the precomputation, for which we refer to [10].

- Precomputation: gene trees are reconciled with the species tree in order to minimize the number of duplications and losses (using DeCo [10]). It consists in annotating each internal node by an ancestral gene, together with the species it belongs to, and the evolutionary event (speciation, duplication, loss) taking place at the bifurcation. This determines a set of ancestral genes for all ancestral species. Gene losses are also annotated in the trees.
- Classify extant adjacencies so that every class can be handled independently. Inside each class, two gene families with two trees are involved and all adjacencies have an extremity in each family.
- For each selected pair of gene families, construct a tree, called the *tree of possible adjacencies*. Its nodes

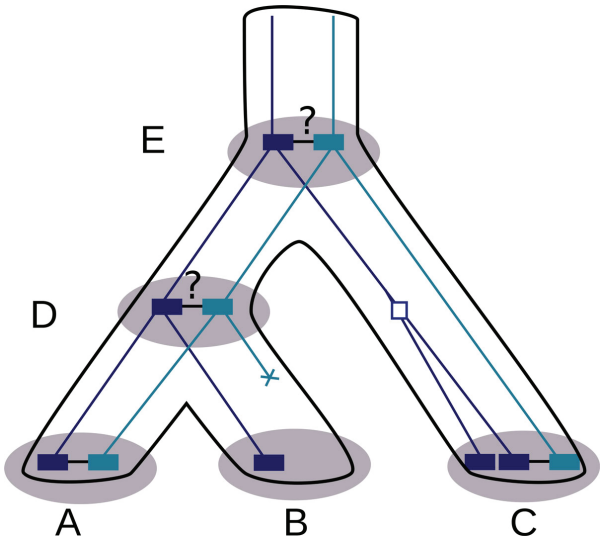


Figure 1 Principle of the method. Given two gene trees (dark blue tree and light blue tree) reconciled within a species tree (black tree), and sharing adjacencies in some extant species (species A and C), we reconstruct hypothetical ancestral adjacencies (in species D and E) using a model of evolution and maximum likelihood algorithm. Our method allows for losses (cross in light blue tree between species B and species D), and duplications (empty square in dark blue tree) of genes.

are all the couples of nodes from each gene tree, which are in the same extant or ancestral species (the speciation nodes), plus some duplication nodes; the leaves are labeled with the pattern of presence/absence of the possible adjacencies in the data.

- Compute, between successive nodes of this tree, the probability of presence or absence of the adjacency using the model of evolution described below.
- Compute the likelihood of the adjacency given the observed adjacencies.
- Compute *a posteriori* probabilities of presence of ancestral adjacencies.

The likelihood computation for one adjacency tree allows to obtain a likelihood for the whole dataset by multiplying all likelihoods, considered as independent, and to optimize parameters. These can concern branch lengths on the species tree, or a law of differential fragility for different genome sites, modeling different susceptibility to rearrangements among chromosomal regions [25,26].

Adjacency classes

We first reduce the problem to two gene trees, without loss of generality, by classifying adjacencies. Reconciled gene trees define ancestral genes of ancestral species. A necessary condition for an adjacency i_1i_2 to be an ancestor of a_1a_2 is that i_1 is an ancestor of a_1 and i_2 an ancestor of a_2. By the same idea a necessary condition for adjacencies a_1a_2 and b_1b_2 to be homologous is that there is a common ancestor i_1 of a_1 and b_1, and a common ancestor i_2 of a_2 and b_2, such that i_1 and i_2 are in the same species. This condition for homology is an equivalence relation on all extant adjacencies, which can be clustered and treated by equivalence classes of homology. To a class we can associate i_1 and i_2 the most ancient distinct common ancestors of all adjacency extremities in the class. So every adjacency in the class has an extremity which is a descendant of i_1 and an extremity which is a descendant of i_2. Without loss of generality we can work with the two sub-trees rooted at i_1 and i_2.

Trees of possible adjacencies

We now suppose that we have G_1 and G_2 two reconciled gene trees with some leaves of G_1 involved in adjacencies with some leaves of G_2. Each node n in G_1 and G_2 is annotated with an event (speciation, duplication, loss) and a species $S(n)$. Take each pair of nodes i_1i_2, where i_1 and i_2 are speciation nodes associated with the same ancestral species s, $i_1 \in G_1$ and $i_2 \in G_2$. Since $S(i_1) = S(i_2)$ and adjacencies exist between leaves of G_1 and leaves of G_2, i_1i_2 is called a *possible adjacency*.

All possible adjacencies define nodes of the tree of possible adjacencies, in which duplication nodes can be added, as explained below.

If i_1i_2 is a possible adjacency such that $S(i_1) = S(i_2) = s$, let s_1 and s_2 be the two children of s in the species tree. There is a descent path in the tree of possible adjacencies from i_1i_2 to all possible adjacencies j_1j_2 in s_1 such that i_1 is an ancestor of j_1 and i_2 is an ancestor of j_2, and a similar independent path from s to s_2. If there is no duplication node between i_1 and j_1 and i_2 and j_2, then this path is a single edge. If there is at least one duplication node between i_1 and j_1 or i_2 and j_2, then the path from i_1i_2 to j_1j_2 has two edges, one between i_1i_2 and d, a new duplication node, and one from d to j_1j_2. The node i_1i_2 always has only two descendants, but the node d can have an arbitrary number, according to the number of duplications in the gene lineages.

Loss of one or both genes involved in the adjacency in a branch leading to a species s' leads to the loss of the adjacency in s'. In this case, a *loss leaf* of the tree of possible adjacencies is constructed. An example of construction of a tree of possible adjacencies for two reconciled gene trees is drawn in Figure 2. Once each pair of nodes i_1i_2 has been considered, the resulting tree is the tree of possible adjacencies for G_1 and G_2 on which we can apply a model of evolution.

Model of evolution

We consider possible adjacencies as evolutionary objects in a binary alphabet. An adjacency can either be present

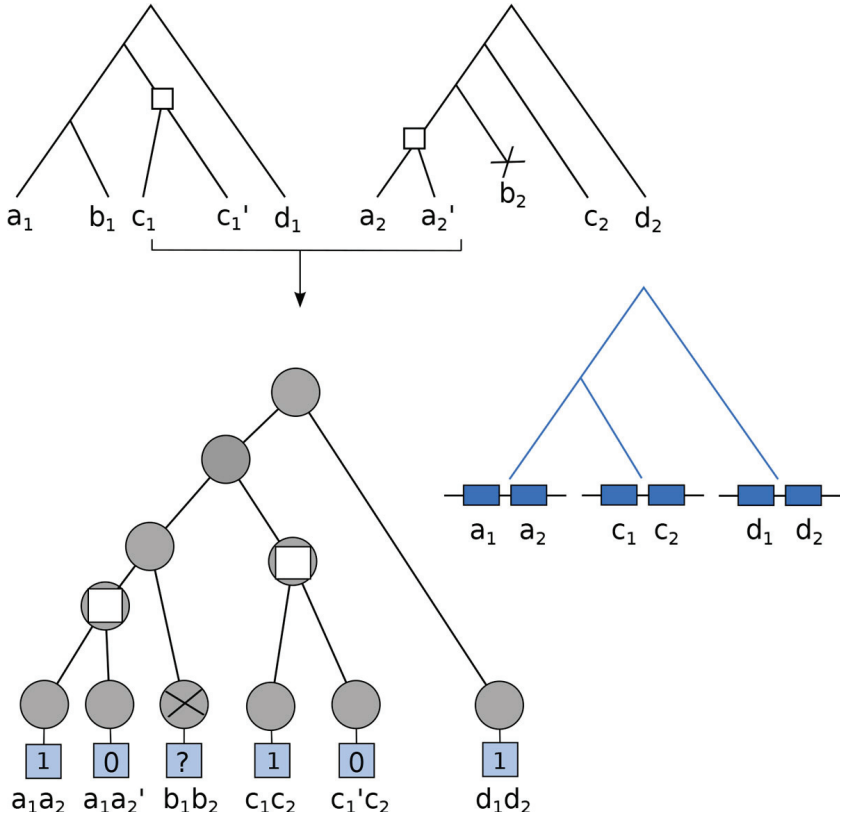


Figure 2 The tree of possible adjacencies. A tree of possible adjacencies is constructed from two reconciled gene trees (top trees). The nodes of the tree are annotated as speciation nodes (grey nodes), duplication nodes (white squares), or losses (crossed grey nodes). A binary state is attributed to each leaf according to the presence/absence pattern of the adjacency in extant species (light blue square). The true evolutionary history of the adjacency is represented on the blue tree. An adjacency exists between genes a_1 and a_2, c_1 and c_2, d_1 and d_2. It is absent between genes a_1 and a_2' and genes c_1' and c_2. If one extremity of the adjacency is lost (species B), the adjacency node is given an undefined state "?".

(state 1) or absent (state 0) in a genome. The transition rate matrix for the birth/death process which describes the evolution of a binary object is:

$$Q = \begin{pmatrix} -\frac{\kappa+1}{2} & \frac{\kappa+1}{2} \\ \frac{\kappa+1}{2\kappa} & -\frac{\kappa+1}{2\kappa} \end{pmatrix} \quad (1)$$

Where κ is the rate of $0 \rightarrow 1$ (gain of an adjacency) over the rate of $1 \rightarrow 0$ (loss of an adjacency). Probabilities of transition between two states separated by a amount t of time can be computed using a classical binary substitution model:

$$P(t) = \begin{pmatrix} \frac{1+\kappa e^{-\lambda t}}{\kappa+1} & \frac{\kappa-\kappa e^{-\lambda t}}{\kappa+1} \\ \frac{1-e^{-\lambda t}}{\kappa+1} & \frac{\kappa+e^{-\lambda rt}}{\kappa+1} \end{pmatrix} \quad (2)$$

Where $\lambda = \frac{(\kappa+1)^2}{2\kappa}$.

In the case when there is no duplication in the two gene trees, likelihoods can be computed directly from the tree of possible adjacencies (which itself has no duplication nodes) with Felsenstein's algorithm [1].

An adjacency can be lost because of a rearrangement ($1 \rightarrow 0$), or because at least one of the two adjacent genes is lost. In the first case, the state of the leaf in the tree of possible adjacency is simply 0. In the second case, we assign an undetermined state ? to the loss leaf in the tree of possible adjacencies to differentiate it from a loss due to a rearrangement. We do not compute probabilities of transition for branches leading to these nodes.

In the case when there are duplication nodes, we write the probabilities according to a model of evolution of adjacencies in presence of duplications: when one gene belonging to an adjacency is duplicated, the adjacency is transmitted to one of the two copies of the gene. This is always verified, whether the duplication is tandem or remote. For example, consider a gene i_2 involved in an

adjacency $i_1 i_2$ in species I with a gene i_1. In species A (descendant of species I), i_1 has one descendant a_1, whereas i_2 is duplicated, giving two copies a_2 and a'_2. If the duplication is in tandem it leads to the gene order $a_1 a_2 a'_2$, and the only adjacency conserved with a_1 is $a_1 a_2$. Otherwise it leads to the gene order $a_1 a_2 \ldots a'_2$ and again only $a_1 a_2$ is conserved. Note nevertheless that the adjacency $a_1 a'_2$ can appear later in the phylogeny following a rearrangement.

Between two speciation events, we have no date for duplication events. We argue that fixing a date, for example with gene branch lengths, would be a mistake as the position of a duplication between two speciations influences the transition probabilities. Besides, the probabilistic approach means that we can account for all possible dates. Hence we compute an average transition probability for the duplicated branch over all the moments on the branch of the species where this duplication could have occurred. To do this, we integrate the transition probabilities $P(t)$ uniformly over the length of this branch. Depending on the date of the duplications, the probabilities of the several resulting adjacencies are more or less linked. Hence, the integrated transition probability is no longer from one adjacency to another adjacency, but from one adjacency to the set of all the possible adjacencies that result from the duplication. In the previous example (one duplication), the transition probability is from $i_1 i_2$ to $((a_1 a_2, a_1 a'_2)$. We can fully model such a process as several processes in parallel. If Q is the generator of the binary model, $Q \otimes Q \otimes \ldots \otimes Q$ is the generator of the whole process, where $\otimes$ is the Kronecker product. Here, from a single Q generator at the beginning of the branch, along the branch each event of duplication gives rise to a larger Kronecker product. From a computational point of view, the whole parallel process is considered all along the branch, but just a subset of the transition probabilities is used.

We restrict here the description of the model to the case when there is at most one duplication node between two speciation nodes in the gene trees, which means that in the tree of possible adjacencies, duplication nodes have at most four descendants (because a gene duplication would have occurred in each gene tree). However, in case of several duplications, the same principle holds, with much more complicated formula.

One duplication

If there is one duplication in one gene tree (from a to a_1 and a_2) and no duplication in the other, then in the non duplicated branch probabilities are settled with the matrix P. The duplicated branch has a length drawn from the uniform distribution on the non duplication branch length, because it starts from the duplication. So the average transition matrix on the duplicated branch is:

$$N^1(t) = \frac{1}{t}\int_0^t P(\tau)\, d\tau \tag{3}$$

As in the duplicated branch there is no adjacency (state 0) at the moment of the duplication, we are only interested by the $(0, z)$ components of $N^1(t)$, $z \in [0, 1]$. Calculating the integral yields:

$$N^1_{0,0}(t) = \frac{\kappa - \kappa e^{-\lambda t} + \lambda t}{(\kappa + 1)\lambda t} \tag{4}$$

$$N^1_{0,1}(t) = \frac{\kappa e^{-\lambda t} - \kappa + \kappa \lambda t}{(\kappa + 1)\lambda t} \tag{5}$$

Let x be the state of adjacency $i_1 i_2$, y the state of $a_1 a_2$ and z the state of $a_1 a'_2$, $(x, y, z) \in [0, 1]^3$. Assuming that $a_1 a'_2$ is on the duplicated branch, the overall transition probabilities from x to y and z are given by $P_{x,y}(t) \times N^1_{0,z}(t)$.

The two choices for the duplicated branch are considered during the computation of the likelihood.

Two duplications

If both i_1 and i_2 are duplicated, we assume that both duplications are independent. Note that with this assumption, we do not model the case of joint duplications, where a fragment of chromosome is duplicated (i.e. several consecutive genes are duplicated following a single duplication event). Without loss of generality, we assume in the computation that one duplication occurs after the other. The average transition matrix integrated uniformly along both branches is:

$$N^{11}(t) = \frac{2}{t^2}\int_{u=0}^{t} P(u) \otimes \int_{v=0}^{u} P(v) \otimes P(v) dv du \tag{6}$$

Since, as before, only one gene pair inherits the adjacency, we are only interested by the $(., 0, 0, 0) \rightarrow (., ., ., .)$ components of $P(t) \otimes N^{11}(t)$ (Figure 3).

Likelihood computation

Likelihood is computed in the rooted tree of possible adjacencies in a bottom-up way. From here, we describe adjacency nodes with single letters for better clarity. We denote as D_i the data that is below node i.

Take a speciation node i in the tree, which descendants are nodes j and k (with branch lengths respectively t_1 and t_2). Let $x, y, z \in [0, 1]$ be the respective states of i, j, k. We compute the partial conditional likelihoods of D_i in the classical way:

$$L(D_i|x) = \left(\sum_{y=0}^{1} P_{xy}(t_1) L(D_j|y)\right) \cdot \left(\sum_{z=0}^{1} P_{xz}(t_2) P(D_k|z)\right) \tag{7}$$

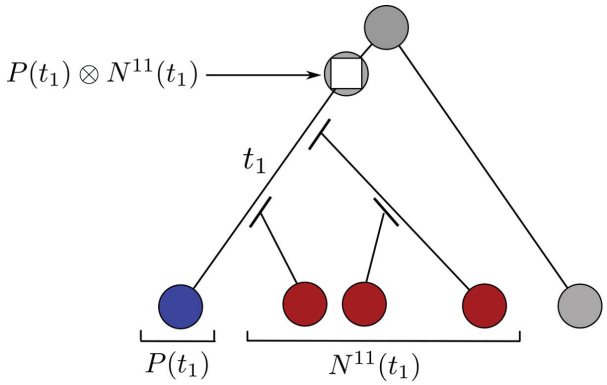


Figure 3 Model of evolution with duplications. A duplication in the same species in each of the two gene trees leads to a duplication node with four children (white square) in the tree of possible adjacencies between the two gene trees. Immediately after a duplication event, the adjacency is broken for the duplicated branch (branch leading to the rightmost red node). The second duplication leads to the simultaneous apparition of two other branches (leading to left and middle red nodes). The adjacency is also broken at the beginning of these two branches. The probabilities of transition between the duplication node and its four children are then given by the (., 0, 0, 0) → (., ., ., .) components of $P(t_1) \otimes N^{11}(t_1)$. In the likelihood computation, all positions for the blue and red nodes are considered.

Now, let i be a duplication node with two children j and k. Since it concerns only one branch in the species tree, there is a unique branch length t involved. We defined the model of evolution such that the contribution of one child is included using the basic transition matrix $P(t)$ and the contribution of the other child (the child on the duplicated branch) is included using the transition matrix $N^1(t)$. The partial likelihoods of i can then be computed by allowing the equal possibility that either j or k is on the duplicated branch:

$$L(D_i|x) = \frac{1}{2}\sum_{yz} L(D_j|y)P_{xy}(t)L(D_k|z)N^1_{0z}(t) + \frac{1}{2}\sum_{yz} L(D_k|y)P_{xy}(t)L(D_j|z)N^1_{0z}(t) \quad (8)$$

If we generalize this problem, computing the partial likelihoods of a duplication node i means exploring the combinatorics of possible states for i's children and the combinatorics of attributing the duplicated branch(es) to the children. Take a duplication node i with n speciation nodes as descendants in the same species. Each node is in a binary state, which means that there are 2^n combinations of states for i's children. We could explore all these combinations to compute i's likelihood but binary characters quickly lead to redundancies in the computation. We can avoid some of these redundancies and reduce the space of exploration by defining *patterns*. A *pattern* is an unordered set of 0s and 1s. There are n+1 possible patterns representing the states of i's children. For each pattern p, we can compute the pattern's pseudo likelihood by exploring all its possible orders (i. e. all the possible ways of ordering the 1s and 0s in the pattern):

$$L(D_i|p) = \sum_Y \prod_{c \le n} L(D_c|Y_c) \quad (9)$$

where Y is one possible order of p. If i has n children, Y is a vector of n binary characters representing the states of the n children. Y_c is thus the c^{th} element of Y and D_c the data below the c^{th} child of i.

We define the *weight* $\omega(p)$ of the pattern p as the number of possible orders for $p : \omega(p) = \binom{n}{N}$, with N the number of 1s in p. We give the generalized formula for computing the partial likelihood of i when i has four children (n = 4, which means that the only concerned integrated transition matrix is $N^{11}(t)$):

$$L(D_i|x) = \sum_p \frac{\omega_p}{2^n}(L(D_i|p)\sum_{Y \in p}(P \otimes N^{11})_{(x,0,0,0)\to Y}(t) \quad (10)$$

This formula is valid for any number of children for the duplication node i, provided N^{11} is replaced by an appropriate matrix.

Ancestral adjacencies reconstruction

Ancestral states, that is, posterior probabilities of presence of adjacencies in the tree of possible adjacencies, are reconstructed by a top-down (from the root to the leaves) algorithm following the the bottom-up likelihood computation algorithm. In the top-down likelihood computation algorithm, we compute the conditional likelihoods of each node i according to the conditional likelihood of the data below it (D_i), and to the conditional likelihood of the data that is on the other part of its father f, D_f , and to the conditional likelihood that is below the brothers of i (say one brother i').

If father f of node i is a speciation node:

$$L(D|y) = L(D_i|y)\sum_{x=0}^{1} P_{xy}(t).L(D_f|x) \cdot \sum_{z=0}^{1} P_{xz}(t')L(D_{i'}|z) \quad (11)$$

where y is the state of i, x is the state of f, z the state of i' and t' the length of the branch from f to i'.

If father f of node i is a duplication node with one duplication (i.e. two sons i and i'), the likelihood of node i is the average of both scenarios:

$$L(D|y) = L(D_i|y).\frac{1}{2}\sum_{x=0}^{1} L(D_f|x).\sum_{z=0}^{1}(P_{xy}(t)N^1_{0z}(t') + P_{xz}(t').N^1_{0y}(t))L(D_{i'}|z) \quad (12)$$

And the equivalent to the case of two duplications in the bottom-up algorithm is achieved by computing i's

partial likelihoods when *i*'s father is a duplication node with four children *i, i', i'', i'''*, and the likelihood is an average of four scenarios:

$$L(D|\gamma) = L(D_i|\gamma).\frac{1}{4}\sum_{x=0}^{1} L(D_f|x). \sum_{wzu} (P_{xy}(t).N^{11}_{0,0,0\to w,z,u}(t) + P_{xw}(t)N^{11}_{0,0,0\to \gamma,z,u}(t) + P_{xz}(t)N^{11}_{0,0,0\to w,\gamma,u}(t) + P_{xu}(t)N^{11}_{0,0,0\to w,z,\gamma}(t)) L(D_{i'}|w).L(D_{i''}|z).L(D_{i'''}|u) \quad (13)$$

From these conditional likelihoods, *a posteriori* probabilities of presence of adjacencies can be computed. The result is, for each ancestral species, a set of adjacencies associated with probabilities of presence. Transforming it into a *bona fide* gene order necessitates finding a subset of probable adjacencies in which one ancestral gene can be adjacent to only two others. Efficient methods exist [20] to do so, but they ignore the main source of possible conflict between adjacencies when they are seen as independently evolving characters: errors in gene trees [27]. So in general we prefer presenting a set of adjacencies associated with probabilities, and leave open the way of choosing among them and/or correcting the input data to avoid conflict.

Implementation and availability

We implemented the model of evolution and the likelihood calculation algorithm in the Bio++ library (http://biopp.univ-montp2.fr/). The algorithm that builds the trees of possible adjacencies was implemented in a separate program which also uses Bio++. Reconciliation was performed with DeCo [10]. All the analytical formulas in our model were computed using Maxima (see Additional file 1). These programs are available upon request to the authors.

Results

Dataset

We selected 12 drosophila species from the Ensembl Metazoa [24] database. We used the species tree from [28], the gene trees and the chromosomal locations from Ensembl Metazoa. We pruned the gene trees to keep only the drosophilae clade, and reconciled them with the species tree using [10]. We reduced the dataset to the 9223 gene trees with at most one duplication between two speciation nodes in reconciled gene trees. We built a set of extant adjacencies by connecting consecutive genes in the reduced dataset, provided they were on the same chromosome or scaffold. We built 13059 trees of possible adjacencies from this set of reconciled gene trees and extant adjacencies. By maximum likelihood, we optimized the branch lengths of the species tree using our model of evolution, from the 3608 trees of possible adjacencies without any duplication (Figure 4). Optimizing branch lengths over many trees remains computationally intensive, especially for trees with several duplications (then the combinatorics increases). The choice of the sample to optimize from was thus a trade-off between

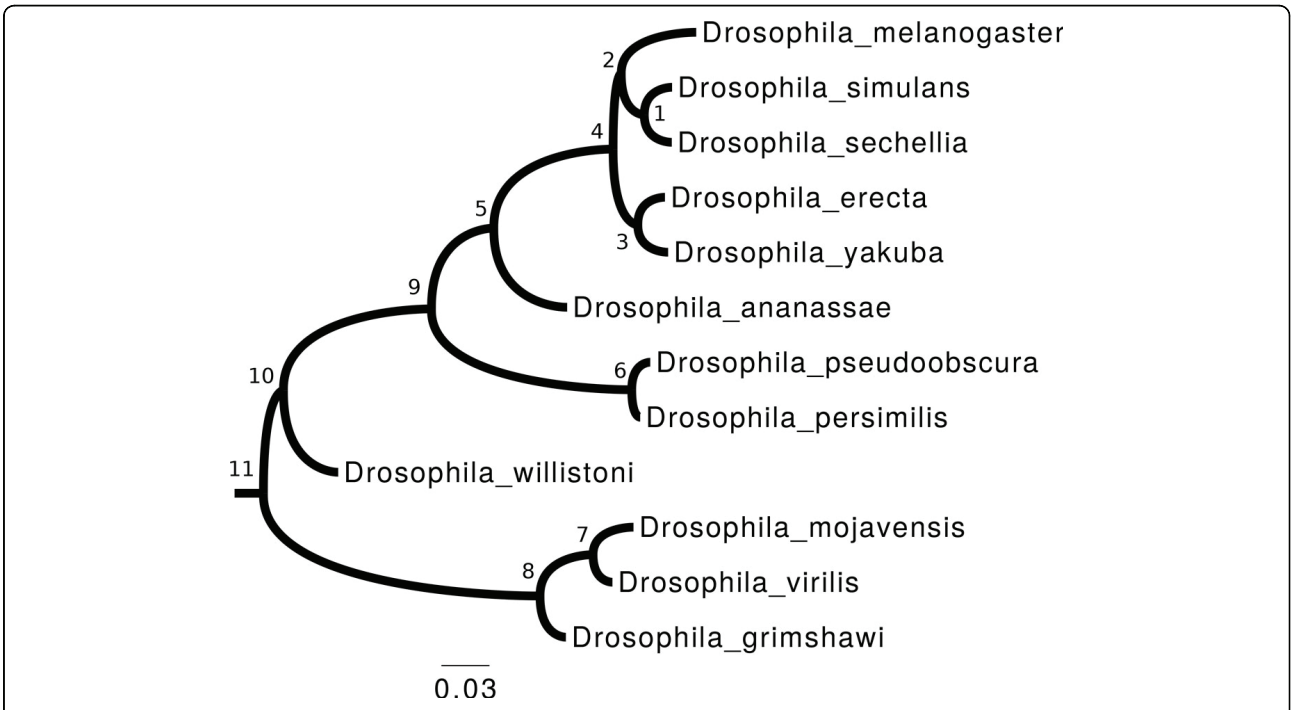


Figure 4 Drosophila phylogeny. The 12 Drosophila species tree with branch lengths optimized according to the model and the synteny data.

accuracy and computational cost. While we optimized branch lengths, we also optimized the model's parameters in a non-stationary way.

Note that the drosophila genomes are not all perfectly assembled and some are fragmented in several hundred contigs. So all the signal does not have to be interpreted as rearrangements, but some of it is due to the absence of adjacencies in extant genomes.

Ancestral adjacencies

We computed posterior probabilities of presence and absence for all possible ancestral adjacencies, given the optimized branch lengths. We report in Table 1 the number of genes and adjacencies in extant and ancestral species. Note that the difference between the number of genes and adjacencies in extant species gives the number of chromosomes or scaffolds. This goes from the well assembled *melanogaster* genomes in 8 scaffolds to *simulans* with 445 scaffolds, with all intermediaries. Despite the fact that assembly is incomplete, we have enough adjacencies in the dataset to make a signal for the reconstruction of ancestral adjacencies. And indeed, 54222 adjacencies with posterior probability > 0.9 are proposed. The signal is weaker for ancient species, as in ANC10, with only 2360 adjacencies for 8026 genes, depicting a very fragmented ancestral genome.

The "degree" column in Table 1 shows that in general less than 4% of the genes harbor a conflicting signal with more than 2 attached adjacencies having posterior probability > 0.9. While this remains a high rate of error, it means that most of the supported signal constitutes linear ancestral contigs or chromosomes. The conflict is variable according to the lineages. A surprisingly high amount of conflict arises for the ancestor of *yacuba* and *erecta*, predicted as recent. Perhaps this reflects an ambiguity in the species tree which precisely at this place is debated [29]. It seems that rearrangements support an alternative topology.

Comparison with parsimony

We compare the results with those obtained by [10] (DeCo software) on the same data (Figure 5). DeCo reconstructs ancestral adjacencies according to a parsimony principle, whereas we reconstruct all possible ancestral adjacencies along with a posterior probability of presence for each one. Most of the adjacencies reconstructed by

Table 1 Statistics of extant and ancestral genomes in the drosophila dataset.

Extant species	genes	adjacencies	coverage
melanogaster	6410	6402	47%
simulans	7195	6750	50%
sechellia	7551	7261	48%
erecta	6961	6910	49%
yacuba	7313	7058	49%
ananassae	6558	6459	47%
pseudoobscura	7280	7007	48%
persimilis	7361	7025	47%
willistoni	6236	6063	43%
mojavensis	6484	6403	48%
virilis	6512	6437	48%
grimsawi	6538	6220	46%
Ancestral species	**genes**	**adjacencies > 0.9**	**genes with more than 2 adjacencies**
ANC1	8054	7164	578
ANC2	8364	5422	164
ANC3	8696	7529	1348
ANC4	9455	3746	113
ANC5	7564	5021	160
ANC6	7242	6117	58
ANC7	6677	6184	210
ANC8	6954	5777	413
ANC9	8816	2872	47
ANC10	8026	2360	24
ANC11	7157	2030	3

Column "genes" is the number of genes in the dataset. Column "coverage" is the proportion of the genes in the dataset to the total number of genes annotated in Ensembl. Column "adjacencies" is the number of adjacencies in an extant genome or adjacencies in ancestral genomes with a posterior probability > 0.9. Column "genes with more than 2 adjacencies" is the number of genes involved in more than 2 adjacencies of posterior probability > 0.9. This value is reported only for ancestral genomes, as in extant all genes have 0, 1 or 2 neighbors by definition.

DeCo are given a high probability of presence according to our model (70% have a support > 0.9). Interestingly, a few of them are given low probabilities of presence (11% have probabilities of presence < 0.5), suggesting that our model could bring a finer understanding of the evolution of these adjacencies. Figure 5 shows the distribution of posterior probabilities, as computed by our model, of all the possible adjacencies (in grey), and of all the adjacencies inferred by parsimony (in red).

We always reconstruct more ancestral adjacencies than DeCo because DeCo reconstructs ancestral adjacencies up to the last common ancestor of an adjacency class, whereas we reconstruct possible ancestral adjacencies up to the most ancient ancestor of an adjacency class. This explains why many possible ancestral adjacencies have low or no support in the presence/absence pattern at the leaves.

Discussion

Probabilistic models of evolution have at least four advantages over parsimony approaches: they provide more accurate results in presence of many mutations; they provide a natural support scheme of the results in the form of a probability of ancestral states; the likelihood is computed by an integration over all scenarios rather than choosing only one, even if optimal; and several models at different scales of the genome can be integrated.

But most probabilistic models of gene order evolution are computationally intractable on large datasets, working with too large state spaces. Coding gene order by binary characters is a solution, like for many characters characterized by their presence or absence. Then it is possible, like in [30], to use a standard model of binary sequence evolution to achieve a probabilistic reconstruction of phylogenies and ancestral gene orders based on the presence/absence of adjacencies in extant species. This way can handle unequal gene content but does not model the processes of joint evolution of gene content and order, and has to simplify the data to make it fit into standard models. As a result a part of the understanding of genome evolution remains out of reach.

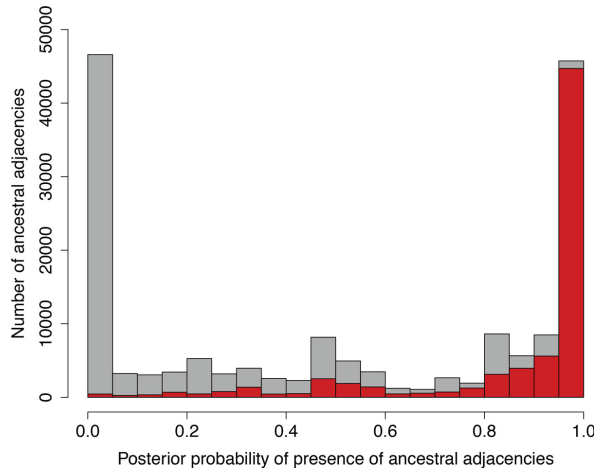


Figure 5 Ancestral adjacencies reconstruction with our method and with DeCo. The posterior probabilities of presence of ancestral adjacencies reconstructed with our model (in grey). In red the part reconstructed by a parsimony method.

This is why we put some efforts in a model of gene neighborhood evolution handling complex histories of genes depicted by their reconciled phylogenies.

We gain several advantages. For example the model allows to follow a pattern of descent of adjacencies. Links between genes evolve, just as genes evolve too. This can be used to detect the positional orthology (orthology of a gene as a locus, in addition to a sequence) when a gene is duplicated in an asymmetric way [31] - not in tandem, so that from the loci point of view, only one duplicate is a descendant of the unique copy before duplication. Here we allow any kind of duplication, symmetric or not, but in any case an adjacency is transmitted to one copy. In the case of a tandem duplication, this does not yield an asymmetry for the genes, because a gene has two adjacencies, and the two can transmit a descendant to a different copy in the case of a tandem duplication. But in the case of an asymmetric duplication, the two adjacencies are transmitted to the same copy of a gene and a positional homolog is detected.

We also keep track of the evolutionary events that can be responsible for the gain and loss of an adjacency. For example an adjacency can be lost because one of the genes is lost, or because of a rearrangement. It is two different reasons for an adjacency to be absent, and we are able with a model to differentiate both cases.

We found that a significant number of adjacencies inferred by parsimony on a drosophila dataset are not supported by a probabilistic model. It corroborates the usual findings in evolutionary models each time reasonably distant species are compared, whether it is sequence evolution [1], gene content evolution [32], or gene order evolution [12].

There are still several limitations to this work. For the moment the computation time is one of them, the efficiency of optimization algorithms coupled with our model allowed us to work only on a small fixed phylogeny. Theoretically we could even infer phylogenies, coupling a model of sequence evolution and such a model of genome organization evolution, but it will necessitate algorithmic progresses. Another limit is that our current implementation only handles independent duplication events, although we are also developing a model for joint duplications. Finally, the possible presence of many duplications

yields intricate integrals difficult to solve analytically, if we want to stick with exact solutions integrating over their position in a branch. Numerical approximations or simplifying hypotheses have to be incorporated. For the moment families with many duplications are filtered out.

Conclusions

The present model is a proof of concept that it is possible to handle whole genomes of dozens of species, including genes with complex histories, into a probabilistic model for gene organization.

We open a path that has many possible continuations:

- Handle joint duplications of two consecutive genes as a single duplication event.
- Handle more than one gene duplication between two gene speciations.
- Handle horizontal gene transfer (a parsimonious framework is available [33]).
- Jointly infer probabilistic presence and absence of genes and gene neighborhoods, using conditional probabilities mixing two models.
- Integrate the model into an integrative probabilistic model of genome evolution, handling both sequence evolution and gene content evolution, like Phyldog [27].
- With this integration the model can be used to infer species phylogenies, or at least in the current state of computational complexity, to test among a small number of species phylogenies. For example we will test two different alternative drosophila species tree topologies according to the likelihood of our model, and according to the coherence of ancestral genomes (the linear organization of genes along chromosomes).
- Use this model to detect highly variable sites by correlating variable rates of adjacency evolution (in a similar framework as for sequence evolution [34]) and intergene sizes, and bring a stone to the study of fragile and solid regions [26].

These constitute our work in progress. We see the model we present here as a decisive step.

Additional material

Additional file 1: Maxima commands for the model of evolution.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MS, ET and LG wrote the model, MS and LG implemented it and MS did the experiments.

Acknowledgements

Publication of this work is funded by the Agence Nationale pour la Recherche, Ancestrome project ANR-10-BINF01-01. This work was performed using the computing facilities of the Computing Center LBBE/PRABI.
This article has been published as part of *BMC Bioinformatics* Volume 16 Supplement 14, 2015: Proceedings of the 13th Annual Research in Computational Molecular Biology (RECOMB) Satellite Workshop on Comparative Genomics: Bioinformatics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcbioinformatics/supplements/16/S14.

Authors' details

[1]Laboratoire de Biométrie et Biologie Évolutive UMR CNRS 5558, Université Claude Bernard Lyon 1, 43 boulevard du 11 novembre 1918, 69622 Villeurbanne, France. [2]INRIA Grenoble Rhône-Alpes, 655 avenue de l'Europe, 38330 Montbonnot, France.

Published: 2 October 2015

References

1. Felsenstein J: **Inferring Phylogenies.** Sinauer Associates, Incorporated, New York; 2004.
2. Szöllősi GJ, Boussau B, Abby SS, Tannier E, Daubin V: **Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations.** *Proc Nat Acad Sci USA* 2012, **109(43)**:17513-17518.
3. Sjöstrand J, Tofigh A, Daubin V, Arvestad L, Sennblad B, Lagergren J: **A bayesian method for analyzing lateral gene transfer.** *Syst Biol* 2014, **63(3)**:409-420.
4. Boussau B, Daubin V: **Genomes as documents of evolutionary history.** *Trends Ecol Evol* 2010, **25(4)**:224-32.
5. Chauve C, El-Mabrouk N, Guéguen L, Semeria M, Tannier E: **Duplication, Rearrangement and Reconciliation: A Follow-Up 13 Years Later.** In *Model Algorithms Genome Evol.* Springer, London;Chauve C, El-Mabrouk N, Tannier E 2013:47-62, Chap. 4.
6. Sturtevant AH, Dobzhansky T: **Inversions in the Third Chromosome of Wild Races of Drosophila Pseudoobscura, and Their Use in the Study of the History of the Species.** *Proc Natl Acad Sci USA* 1936, **22(7)**:448-50.
7. Sankoff D: **Mechanisms of genome evolution: models and inference.** *Bulletin of international statistical institute* 1989, **47**:461-475.
8. Bourque G, Pevzner PA, Tesler G: **Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes.** *Genome Res* 2004, **14(4)**:507-16.
9. Muffato M, Louis A, Poisnel C-E, Roest Crollius H: **Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes.** *Bioinformatics* 2010, **26(8)**:1119-21.
10. Bérard S, Galien C, Boussau B, Szollosi G, Daubin V, Tannier E: **Evolution of gene neighborhoods within reconciled phylogenies.** *Bioinformatics* 2012, **28(18)**:382-388.
11. Gagnon Y, Blanchette M, El-Mabrouk N: **A flexible ancestral genome reconstruction method based on gapped adjacencies.** *BMC Bioinformatics* 2012, **13 Suppl 1(Suppl 19)**:4.
12. Durrett R, Nielsen R, York TL: **Bayesian estimation of genomic distance.** *Genetics* 2004, **166(1)**:621-629.
13. Larget B, Kadane J, Simon D: **A bayesian approach to the estimation of ancestral genome arrangements.** *Molecular phylogenetics and evolution* 2005, **36(2)**:214-223.
14. Ma J: **A probabilistic framework for inferring ancestral genomic orders.** *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 2010, 179-184.
15. Zhang Y, Hu F, Tang J: **A mixture framework for inferring ancestral gene orders.** *BMC Genomics* 2012, **13(Suppl 1)**:7.
16. Lin Y, Hu F, Tang J, Moret B: **Maximum likelihood phylogenetic reconstruction from high-resolution whole-genome data and a tree of 68 eukaryotes.** *Proc 18th Pacific Symp Bio* 2013, 285-96.
17. Yang N, Hu F, Zhou L, Tang J: **Reconstruction of ancestral gene orders using probabilistic and gene encoding approaches.** *PLoS One* 2014, **9(10)**:108796.
18. Blin G, Chauve C, Fertin G, Rizzi R, Vialette S: **Comparing genomes with duplications: a computational complexity point of view.** *IEEE/ACM Trans Comput Biol Bioinform* 2007, **4(4)**:523-534.

19. Gallut C, Barriel V: **Cladistic coding of genomic maps.** *Cladistics* 2002, **18(5)**:526-536.
20. Maňuch J, Patterson M, Wittler R, Chauve C, Tannier E: **Linearization of ancestral multichromosomal genomes.** *BMC Bioinformatics* 2012, **13(Suppl 19)**:11.
21. Wu Y, Rasmussen M, Kellis M: **Evolution at the subgene level: domain rearrangements in the drosophila phylogeny.** *Mol Biol Evol* 2012, **29(2)**:689-705.
22. Ma J, Ratan A, Raney B: **DUPCAR: reconstructing contiguous ancestral regions with duplications.** *Journal of computational biology* 2008, **15(8)**:1007-1027.
23. Guéguen L, Gaillard S, Boussau B, Gouy M, Groussin M, Rochette NC, Bigot T, Fournier D, Pouyet F, Cahais V, Bernard A, Scornavacca C, Nabholz B, Haudry A, Dachary L, Galtier N, Belkhir K, Dutheil JY: **Bio++: efficient extensible libraries and tools for computational molecular evolution.** *Mol Biol Evol* 2013, **30(8)**:1745-1750.
24. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, Gir´on CG, Gordon L, Hourlier T, Hunt SE, Janacek SH, Johnson N, Juettemann T, Kähäri AK, Keenan S, Martin FJ, Maurel T, McLaren W, Murphy DN, Nag R, Overduin B, Parker A, Patricio M, Perry E, Pignatelli M, Riat HS, Sheppard D, Taylor K, Thormann A, Vullo A, Wilder SP, Zadissa A, Aken BL, Birney E, Harrow J, Kinsella R, Muffato M, Ruffier M, Searle SMJ, Spudich G, Trevanion SJ, Yates A, Zerbino DR, Flicek P: **Ensembl 2015.** *Nucleic Acids Research* 2015, 662-669.
25. Peng Q, Pevzner PA, Tesler G: **The fragile breakage versus random breakage models of chromosome evolution.** *PLoS Comput Biol* 2006, **2(2)**:14.
26. Berthelot C, Muffato M, Abecassis J, Roest Crollius H: **The 3D Organization of Chromatin Explains Evolutionary Fragile Genomic Regions.** *Cell Reports* 2015, **10**:1-12.
27. Boussau B, Szöllősi GJ, Duret L, Gouy M, Daubin V: **Genome-scale coestimation of species and gene trees.** *Genome Research* 2013, **23**:323-330.
28. **Evolution of genes and genomes on the drosophila phylogeny.** *Nature* 2007, **450(7167)**:203-218, Drosophila 12 Genomes Consortium.
29. Pollard DA, Iyer VN, Moses AM, Eisen MB: **Widespread discordance of gene trees with species tree in drosophila: evidence for incomplete lineage sorting.** *PLoS Genet* 2006, **2**:173.
30. Hu F, Zhou J, Zhou L, Tang J: **Probabilistic Reconstruction of Ancestral Gene Orders with Insertions and Deletions.** *IEEE/ACM Trans Comput Biol Bioinforma* 2014, **5963(c)**:1-1.
31. Dewey CN: **Positional orthology: putting genomic evolutionary relationships into context.** *Brief. Bioinform.* 2011, **12(5)**:401-12.
32. Mahmudi O, Sjostrand J, Sennblad B, Lagergren J: **Genome-wide probabilistic reconciliation analysis across vertebrates.** *BMC Bioinformatics* 2013, **14(Suppl 15)**:10.
33. Patterson M, Szöllősi G, Daubin V, Tannier E: **Lateral gene transfer, rearrangement, reconciliation.** *BMC bioinformatics* 2013, **14(Suppl 15)**:4.
34. Yang Z: **Among-site rate variation and its impact on phylogenetic analyses.** *Trends Ecol Evol* 1996, **11**:367-372.

doi:10.1186/1471-2105-16-S14-S5
Cite this article as: Semeria *et al.*: Probabilistic modeling of the evolution of gene synteny within reconciled phylogenies. *BMC Bioinformatics* 2015 **16**(Suppl 14):S5.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

2.3 Implémentation

2.3.1 Bio++ et la programmation orientée objet

Harpi a été implémentée en utilisant et en ajoutant des nouvelles fonctions à la bibliothèque Bio++ [61, 45]. Le projet Bio++ vise à faciliter l'utilisation et l'implémentation en C++ de méthodes concernant l'analyse de séquences, la phylogénie, la génétique des populations et l'évolution moléculaire. Il comprend une bibliothèque et une suite de programmes développés à partir de la bibliothèque. La bibliothèque est construite à partir du principe de la programmation objet, dont les trois aspects principaux sont :

- L'encapsulation : des variables et des fonctions peuvent être regroupées dans une même entité, appelée *classe*. On les appelle alors respectivement les attributs et les méthodes de la classe. L'accès aux attributs et méthodes d'une classe peut être réglementé.

- L'héritage : On peut définir une hiérarchie des classes. Une classe fille hérite des méthodes et des données de la classe mère.

- Le polymorphisme qui permet de définir plusieurs fonctions portant le même nom mais dont les paramètres et l'implémentation diffèrent.

La programmation orientée objet est un outil très puissant car elle permet d'organiser le code manière lisible et modulaire, de ré-utiliser les fonctionnalités déjà implémentées et d'en ajouter des nouvelles facilement. Le choix de Bio++ pour implémenter Harpi a donc permis d'utiliser les structures et les méthodes déjà disponibles. Parmi celles-ci, la structure d'arbre phylogénétique et les méthodes permettant de manipuler cette structure sont au cœur de tout programme de phylogénie, et il aurait été dommage de les ré-implémenter. Dans Bio++, un *arbre* est constitué d'objets basiques appelés *nœuds*. Chaque nœud possède les attributs suivants : un identifiant, un nom, un pointeur vers un nœud père, des pointeurs vers des nœud fils, une distance le séparant du nœud père, et des propriétés supplémentaires qui peuvent être définies par le développeur. Dans les arbres de gènes réconciliés utilisés par Harpi, chaque nœud est ainsi annoté avec l'identifiant de l'espèce à laquelle il correspond et un événement évolutif (spéciation,

duplication, ou perte de gène). L'objet *arbre* possède des méthodes permettant d'accéder aux nœuds, d'obtenir les valeurs de leurs paramètres et de les manipuler. L'attribut principal d'un arbre est un pointeur vers son nœud racine, qui donne le sens de lecture de l'arbre. Les méthodes impliquant un parcours de l'arbre sont généralement récursives. Prenons par exemple la méthode `hasNode(id)` qui permet de vérifier si l'arbre contient un nœud d'identifiant donné. Cette fonction va vérifier si la racine est le nœud recherché, et si ça n'est pas le cas elle va appliquer la même procédure pour les deux nœuds fils, puis pour leurs fils, jusqu'à trouver le nœud recherché, où épuiser tous les nœuds de l'arbre.

2.3.2 Composantes de Harpi

La première étape de Harpi consiste à construire des arbres d'adjacences possibles à partir d'arbres de gènes réconciliés et d'une liste d'adjacence. On utilise alors directement les méthodes de manipulation de nœuds et d'arbres de Bio++ pour créer individuellement chaque nœud représentant une adjacence possible et établir les liens de parenté entre les nœuds (en gérant les pointeurs vers les nœuds père et fils). Les nœuds sont annotés avec un événement évolutif et un identifiant d'espèce. La gestion des fichiers d'entrée et de sortie est aussi réalisée grâce à des fonctions de Bio++. Par exemple, une fonction très utile permet de lire en fichier d'arbres en format texte et de créer les objets correspondant. Une autre permet de réaliser l'opération inverse. À l'issue de cette première étape, on écrit ainsi les arbres d'adjacences possibles construits, et le motif binaire de présence et d'absence des adjacences aux feuilles pour chaque arbre.

La deuxième étape consiste à calculer une vraisemblance sur un arbre d'adjacences possibles en tenant compte des événements évolutifs associés aux nœuds. Le calcul de vraisemblance était déjà implémenté dans Bio++, et disponible sous la forme d'un programme déjà bien optimisé : `bppml`. Ce programme a été conçu pour permettre à l'utilisateur de choisir son modèle d'évolution parmi les modèles d'évolutions implémentés dans Bio++. La grande majorité des modèles sont des modèles de substitutions de nucléotides, acides aminés ou codons. Il nous restait alors à intégrer dans Bio++ le modèle d'évolution présenté dans la partie 2.2 et à adapter les fonctions du calcul de vraisemblance pour qu'elles prennent en compte l'événement évolutif associé à un nœud et appliquent un calcul spécifique au type d'événement. Ces développements ont

donc consisté à ajouter des nouvelles fonctionnalités à Bio++.

Une troisième étape éventuelle est l'inférence des états des adjacences ancestrales via le calcul de probabilités a posteriori. Là aussi, il existait déjà un programme de la suite Bio++ dédié à l'inférence d'états ancestraux : `bppancestor`. Ma contribution a alors été, d'adapter les formules de calcul de la vraisemblance utilisées par `bppancestor` pour prendre en compte les événements évolutifs et mon modèle d'évolution.

2.3.3 Extensibilité et disponibilité

Harpi a été pensée pour s'intégrer à l'architecture de Bio++ et pouvoir utiliser deux programmes de la suite Bio++ : `bppml` et `bppancestor`. Ces deux programmes ont été développés pour calculer des vraisemblances (`bppml`) et reconstruire des états ancestraux (`bppancestor`) à partir de séquences et d'arbres de gènes. L'idée derrière Harpi était donc de faire fonctionner ces programmes avec d'autres objets que ceux prévus initialement. Pour cela, j'ai écrit et implémenté un algorithme permettant de construire un arbre des adjacences possibles à partir de deux arbres de gènes réconciliés. Une fois construit, l'arbre des adjacences possibles a la même structure et peut utiliser les mêmes fonctions que n'importe quel arbre implémenté dans Bio++. `bppml` et `bppancestor` nécessitent aussi des alignements de séquences. Dans le cadre de Harpi, on fournit pour chaque arbre des adjacences possibles le motif binaire correspondant à la présence et à l'absence de l'adjacence aux feuilles. Comme la plupart des programmes de la suite Bio++, `bppml` et `bppancestor` ont été conçus pour fonctionner avec différents modèles d'évolution de séquences, le choix du modèle revenant à l'utilisateur. Les modèles déjà implémentés dans Bio++ étaient majoritairement des modèles de substitution de nucléotides, d'acides aminés ou de codons. L'idée était alors d'implémenter un nouveau modèle, spécifique à l'évolution d'adjacences, et de l'utiliser via `bppml` et `bppancestor`. Pour utiliser ce nouveau modèle, il a ensuite fallu adapter les fonctions du calcul de vraisemblance de Bio++.

Harpi a donc été implémentée pour s'adapter aux structures et au programmes de Bio++. Ce choix d'implémentation présente de grands avantages. Notamment, les fonctionnalités existantes de Bio++, telles que l'optimisation de paramètres lors du maximum de vraisemblance fonctionnent avec Harpi sans aucune adaptation. L'insertion dans Bio++ facilite aussi l'accès au code de Harpi pour tous les utilisateurs et les

développeurs de la bibliothèque.

3

Correction d'arbres de gènes avec de la synténie

3.1 Introduction

Le résultat d'une reconstruction parcimonieuse basée sur l'évolution indépendante d'adjacences est un motif binaire représentant la présence et l'absence des ancêtres des adjacences actuelles dans les génomes ancestraux (voir 1.1.7.3). Dans le cas de Harpi il s'agit des probabilités postérieures de présence et d'absence des adjacences dans les génomes ancestraux. Pour achever la reconstruction de chromosomes ancestraux, il faut alors assembler les adjacences présentes dans chaque génome ancestral, les gènes ancestraux ayant été déterminés au préalable, par la réconciliation par exemple. Reprenons l'image du génome comme un graphe dans lequel les sommets sont les gènes et les arêtes sont les adjacences. À l'issue de la reconstruction, on dispose pour le génome ancestral recherché d'un ensemble de sommets (les gènes ancestraux inférés) et d'un ensemble d'arêtes (les adjacences ancestrales inférées). Pour visualiser le génome ancestral entier, il faut alors placer les arêtes entre les sommets.

En réalité, il n'est presque jamais possible d'assembler parfaitement toutes les adjacences ancestrales et de reconstituer le caryotype hypothétique du génome ancestral. En effet, l'inférence des adjacences ancestrales, et des gènes ancestraux dépend de la qualité des données, ainsi que des spécificités et performances de la méthode ou des méthodes utilisées. Un problème rencontré fréquemment est que les génomes étudiés

ne sont pas complètement assemblés. Les génomes mal assemblés se présentent alors non pas sous la forme de chromosomes bien identifiés mais d'un certain nombre de morceaux de chromosomes qu'on ne sait comment mettre bout à bout : les adjacences devant relier les morceaux manquent dans le jeu de données. On pourra alors difficilement déterminer si elles étaient présentes chez l'ancêtre et aboutir à un génome ancestral parfaitement assemblé. Très récemment, Anselmetti et al. ont montré qu'il était cependant possible d'utiliser la reconstruction d'adjacences ancestrales avec DeCo pour améliorer l'assemblage de génomes d'espèces actuelles [7].

Un autre problème peut être la saturation du signal phylogénétique. En présence de taux de réarrangement trop élevés, il peut ainsi être impossible de reconstruire l'histoire des réarrangements.

L'inférence des gènes ancestraux peut aussi poser problème. Dans le cas de DeCo et Harpi, les gènes ancestraux sont déterminés au cours d'une étape de réconciliation. Mais si les arbres de gènes, ou l'arbre des espèces utilisés pour la réconciliation sont erronés (pour des raisons qu'on développera plus loin dans ce chapitre), l'inférence des gènes ancestraux est faussée, ce qui va se ressentir lors de l'étape d'assemblage des adjacences ancestrales.

Enfin, les morceaux obtenus après l'assemblage des adjacences ancestrales ne sont pas forcément linéaires ou circulaires, comme attendu pour des chromosomes. En raison de l'hypothèse d'indépendance des adjacences, la linéarité des génomes est loin d'être garantie. Un exemple de reconstruction conduisant à un génome ancestral non linéaire est illustré par la figure 3.1. Dans cet exemple, les génomes étudiés se différencient par l'inversion de l'ordre des gènes 3, 4, et 5 dans les génomes D et E (Figure 3.1a). On considère que la topologie de l'arbre d'espèce et des arbres de gènes sont celles représentées sur la figure 3.1a. Pour chaque adjacence présente dans les génomes A, B, C, D, E, on reconstruit un arbre d'adjacences et on infère les états (présence/absence) de l'adjacence dans les génomes ancestraux F, G, H, I selon une approche parcimonieuse. On considère que les adjacences ne sont pas orientées, c'est-à-dire que les adjacences 1-2 et 2-1 correspondent à une unique adjacence. Avec les états observés aux feuilles, et la topologie donnée, il y a une solution parcimonieuse pour inférer les états ancestraux des adjacences 1-2, 3-4, et 4-5 (Figure 3.1b) et deux solutions également parcimonieuses pour les adjacences 2-3, 2-5, 3-6 et 5-6 (Figures 3.1c et 3.1d). Si on choisit systématiquement la solution qui donne l'adjacence présente dans le génome I, on obtient un

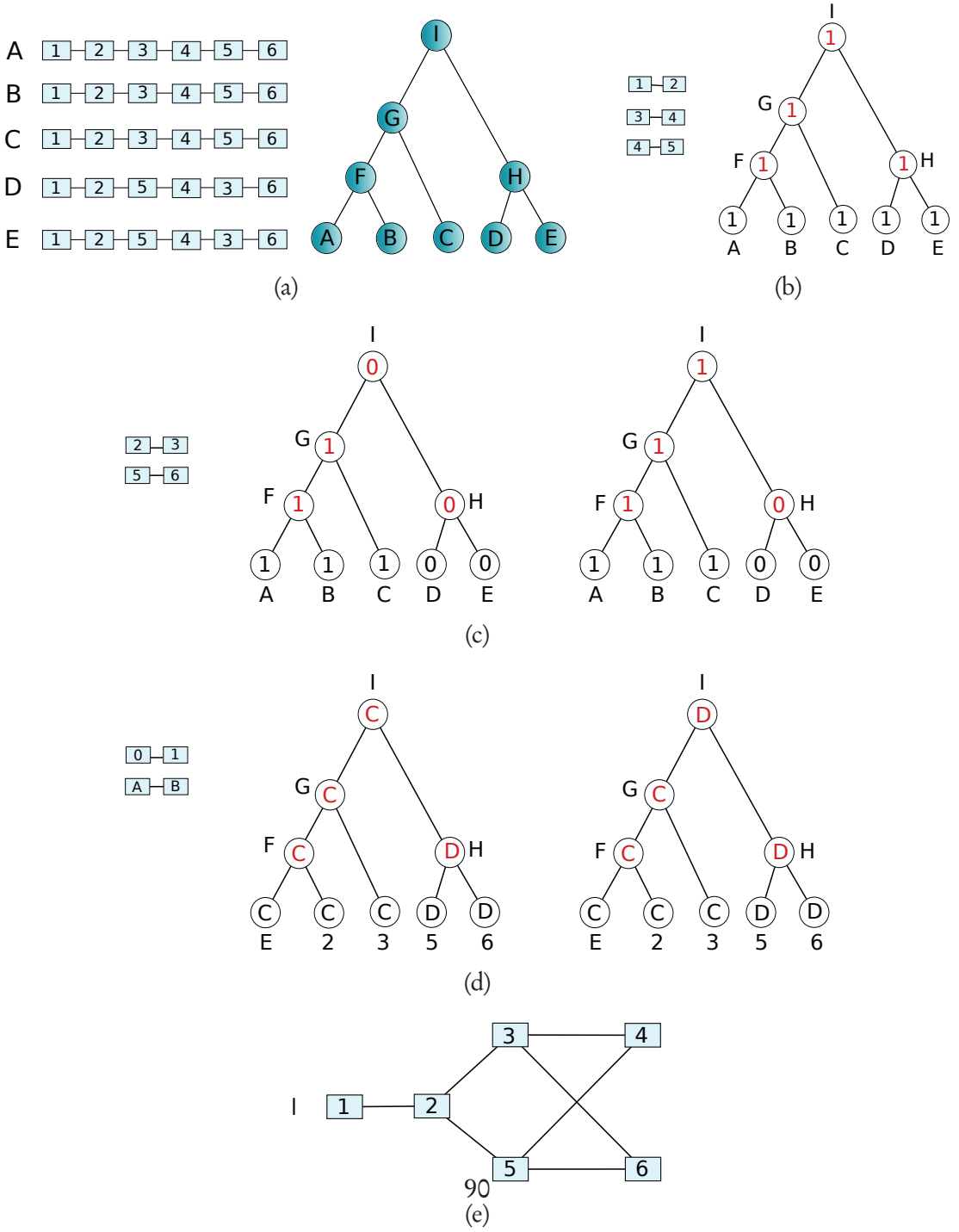
génome non linéaire pour l'espèce ancestrale I (Figure 3.1e).

Si l'on utilise DeCo pour reconstruire des adjacences ancestrales à partir des gènes de la version 70 d'Ensembl Compara [137], on obtient une majorité de gènes ancestraux impliqués dans une ou deux adjacences (respectivement 43% et 45%), mais une proportion non négligeable (12%) de gènes ancestraux impliqués dans trois adjacences ou plus. Pour avoir des génomes ancestraux linéaires, il faudrait que tous les gènes ancestraux soient impliqués dans deux adjacences ou moins.

Les génomes ancestraux obtenus à ce stade sont donc imparfaits : des adjacences manquent, certains gènes sont reliés à trois voisins ou plus, certains gènes et certaines adjacences sont erronées. Comme souligné précédemment, ces imperfections sont le signe d'erreurs dans la reconstruction ou de problèmes liés à la qualité des données. On peut chercher à améliorer les génomes ancestraux obtenus, notamment en résolvant les conflits qui donnent lieu à des gènes avec de trop nombreux voisins. On parle alors de linéarisation des génomes. Il existe plusieurs approches pour linéariser les génomes ancestraux. Une approche populaire consiste à éliminer des adjacences impliquées dans des conflits en résolvant un problème du type du problème du voyageur de commerce : étant donné un ensemble de sommets reliés par des arrêtes, trouver le plus court chemin passant par tous les points une seule et unique fois. Appliqué à notre problème, cela donne : étant donné un ensemble de gènes appartenant à un génome ancestral, trouver la solution qui permet de relier chaque gène à au plus deux voisins. Ce type de problème étant NP-complet, les auteurs choisissant cette solution utilisent des heuristiques [93, 31]. Cependant, cette approche n'exploite pas l'information potentiellement contenue dans les adjacences conflictuelles. Les conflits sont pourtant révélateurs de problèmes d'assemblage et d'annotation dans le jeu de données, et de reconstruction phylogénétique. Dans ce chapitre nous explorerons donc une seconde approche, qui consiste à modifier les arbres de gènes à l'origine des problèmes de linéarité pour résoudre les conflits.

Nous montrerons ainsi que la synténie ancestrale peut être utilisée pour détecter ces problèmes, et pour corriger certaines topologies erronées dans les arbres de gènes. Cette étude est antérieure à celles décrites dans les chapitres précédents. Elle a donc été réalisée à partir du logiciel DeCo [12], et non de Harpi. La méthode développée a été utilisée dans trois articles écrits en collaboration avec Bastien Boussau, Cédric Chauve, Laurent Guéguen, Manuel Lafond, Nadia El-Mabrouk,Emmanuel Noutahi, Jonathan

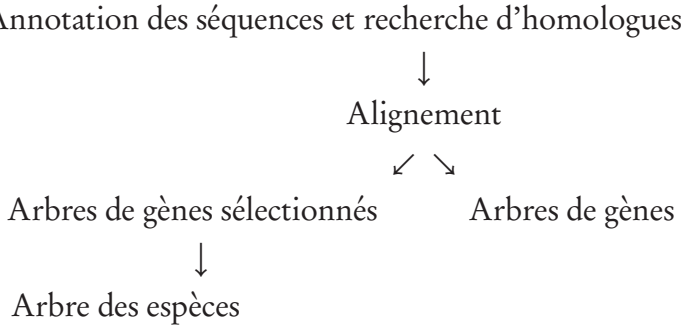
FIGURE 3.1 – Inférence d'adjacences ancestrales conduisant à un génome non linéaire. (a) Arbre d'espèces et ordre des gènes dans les espèces actuelles et ancestrales. (b) Reconstruction parcimonieuse pour les adjacences 1-2, 3-4, 4-5. (c) Reconstructions équiparcimonieuses pour 2-3 et 5-6. (d) Pour 2-5 et 3-6. (e) Génome inféré pour l'espèce ancestrale I si on choisit systématiquement les solutions qui donnent les adjacences présentes à la racine.



Seguin, Krister Swenson, et Eric Tannier (voir Appendix A).

3.2 Sources d'erreurs dans les arbres de gènes réconciliés

Les arbres de gènes et les arbres d'espèces sont le résultat de plusieurs étapes, généralement réalisées de manière séquentielle [22, 40] (voir section 1.1.4) :



À chaque étape de ce parcours, des erreurs peuvent survenir. En amont de l'alignement, le regroupement de séquences similaires en familles d'homologues peut aussi poser problème. En effet, la recherche d'homologues par similarité des séquences ne permet pas de détecter des événements de recombinaison homologue ou de fusion de gènes. La recombinaison homologue, par laquelle une partie d'une séquence est remplacée par une séquence homologue peut être à l'origine d'erreurs dans l'alignement et la phylogénie car la similarité des séquences n'est alors pas proportionnelle au temps de divergence des séquences. La fusion de gènes, par laquelle une partie d'un gène est remplacée par une séquence non homologue, brouille encore plus le signal phylogénétique et remet même en cause la notion de famille d'homologues : peut-on on encore considérer que le gène entier est homologue avec les autres gènes de la famille ?

Par exemple, la plupart des méthodes d'alignement utilisent des heuristiques et des pénalités arbitraires pour le placement des gaps. L'alignement obtenu est alors le meilleur alignement étant donné les pénalités utilisées et l'espace des solutions explorées par l'heuristique. En admettant que les pénalités soient adaptées à la famille de séquences étudiée, il est toujours possible que l'alignement obtenu, même s'il est optimal, ne représente pas l'histoire évolutive réelle de cette famille [22]. Lors de la reconstruction phylogénétique, l'alignement est rarement remis en cause. Pourtant, les erreurs et

biais des méthodes d'alignement se propagent à la phylogénie. Wong et al. montrent l'exemple d'une famille particulièrement difficile à aligner, où sept méthodes d'alignements différentes conduisent à six topologies différentes avec une même méthode de reconstruction phylogénétique [140]. A l'échelle d'analyses génomiques, où il est difficile de corriger les biais et erreurs spécifiques à chaque famille, on peut donc s'attendre à une proportion non négligeable de phylogénies erronées dues à des problèmes d'alignement [140][74]. Le calcul d'un soutien statistique tel que le bootstrap au moment de la reconstruction permet d'obtenir une mesure de la fiabilité des bifurcations constituant l'arbre de gènes. Cependant, le calcul du bootstrap est réalisé à partir d'un alignement fixé : les colonnes de l'alignement sont ré-échantillonnées mais pas modifiées. Le bootstrap ne fournit donc aucune information sur la variabilité de la reconstruction qui serait due à la variabilité de l'alignement. Seules quelques études utilisent vraiment l'incertitude de l'alignement pour reconstruire des phylogénies [88, 116, 84, 85, 66] . De même, les tests de vraisemblance tels que le KH, SH et AU peuvent être utilisés pour comparer les vraisemblances d'arbres de gènes. Mais, comme montré par Levy Karin et al. dans un récent article [81], le test KH est biaisé lorsque l'alignement contient des erreurs.

Au niveau de l'inférence phylogénétique proprement dite, la recherche du meilleur arbre étant donné un alignement passe souvent par des mouvements locaux aléatoires tel que le *Nearest Neighbour Interchange* (NNI) ou le *Subtree Pruning and Regrafting* (SPR). Comme dans le cas de la recherche du meilleur alignement, le résultat obtenu n'est pas forcément la solution optimale. Si les phylogénies des espèces sont construites à partir de familles de gènes aux histoires évolutives relativement simples et consensuelles, les arbres de gènes gagnent en retour à être interprétés à la lumière de l'arbre des espèces. C'est le principe derrière les méthodes de réconciliation, qui consistent à annoter les arbres de gènes avec des événements de spéciation, de duplication, de perte, voire de transfert horizontal avec l'aide d'un arbre des espèces.

Arbre des espèces Arbres de gènes

↘ ↙

Arbres de gènes réconciliés

Mais, là aussi, ces méthodes ne sont pas exemptes de biais. Le problème majeur de la plupart des méthodes de réconciliation est qu'elles font l'hypothèse que les arbres de

gènes et l'arbre d'espèces ne contiennent pas d'erreurs. Or, si les phylogénies de certains ensembles d'espèces ont pu être établies avec une grande certitude, de nombreux embranchements de l'arbre du vivant sont encore incertains. Et en ce qui concerne les arbres de gènes, inférés individuellement avec un nombre de sites limité, et avec les biais de l'annotation et de l'alignement, il est rare de pouvoir se fier entièrement à une topologie. Certaines méthodes prennent en compte les incertitudes sur les arbres de gènes, par exemple en contractant les nœuds à faible support [13, 101], ou en permettant l'exploration d'une distribution d'arbres de gènes issue d'une méthode bayésienne [121] ou probabiliste [133].

Plus ambitieux mais beaucoup plus coûteux en temps de calcul, la méthode développée par Boussau et al. recherche conjointement l'arbre des espèces et les arbres de gènes réconciliés optimaux avec un algorithme de maximum de vraisemblance, et tient ainsi compte des incertitudes sur l'arbre des espèces et sur les arbres de gènes [24]. Cependant, même les méthodes de réconciliation les plus élaborées restent sensibles à plusieurs biais. Ainsi, un arbre de gènes peut être en désaccord avec l'arbre d'espèces pour diverses raisons, telles qu'un tri de lignées incomplet ou un événement de recombinaison. La réconciliation tend alors à ajouter une duplication erronée en amont du point de désaccord entre les deux arbres, et des pertes supplémentaires pour équilibrer l'arbre en dessous de la duplication (Figure 3.2). Ceci se traduit par un biais qui consiste à surestimer le nombre de duplications et de pertes dans les arbres réconciliés. Un corollaire de ce biais est que les duplications erronées sont placées relativement profondément dans l'arbre (car la topologie d'un sous-arbre ne peut être incertaine que si ce sous-arbre contient plus de deux feuilles) [64].

Avec ce biais, on peut remettre en question les conclusions de certaines analyses phylogénétiques à l'échelles des génomes : Blomme et al. [18] trouvent ainsi de nombreuses duplications anciennes, et pertes récentes chez les vertébrés, ce qui va dans le sens du biais de réconciliation [64].

Malgré ce scénario catastrophe, les biais des différentes étapes de l'analyse phylogénétique ne sont pas insurmontables. Les méthodes qui intègrent plusieurs étapes de l'analyse en prenant en compte les incertitudes liées à chaque étape semblent prometteuses [22]. Dans ce chapitre, nous proposons d'utiliser une méthode de contournement pour repérer et corriger les parties erronées d'arbres de gènes grâce à la reconstruction de la synténie ancestrale avec DeCo.

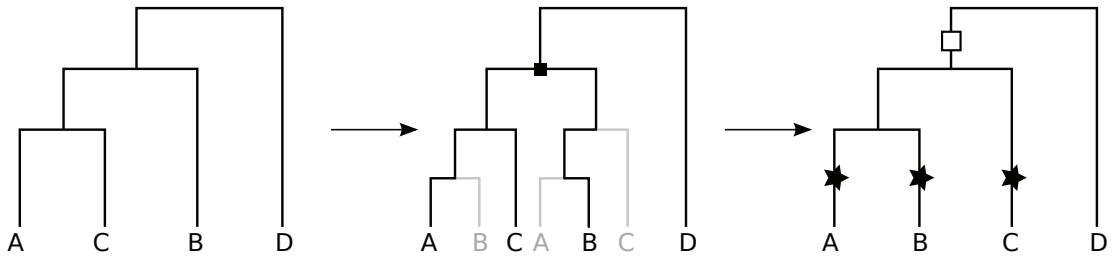


FIGURE 3.2 – Dans cet exemple, l'arbre de gène (à gauche) n'a pas la même topologie que l'arbre des espèces (à droite). La réconciliation va alors ajouter un nœud de duplication au niveau du branchement ((A,C),B) et une perte dans chacune des espèces A,B et C pour résoudre l'incongruence. Cela s'interprète dans l'arbre des espèces par une duplication chez l'ancêtre de A, B, C (carré blanc) suivie de trois pertes (étoiles noires).

3.3 Détection d'erreurs et correction avec la synténie

Cette étude a été réalisée à partir des données de la version 70 d'Ensembl Compara [137], restreintes à 41 espèces de mammifères (Figure 3.3). Ce jeu de données restreint contient 13132 arbres de gènes, réconciliés avec DeCo. Après avoir établi l'ordre des gènes dans les espèces actuelles à partir des coordonnées chromosomiques disponibles dans Ensembl, j'ai reconstruit les adjacences ancestrales du jeu de données avec DeCo. Sur les 957547 gènes ancestraux dans les arbres réconciliés, 12% forment plus de 2 adjacences avec d'autres gènes et 43% ne forment qu'une unique adjacence. Certains de ces gènes à un seul voisin peuvent correspondre à des extrémités de chromosomes ancestraux mais dans la plupart des cas il s'agit d'adjacences manquantes. Une des raisons pour ce manque de puissance de DeCo est la qualité de l'assemblage des génomes étudiés. En effet si le génome humain est constitué de 20383 adjacences pour 20429 gènes (soit 46 scaffolds), le génome du wallaby (Macropus eugenii) ne contient que 2419 adjacences pour 15190 gène, ce qui correspond à 12771 scaffolds. On trouve ainsi un grand nombre d'adjacences manquantes chez les ancêtres des génomes mal assemblés. Par exemple, chez l'ancêtre du tatou (Dasypus novemcinctus) et du paresseux (Choloepus hoffmanni), 80% des gènes n'ont qu'un unique voisin.

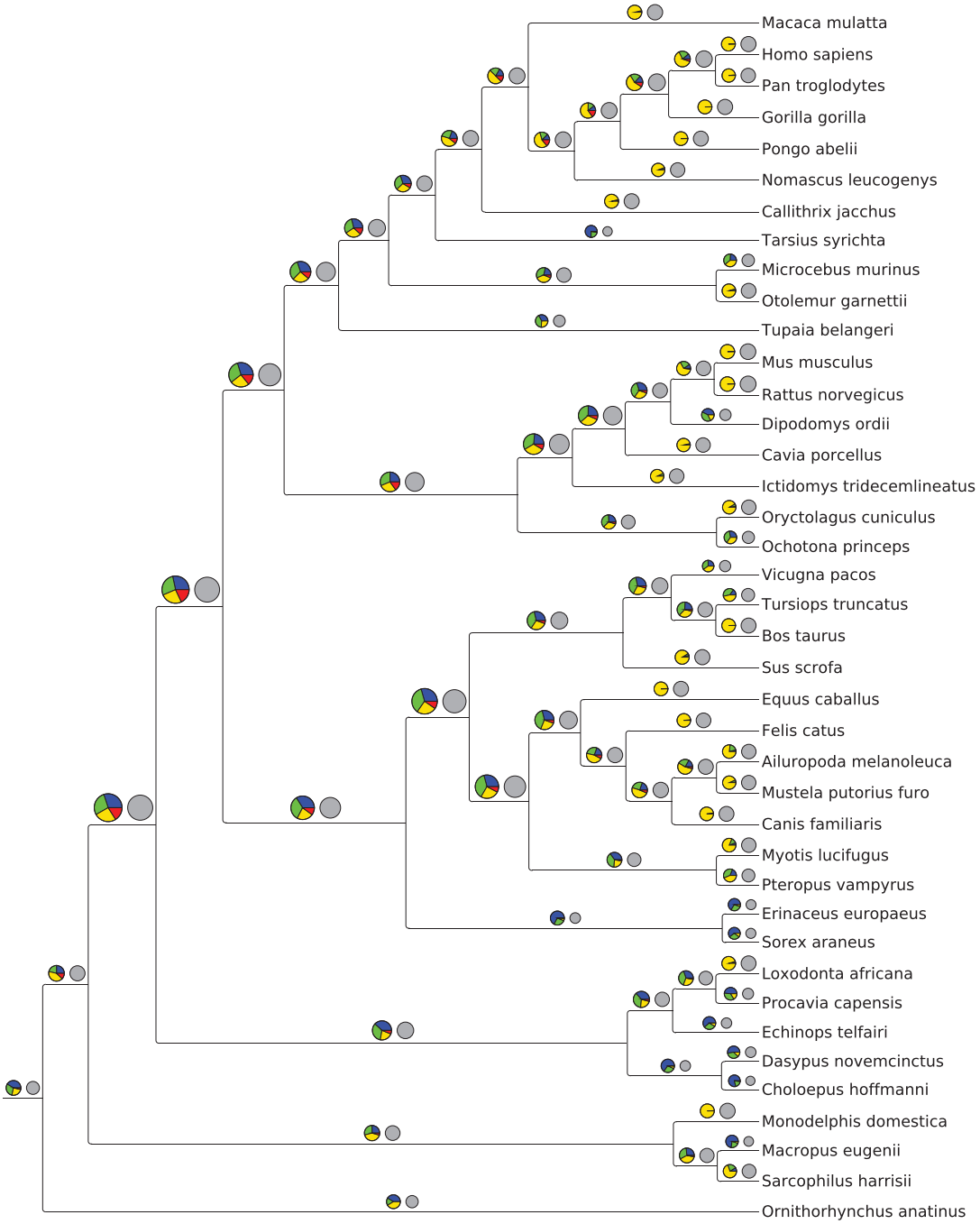


FIGURE 3.3 – L'arbre d'espèces utilisé. Le diagramme en secteurs représente les proportions de gènes n'ayant aucun voisin (bleu), ayant un voisin (vert), deux voisins (jaune) ou plus de deux voisins (rouge). La taille des diagrammes en secteurs et des ronds gris représentent respectivement le nombre de gènes et le nombre d'adjacences dans chaque espèce.

Les cas des gènes ayant plus de deux voisins reflètent des erreurs d'annotation ou de topologies des arbres de gènes. Ce qui pouvait être vu comme un défaut de la méthode (DeCo ne produit pas des génomes linéaires) peut alors être utilisé pour évaluer la qualité des arbres de gènes utilisés en entrée. Il est cependant difficile de faire émerger un unique scénario qui conduirait DeCo à inférer plus de deux voisins pour un gène ancestral. Dans de nombreux cas, les erreurs sont dues à une mauvaise annotation des gènes dans Ensembl qui doit être repérée manuellement. Dans d'autres cas, les conflits de linéarité sont dûs à des erreurs de reconstruction de DeCo. L'approche parcimonieuse présente en effet quelques défauts. Lorsqu'il y a plusieurs solutions équiparcimonieuses pour reconstruire les adjacences ancestrales, DeCo en sélectionne une au hasard. Et, plus généralement, la parcimonie peut être biaisée en présence de convergence évolutive ou de taux de mutation trop élevés (voir chapitre 2).

3.3.1 Principe

Un scénario de conflit relativement fréquent est le cas où le placement d'une duplication trop profondément dans un arbre de gène conduit à inférer trop d'adjacences pour un gène ancestral. Dans ce scénario, illustré sur la Figure 3.4, le placement de la duplication revient à inférer deux gènes chez l'espèce E dans l'arbre G_1. Si des adjacences existent ou sont inférées entre les gènes des espèces A et B des arbres G_1 et G_2, le scénario le plus parcimonieux pour l'histoire évolutive des adjacences entre les deux arbres consiste à mettre une adjacence entre E_1 et E_2 et une adjacence entre E_1' et E_2. Mais, l'histoire évolutive des adjacences entre chaque couple d'arbres étant reconstruite indépendamment des autres, il est possible qu'une adjacence soit inférée entre E_2 et un gène E_3 appartenant à une troisième famille. E_2 se retrouve alors avec trois voisins. On peut résoudre le conflit d'adjacences en modifiant l'arbre G_1. La solution est de ne mettre qu'un seul gène d'espèce E, ce qui revient à faire descendre la duplication dans les espèces A et B en modifiant G_1.

Cette méthode de correction est décrite plus en détail dans l'article A.4 [76]. Elle est appelée Unduplicator.

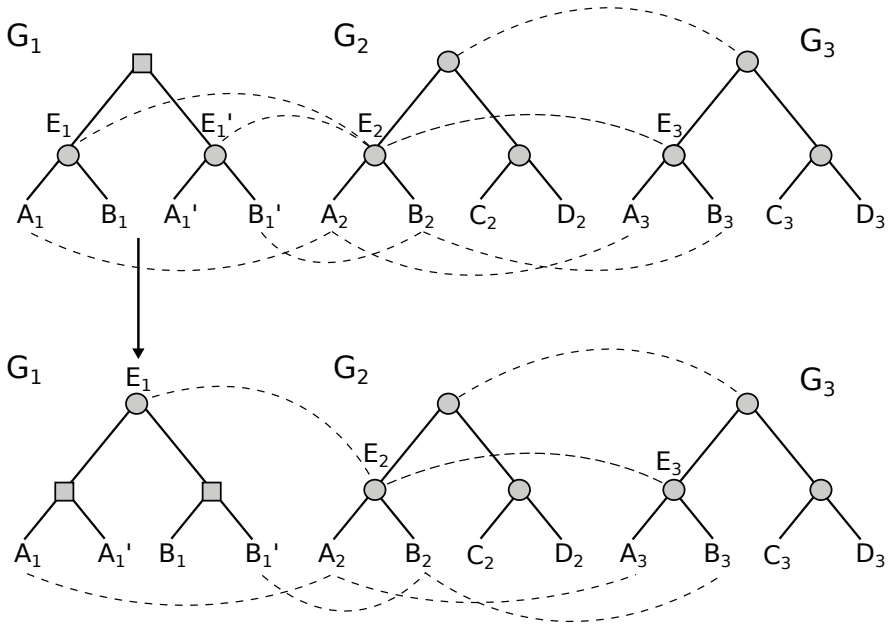


FIGURE 3.4 – Exemple d'un conflit d'adjacence et de la correction proposée. Ici, une duplication potentiellement erronée dans l'arbre de gènes G_1 se traduit par l'inférence de trois adjacences pour le gène ancestral E_2 (lignes en pointillés). Un réarrangement des clades sous la duplication permet de résoudre le conflit d'adjacences pour E_2. NB : on ne représente pas d'adjacence pour les nœuds de duplication.

3.3.2 Un exemple sur des données réelles

Prenons un arbre dans ce cas de figure. Sur le sous-arbre de la Figure 3.5, la duplication chez les euthériens, et les adjacences aux feuilles conduisent à l'inférence d'une adjacence entre chacun les deux enfants de la duplication et un même gène d'une famille voisine possédant par ailleurs une autre adjacence avec une autre famille. Modifier ce sous-arbre d'après la correction proposée en Figure 3.4 (Unduplicator) revient à rassembler les deux groupes de boréoeuthériens sous une duplication spécifique d'une part et les groupes xénarthres et les afrothériens sous une duplication spécifique. Ce réarrangement est illustré sur la Figure 3.6. Si ce réarrangement permet de rétablir localement la linéarité des génomes, il revient aussi à corriger l'arbre de gène en se rapprochant de la phylogénie des espèces d'Ensembl, qui place les xénarthre et les afrothériens en groupes frères (Figure 3.7). Notons cependant que ce branchement a longtemps été dé-

battu [127] et que le résultat de la réconciliation et a fortiori de la reconstruction des adjacences ancestrales avec DeCo dépend de l'arbre d'espèces utilisé.

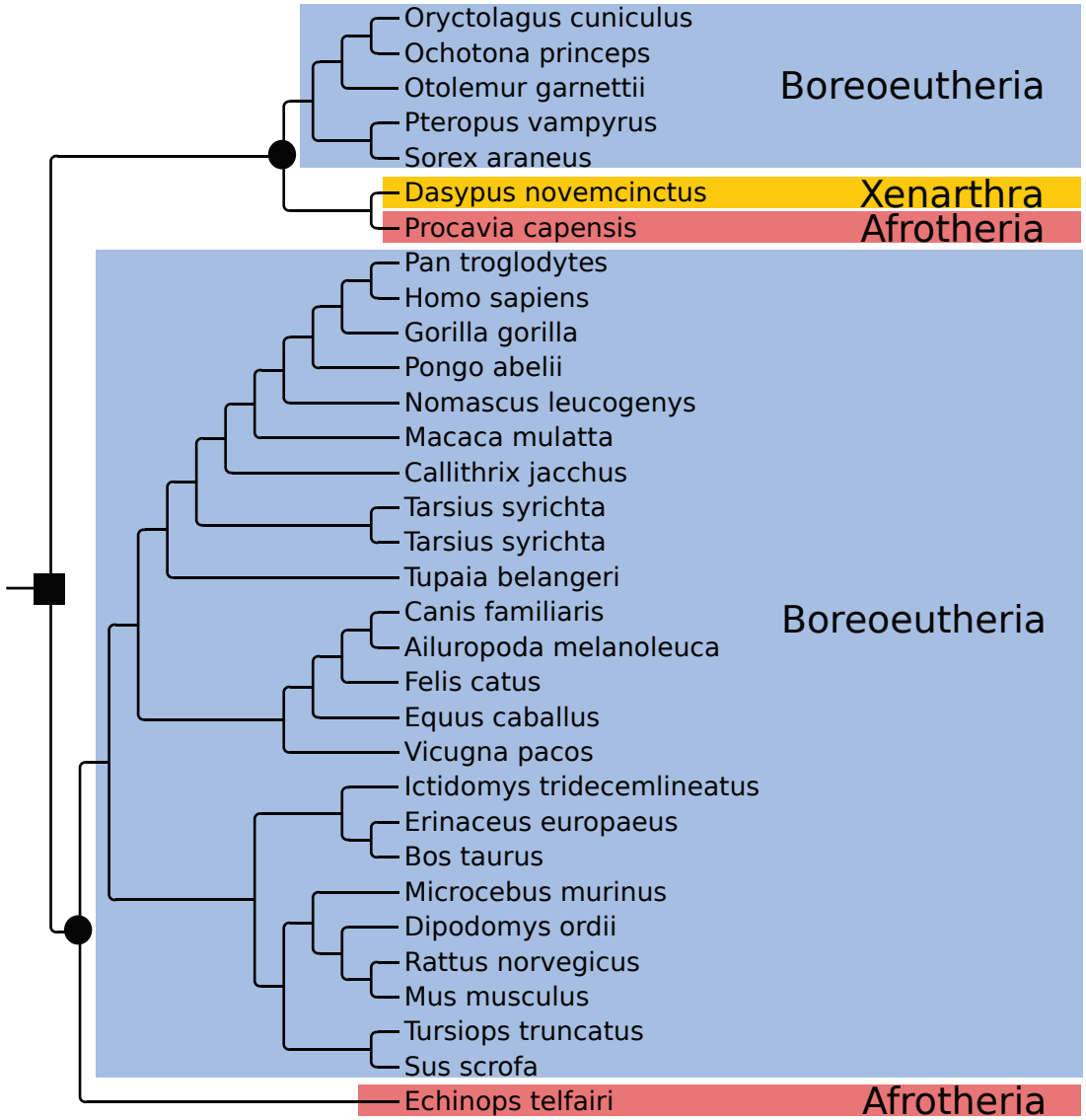


FIGURE 3.5 – Arbre de gènes à l'origine d'adjacences conflictuelles. On peut résoudre le conflit de linéarité en modifiant le sous arbre sous la duplication (carré noir).

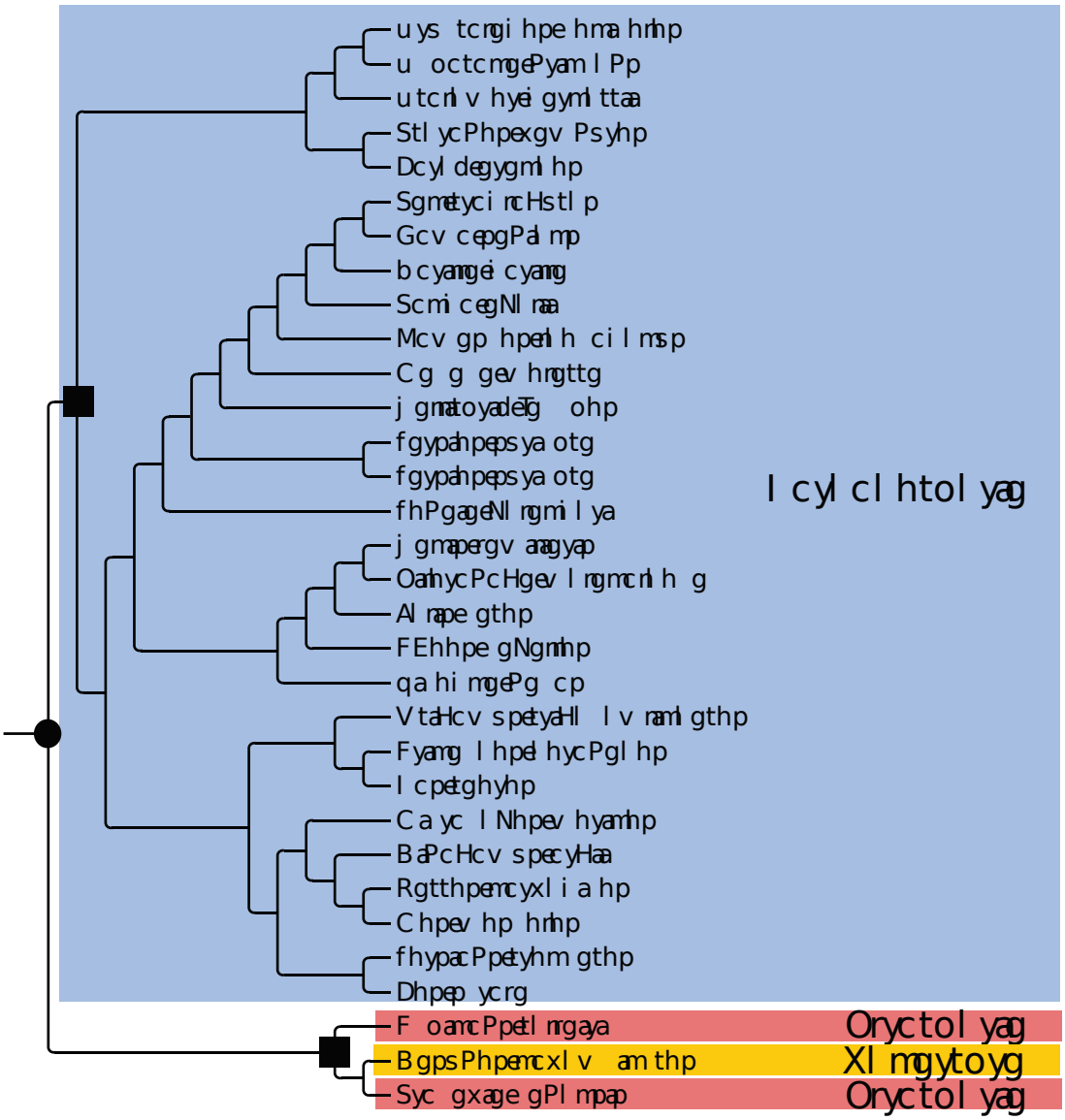


FIGURE 3.6 – Arbre de gènes corrigé. La correction résout le conflit de linéarité.

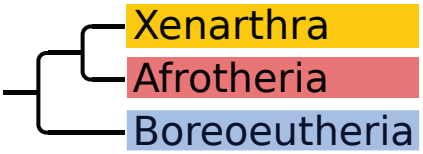


FIGURE 3.7 – Arbre d'espèces d'Ensembl simplifié.

La topologie de l'arbre corrigé n'est pas entièrement satisfaisante. On aurait en effet envie de regrouper les afrothériens pour éliminer la duplication chez l'ancêtre de *Echinops telfairi*, *Procavia capensis* et *Dasypus novemcinctus*, car cette duplication n'est pas soutenue par l'existence de plusieurs gènes chez les descendants. De même, chez les boréoeuthériens, il reste plusieurs duplications qu'on pourrait envisager d'éliminer. En appliquant la correction d'Unduplicator à cet arbre de gène, nous avons donc réalisé une itération vers une topologie qui paraît plus raisonnable compte tenu de la synténie ancestrale inférée avec DeCo et de l'arbre des espèce. Mais la topologie obtenue pourrait être encore modifiée, et idéalement, encore améliorée. Une manière possible serait de reconstruire les adjacences ancestrales avec cette topologie et de voir si d'autres conflits de linéarité peuvent être corrigés.

Par ailleurs, il aurait été utile de disposer d'un support statistique de type bootstrap pour les embranchements de l'arbre initial. On aurait alors pu savoir si les branches cassées par la correction étaient soutenues par un fort signal de l'alignement [101, 13]. Cependant, on ne trouve pas systématiquement de supports sur les branches des arbres d'Ensembl (Ensembl propose tout de même un support pour les nœuds de duplication, voir section 3.4).

3.3.3 Description formelle de la méthode Clade Orthology Correction

On répertorie tous les cas où :

- On infère avec DeCo plus de deux voisins pour un gène ancestral E_2 dans un arbre G_2.
- E_2 forme une adjacence avec un gène E_1 et une adjacence avec un gène E_1', E_1 et E_1' appartenant à un même arbre G_1.

- E_1 et E'_1 sont les enfants d'une duplication.
- Soit A_1 et B_1 les descendants de E_1. A_1 est une clade contenant des gènes appartenant à une espèce descendante de l'espèce A. B_1 est une clade contenant des gènes appartenant à une espèce descendante de l'espèce B. Dans l'arbre des espèces, A et B sont les deux enfants d'une même espèce E. De même, soit A'_1 et B'_1 les clades descendant de E'_1.

On applique pour chaque cas la correction illustrée sur la figure 3.4 (Unduplicator) :

- Dans l'arbre G_1, on échange les clades A'_1 et B_1.

3.3.4 Analyse à large échelle

Parmi les 13132 arbres de gènes d'Ensembl contenant des mammifères, on trouve 1519 arbres pouvant correspondre au problème décrit ci-dessus. On applique donc Unduplicator à tous ces arbres. La question est ensuite de savoir si les topologies modifiées, meilleures d'après un critère de linéarité des génomes, sont équivalentes aux topologies initiales selon un critère de vraisemblance. Pour répondre à cette question, j'ai calculé la vraisemblance des 1519 arbres initiaux et de ces mêmes arbres corrigés d'après les alignements disponibles dans Ensembl et le logiciel PhyML [62]. J'ai ensuite évalué l'impact de la correction sur la vraisemblance des arbres en comparant les deux vraisemblances au moyen d'un test AU [123], implémenté dans le logiciel CONSEL [125] (voir section 1). Pour chaque arbre T, on cherche ainsi à comparer les vraisemblance de la topologie initiale T_i et de la topologie corrigée T_c. Le résultat du test AU consiste en deux probabilités : une p-value p_i correspondant à l'hypothèse nulle H_0 : "T_i est meilleure que T_c" et à l'hypothèse alternative H_1 : "T_c est meilleure que T_i", et un p-value p_c correspondant à l'hypothèse nulle H_0 : "T_c est meilleure que T_i" et à l'hypothèse alternative H_1 : "T_i est meilleure que T_c". La figure 3.8 représente la distribution des p-values p_c correspondant à l'hypothèse nulle H_0 : "T_i est meilleure que T_c" obtenues avec des test AU pour les 1519 arbres tirés d'Ensembl. Lorsque la valeur de p_c est faible (à l'extrémité gauche de la distribution), on rejette H_0 : "T_c est meilleure que T_i". Autrement dit, on a eu tort de faire corriger ces arbres car la topologie initiale est meilleure que la topologie corrigée. À l'inverse, pour le reste de la distribution, les valeurs de p_c ne sont pas significatives. On considère donc que la correction est valable

pour ces arbres car la topologie corrigée est meilleure que ou équivalente à la topologie initiale. Le nombre d'arbres pour lesquels la topologie corrigée est meilleure que ou équivalente à la topologie initiale dépend alors du seuil qu'on considère. Si l'on choisit de rejeter la correction pour les arbres avec une probabilité < 0.05, en corrigeant pour les tests multiples avec la méthode de Bonferroni, la correction n'est rejetée que dans 4% des cas. Ceci prouve qu'il est possible d'utiliser une méthode de reconstruction de la synténie ancestrale pour proposer des topologies alternatives. De même notre méthode de correction montre qu'on peut utiliser la synténie ancestrale pour discriminer deux topologies statistiquement équivalentes si on ne prend en compte que les séquences. Selon ce principe, si deux topologies sont jugées statistiquement équivalentes par un test de vraisemblance, mais qu'une des topologies permet d'inférer un meilleur ordre des gènes (i.e. un génome ancestral plus linéaire), on peut préférer la topologie qui satisfait le plus de critères : signal des séquences et linéarité des génomes ancestraux.

3.4 Fiabilité des topologies et autres méthodes de correction

L'idée d'intégrer la synténie dans des analyses phylogénétiques n'est pas nouvelle. En effet, la synténie est fréquemment utilisée pour inférer des relations d'homologie entre des gènes [5, 70] en complément ou à la place des méthodes basées sur la similarité des séquences. L'idée sous-jacente est que la conservation de l'ordre d'une séquence de gènes dans un groupe d'espèces peut refléter une histoire évolutive commune. Des méthodes, telles que [87] ont ainsi été développées pour déterminer des *blocs de synténie* conservés à travers les espèces et tenter de reconstruire les relations évolutives entre les gènes de ces blocs.

Il existe un grand nombre de méthodes de correction d'arbre de gènes. J'ai choisi de discuter plus particulièrement ici des méthodes développées par nos collaborateurs Manuel Lafond, Nadia El-Mabrouk et Krister Swenson, et de celles que nous avons développées en collaboration avec eux. Les articles A.4 et A.5 écrits au cours de ma thèse sont le fruit de cette collaboration. Dans l'article "Gene tree correction guided by orthology" A.4 [76], nous proposons ainsi deux méthodes de correction similaires : la Unduplicator présentée dans la section 3.3, et ParalogyCorrector développée par nos

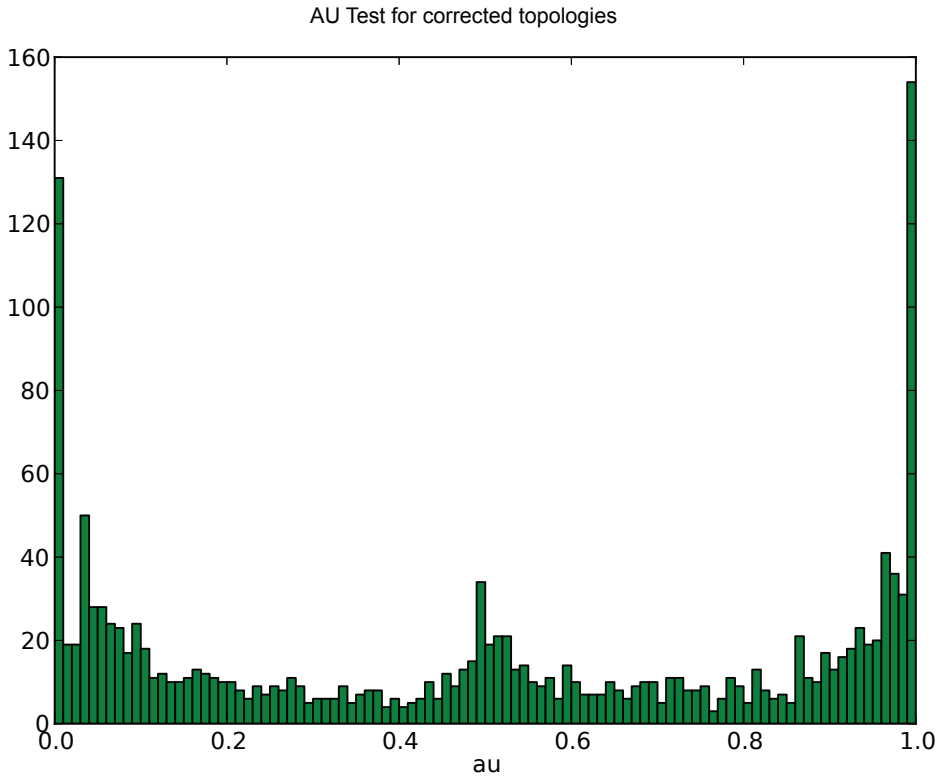


FIGURE 3.8 – Distribution des p-value correspondant au test H_0 : "la topologie corrigée est meilleure que la topologie initiale" obtenues avec des test AU pour les 1519 arbres tirés d'Ensembl.

collaborateurs. Dans ParalogyCorrector, la correction consiste à proposer la topologie la plus proche possible de la topologie originale qui respecte des relations d'orthologie inférées à partir de blocs de synténie. Plus exactement :

- On prend en entrée :
  - Un arbre de gènes.
  - Un arbre d'espèces.
  - Des relations d'orthologies deux à deux parmi les gènes aux feuilles de l'arbre.
- On obtient :

– L'arbre de gène le plus proche possible (selon la distance de Robinson-Foulds) de l'arbre donné en entrée, avec la contrainte que l'ancêtre commun le plus récent des couples de gènes apparaissant dans les relations d'orthologies soit une spéciation.

Ce problème a une solution en temps polynomial (voir A.4).

Une limite des méthodes Unduplicator et ParalogyCorrector est qu'elles ne prennent pas en compte de support des nœuds modifiés par la correction. Il est donc possible de corriger des nœuds fortement soutenus par ailleurs, ce qui peut conduire à des topologies statistiquement moins bonnes que les topologies initiales. Ensembl fournit un support pour les nœuds de duplication dans ses arbres de gènes. Ce support est calculé à partir du nombre de gènes dupliqués observables chez les descendants de la duplication (Figure 3.9). Une duplication se voit ainsi attribuer un score de 0 s'il n'existe aucun gène dupliqué observable chez ses descendants. Ces duplications peu fiables sont annotées comme douteuses ("Dubious") dans les arbres d'Ensembl, et correspondent aux duplications non apparentes (NAD) étudiées par [75, 131].

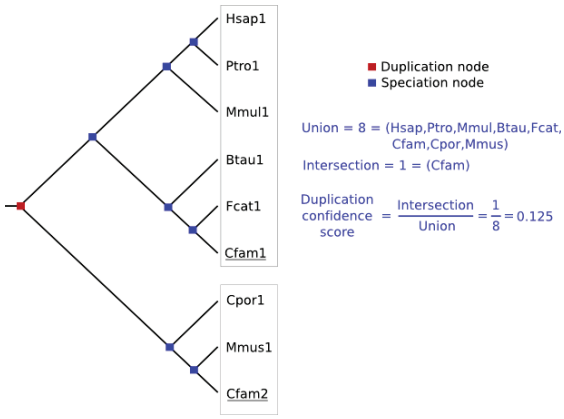


FIGURE 3.9 – Calcul du score d'une duplication dans Ensembl. Tiré du site d'Ensembl.

Nous avons comparé la détection d'erreur par la synténie avec les duplications non apparentes d'Ensembl (Figure 3.10). 85% des duplications corrigées avec Unduplicator correspondent à des NAD, ce qui confirme le caractère douteux de ces duplications. Si on regarde les 15% d'arbres restant, la correction reste acceptable dans la plupart des cas, mais une proportion plus élevée de topologies corrigées sont rejetées (11%).

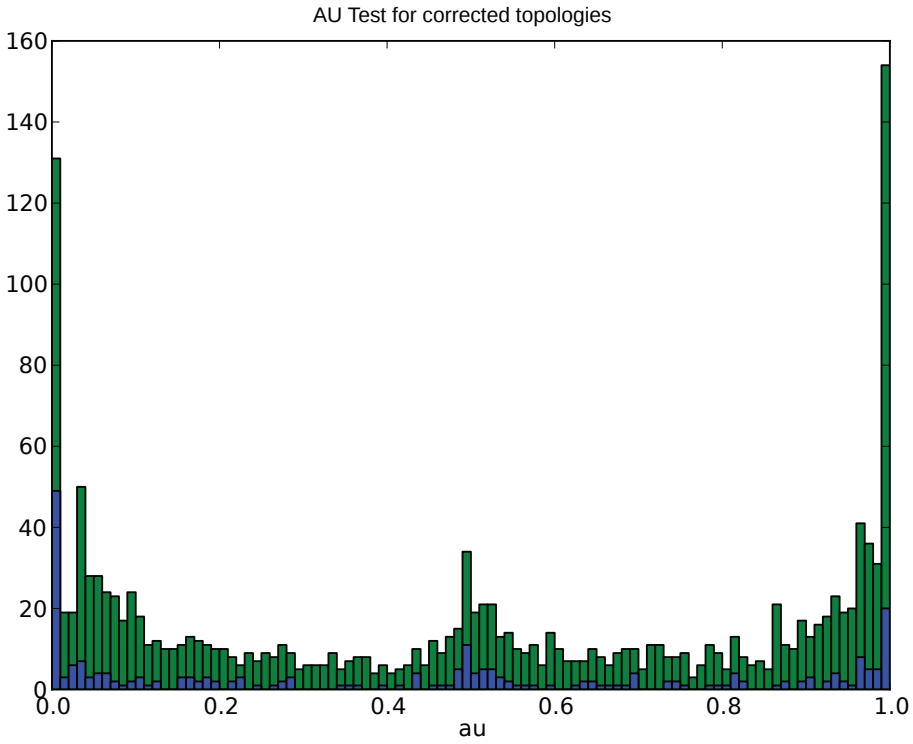


FIGURE 3.10 – Distribution des p-value correspondant au test H_0 : "la topologie corrigée est meilleure que la topologie initiale" obtenues avec des test AU pour les 1519 arbres tirés d'Ensembl. En bleu : les arbres pour lesquels la correction ne concerne pas une NAD.

Dans un autre registre, les supports tels que le bootstrap peuvent être utilisés par des méthodes de correction. Certaines méthodes de réconciliation recherchent ainsi le meilleur arbre réconcilié en effectuant des réarrangement locaux de type NNI autour des nœuds à faible support [77, 131, 29, 101]. La méthode ProfileNJ, décrite dans l'article écrit en collaboration avec Emmanuel Noutahi, Manuel Lafond, Nadia El-Mabrouk, et Bastien Boussau [103] (Appendix A), se base sur ces méthodes mais propose une correction guidée par l'arbre des espèces. Cette méthode de correction consiste à contracter les nœuds à faible support dans les arbres de gènes, puis à résoudre les polytomies ainsi crées en se basant sur l'arbre des espèces et un algorithme de Neighbour Joining pour résoudre la topologie dans la cas de gènes dupliqués. Une

description plus exacte de ProfileNJ est :

- On prend en entrée :
  - Un arbre de gènes avec des supports statistiques.
  - Un arbre d'espèces.
  - Un seuil de support.
  - Une matrice de distance.
- On obtient :
  - Un arbre de gène corrigé dans lequel toutes les branches de l'arbre donné en entrée soutenues au-delà du support sont présentes.
  - Parmi toutes les arbres possibles, on choisit celui qui minimise le nombre de duplications et de pertes.
  - Parmi toutes les arbres qui conduisent au nombre minimum de duplications et de pertes, on choisit celui dont les distances sont les plus proches possible de celles de la matrice.

Ce problème a une solution en temps polynomial, sauf en ce qui concerne le critère de distance qui est approximé par un Neighbour Joining.

L'algorithme de ProfileNJ a été développé par nos collaborateurs Emmanuel Noutahi, Manuel Lafond et Nadia El-Mabrouk. Nous avons ensuite, avec ces même collaborateurs, intégré ProfileNJ, ParalogyCorrector et Unduplicator dans une pipeline de correction d'arbres de gènes appelée RefineTree et présentée dans l'article A.5 [103].

ProfileNJ présente l'avantage de ne modifier que des sous-arbres faiblement soutenus par l'alignement, et comparée à Unduplicator, réduit donc le risque de faire chuter la vraisemblance de l'arbre en effectuant la correction. En réalité, les topologies corrigées par ProfileNJ ont souvent une meilleure vraisemblance que les topologies initiales, ou les topologies corrigées par d'autres méthodes (Figure 3.11).

ProfileNJ présente toutefois un léger inconvénient lié à l'utilisation de supports statistiques. Il faut en effet définir un seuil en dessous duquel le support sera considéré comme faible et la correction pourra être appliquée. Or, définir un tel seuil n'est pas trivial [109]. De plus, un fort soutien et une vraisemblance élevée ne permettent

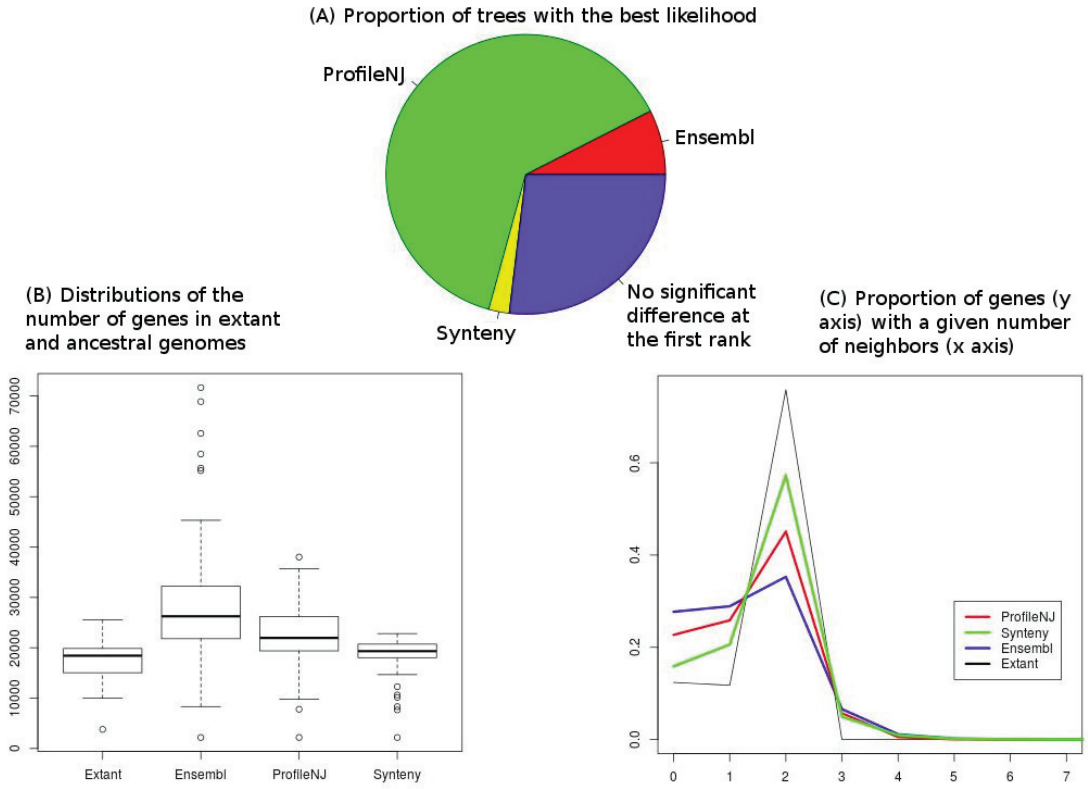


FIGURE 3.11 – (A) Vraisemblance des séquences étant données trois topologies pour chaque famille : la topologie d'Ensembl, la topologie corrigée avec ProfileNJ, et la topologie corrigée avec une méthode de synténie (Unduplicator ou ParalogyCorrector selon les arbres). On représente sur le diagramme la meilleure topologie d'après un test AU. (B) Le nombre de gènes ancestraux inférés avec DeCo. Le nombre de gènes dans les génomes actuels est donné comme référence ("Extant"). (C) La linéarité des génomes ancestraux d'après les adjacences ancestrales inférées avec DeCo. En abscisse, le nombre de voisins possibles pour un gène. En ordonnée, la proportion de gènes ancestraux pour lesquels on infère ce nombre de voisins. On donne les génomes actuels ("Extant") comme référence. Pour (B) et (C), on fait l'hypothèse que les statistiques étudiées sont meilleures lorsqu'elles sont proches des valeurs des génomes actuels. Tiré de [103].

pas forcément de dire que l'arbre obtenu reflète l'histoire évolutive de la famille de gènes. Cette difficulté est due au fait qu'il peut exister plusieurs histoires évolutives pour une même famille de gènes. Ainsi, dans l'arbre de la figure 3.12, l'alignement des

séquences conduit à fort soutien statistique pour le groupe chimpanzé-gorille (gènes ENSPTRP00000033018 et ENSGGOP00000011432). Cette topologie ne serait donc pas remise en question avec ProfileNJ. Cependant, les contraintes d'orthologies établies avec la méthode PhylDiag [87] conduisent ParalogyCorrector à corriger cette topologie en branchant ces deux gènes successivement au-dessus des deux gènes humains ENSP00000414208 et ENSP00000378687. On a donc une contradiction entre les deux méthodes de correction concernant cet endroit de l'arbre : si on veut respecter les contraintes d'orthologie (ParalogyCorrector), il faut casser un branchement associé à un fort soutien statistique, ce qui n'est pas autorisé avec ProfileNJ. cette contradiction peut s'expliquer si le regroupement des gènes ENSPTRP00000033018 et ENSGGOP00000011432 dans l'arbre initial est dû à un tri des lignées incomplet (voir section 1.1.5.3).

FIGURE 3.12 – Potentiel exemple de tri de lignées incomplet. À gauche, un arbre de gènes avec ses supports statistiques (aLRT). Le groupe gorille-chimpanzé est fortement soutenu (en rouge). À droite l'arbre corrigé avec ParalogyCorrector. Le groupe gorille-chimpanzé a été cassé pour respecter les contraintes d'orthologie déterminées avec PhylDiag (en rouge). Tiré de [103].

En présence de tri de lignées incomplet, ou de recombinaison, deux histoires évolutives distinctes peuvent co-exister pour une famille de gènes. D'une part, l'histoire évolutive des séquences, soumise au tri de lignées incomplet et à la recombinaison. D'autre part, l'histoire évolutive du locus. Ces deux histoires évolutives peuvent conduire à des topologies différentes tout en étant toutes les deux correctes. Cette contradiction apparente s'explique en fait par l'ambiguïté de la définition d'un gène comme une séquence

ADN ou protéique, ou comme une entité indivisible liée à un locus. Rasmussen et al. [115] modélisent cette ambiguïté en présence de tri de lignées incomplet en construisant séparément un arbre pour les séquences et un arbre pour le locus. Les méthodes de correction basées sur la synténie peuvent être vues comme un effort pour intégrer l'histoire évolutive du locus dans l'arbre de gène. La prochaine étape semble alors être l'intégration complète d'un critère de synténie dans un algorithme de réconciliation. Pour l'instant, les méthodes de réconciliation qui tentent d'intégrer de la synténie sont peu nombreuses [138]. Mais avec les efforts réalisés pour intégrer la réconciliation avec les étapes en amont, notamment par Boussau et al. [24], on peut espérer réaliser un jour une réconciliation qui maximiserait conjointement la vraisemblance de la topologie d'après les séquences et un critère de synténie.

4 Conclusion

Une thématique centrale de ma thèse a été l'articulation de plusieurs échelles d'évolution, en particulier l'évolution de la synténie, des familles de gènes, des séquences géniques et des espèces. Avec Harpi, on modélise l'évolution d'adjacences à partir d'arbres de gènes réconciliant l'évolution des gènes et des espèces (voir chapitre 2). Avec Unduplicator, la méthode de correction d'arbres de gènes présentée dans le chapitre 3, on utilise l'inférence d'adjacences ancestrales pour détecter des erreurs dans les topologies des arbres de gènes et proposer des topologies alternatives. Ces deux approches utilisent l'information synténique à deux échelles différentes : l'une au niveau d'une adjacence et l'autre au niveau des relations entre les adjacences. Chaque approche correspond à un moyen différent d'intégrer la synténie à la reconstruction phylogénétique. J'ai envisagé ou essayé d'intégrer chacune de ces approches dans la méthode de reconstruction phylogénétique multi-échelle Phyldog [24].

Pour Harpi, l'intégration envisagée serait de regrouper les modèles d'évolution, comme Phyldog le fait déjà pour les séquences et le contenu en gènes. En effet, l'intégration de plusieurs échelles permet de reconstruire l'histoire évolutive des génomes de manière plus complète, mais aussi de détecter et de limiter les biais et erreurs lors d'une reconstruction phylogénétique [22]. Boussau et al. ont par exemple montré qu'il était possible de reconstruire de meilleurs arbres de gènes et d'espèces en intégrant la réconciliation à la reconstruction phylogénétique dans un modèle probabiliste [24]. L'approche probabiliste est attrayante pour la modélisation multi-échelle de l'évolution. D'une part, il est possible de reconstruire une histoire évolutive en prenant en compte

toutes les solutions possibles, au lieu d'une seule solution parmi de nombreuses équivalentes pour les méthodes de parcimonie. D'autre part, il est possible de combiner des vraisemblances calculées à partir de plusieurs objets évolutifs. Lorsqu'on utilise Phyml [63] ou RAxML [128], on calcule ainsi la vraisemblance d'alignements multiples étant donné un modèle d'évolution de séquences. Avec Harpi, on calcule la vraisemblance d'adjacences étant donnés un modèle d'évolution pour les duplications et les pertes de gènes et d'adjacences, des arbres réconciliés, et des longueurs de branches. Avec Phyldog, Boussau et al. calculent pour chaque arbre de gènes une vraisemblance correspondant à l'alignement des séquences et une vraisemblance correspondant à la réconciliation avec l'arbre des espèces [24]. La vraisemblance globale d'un arbre de gènes est ensuite le produit de ces deux vraisemblances.

Dans le même esprit, on peut imaginer intégrer la vraisemblance des adjacences calculée par Harpi dans le calcul d'une vraisemblance globale. En s'inspirant de la vraisemblance calculée par Phyldog, on pourrait ainsi calculer une vraisemblance globale correspondant au produit d'une vraisemblance basées sur les séquences, d'une vraisemblance basée sur la réconciliation, et d'une vraisemblance basée sur les adjacences (Figure 4.1). En étant capable de calculer une telle vraisemblance, on pourrait ensuite effectuer de manière conjointe la reconstruction des arbres de gènes, de la synténie ancestrale, et la réconciliation. Cette démarche n'est cependant pas si évidente, notamment parce que les ordres de grandeurs et la variabilité des vraisemblances calculées à partir d'objets évolutifs différents peuvent varier considérablement. Comparer et multiplier des vraisemblances correspondant à des objets et des modèles différents n'est alors peut-être pas la solution la plus pertinente. Une meilleure solution pourrait être de multiplier des probabilités *a posteriori* correspondant aux différents objets. Une autre difficulté est le temps et la mémoire requis pour calculer et maximiser les vraisemblances. Cette difficulté contraint déjà l'utilisation de Phyldog. L'intégration de l'évolution de la synténie, des familles de gènes et des séquences dans un unique modèle probabiliste n'est ainsi pas encore tout à fait à notre portée.

Les méthodes de correction d'arbres de gènes peuvent aussi s'intégrer à Phyldog. Dans ce cas, elles servent à guider la recherche de topologie et à améliorer les performances de Phyldog. Nous avons par exemple montré qu'il est possible d'améliorer les arbres de gènes reconstruits avec Phyldog en proposant en entrée des arbres reconstruits avec la méthode ProfileNJ. Les figures 4.2 et 4.3 montrent les résultats d'une

analyse effectuée sur un échantillon de 1942 arbres de gènes tirés d' Ensembl. À chacune de ces familles de gènes, nous avons associé trois topologies :

- La topologie disponible dans Ensembl.
- La topologie corrigée avec ProfileNJ.
- La topologie reconstruite avec Phyldog, avec comme point de départ pour la recherche locale la topologie ProfileNJ.

Nous avons ensuite inféré les adjacences ancestrales avec DeCo pour tous les arbres de gènes, et ce pour les trois jeux de données correspondant aux trois types de topologies. Ces résultats préliminaires montrent que la linéarité est meilleure lorsqu'on reconstruit les arbres de gènes avec Phyldog, en donnant les arbres ProfileNJ comme point de départ pour la recherche locale (Figure 4.2). En ce qui concerne les contenus en gènes des génomes ancestraux, on observe pour les arbres d'Ensembl une grande variabilité dans les tailles (le nombre de gènes) des génomes ancestraux après la réconciliation. Cette variabilité est moindre pour les arbres reconstruits avec ProfileNJ et les arbres reconstruits avec Phyldog à partir des arbres ProfileNJ, et plus proche de la variabilité observée dans les génomes actuels (Figure 4.3).

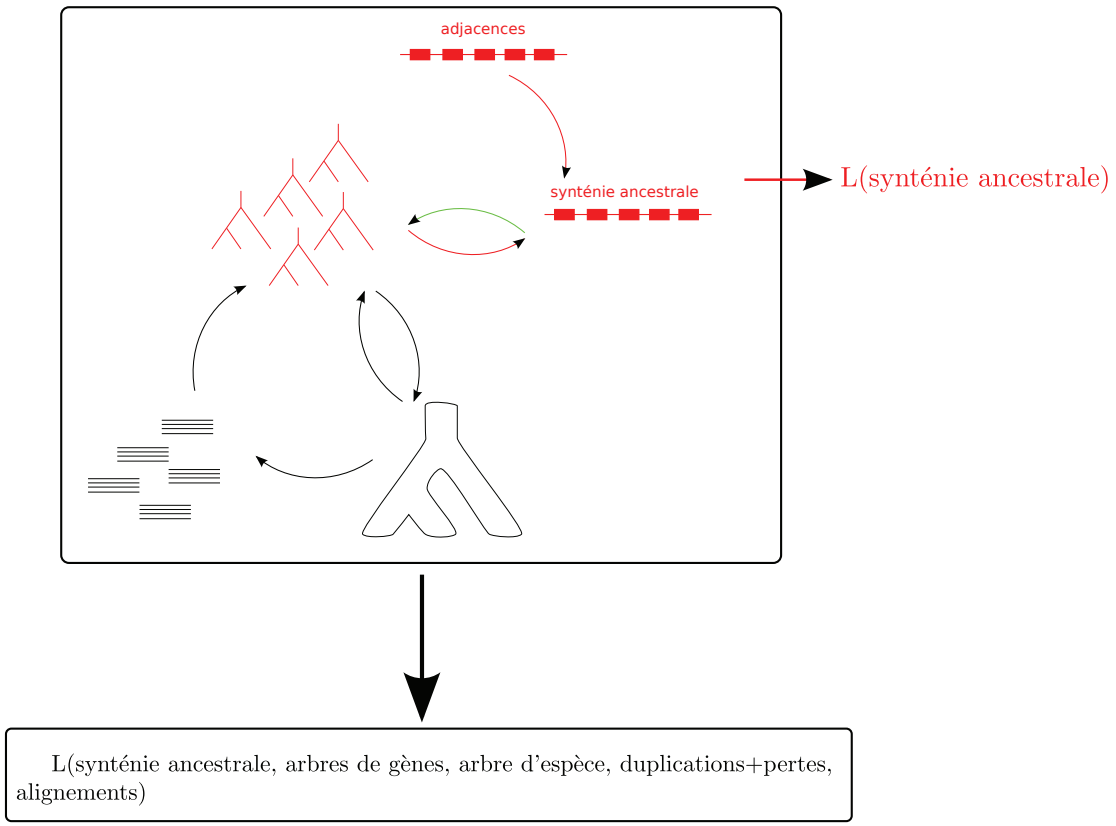


FIGURE 4.1 – Une manière possible d'intégrer un modèle d'évolution de synténie dans Phyldog. Les flèches représentent les étapes de l'inférence. En rouge les données utilisées en entrée et l'inférence de la synténie ancestrale par Harpi. En vert, les méthodes de corrections, ou le calcul d'un vraisemblance prenant en compte celle calculée par Harpi. $L(...)$ représente une vraisemblance.

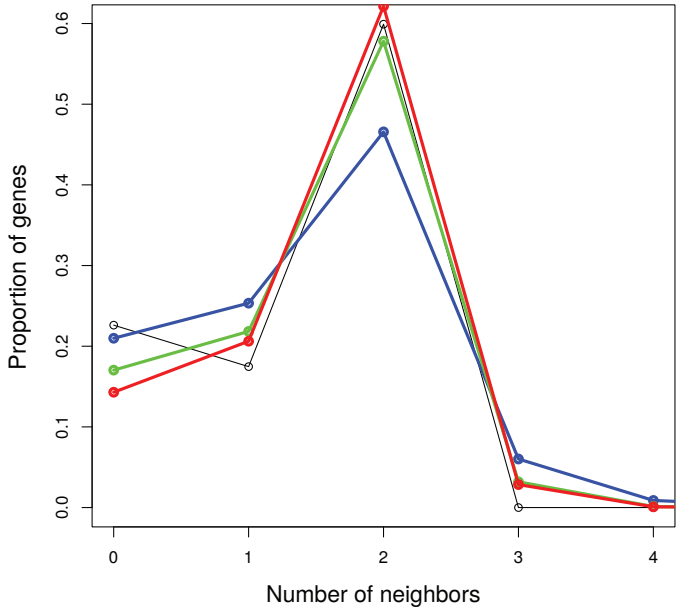


FIGURE 4.2 – Nombres de voisins pour les gènes ancestraux, inférés avec DeCo. En noir : les génomes actuels (la référence). En bleu, les topologies d'Ensembl. En vert : les topologies corrigées avec ProfileNJ, données comme arbres de départ pour Phyldog. En rouge : les topologies obtenues après optimisation et réconciliation avec Phyldog des arbres ProfileNJ. Figure réalisée par Thomas Bigot.

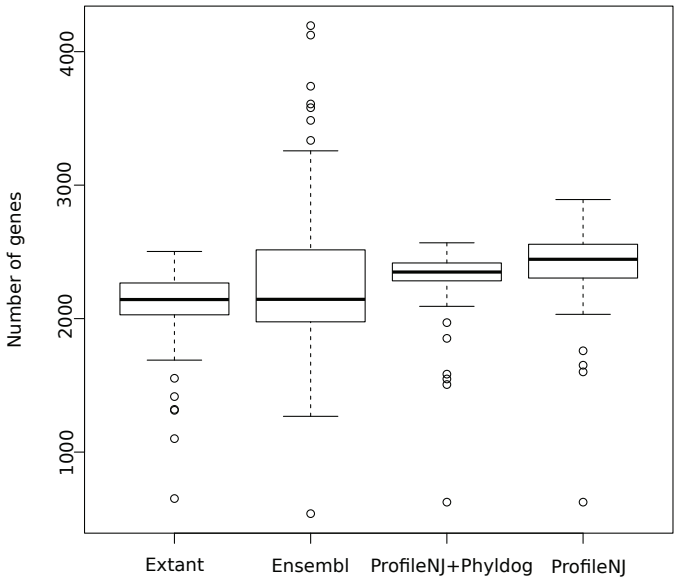


FIGURE 4.3 – Nombres de gènes ancestraux pour chaque topologie après réconciliation. "Extant" correspond aux génomes actuels, donnés comme référence. "Ensembl" correspond aux topologies d'Ensembl, "ProfileNJ" aux topologies corrigées par ProfileNJ, et "ProfileNJ+Phyldog" aux arbres reconstruits avec Phyldog, en partant des arbres ProfileNJ. Figure réalisée par Thomas Bigot.

Bibliographie

[1] Sophie S ABBY, Eric TANNIER, Manolo GOUY et Vincent DAUBIN : Lateral gene transfer as a support for the tree of life. *Proc Natl Acad Sci USA*, 109(13): 4962–4967, 3 2012.

[2] Orjan AKERBORG, Bengt SENNBLAD, Lars ARVESTAD et Jens LAGERGREN : Simultaneous bayesian gene tree reconstruction and reconciliation analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 106(14):5714–9, 4 2009.

[3] John ALDRICH : R.a. fisher and the making of maximum likelihood 1912-1922. *Statistical Science*, 12(3):162–176, 9 1997.

[4] Max A ALEKSEYEV et Pavel A PEVZNER : Are there rearrangement hotspots in the human genome ? *PLoS computational biology*, 3(11):e209, 11 2007.

[5] Raja ALI, Sayyed MUHAMMAD, Mehmood KHAN et Lars ARVESTAD : Quantitative synteny scoring improves homology inference and partitioning of gene families. *BMC Bioinformatics*, 14(Suppl 15):S12, 2013.

[6] S F ALTSCHUL, W GISH, W MILLER, E W MYERS et D J LIPMAN : Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–10, 10 1990.

[7] Yoann ANSELMETTI, Vincent BERRY, Cedric CHAUVE, Annie CHATEAU, Eric TANNIER et Sèverine BÉRARD : Ancestral gene synteny reconstruction improves extant species scaffolding. *BMC Genomics*, 16(Suppl 10):S11, 2015.

[8] L ARVESTAD, A C BERGLUND, J LAGERGREN et B SENNBLAD : Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. *In RECOMB*, pages 326\–335, 2004.

[9] Lars ARVESTAD, Ann-Charlotte BERGLUND, Jens LAGERGREN et Bengt SENNBLAD : Bayesian gene/species tree reconciliation and orthology analysis using mcmc. *Bioinformatics (Oxford, England)*, 19 Suppl 1:7–15, 1 2003.

[10] Doris BACHTROG : Y-chromosome evolution : emerging insights into processes of y-chromosome degeneration. *Nature reviews. Genetics*, 14(2):113–24, 3 2013.

[11] Eric BAPTESTE, Maureen A O'MALLEY, Robert G BEIKO, Marc ERESHEFSKY, J Peter GOGARTEN, Laura FRANKLIN-HALL, François-Joseph LAPOINTE, John DUPRÉ, Tal DAGAN, Yan BOUCHER et William MARTIN : Prokaryotic evolution and the tree of life are two different things. *Biology direct*, 4:34, 1 2009.

[12] Sèverine BÉRARD, Coralie GALLIEN, Bastien BOUSSAU, Gergely J SZÖLLŐSI, Vincent DAUBIN et Eric TANNIER : Evolution of gene neighborhoods within reconciled phylogenies. *Bioinformatics (Oxford, England)*, 28(18):i382–i388, 9 2012.

[13] Ann Charlotte BERGLUND-SONNHAMMER, Pär STEFFANSSON, Matthew J. BETTS et David A. LIBERLES : Optimal gene trees from sequences and species trees using a soft interpretation of parsimony. *Journal of Molecular Evolution*, 63(2):240–250, 2006.

[14] Ulfar BERGTHORSSON, Dan I ANDERSSON et John R ROTH : Ohno's dilemma : evolution of new genes under continuous selection. *Proceedings of the National Academy of Sciences of the United States of America*, 104(43):17004–9, 10 2007.

[15] Arjun BHUTKAR, William M GELBART et Temple F SMITH : Inferring genome-scale rearrangement phylogeny and ancestral gene order : a drosophila case study. *Genome biology*, 8(11):R236, 1 2007.

[16] Olaf R. P. BININDA-EMONDS, John L. GITTLEMAN et Mike A STEEL : The (super)tree of life : Procedures, problems, and prospects. *Annual Review of Ecology and Systematics*, 33:265–289, 2002.

[17] M BLANCHETTE, G BOURQUE et D SANKOFF : Breakpoint phylogenies. *Genome informatics. Workshop on Genome Informatics*, 8:25–34, 1 1997.

[18] Tine BLOMME, Klaas VANDEPOELE, Stefanie DE BODT, Cedric SIMILLION, Steven MAERE et Yves Van de PEER : The gain and loss of genes during 600 million years of vertebrate evolution. *Genome biology*, 7(5):R43, 1 2006.

[19] Digamber S BORGAONKAR : Chromosomal variation in man, 1975.

[20] Guillaume BOURQUE et Pavel A. PEVZNER : Genome-scale evolution : Reconstructing gene orders in the ancestral species. *Genome Res.*, 12(1):26–36, 1 2002.

[21] Guillaume BOURQUE, Pavel A PEVZNER et Glenn TESLER : Reconstructing the genomic architecture of ancestral mammals : lessons from human, mouse, and rat genomes. *Genome research*, 14(4):507–16, 4 2004.

[22] Bastien BOUSSAU et Vincent DAUBIN : Genomes as documents of evolutionary history. *Trends in ecology & evolution*, 25(4):224–32, 4 2010.

[23] Bastien BOUSSAU et Manolo GOUY : Efficient likelihood computations with nonreversible models of evolution. *Systematic biology*, 55(5):756–68, 10 2006.

[24] Bastien BOUSSAU, Gergely J SZÖLLOSI, Laurent DURET, Manolo GOUY, Eric TANNIER et Vincent DAUBIN : Genome-scale coestimation of species and gene trees. *Genome research*, 23(2):323–30, 2 2013.

[25] Céline BROCHIER, Eric BAPTESTE, David MOREIRA et Hervé PHILIPPE : Eubacterial phylogeny based on translational apparatus proteins. *Trends in genetics : TIG*, 18(1):1–5, 1 2002.

[26] D W BURT, C BRULEY, I C DUNN, C T JONES, A RAMAGE, A S LAW, D R MORRICE, I R PATON, J SMITH, D WINDSOR, A SAZANOV, R FRIES et D WADDINGTON : The dynamics of chromosome evolution in birds and mammals. *Nature*, 402(6760):411–3, 11 1999.

[27] G L BUSH, S M CASE, A C WILSON et J L PATTON : Rapid speciation and chromosomal evolution in mammals. *Proceedings of the National Academy of Sciences of the United States of America*, 74(9):3942–6, 9 1977.

[28] Roger BUTLIN, Allan DEBELLE, Claudius KERTH, Rhonda R SNOOK, Leo W BEUKEBOOM, Ruth F CASTILLO CAJAS, Wenwen DIAO, Martine E MAAN, Silvia PAOLUCCI, Franz J WEISSING, Louis van de ZANDE, Anneli HOIKKALA, Elzemiek GEUVERINK, Jackson JENNINGS, Maaria KANKARE, K Emily KNOTT, Venera I TYUKMAEVA, Christos ZOUMADAKIS, Michael G RITCHIE, Daniel BARKER, Elina IMMONEN, Mark KIRKPATRICK, Mohamed NOOR, Constantino MACIAS GARCIA, Thomas SCHMITT et Menno SCHILTHUIZEN : What do we need to know about speciation ? *Trends in ecology & evolution*, 27(1):27–39, 1 2012.

[29] Wen-Chieh CHANG et Oliver EULENSTEIN : Reconciling gene trees with apparent polytomies. *Computer Science Technical Reports*, 2006.

[30] Cedric CHAUVE, Nadia EL-MABROUK, Laurent GUÉGUEN, Magali SEMERIA et Eric TANNIER : Duplication, rearrangement and reconciliation : A follow-up 13 years later. *In* Cedric CHAUVE, Nadia EL-MABROUK et Eric TANNIER, éditeurs : *Models and Algorithms for Genome Evolution*, chapitre 4, pages 47–62. Springer, 2013.

[31] Cedric CHAUVE et Eric TANNIER : A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes. *PLoS computational biology*, 4(11):e1000234, 11 2008.

[32] Avril COGHLAN, Evan E EICHLER, Stephen G OLIVER, Andrew H PATERSON et Lincoln STEIN : Chromosome evolution in eukaryotes : a multi-kingdom perspective. *Trends in genetics : TIG*, 21(12):673–82, 12 2005.

[33] Avril COGHLAN et Kenneth H WOLFE : Fourfold faster rate of genome rearrangement in nematodes than in drosophila. *Genome research*, 12(6):857–67, 6 2002.

[34] Stephen A. COOK : The complexity of theorem-proving procedures.

[35] Tal DAGAN et William MARTIN : The tree of one percent. *Genome biology*, 7(10):118, 1 2006.

[36] Charles DARWIN : *C. Darwin, The Origin of Species by Means of Natural Selection*. Murray, 1859.

[37] Constantinos DASKALAKIS et Sebastien ROCH : Species trees from gene trees despite a high rate of lateral genetic transfer : A tight bound. *arXiv*, 1508.01962, 8 2015.

[38] Vincent DAUBIN, Manolo GOUY et Guy PERRIÈRE : A phylogenomic approach to bacterial phylogeny : evidence of a core of genes sharing a common history. *Genome research*, 12(7):1080–90, 7 2002.

[39] Julian DAVIES et Dorothy DAVIES : Origins and evolution of antibiotic resistance. *Microbiology and molecular biology reviews : MMBR*, 74(3):417–33, 9 2010.

[40] Frédéric DELSUC, Henner BRINKMANN et Hervé PHILIPPE : Phylogenomics and the reconstruction of the tree of life. *Nature reviews. Genetics*, 6(5):361–75, 5 2005.

[41] A DEQUEIROZ et J GATESY : The supermatrix approach to systematics. *Trends in Ecology & Evolution*, 22(1):34–41, 1 2007.

[42] W. F. DOOLITTLE : Phylogenetic classification and the universal tree. *Science*, 284(5423):2124–2128, 6 1999.

[43] Jean-Philippe DOYON, Céline SCORNAVACCA, K. Yu. GORBUNOV, Gergely J SZÖLLŐSI, Vincent RANWEZ et Vincent BERRY : Comparative genomics. *In* Eric TANNIER, éditeur : *Comparative Genomics*, volume 6398 de *Lecture Notes in Computer Science*, pages 93–108. Springer Berlin Heidelberg, 2010.

[44] Laurent DURET, Corinne CHUREAU, Sylvie SAMAIN, Jean WEISSENBACH et Philip AVNER : The xist rna gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science (New York, N.Y.)*, 312(5780):1653–5, 6 2006.

[45] Julien DUTHEIL, Sylvain GAILLARD, Eric BAZIN, Sylvain GLÉMIN, Vincent RANWEZ, Nicolas GALTIER et Khalid BELKHIR : Bio++ : a set of c++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. *BMC bioinformatics*, 7:188, 1 2006.

[46] Evan E EICHLER et David SANKOFF : Structural dynamics of eukaryotic chromosome evolution. *Science (New York, N.Y.)*, 301(5634):793–7, 8 2003.

[47] J FELSENSTEIN : Evolurionary trees from {dna} sequences : a maximum likelihood approach. *Journal of Molecular Evolution*, 17:368\–376, 1981.

[48] Joseph FELSENSTEIN : Confidence limits on phylogenies : An approach using the bootstrap. *Evolution*, pages 783–791, 1985.

[49] Joseph FELSENSTEIN : *Inferring Phylogenies*. Sinauer Associates, Incorporated, 2004.

[50] W. M. FITCH : Toward defining the course of evolution : Minimum change for a specific tree topology. *Systematic Biology*, 20(4):406–416, 12 1971.

[51] A FORCE, M LYNCH, F B PICKETT, A AMORES, Y L YAN et J POSTLETHWAIT : Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4):1531–45, 4 1999.

[52] Yoshikazu FURUTA, Mikihiko KAWAI, Koji YAHARA, Noriko TAKAHASHI, Naofumi HANDA, Takeshi TSURU, Kenshiro OSHIMA, Masaru YOSHIDA, Takeshi AZUMA, Masahira HATTORI, Ikuo UCHIYAMA et Ichizo KOBAYASHI : Birth and death of genes linked to chromosomal inversion. *Proceedings of the National Academy of Sciences of the United States of America*, 108(4):1501–6, 1 2011.

[53] Yves GAGNON, Mathieu BLANCHETTE et Nadia EL-MABROUK : A flexible ancestral genome reconstruction method based on gapped adjacencies. *BMC bioinformatics*, 13 Suppl 1(Suppl 19):S4, 1 2012.

[54] Nicolas GALTIER et Vincent DAUBIN : Dealing with incongruence in phylogenomic analyses. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 363(1512):4023–9, 12 2008.

[55] Sergey GAVRILETS : Models of speciation : where are we now ? *The Journal of heredity*, 105 Suppl(S1):743–55, 1 2014.

[56] S. K. GIRE, A. GOBA, K. G. ANDERSEN, R. S. G. SEALFON, D. J. PARK, L. KANNEH, S. JALLOH, M. MOMOH, M. FULLAH, G. DUDAS, S. WOHL, L. M. MOSES, N. L. YOZWIAK, S. WINNICKI, C. B. MATRANGA, C. M. MALBOEUF, J. QU, A. D. GLADDEN, S. F. SCHAFFNER, X. YANG, P.-P. JIANG, M. NEKOUI, A. COLUBRI, M. R. COOMBER, M. FONNIE, A. MOIGBOI, M. GBAKIE, F. K. KAMARA, V. TUCKER, E. KONUWA, S. SAFFA, J. SELLU, A. A. JALLOH, A. KOVOMA, J. KONINGA, I. MUSTAPHA, K. KARGBO, M. FODAY, M. YILLAH, F. KANNEH, W. ROBERT, J. L. B. MASSALLY, S. B. CHAPMAN, J. BOCHICCHIO, C. MURPHY, C. NUSBAUM, S. YOUNG, B. W. BIRREN, D. S. GRANT, J. S. SCHEIFFELIN, E. S. LANDER, C. HAPPI, S. M. GEVAO, A. GNIRKE, A. RAMBAUT, R. F. GARRY, S. H. KHAN et P. C. SABETI : Genomic surveillance elucidates ebola virus origin and transmission during the 2014 outbreak. *Science*, 345(6202):1369–72, 8 2014.

[57] Allen G. Nick GOLDMAN, Jon P. Anderson : Likelihood-based tests of topologies in phylogenetics. *Systematic Biology*, 49(4):652–670, 12 2000.

[58] M. GOODMAN, J. CZELUSNIAK, G. W. MOORE, A. E. ROMERO-HERRERA et G. MATSUDA : Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Biology*, 28(2):132–163, 6 1979.

[59] Leo GOODSTADT et Chris P PONTING : Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS computational biology*, 2(9):e133, 9 2006.

[60] F GRIFFITH : The significance of pneumococcal types. *The Journal of hygiene*, 27(2):113–59, 1 1928.

[61] Laurent GUÉGUEN, Sylvain GAILLARD, Bastien BOUSSAU, Manolo GOUY, Mathieu GROUSSIN, Nicolas C ROCHETTE, Thomas BIGOT, David FOURNIER, Fanny POUYET, Vincent CAHAIS, Aurélien BERNARD, Céline SCORNAVACCA, Benoît NABHOLZ, Annabelle HAUDRY, Loïc DACHARY, Nicolas GALTIER, Khalid BELKHIR et Julien Y DUTHEIL : Bio++ : efficient extensible libraries and tools for computational molecular evolution. *Molecular biology and evolution*, 30(8):1745–50, 8 2013.

[62] S GUINDON et O GASCUEL : A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52:696\– 704, 2003.

[63] Stéphane GUINDON, Frédéric DELSUC, Jean-François DUFAYARD et Olivier GASCUEL : *Bioinformatics for DNA Sequence Analysis*, volume 537 de *Methods in Molecular Biology*. Humana Press, 1 2009.

[64] M W HAHN : Bias in phylogenetic tree reconciliation methods : implications for vertebrate genome evolution. *Genome Biology*, 8(R141), 2007.

[65] Matthew W HAHN : Distinguishing among evolutionary models for the maintenance of gene duplicates. *The Journal of heredity*, 100(5):605–17, 1.

[66] Joseph L HERMAN, Ádám NOVÁK, Rune LYNGSØ, Adrienn SZABÓ, István MIKLÓS et Jotun HEIN : Efficient representation of uncertainty in multiple sequence alignments using directed acyclic graphs. *BMC bioinformatics*, 16(1): 108, 1 2015.

[67] Fei HU, Jun ZHOU, Lingxi ZHOU et Jijun TANG : Probabilistic reconstruction of ancestral gene orders with insertions and deletions. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 5963(c):1–1, 2014.

[68] A L HUGHES : The evolution of functionally novel proteins after gene duplication. *Proceedings. Biological sciences / The Royal Society*, 256(1346):119–24, 5 1994.

[69] Hideki INNAN et Fyodor KONDRASHOV : The evolution of gene duplications : classifying and distinguishing between models. *Nature Reviews Genetics*, 11(4):4, 1 2010.

[70] Jin JUN, Ion I MANDOIU et Craig E NELSON : Identification of mammalian orthologs using local synteny. *BMC genomics*, 10(1):630, 1 2009.

[71] R. M. KARP : Reducibility among combinatorial problems. *In* R. E. MILLER et J.W THATCHER, éditeurs : *Complexity of Computer Computations*, pages 85–103. Plenum Press, 1972.

[72] H KISHINO et M HASEGAWA : Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from dna sequence data, and the branching order in hominoidea. *Journal of molecular evolution*, 29(2):170–9, 8 1989.

[73] A. KONRAD, A. I. TEUFEL, J. A. GRAHNEN et D. A. LIBERLES : Toward a general model for the evolutionary dynamics of gene duplicates. *Genome Biology and Evolution*, 3:1197–1209, 9 2011.

[74] Sudhir KUMAR et Alan FILIPSKI : Multiple sequence alignment : in pursuit of homologous dna positions. *Genome research*, 17(2):127–35, 2 2007.

[75] Manuel LAFOND, Cedric CHAUVE, Riccardo DONDI et Nadia EL-MABROUK : Polytomy refinement for the correction of dubious duplications in gene trees. *Bioinformatics (Oxford, England)*, 30(17):519–26, 9 2014.

[76] Manuel LAFOND, Magali SEMERIA, Krister M SWENSON, Eric TANNIER et Nadia EL-MABROUK : Gene tree correction guided by orthology. *BMC Bioinformatics*, 14(Suppl 15):S5, 2013.

[77] Manuel LAFOND, K M SWENSON et Nadia EL-MABROUK : An optimal reconciliation algorithm for gene trees with polytomies. *In WABI*, pages 106–122, 2012.

[78] Joshua LEDERBERG et E TATUM : Gene recombination in escherichia coli. *Nature*, 1946.

[79] Claire LEMAITRE, Marilia D V BRAGA, Christian GAUTIER, Marie-France SAGOT, Eric TANNIER et Gabriel A B MARAIS : Footprints of inversions at present and past pseudoautosomal boundaries in human sex chromosomes. *Genome biology and evolution*, 1(0):56–66, 1 2009.

[80] Christophe LETT, Magali SEMERIA, Andréa THIEBAULT et Yann TREMBLAY : Effects of successive predator attacks on prey aggregations. *Theoretical Ecology*, 1 2014.

[81] E. LEVY KARIN, E. SUSKO et T. PUPKO : Alignment errors strongly impact likelihood-based tests for comparing topologies. *Molecular Biology and Evolution*, 31(11):3057–3067, 8 2014.

[82] Ying LIANG, Xuexin HOU, Yanhua WANG, Zhigang CUI, Zhikai ZHANG, Xiaoyu ZHU, Lianxu XIA, Xiaona SHEN, Hong CAI, Jian WANG, Donglei XU, Enmin ZHANG, Huijuan ZHANG, Jianchun WEI, Jinrong HE, Zhizhong SONG, Xue-jie YU, Dongzheng YU et Rong HAI : Genome rearrangements of completely sequenced strains of yersinia pestis. *Journal of clinical microbiology*, 48(5):1619–23, 5 2010.

[83] Yu LIN, Fei HU, Jijun TANG et B MORET : Maximum likelihood phylogenetic reconstruction from high-resolution whole-genome data and a tree of 68 eukaryotes. *Pac Symp Biocomput*, pages 285–96, 2013.

[84] Kevin LIU, Sindhu RAGHAVAN, Serita NELESEN, C Randal LINDER et Tandy WARNOW : Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science (New York, N.Y.)*, 324(5934):1561–4, 6 2009.

[85] Kevin LIU, Tandy J WARNOW, Mark T HOLDER, Serita M NELESEN, Jiaye YU, Alexandros P STAMATAKIS et C Randal LINDER : Sate-ii : very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Systematic biology*, 61(1):90–106, 1 2012.

[86] Liang LIU et Dennis K PEARL : Species trees from gene trees : reconstructing bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Systematic biology*, 56(3):504–14, 6 2007.

[87] Joseph Mex LUCAS, Matthieu MUFFATO et Hugues ROEST CROLLIUS : Phyldiag : identifying complex synteny blocks that include tandem duplications using phylogenetic gene trees. *BMC Bioinformatics*, 15(1):268, 2014.

[88] Gerton LUNTER, István MIKLÓS, Alexei DRUMMOND, Jens Ledet JENSEN et Jotun HEIN : Bayesian coestimation of phylogeny and sequence alignment. *BMC bioinformatics*, 6:83, 1 2005.

[89] Haiwei LUO, William ARNDT, Yiwei ZHANG, Guanqun SHI, Max A ALEKSEYEV, Jijun TANG, Austin L HUGHES et Robert FRIEDMAN : Phylogenetic analysis of genome rearrangements among five mammalian orders. *Molecular phylogenetics and evolution*, 65(3):871–82, 12 2012.

[90] M LYNCH et A FORCE : The probability of duplicate gene preservation by subfunctionalization. *Genetics*, 154(1):459–73, 1 2000.

[91] Jian MA : A probabilistic framework for inferring ancestral genomic orders. *In 2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 179–184. IEEE, 12 2010.

[92] Jian MA, Aakrosh RATAN et BJ RANEY : Dupcar : reconstructing contiguous ancestral regions with duplications. *Journal of Computational Biology*, 15(8): 1007–1027, 10 2008.

[93] Jian MA, Louxin ZHANG et BB SUH : Reconstructing contiguous regions of an ancestral genome. *Genome research*, 16(12):1557–1565, 12 2006.

[94] W. P. MADDISON : Gene trees in species trees. *Systematic Biology*, 46(3):523–536, 9 1997.

[95] Wayne P MADDISON et L Lacey KNOWLES : Inferring phylogeny despite incomplete lineage sorting. *Systematic biology*, 55(1):21–30, 2 2006.

[96] B M MORET, L S WANG, T WARNOW et S K WYMAN : New approaches for reconstructing phylogenies from gene order data. *Bioinformatics (Oxford, England)*, 17 Suppl 1:S165–S173, 2001.

[97] B M MORET, S WYMAN, D A BADER, T WARNOW et M YAN : A new implementation and detailed study of breakpoint analysis. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 583–94, 1 2001.

[98] 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 (New York, N.Y.)*, 309(5734):613–7, 7 2005.

[99] J H NADEAU et B A TAYLOR : Lengths of chromosomal segments conserved since divergence of man and mouse. *Proceedings of the National Academy of Sciences of the United States of America*, 81(3):814–8, 2 1984.

[100] Magali NAVILLE, Minaka ISHIBASHI, Marco FERG, Hemant BENGANI, Silke RINKWITZ, Monika KRECSMARIK, Thomas A HAWKINS, Stephen W WILSON, Elizabeth MANNING, Chandra S R CHILAMAKURI, David I WILSON, Alexandra LOUIS, F LUCY RAYMOND, Sepand RASTEGAR, Uwe STRÄHLE, Boris LENHARD, Laure BALLY-CUIF, Veronica van HEYNINGEN, David R FITZPATRICK, Thomas S BECKER et Hugues ROEST CROLLIUS : Long-range evolutionary constraints reveal cis-regulatory interactions on the human x chromosome. *Nature communications*, 6:6904, 1 2015.

[101] Thi-Hau NGUYEN, Jean-Philippe DOYON, Stéphanie POINTET, Anne-Muriel ARIGON, Vincent RANWEZ et Vincent BERRY : Accounting for gene tree uncertainties improves gene trees and reconciliation inference. *In* Ben RAPHAEL et Jijun TANG, éditeurs : *Algorithms in Bioinformatics*, volume 7534 de *Lecture Notes in Computer Science*, pages 123–134. Springer Berlin Heidelberg, 2012.

[102] M A NOOR, K L GRAMS, L A BERTUCCI et J REILAND : Chromosomal inversions and the reproductive isolation of species. *Proceedings of the National Academy of Sciences of the United States of America*, 98(21):12084–8, 10 2001.

[103] Emmanuel NOUTAHI, Magali SEMERIA, Manuel LAFOND, Jonathan SEGUIN, Bastien BOUSSAU, Laurent GUÉGUEN, Nadia EL-MABROUK et Eric TANNIER : Genome evolution aware gene trees, 6 2015.

[104] H OCHMAN, J G LAWRENCE et E A GROISMAN : Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784):299–304, 5 2000.

[105] Susumu OHNO : *Evolution by Gene Duplication*. Springer Science & Business Media, 2013.

[106] E PASSARGE, B HORSTHEMKE et R A FARBER : Incorrect use of the term synteny. *Nature genetics*, 23(4):387, 12 1999.

[107] Murray PATTERSON, Gergely SZÖLLŐSI, Vincent DAUBIN et Eric TANNIER : Lateral gene transfer, rearrangement, reconciliation. *BMC Bioinformatics*, 14(Suppl 15):S4, 2013.

[108] Vinita PERIWAL et Vinod SCARIA : Insights into structural variations and genome rearrangements in prokaryotic genomes. *Bioinformatics (Oxford, England)*, 31(1):1–9, 1 2015.

[109] Guy PERRIÈRE et Céline BROCHIER-ARMANET : *Concepts et méthodes en phylogénie moléculaire*. Springer, 2010.

[110] P. PEVZNER : Genome rearrangements in mammalian evolution : Lessons from human and mouse genomes. *Genome Research*, 13(1):37–45, 12 2002.

[111] Pavel A. PEVZNER et Glenn TESLER : P. pevzner, g. tesler, , organized by the max planck institute for molecular genetics and the berlin center for genome-based bioinformatics, 10-13 april 2003, berlin, germany. *In RECOMB 2003 : The Seventh Annual International Conference on Research in Computational Molecular Biology*, 2003.

[112] P PFEIFFER, W GOEDECKE et G OBE : Mechanisms of dna double-strand break repair and their potential to induce chromosomal aberrations. *Mutagenesis*, 15(4):289–302, 7 2000.

[113] Hervé PHILIPPE et Christophe J DOUADY : Horizontal gene transfer and phylogenetics. *Current Opinion in Microbiology*, 6(5):498–505, 10 2003.

[114] J M RANZ, F CASALS et A RUIZ : How malleable is the eukaryotic genome ? extreme rate of chromosomal rearrangement in the genus drosophila. *Genome research*, 11(2):230–9, 2 2001.

[115] Matthew D RASMUSSEN et Manolis KELLIS : Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Res*, 22(4):755–765, 2012.

[116] Benjamin D REDELINGS et Marc A SUCHARD : Joint bayesian estimation of alignment and phylogeny. *Systematic biology*, 54(3):401–18, 6 2005.

[117] M J SANDERSON, A PURVIS et C HENZE : Phylogenetic supertrees : Assembling the trees of life. *Trends in ecology & evolution*, 13(3):105–9, 3 1998.

[118] David SANKOFF : Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics*, 28(1):35–42, 1975.

[119] George SAWA, Jo DICKS et Ian N ROBERTS : Current approaches to whole genome phylogenetic analysis. *Briefings in bioinformatics*, 4(1):63–74, 3 2003.

[120] Aylwyn SCALLY, Julien Y DUTHEIL, LaDeana W HILLIER, Gregory E JORDAN, Ian GOODHEAD, Javier HERRERO, Asger HOBOLTH, Tuuli LAPPALAINEN, Thomas MAILUND, Tomas MARQUES-BONET, Shane MCCARTHY, Stephen H MONTGOMERY, Petra C SCHWALIE, Y Amy TANG, Michelle C WARD, Yali XUE, Bryndis YNGVADOTTIR, Can ALKAN, Lars N ANDERSEN, Qasim AYUB, Edward V BALL, Kathryn BEAL, Brenda J BRADLEY, Yuan CHEN, Chris M CLEE, Stephen FITZGERALD, Tina A GRAVES, Yong GU, Paul HEATH, Andreas HEGER, Emre KARAKOC, Anja KOLB-KOKOCINSKI, Gavin K LAIRD, Gerton LUNTER, Stephen MEADER, Matthew

MORT, James C MULLIKIN, Kasper MUNCH, Timothy D O'CONNOR, Andrew D PHILLIPS, Javier PRADO-MARTINEZ, Anthony S ROGERS, Saba SAJJADIAN, Dominic SCHMIDT, Katy SHAW, Jared T SIMPSON, Peter D STENSON, Daniel J TURNER, Linda VIGILANT, Albert J VILELLA, Weldon WHITENER, Baoli ZHU, David N COOPER, Pieter de JONG, Emmanouil T DERMITZAKIS, Evan E EICHLER, Paul FLICEK, Nick GOLDMAN, Nicholas I MUNDY, Zemin NING, Duncan T ODOM, Chris P PONTING, Michael A QUAIL, Oliver A RYDER, Stephen M SEARLE, Wesley C WARREN, Richard K WILSON, Mikkel H SCHIERUP, Jane ROGERS, Chris TYLER-SMITH et Richard DURBIN : Insights into hominid evolution from the gorilla genome sequence. *Nature*, 483(7388):169–75, 3 2012.

[121] Celine SCORNAVACCA, Edwin JACOX et Gergely J SZÖLLŐSI : Joint amalgamation of most parsimonious reconciled gene trees. *Bioinformatics (Oxford, England)*, 31(6):841–8, 3 2015.

[122] Magali SEMERIA, Eric TANNIER et Laurent GUÉGUEN : Probabilistic modeling of the evolution of gene synteny within reconciled phylogenies. *BMC Bioinformatics*, 16(Suppl 14):S5, 2015.

[123] Hidetoshi SHIMODAIRA : An approximately unbiased test of phylogenetic tree selection. *Systematic biology*, 51(3):492–508, 6 2002.

[124] Hidetoshi SHIMODAIRA et Masami HASEGAWA : Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular biology and evolution*, 16:1114–1116, 1999.

[125] Hidetoshi SHIMODAIRA et Masami HASEGAWA : Consel : for assessing the confidence of phylogenetic tree selection. *Bioinformatics*, 17(12):1246–1247, 2001.

[126] Amit U SINHA et Jaroslaw MELLER : Cinteny : flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. *BMC bioinformatics*, 8(1):82, 1 2007.

[127] Sen SONG, Liang LIU, Scott V EDWARDS et Shaoyuan WU : Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proceedings of the National Academy of Sciences of the United States of America*, 109(37):14942–7, 9 2012.

[128] Alexandros STAMATAKIS : Raxml version 8 : a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)*, 30(9): 1312–3, 5 2014.

[129] A H STURTEVANT : A case of rearrangement of genes in drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, 7(8):235–7, 8 1921.

[130] A H STURTEVANT et T DOBZHANSKY : Inversions in the third chromosome of wild races of drosophila pseudoobscura, and their use in the study of the history of the species. *Proceedings of the National Academy of Sciences of the United States of America*, 22(7):448–50, 7 1936.

[131] Krister M SWENSON, Andrea DOROFTEI et Nadia EL-MABROUK : Gene tree correction for reconciliation and species tree inference. *Algorithms for molecular biology : AMB*, 7(1):31, 1 2012.

[132] Gergely J SZÖLLŐSI, Eric TANNIER, Vincent DAUBIN et Bastien BOUSSAU : The inference of gene trees with species trees. *Syst Biol*, 64(1):42–62, 2015.

[133] Gergely J SZÖLLŐSI, Wojciech ROSIKIEWICZ, Bastien BOUSSAU, Eric TANNIER et Vincent DAUBIN : Efficient exploration of the space of reconciled gene trees. *Systematic biology*, 62(6):901–12, 11 2013.

[134] Anastasia THANUKOS : A name by any other tree. *Evolution : Education and Outreach*, 2(2):303–309, 4 2009.

[135] Andréa THIEBAULT, Magali SEMERIA, Christophe LETT et Yann TREMBLAY : How to capture fish in a school ? effect of successive predator attacks on seabird feeding success. *Journal of Animal Ecology*, pages n/a–n/a, 10 2015.

[136] Christopher M THOMAS et Kaare M NIELSEN : Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature reviews. Microbiology*, 3(9): 711–21, 9 2005.

[137] 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 research*, 19(2):327–335, 2 2009.

[138] Ilan WAPINSKI, Avi PFEFFER, Nir FRIEDMAN et Aviv REGEV : Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics (Oxford, England)*, 23(13):549–58, 7 2007.

[139] C R WOESE et G E FOX : Phylogenetic structure of the prokaryotic domain : the primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America*, 74(11):5088–90, 11 1977.

[140] K M WONG, M A SUCHARD et J P HUELSENBECK : Alignment uncertainty and genomic analysis. *Science*, 319:473\–476, 2008.

[141] Yi-Chieh WU, Matthew D RASMUSSEN, Mukul S BANSAL et Manolis KELLIS : Most parsimonious reconciliation in the presence of gene duplication, loss, and deep coalescence using labeled coalescent trees. *Genome research*, 24(3):475–86, 3 2014.

[142] Jianzhi ZHANG : Evolution by gene duplication : an update. *Trends in Ecology & Evolution*, 18(6):292–298, 6 2003.

[143] N D ZINDER et J LEDERBERG : Genetic exchange in salmonella. *Journal of bacteriology*, 64(5):679–99, 11 1952.

[144] Emile ZUCKERKANDL et Linus PAULING : Molecules as documents of evolutionary history. *Journal of Theoretical Biology*, 8(2):357–366, 3 1965.

A

Articles publiés

A.1 Présentation des articles

Cette section contient la liste de mes publications. L'article A.2, inclus dans la partie 2.2, présente Harpi. Il a été publié en octobre 2015, à l'occasion de la conférence Recomb-CG. Le chapitre de livre présenté dans la partie A.3 est le résultat d'une collaboration avec Cedric Chauve et Nadia El-Mabrouk. Il contient une discussion sur l'intégration de différentes échelles d'évolution. Ma contribution à cette discussion a été de montrer que la synténie pouvait être utilisée comme une source d'information pour corriger des arbres de gènes. Il a été publié a l'automne 2013. La réflexion sur les méthode permettant d'intégrer la modélisation de la synténie et la reconstruction d'arbres de gènes nous a peu de temps après permis de publier l'article A.4, en collaboration avec Nadia El-Mabrouk, Manuel Lafond et Krister M Swenson. Dans cet article, nous présentons formellement la méthode de correction introduite en A.3 (Unduplicator), et une méthode de correction similaire développée par Manuel Lafond et Nadia El-Mabrouk (ParalogyCorrector). Une autre méthode de correction d'arbres de gènes, ProfileNJ, a ensuite été développée dans l'équipe de Nadia El-Mabrouk (principalement par Emmanuel Noutahi) dans la continuité de cette thématique. Les 3 méthodes de correction (ProfileNJ, ParalogyCorrector et Unduplicator) ont ensuite été intégrées dans une pipeline de correction d'arbres de gènes appelée RefineTree. L'article A.5, écrit avec Emmanuel Noutahi, Manuel Lafond, Jonathan Seguin, Bastien Boussau, et Nadia El-Mabrouk présente RefineTree, décrit l'algorithme de ProfileNJ, compare les résultats obtenus avec les trois méthodes de correction, et montre que l'association de ces trois méthodes permet d'améliorer les arbres de gènes disponibles dans la base de données Ensembl. Cet article est en cours de préparation.

Les articles A.6 et A.7 concernent le travail que j'ai réalisé au cours d'un stage à l'IRD de Sète en 2011. Mon sujet de recherche était alors très différent de mon sujet de thèse : je cherchais à modéliser les réactions d'un banc de poissons attaqué par un groupe d'oiseaux. L'article A.6 présente le modèle individu-centré que j'ai développé au cours de mon stage et montre qu'il est possible de simuler les perturbations d'un banc de poisson provoquées par des attaques successives de prédateurs avec un modèle simple. Avec ces simulations, nous avons montré que la fréquence des attaques influence fortement le comportement du banc, confirmant ainsi les observations d'Andréa Thiebault et de Yann Tremblay. L'article A.7 reprend les résultats obtenus avec

le modèle individu-centré et met l'accent sur les observations du comportement des oiseaux prédateurs.

A.2 Probabilistic modeling of the evolution of gene synteny within reconciled phylogenies

Auteurs : Magali Semeria, Eric Tannier, Laurent Guéguen

Revue : BMC Bioinformatics

Statut : publié

URL : `http://www.biomedcentral.com/1471-2105/16/S14/S5`

Référence Bibliographique : [122]

Cet article est inclus dans la section 2.2.

A.3 Duplication, Rearrangement and Reconciliation :a follow-up 13 years later

Auteurs : Cedric Chauve, Nadia El-Mabrouk, Laurent Guéguen, Magali Semeria, Eric Tannier

Livre : Models and Algorithms for Genome Evolution

Statut : publié

URL : `http://link.springer.com/chapter/10.1007%2F978-1-4471-5298-9_4`

Référence Bibliographique : [30]

Le contrat d'édition de *Springer* ne permet pas la libre diffusion de cet article.

A.4 Gene tree correction guided by orthology

Auteurs : Manuel Lafond, Magali Semeria, Krister M Swenson, Eric Tannier, Nadia El-Mabrouk

Revue : BMC Bioinformatics

Statut : publié

URL : `http://www.biomedcentral.com/1471-2105/14/S15/S5`

Référence Bibliographique : [76]

Lafond et al. BMC Bioinformatics 2013, 14(Suppl 15):S5
http://www.biomedcentral.com/1471-2105/14/S15/S5

BMC Bioinformatics

PROCEEDINGS Open Access

Gene tree correction guided by orthology

Manuel Lafond[1*], Magali Semeria[2], Krister M Swenson[1,4], Eric Tannier[2,3], Nadia El-Mabrouk[1]

From Eleventh Annual Research in Computational Molecular Biology (RECOMB) Satellite Workshop on Comparative Genomics
Lyon, France. 17-19 October 2013

Abstract

Background: Reconciled gene trees yield orthology and paralogy relationships between genes. This information may however contradict other information on orthology and paralogy provided by other footprints of evolution, such as conserved synteny.

Results: We explore a way to include external information on orthology in the process of gene tree construction. Given an initial gene tree and a set of orthology constraints on pairs of genes or on clades, we give polynomial-time algorithms for producing a modified gene tree satisfying the set of constraints, that is as close as possible to the original one according to the Robinson-Foulds distance. We assess the validity of the modifications we propose by computing the likelihood ratio between initial and modified trees according to sequence alignments on Ensembl trees, showing that often the two trees are statistically equivalent.

Availability: Software and data available upon request to the corresponding author.

Introduction

A gene tree represents the evolutionary relationships between a set of homologous genes. Gene trees are useful to unveil the molecular evolutionary events that have shaped today's genomes. They are traditionally constructed from sequence alignments [1], while recent methods also use the information from species phylogenies through reconciliation [2-8]. But constructing good gene trees is still challenging: for example, while they yield orthology and paralogy relationships between genes, often alternative or additional information, such as conserved synteny, is used to provide or confirm orthology [9].

The orthology information suggested by gene tree reconciliation may be contradictory with that suggested by an external source, such as conserved synteny [10,11]. We explore a way to reconcile them by performing slight modifications to a given gene tree in order to fit external information on orthology.

We propose two kinds of gene tree modification, which consist in computing a gene tree as close as possible to the initial one, satisfying two kinds of constraints. One kind is a set of pairs of genes that should be orthologous but are seen as paralogous in the initial tree. This occurs when orthologs are computed with synteny for example [11]. The other kind is a set of clades that should be rooted by speciation nodes but are rooted by duplication nodes in the initial tree. This occurs when dubious duplications are detected because of the absence of extant support for a duplication, or because of ancestral synteny information [10]. We give polynomial-time algorithms for both problems under the Robinson-Foulds distance, thus proposing several ways to improve gene trees according to external information.

There are very few gene tree reconstruction methods including synteny information [12], whereas integrating this information could be valuable [13]. The modifications we propose could be included in a local search framework as other kinds of modifications based on duplications and losses [14-17]. We assess the validity of the modifications we propose by computing the likelihood ratio between initial and modified trees according

* Correspondence: lafonman@iro.umontreal.ca
[1]Département d'Informatique (DIRO), Université de Montréal, H3C3J7, Canada
Full list of author information is available at the end of the article

BioMed Central

© 2013 Lafond et al; licensee BioMed Central Ltd. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

to sequence alignments on Ensembl trees [18], showing that often the two trees are statistically equivalent.

Different gene tree corrections

Phylogenies

A *phylogeny* is a rooted binary tree which represents the evolutionary relationships between the nodes. Internal nodes are extinct ancestors, leaves are extant elements and edges represent direct descents between parents and children. Given a node x of a phylogeny T, we call an *ancestor* of x any node on the path from the root (inclusively) of T to the parent of x. For a leaf-subset X of T, $\text{lca}_T(X)$, the *lowest common ancestor* of X, denotes the farthest node from the root which is an ancestor of all elements of X. We use the notation $l(x)$, and call the *clade* of x, the set of leaves which are descendant from an internal node x. We also denote by $l(T)$ the set of leaves, and by $V(T)$ the set of nodes of T.

We define two kinds of phylogenies: species trees and gene trees. Species are identified with *genomes*. For our purpose, genomes are simply sets of genes. Therefore, each gene g, extant or ancestral, belongs to a species $s(g)$. We then have one species tree S, where nodes are identified with species, and many gene trees, where nodes are identified with genes. The set of genes in a gene tree is called a *gene family*.

A *reconciliation* between a gene tree G and a species tree S consists in assigning to each gene g of G (both extant and ancestral) the species $s(g)$ corresponding to the lowest common ancestor in S of the set $\{s(l)$, for all $l \lfloor l(g)\}$. Every internal node g of G is labeled by an *event* $E(g)$, verifying $E(g)$ = speciation if $s(g)$ is different from $s(g_\ell)$ and $s(g_r)$ where g_ℓ and g_r are the two children of g, and $E(g)$ = duplication otherwise.

The reconciliation of G and S gives all informations about the gene family history. In particular it defines the gene content of an ancestral species at the time of speciation. A reconciliation also implies the orthology and paralogy relationships between genes: two genes g and g' of T are said to be *orthologous* if $E(\text{lca}_T(g, g'))$ = speciation; g and g' are *paralogous* if $E(\text{lca}_T(g, g'))$ = duplication. For example, Figure 1(1) shows a gene tree reconciled with a species tree. In this gene tree a_1 and b_1 are paralogous as their lowest common ancestor is d which is a duplication node, while $a2$ and $b2$ are orthologous. The number of dots inside big circles represents the number of genes in the corresponding genome (each big circle represents a species).

The Robinson-Fould (RF) distance

The RF distance $RF(G, G')$ between two phylogenies G and G' is the cardinality of the symmetric difference between the clade-sets of the two trees. In other words, denote by $c(G, G')$ the number of clades that are in G but not in G'. Then $RF(G, G') = c(G, G') + c(G', G)$.

In this paper, since we only compare rooted binary trees sharing the same leaf-sets, they always have the same number of internal nodes, and hence the same number of clades. Therefore $c(G, G') = c(G', G)$, and $RF(G, G') = 2c(G, G')$.

Two correction problems

Suppose that in addition to a species tree and a set of reconciled gene trees, we are given additional information of two kinds:

- Pairs of genes that we know are orthologous;
- Duplication nodes of some gene trees that we suspect to be false.

Constraints of orthology on pairs of genes may for example be generated from synteny analysis [9,11]. Some pairs may contradict the information given by the gene tree. Let P be a set of pairs (g_1, g_2) of orthologous extant genes (verifying $s(g_1) \neq s(g_2)$). A gene tree G is said to *satisfy* a set P if, for any pair $(g_1, g_2) \in P$, $\text{lca}_G(g_1, g_2)$ is a speciation node.

Problem 1 ***Gene Orthology Correction [GOC] Problem***

Input: *A gene tree G reconciled with a species tree S, and a set P of gene pairs that are required to be orthologous;*

Output: *A corrected gene tree G_P satisfying P, such that $RF(G, G_P)$ is minimum among all possible solutions.*

An example is given in Figure 1: (1) is the initial tree, and (2) depicts two syntenic regions of size 3 surrounding genes $b1$ and $a1$. In general (if we neglect the effect of gene conversion) genes in two syntenic regions should be either all pairwise orthologous or all pairwise paralogous [11]. Consequently, if the two neighbors of $b1$ on genome B and of $a1$ on genome C are inferred to be orthologous (according to their lowest common ancestor in their respective gene trees), then an orthology constraint should be imposed on the pair $(b1, a1)$. Figure 1. This principle is usually considered as one of the most efficient method to detect orthologies [9]. (3) is a corrected tree.

On the other hand, duplication nodes of a gene tree can be considered dubious for different reasons. For example, in Ensembl [19], "dubious" is a label assigned to the non-apparent duplication nodes [20,21] pointing to an incongruence between the gene tree and the species tree. Alternatively, inferred ancestral synteny may also point to dubious duplication nodes [10]. Formally, clades corresponding to some duplication nodes may erroneously be considered as sets of paralogous genes, and should rather be considered as orthologous.

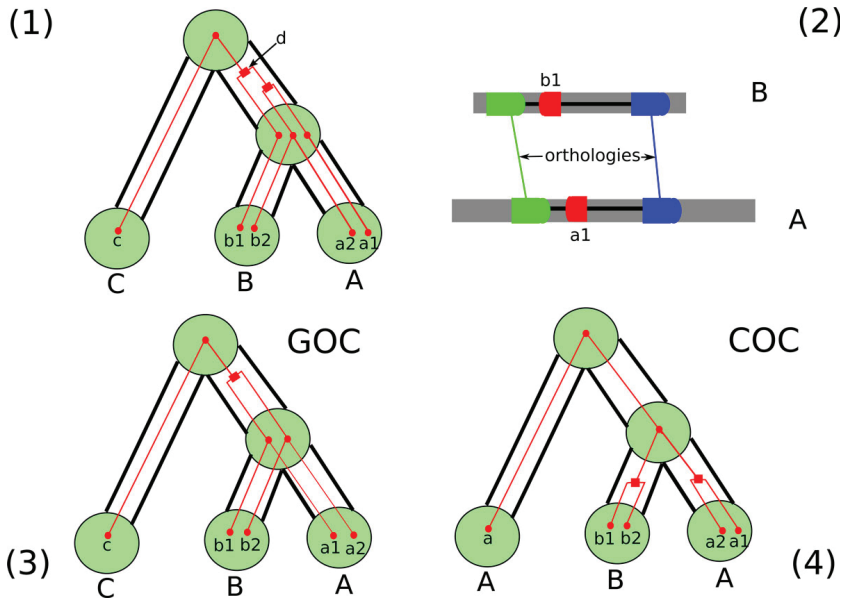


Figure 1 Description of the two problems. (1) A gene tree (the "initial tree") for the gene family {*c*, *b*1, *b*2, *a*1, *a*2} is shown with small red nodes and single thin red edges. It is reconciled with the phylogeny of the three species *A*, *B* and *C* shown with large green nodes and hollow edges represented by a pair of parallel black lines. Duplication nodes of the reconciled gene tree are squared, while speciation nodes and leaves are dots. (2) The two neighbors of *b*1 on genome *B* and of *a*1 on genome *A* are inferred to be orthologous according to their lowest common ancestor in their respective gene trees (not shown). This is an argument for infering orthology between *b*1 and *a*1, which is in contradiction with the information provided by the initial tree: their lowest common ancestor is a duplication, and thus they are inferred to be paralogous. (3) A solution to the GOC problem, that is a gene tree of minimum RF distance with the initial tree verifying the constraint of *b*1 and *a*1 being orthologous. (4) A solution to the COC problem, that is a reconciled tree in which the clade {*b*1, *b*2, *a*1, *a*2} of *d* in the initial tree is rather rooted by a speciation node in the corrected tree. This is an example where the optimal solutions to the two problems differ.

A gene tree G is said to *satisfy* a set C of its clades if $E(\text{lca}_G(c))$ = speciation for all $c \sqsubset C$.

Problem 2 ***Clade Orthology Correction*** **[*COC*]** ***Problem***

Input: *A gene tree G reconciled with a species tree S, and a set C of clades of G assigned to duplication nodes;*

Output: *A corrected tree GC satisfying C, such that RF* (G, G_C) *is minimum among all possible solutions.*

The two problems are different, as exemplified by Figure 1, where (3) is an optimal solution to GOC while (4) is an optimal solution to COC, the latter more distant to the initial tree.

In the next two sections, we use S for the species tree name, G for the reconciled gene tree, and we give efficient solutions to these two problems.

The Gene Orthology Correction Problem

Notice that for any instance of the GOC problem, a corrected tree satisfying P always exists. Indeed, for any extant species x of S, one can make a tree whose leaf-set is all the extant genes g of G for which $s(g) = x$. Doing this for every species yields a forest whose roots can be reconnected by matching the topology of S, ensuring that any pair of genes not in the same species are orthologous. However, the obtained tree can be very far from the original.

Let P be a set of gene pairs (which are leaves of G) required to be orthologous. Notice that if $(a, b) \in P$, then we also have $(b, a) \in P$. For any pair $(a, b) \in P$, if $\text{lca}_G(a, b)$ is a duplication in G, then (a, b) is a pair of *false paralogs*. The set $P_f \subseteq P$ denotes the set of all false paralogous pairs of P.

Given two distinct leaves a and b of G, we set $r_{a,b} = \text{lca}_G(a, b)$, $s_{a,b} = \text{lca}_S(s(a), s(b))$, and define $h_{a,b}$ as the highest node (closest to the root) on the path from a to $r_{a,b}$ such that $s(h_{a,b})$ is a descendant of $s_{a,b}$. Notice that $h_{a,b}$ can be a itself, but not $r_{a,b}$.

For instance on Figure 2(1), a_1, c_2 are false paralogs with $r_{a_1,c_2} = e_3$ and $s_{a_1,c_2} = E$. From this, one can deduce that $h_{a_1,c_2} = d_2$ and $h_{c_2,a_1} = c_2$. We show below that, for any pair (a, b) of false paralogs, $h_{a,b}$ is the highest node on the path from a to $r_{a,b}$ over which we can move b to make $\text{lca}_G(a, b)$ a speciation node. The reason for moving b as high as possible is to preserve as many clades as possible, allowing a minimum RF distance between the initial and corrected tree.

Lemma 1 *Let* (a, b) *be a pair of false paralogs in G, and let G' be a tree in which a and b are orthologous. If*

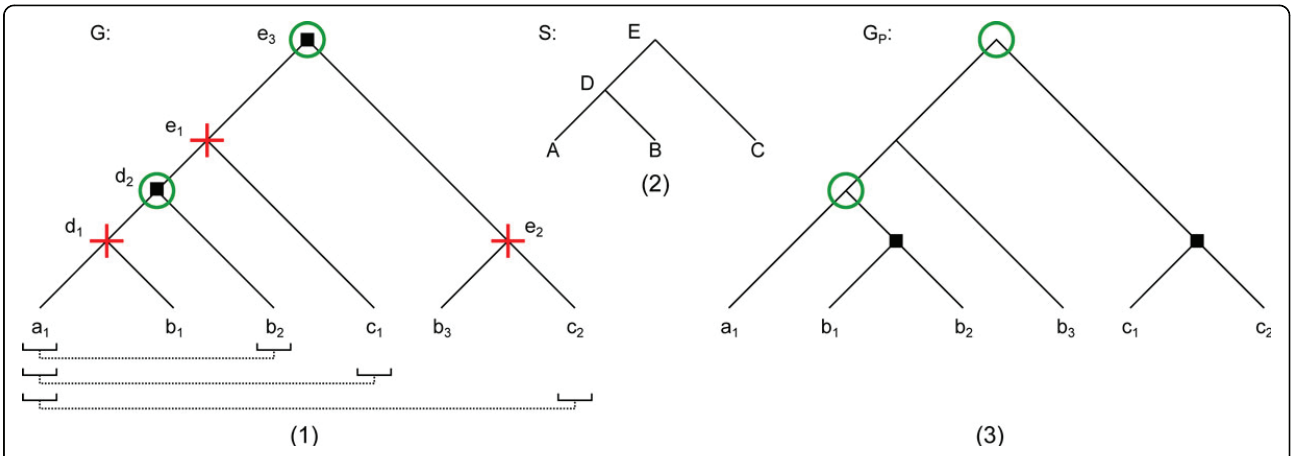


Figure 2 GOC Procedure. (1) A gene tree G reconciled with species tree S. Duplication nodes are denoted by a black square. The leaves and internal nodes of G are labeled with the letter of their corresponding species. Brackets denote the required orthologs given by the input set $P = \{(a_1, b_2), (a_1, c_1), (a_1, c_2)\}$. The non-preservable nodes (nodes of H) are depicted by red crosses, while preservable nodes are circled in green. (2) The species tree associated with G. (3) The tree G_P, a solution to the GOC problem, which preserves every possible clade.

x is an ancestor of $h_{a,b}$ and a descendant of $r_{a,b}$, then the clade of x is not in G'.

Proof: Suppose otherwise that there is some $x' \in V(G')$ with the same clade as x (and hence $s(x) = s(x')$). Let $r'_{a,b} = \text{lca}_{G'}(a, b)$, which should be a speciation. Since b was not in the clade of x, it cannot be in the clade of x' either, implying that $r'_{a,b}$ is an ancestor of x'. Also, since $s(x') = s(x)$ and x is above $h_{a,b}$ in G, we have that $s(x')$ is $s_{a,b}$ or one of its ancestors (otherwise we would have picked x to be $h_{a,b}$). But r' has x' in one of its subtrees, and b in the other, implying that $r'_{a,b}$ is a duplication: contradiction. □

We now have a way to identify a set of clades that cannot be in G_P. For any $(a, b) \in P_f$, denote by $H_{a,b}$ the set of ancestors of $h_{a,b}$ that are descendants of $r_{a,b}$. If G_P satisfies the set P_f, G_P cannot contain any clade from the set $H = \cup_{(a,b)\in P_f} H_{a,b}$. It follows that a minimum of $|H|$ clades of G are missing in G_P. We claim that a solution G_P to the GOC problem is obtained by modifying exactly $c(G, G_P) = |H|$ clades.

Theorem 1 *Let G_P be a solution to the GOC problem. Then $RF(G, G_P) = 2|H|$.*

In what follows, we give a constructive proof of Theorem 1 by describing an algorithm for solving the GOC problem.

An algorithm for the GOC problem

Call $V(G) \backslash H$ the set of *preservable nodes* of G (those that we hope to preserve). For example in Figure 2(1), $H = H_{a_1,c_2} \cup H_{c_2,a_1} \cup H_{a_1,c_1} \cup H_{c_1,a_1} \cup H_{a_1,b_2} \cup H_{b_2,a_1} = \{e_1\} \cup \{e_2\} \cup \emptyset \cup \emptyset \cup \{d_1\} \cup \emptyset = \{e_1, e_2, d_1\}$. The nodes of H are represented by red crosses, while the preservable nodes are circled in green. Notice that the root r of G is preservable, since any solution G_P to the GOC problem should share the same leaf-set as G. Consider the set $\mathcal{G}$ of subtrees of G rooted on the *highest preservable descendants* of r, i.e. preservable nodes for which r is the unique preservable ancestor. Observe that since any leaf of G is preservable, we have $\cup_{G_x \in \mathcal{G}} l(G_x) = l(G)$. If, for some $(g_1, g_2) \in P$, g_1 and g_2 are scattered across two subtrees of G, we call these subtrees *required orthologous subtrees* (or simply *required orthologs* when the context is clear as to whether we are comparing genes or subtrees). For example in the tree G of Figure 2(1), G is the set of subtrees rooted at d_2, c_1, b_3 and c_2 (the last four restricted to a single leaf), and the subtrees rooted at d_2 and c_1 are required orthologs, as well as those rooted at d_2 and c_2. However, connecting two subtrees under a speciation might not always be feasible. A definition of *possible orthologs* follows.

Definition 1 (Possible orthologs) *Two subtrees G_1, $G_2 \in \mathcal{G}$ rooted at x_1, x_2 respectively are* possible orthologs *if and only if $s(x_1)$ and $s(x_2)$ are unrelated, i.e. neither is an ancestor of the other in S.*

The following lemma ensures that the roots of required orthologous subtrees can actually be joined under a common parent which is a speciation.

Lemma 2 *Let G_1, $G_2 \in \mathcal{G}$ be required orthologs. Then G_1 and G_2 are possible orthologs.*

Proof: Let x_1, x_2 be the roots of G_1, G_2 respectively, and let $(g_1, g_2) \in P$ such that $g_1 \in l(G_1)$ and $g_2 \in l(G_2)$. Let s_ℓ, s_r be the left and right children of s_{g_1,g_2}, and denote by S_ℓ and S_r the subtrees of S rooted at s_ℓ and s_r respectively. Suppose without loss of generality that $s(g_1)$ is in $l(S_\ell)$ and $s(g_2)$ is in $l(S_r)$. Since x_1 is preservable and on the path between g_1 and r_{g_1,g_2}, we have $x_1 \notin H_{g_1,g_2}$ and thus $s(x_1) \in V(S_\ell)$. Similarly, $s(x_2) \in V(S_r)$. Therefore $s(x_1)$ and $s(x_2)$ are unrelated and possible orthologs.

The problem, formally defined in the sequel as the *maximum orthology tree*, consists in joining all trees of $\mathcal{G}$ into a single tree G' in a way ensuring that each pair of possible orthologs is joined under a speciation. More precisely, for some possible orthologs $G_1, G_2 \in \mathcal{G}$ rooted at nodes x_1, x_2, we get that $\text{lca}_{G'}(x1, x_2)$ is a speciation, with G_1, G_2 being unchanged.

We begin by giving an overview of the whole algorithm.

Algorithm Outline:

1. Compute the set $H = \cup_{(a,b) \in P_f} H_{a,b}$ of internal nodes of G corresponding to clades that cannot be in G_P;

2. Compute the set $\mathcal{G}$ of subtrees rooted at the highest preservable descendants of the root of G. If $\mathcal{G}$ is empty, return G and terminate;

3. Construct a tree G' by joining all trees of $\mathcal{G}$ in a way ensuring that possible orthologs are joined under speciation. We call G' the *maximum orthology tree* for $\mathcal{G}$;

4. For every tree $G_x \in \mathcal{G}$, construct $G_{x,P}$ by recursively repeating Steps 2 to 4 with G being G_x, and replace the G_x subtree of G' by $G_{x,P}$.

The tree obtained corresponds to the corrected tree G_P we want. Running this algorithm on the G tree of Figure 2 yields the corrected tree G_P. This algorithm terminates, since we eventually reach all the leaves of G, which correspond to terminal cases in the recursion. Implementing step 1 is straightforward, while step 2 can be done by performing a depth-first search from the root, in which upon visiting a preservable node, we add it to $\mathcal{G}$ and continue the search without visiting its children. Step 3 is the purpose of the next section, so assume for now that it can be performed correctly as stated. This algorithm can be implemented to run in $O(|P| \times |V(G)|)$ steps in the worst case, the main bottleneck being the computation of H. The algorithm correctness follows from the two lemmas below.

Lemma 3 *Any preservable node x of G is* preserved *in G_P, meaning that the clade of G rooted at x is a clade of G_P.*

Proof: Let x be a preservable node of G and G_x be the subtree rooted at x. It is not hard to see that eventually, steps 2-4 will be run on G_x and return a tree $G_{x,P}$, which will itself be a subtree of the final corrected tree G_P. As the algorithm only moves and reconnects subtrees of G_x, we have that $l(G_x) = l(G_{x,P})$. Since $G_{x,P}$ is a subtree of G_P, it follows that the clade of x is preserved in G_P.

Lemma 4 *Let $(g_1, g_2) \in P$. Then g_1 and g_2 are orthologs in G_P.*

Proof: Denote by G_v the subtree rooted at v, for some $v \in V(G)$. Let x be a preservable node and $G_{x,P}$ be the subtree produced after running steps 2-4 on G_x. Let D be the set of highest preservable descendants of x. We say that a gene pair (g_1, g_2) is *contained* in G_x if $g_1, g_2 \in l(G_x)$. We use induction on the height of the tree to show that all gene pairs in P that are contained in G_x are orthologous in $G_{x,P}$ (which proves the lemma since x can be the root). This is trivially true for leaves as they are preservable and contain no gene pairs. We thus suppose by induction that for any $d \in D$, gene pairs in P that are contained in G_d are orthologous in $G_{d,P}$. Let $(g_1, g_2) \in P$ such that (g_1, g_2) is contained in G_x, but there is no $d \in D$ such that G_d contains (g_1, g_2). What is left to prove is that g_1 and g_2 are orthologous in $G_{x,P}$.

We first observe that g_1, g_2 belong to two different subtrees G_{d_1}, G_{d_2}, where $d_1, d_2 \in D$. Otherwise $G_{d_1} = G_{d_2}$, implying that (g_1, g_2) is contained in G_{d_1} and we are done. Therefore, G_{d_1}, G_{d_2} *are required orthologs, and hence possible orthologs. Since we may assume that* G_{d1} and G_{d2} are joined under a speciation in $G_{x,P}$, we get that $\text{lca}_{G_{x,P}}(g_1, g_2)$ is a speciation. The result follows from observing that $G_{x,P}$ is a subtree of G_P.

Maximum orthology tree

We now describe a solution to the maximum orthology tree problem. Formally, given a set of k possible orthologous subtrees of G rooted on a set of nodes $X = \{x_1, \ldots, x_k\}$, the problem is to construct a tree F with $l(F) = X$, such that for each pair $x_i, x_j \in X$ that correspond to roots of possible orthologs, x_i and x_j are orthologous in F.

Roughly speaking, the algorithm proceeds as follows: start with F_0 being a copy of S. Iterate over i from 1 to k, at each step constructing F_i by grafting x_i on F_{i-1} right above the node $v \in V(F_0)$ such that $s(v) = s(x_i)$. Proceeding this way, we show in Lemma 5 that nodes of $V(F_0)$ are ensured to remain speciation nodes all over the procedure, and in lemma 6 that the lowest common ancestor of two possible orthologs belongs to $V(F_0)$, leading to corollary 1 stating that possible orthologs are in fact orthologous in the output tree. Finally remove the leaves artificially introduced by F_0 and *standardize* the tree, which means

- remove all nodes with no descendant labeled with extant genes;
- contract non-root degree 2 nodes, then contract the root if it is of degree one.

Starting with F_0 being a copy of S is a step that might be omitted, but the set of nodes $V(F_0)$ serves as a skeleton around which we graft our x_i's, making it both easily implementable and provable. Figure 3 shows how the algorithm proceeds on the set of highest preservable descendants of the root of the tree G in Figure 2(1).

Algorithm 1 *findMaxOrthology*$(S, X = \{x_1, \ldots, x_k\})$

$F_0 \leftarrow$ A copy of S

$V_0 \leftarrow V(F_0)$

$L \leftarrow l(F_0)$

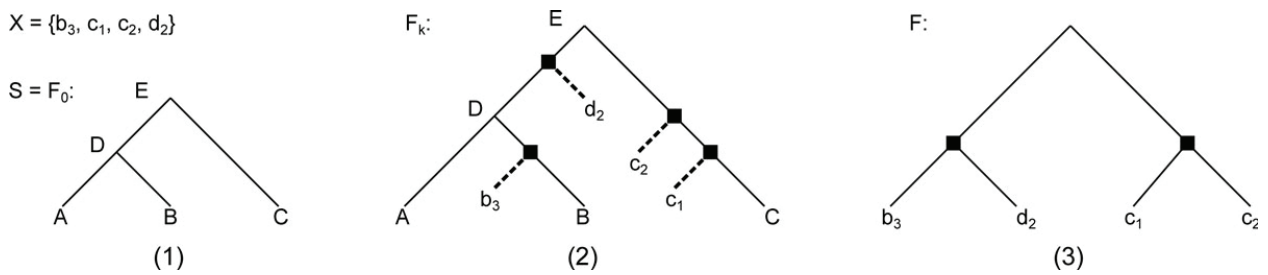


Figure 3 The Max Orthology problem. An instance of the *max orthology problem*, with X being the highest preservable descendants of the root of G in figure 2. (1) The starting tree F_0, which is a copy of S. (2) The F_k tree, which depicts the tree obtained after grafting every node of X. (3) The final tree F, obtained by removing the leaves initially in F_0 and standardizing.

for $i = 1 \rightarrow k$ **do**

Find the unique node $v \in V_0$ such that $s(v) = s(x_i)$

F_i $\Re$ a copy of F_{i-1} on which we graft x_i on the edge linking v to its parent node (or if v is the root of F_{i-1}, create a new root with children v and x_i)

end for

F $\Re F_k$ on which we remove L and stardardize

Lemma 5 *If $r \in V(F_0) \cap V(F)$, then r is a speciation.*

Proof: Since F_0 is a copy of S, all nodes of $V(F_0)$ are initially speciation nodes. We show that each grafting operation does not change the event corresponding to these nodes. Say that at iteration i, we graft x_i on the edge linking v to its parent node p. We first observe that the only nodes that can be transformed from speciation in $F_{i-1}$to duplication in F_i are on the path from p to the root of F_{i-1}. Suppose without loss of generality that v is the left child of p in F_{i-1}, and let w be the newly created node between p and v in F_i. Thus w has children xi and v, and since $s(x_i) = s(v)$, we get that $s(w) = s(v)$. It follows that if p was a speciation in F_{i-1}, it remains a speciation in F_i. Moreover, this implies that $s(p)$ is left unchanged in F_i, implying in turn that any ancestor of p cannot change from speciation to duplication. Therefore, no grafting operation can affect speciation of any vertex in $V(F_{i-1})$. Finally, we note that removing leaves or deleting degree two nodes in F also cannot affect speciation nodes.

Lemma 6 *Let $x_i, x_j \in X$ be the roots of possible orthologous subtrees. Then,* $\text{lca}_F(x_i, x_j) \in V(F_0)$.

Proof: First recall that if x_i,x_j are the roots of possible ortholog subtrees, then there is some $s \in V(S)$ such that $s(x_i)$ and $s(x_j)$ are in the left and right subtrees of s, respectively. Now, let r be the unique node in $V(F_0)$ such that $s(r) = s$, and let $v_i, v_j \in V(F_0)$ such that $s(v_i) = s(x_i)$ and $s(v_j) = s(x_j)$. It is clear that in F_0, $\text{lca}(v_i,v_j) = r$. This also holds for any F_i by observing that grafting nodes cannot change the lca relationship. Since x_i is grafted on some edge between v_i and r, and x_j between v_j and r, it follows that $\text{lca}(x_i, x_j) = r \in V(F_0)$.

Corollary 1 *Let $x_i, x_j \in X$ be the roots of possible orthologs. Then they are orthologous in F.*

The Clade Orthology Correction Problem

We prove several results characterizing the solutions to the COC problem. Let C be a set of clades that has to be satisfied. For a clade $c \in C$, we denote by $s(c)$ the value of $s(r(c))$ where $r(c)$ is the root of c, and by $E(c)$ the value of $E(r(c))$ that we call *the label of c*.

First, unlike in the GOC problem, a solution to the COC problem does not always exist. Indeed, it is possible that no gene tree has all clades in C labeled by speciations. We give a necessary and sufficient condition for the existence of a solution. The following lemma is obvious from the definition of reconciliation, and will be used in several proofs.

Lemma 7 *For a reconciled gene tree G, if a node x is an ancestor of a node y and $s(x) = s(y)$ then $E(x) =$ duplication.*

Theorem 2 *There is a solution to the COC problem if and only if for every clade $c \in C$, $s(c)$ is not a leaf of S, and if for every pair $c_1, c_2 \in C$, either c_1 and c_2 are disjoint sets of leaves, or $s(c_1) \neq s(c_2)$.*

The necessity of these conditions directly follow from Lemma 7, since $s(c_1)$, $s(c_2)$ and the ancestry relationship between c_1 and c_2 remain unchanged in a solution. Their sufficiency will be constructively demonstrated in the sequel. Suppose that the conditions are satisfied. We give a way of finding all optimal solutions according to the RF distance, followed by two ways of finding an optimal one optimizing other criteria in addition.

Given a duplication node x of G, *pushing x by multifurcation* means applying the following procedure:

- Let $s = s(x)$, and A and B be the two children of s in S.
- Let T^A be the set of maximal subtrees of the subtree of G rooted at x, such that all their leaves l verify that $s(l)$ is a descendant of A (including A itself).

Let $G^A[x]$ be the multifurcated tree obtained by joining all roots of trees in T^A under a common root.

• Let symmetrically T^B be the set of maximal subtrees of the subtree of G rooted at x, such that all their leaves l verify that $s(l)$ is a descendant of B (including B itself). Let $G^B[x]$ be the multifurcated tree obtained by joining all roots of trees in T^B under a common root.

• Let G' be obtained from G by replacing the clade rooted at x by a new subtree, obtained by joining $G^A[x]$ and $G^B[x]$ under a common root.

This rearrangement is described in [16] and applied to dubious duplications as a preprocessing step for ancestral genome reconstruction.

A *binary resolution* G_b of a multifurcated tree G is a binary tree in which all the clades of G are in G^b.

Theorem 3 *If there is a solution to the COC problem, then a binary gene tree is an optimal solution if and only if it is a binary resolution of the multifurcated tree obtained by pushing the roots of the elements of C by multifurcation (in any order).*

Proof: It is clear that a binary resolution is a solution, provided that the conditions for the existence of a solution are satisfied. Indeed any clade is preserved through pushing a duplication node, so this operation can be done for all clades in C independently. This proves the converse part of Theorem 2.

Then it is an optimal solution because by Lemma 7, no clade x which is a descendant of the pushed clade c such that $s(c) = s(x)$ may be conserved if we want c to be a speciation node. And by construction all clades such that $s(c) \neq s(x)$ are preserved by this operation.

Binary resolutions which minimize the number of duplications and losses are studied by [22] and may be applied to provide *bona fide* phylogenies. We describe an alternative maximizing the number of common triplets. A *triplet* in a tree G is a set of three leaves $((a, b), c)$ of G, such that the LCA of the three is strictly more ancient than the LCA of the first two.

Given a species tree S, a reconciled gene tree G and one of its duplication nodes x, *pushing x by tree duplication* means applying the following procedure, illustrated in Figure 4:

• Let $s = s(x)$, and A and B be the two children of s in S.

• Let $G^A[x]$ be a tree obtained from the subtree of G rooted at x, by deleting all leaves l with $s(l)$ being a descendant of A, and standardizing it, which as in the previous sections, means

- removing all nodes with no descendant labeled with extant genes;

- contracting non-root degree 2 nodes, then contracting the root if it is of degree one.

• Let symmetrically $G^B[x]$ be a tree obtained from the subtree of G rooted at x, by deleting all leaves l with $s(l)$ being a descendant of B, and standardizing it.

• Let G' be obtained from G by replacing the clade rooted at x by a new subtree, obtained by joining $G^A[x]$ and $G^B[x]$ under a common root.

Note that if a clade y is disjoint from x or assigned to a different species, then pushing x by tree duplication does not affect the subtree rooted at y. In consequence, pushing several clades by tree duplications in any order gives a unique solution if the clades satisfy the properties of Lemma 2.

Theorem 4 *If there is a solution to the Clade Orthology Correction problem, the gene tree obtained by successively pushing the roots of the elements of C by tree duplication (in any order) is an optimal solution. Among all optimal solutions, it maximizes the number of common triplets with G.*

Proof: As already noticed pushing a duplication by multifurcation preserves all clades assigned to species which are different from the species assigned to the pushed node. So it is an optimal solution.

Now we have to prove that none of the triplets that are in G but not in G' can be preserved in any other optimal solution. For this we characterize the triplets that can be preserved. For a triplet $((a, b), c)$ of G, let $T_{((a,b),c)}$ be the rooted phylogeny with three leaves and two internal nodes containing the triplet. If the leaves a, b, c are in the pushed clade x, then the triplet can be preserved only if in the reconciliation of $T_{((a,b),c)}$, the lowest internal node is not mapped to $s(x)$. Otherwise by Lemma 7, the root node of the triplet cannot be a speciation.

Let $((a, b), c)$ be a triplet such that in the reconciliation of $T_{((a,b),c)}$, the lowest internal node is not mapped to $s(x)$. This triplet is entirely included in $G^1[x]$ or $G^2[x]$. So it is preserved. In consequence all triplets possibly preserved are indeed preserved by the operation, showing the optimality of the procedure reguarding the number of common triplets.

Now if there is no solution to the Clade Orthology problem, we advice to push duplication nodes in C starting from the highest ones, without having formalized why we find this solution adequate.

Fish gene trees

Using synteny as evidence of orthology, we wanted to test the ability of our algorithm designed for the GOC problem to correct gene trees. To this end, we considered

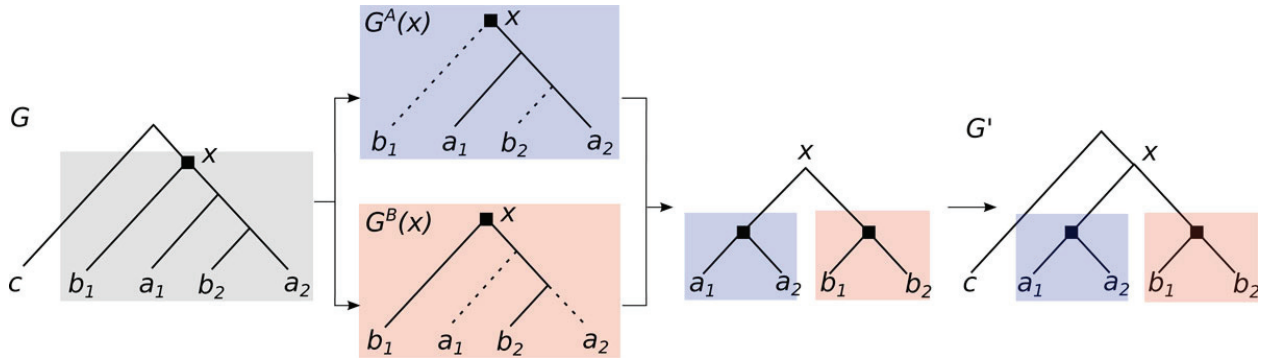


Figure 4 COC Procedure. An instance of the COC problem. In the gene tree G, where $s(a_1) = s(a_2) = A$, $s(b_1) = s(b_2) = B$ and $s(c) = C$, extract and copy the subtree rooted at × to get the subtrees $G^A(x)$ and $G^B(x)$. Remove b_1 and b_2 from $G^A(x)$ and a_1 and a_2 from $G^B(x)$. Join $G^A(x)$ and $G^B(x)$ under a common root and replace G(x) by the new subtree in the gene tree G'.

the four fish genomes *Gasterosteus aculeatus* (Stickleback), *Oryzias latipes* (Medaka), *Tetraodon nigroviridis*, and *Danio rerio* (Zebrafish) with human and mouse as outgroups. We used the *Ensembl Genome Browser* to collect all available gene trees, and filtered each tree to preserve only genes from the taxa of interest. We then reconciled the trees with the known species trees, and identified duplication and speciation nodes. Following our methodology in [11], a region surrounding a gene is defined as the substring containing the gene and both its left and right adjacencies, and two regions are considered syntenic if they contain homologous genes in the same order. We observed in [11] that more than 22% of the 6241 collected gene trees contain at least one false paralogy, that is a pair of genes required from synteny to be orthologous, but the LCA of the corresponding leaves being a duplication rather than a speciation node.

For 1000 of the trees containing at least one false paralogy, we applied the correction procedure previously described, and retrieved the gene family alignment from Ensembl. With PhyML [23], we computed the likelihood of the initial and corrected tree, given the alignment. These two likelihood values were compared with Consel [24]. For only 17.7% of the trees, the correction was rejected by the AU test. In other words, the correction algorithm is valid for a vast majority (82.3%) of the tested trees. Moreover, the likelihood of the corrected tree is higher than the original for 44.4% of the trees. Interestingly, 14.8% of the original Ensembl gene trees were rejected when compared to the corrected trees.

The correction of the gene tree for the *ZNF800* gene family, which is related to transcriptional regulation, is given as an example in Figure 5. The corrected tree was highly favored by the AU Test, giving it a statistical support advantage with a p-value below 0.001. Furthermore, the non-apparent duplication of G, located at the root of the (m_1, t_1, s_1) subtree, was eliminated, resulting in one less duplication in G_P.

Conclusion

We give two efficient algorithms for two new gene tree rearrangement problems, related to the correction of a

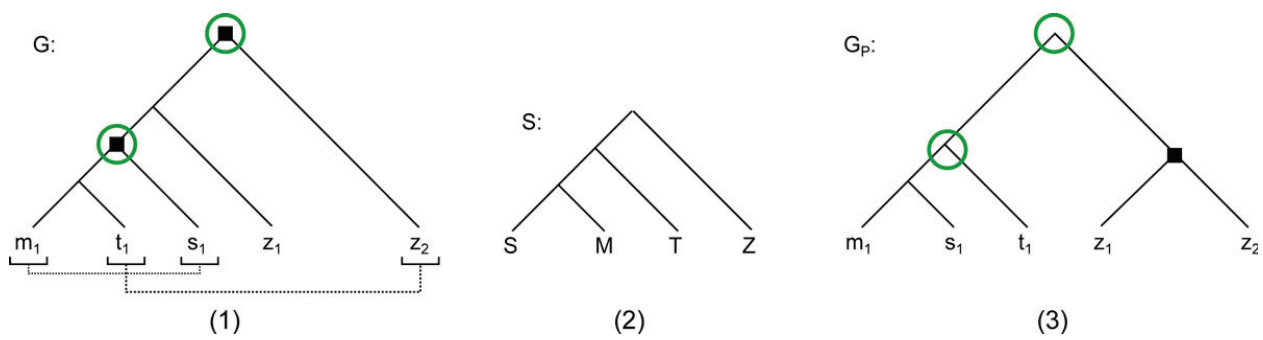


Figure 5 An example of corrected fish tree. The tree for the ZNF800 gene family before and after correction, restricted to the species Stickleback (S), Medaka (M), Tetraodon (T) and Zebrafish (Z). (1) The original gene tree G given by Ensembl, using the same notation as in figure 2 for duplications, preservable nodes and required orthologs. Gene region analysis gave us the required orthologs $P = \{(m_1, s_1), (t_1, d_2)\}$. (2) The species tree associated with the four species. (3) The gene tree given by our correction algorithm.

gene tree according to some external information on orthology. The rearrangements are modifications that are as small as possible, given some distance criterion (namely the RF distance), but can be more significant according to other distances such as the usual NNI (nearest neighbor interchange) distance. We show that for fish genomes, the rearrangements we define can be efficient to explore statistically equivalent gene trees when sequence alignement is used to compute likelihood. As corrected trees satisfy synteny contraints, we can be confident enough that they describe the gene family evolution better.

Many algorithmic and theoretical problems remain open. For example, is there a similar way for handling paralogy constraints? What about having both orthology and paralogy constraints? It can be shown that there exist sets of constraints with both types that cannot be satisfied. What are the conditions for a set of orthology/paralogy constraints to be satisfiable?

These algorithms may be used in a global framework to contruct large gene tree sets which are arguably better than those found in standard databases. The implementation of such a framework is an on-going work.

Competing interests
None.

Authors' contributions
ML, MS, KS, ET, NE modeled the problem, devised the algorithms and wrote the paper. ML implemented the software.

Declarations
This work is funded by the Natural Sciences and Engineering Reserach Council of Canada (NSERC), Fonds de Recherche Nature et technologies of Quebec, Agence Nationale pour la Recherche and Ancestrome project ANR-10-BINF-01-01.
This article has been published as part of *BMC Bioinformatics* Volume 14 Supplement 15, 2013: Proceedings from the Eleventh Annual Research in Computational Molecular Biology (RECOMB) Satellite Workshop on Comparative Genomics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcbioinformatics/supplements/14/S15.

Authors' details
[1]Département d'Informatique (DIRO), Université de Montréal, H3C3J7, Canada. [2]Laboratoire de Biométrie et Biologie Évolutive, UMR CNRS 5558, Université Lyon I, F-69622 Villeurbanne, France. [3]INRIA Grenoble Rhône-Alpes, F-38334 Montbonnot, France. [4]McGill Center for Bioinformatics, McGill University, H3C2B4, Canada.

Published: 15 October 2013

References

1. Felsenstein J: **Evolurionary trees from DNA sequences: a maximum likelihood approach.** *Journal of Molecular Evolution* 1981, **17**:368-376.
2. Akerborg O, Sennblad B, Arvestad L, Lagergren J: **Simultaneous Bayesian gene tree reconstruction and reconciliation analysis.** *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106(14)**:5714-5719.
3. Berglund-Sonnhammer A, Steffansson P, Betts M, Liberles D: **Optimal gene trees from sequences and species trees using a soft interpretation of parsimony.** *Journal of Molecular Evolution* 2006, **63**:240-250.
4. Boussau B, Szöllosi GJ, Duret L, Gouy M, Tannier E, Daubin V: **Genome-scale coestimation of species and gene trees.** *Genome Research* 2013, **23**:323-330.
5. Gorecki P, Eulenstein O: **A linear-time algorithm for error-corrected reconciliation of unrooted gene trees.** *ISBRA, Volume 6674 of LNBI* Springer-Verlag; 2011, 148-159.
6. Rasmussen MD, Kellis M: **A Bayesian approach for fast and accurate gene tree reconstruction.** *Molecular Biology and Evolution* 2011, **28**:273-290.
7. Szöllosi GJ, Rosikiewicz W, Bousseau B, Tannier E, Daubin V: **Efficient Exploration of the Space of Reconciled Gene Trees 2013. [Submitted].**.
8. Thomas P: **GIGA: a simple, efficient algorithm for gene tree inference in the genomic age.** *BMC Bioinformatics* 2010, **11**:312.
9. Jun J, Mandoiu II, Nelson CE: **Identification of mammalian orthologs using local synteny.** *BMC Genomics* 2009, **10**:630 [http://dx.doi.org/10.1186/1471-2164-10-630].
10. Chauve C, El-Mabrouk N, Gueguen L, Semeria M, Tannier E: *Models and algorithms for genome evolution* Springer; 2013, chap. Duplication, rearrangement and reconciliation: a follow-up 13 years later.
11. Lafond M, Swenson K, El-Mabrouk N: *Models and algorithms for genome evolution* Springer; 2013, chap. Error detection and correction of gene trees.
12. Wapinski I, Pfeffer A, Friedman N, Regev A: **Automatic genome-wide reconstruction of phylogenetic gene trees.** *Bioinformatics* 2007, **23(13)**: i549-i558 [http://bioinformatics.oxfordjournals.org/content/23/13/i549.abstract].
13. Bérard S, Gallien C, Boussau B, Szöllosi GJ, Daubin V, Tannier E: **Evolution of gene neighborhoods within reconciled phylogenies.** *Bioinformatics* 2012, **28(18)**:i382-i388.
14. Chaudhary R, Burleigh J, Eulenstein O: **Efficient error correction algorithms for gene tree reconciliation based on duplication, duplication and loss, and deep coalescence.** *BMC-Bioinformatics* 2011, **13(Supp 10)**:S11.
15. Górecki P, Eulenstein O: **Algorithms: simultaneous error-correction and rooting for gene tree reconciliation and the gene duplication problem.** *BMC Bioinformatics* 2012, **13(Suppl 10)**:S14.
16. Muffato M, Louis A, Poisnel CE, Crollius HR: **Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes.** *Bioinformatics* 2010, **26(8)**:1119-1121.
17. Nguyen T, Ranwez V, Pointet S, Chifolleau AMA, Doyon JP, Berry V: **Reconciliation and local gene tree rearrangement can be of mutual profit.** *Algorithms Mol Biol* 2013, **8**:12 [http://dx.doi.org/10.1186/17487188-8-12].
18. Flicek P: **Ensembl 2012.** *Nucleic Acids Research* 2012, **40(Database)**:D84-D90.
19. Vilella A, Severin J, Ureta-Vidal A, Heng L, Birney E: **EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates.** *Genome Research* 2009, **19(2)**:327-335.
20. Chauve C, El-Mabrouk N: **New perspectives on gene family evolution: losses in reconciliation and a link with supertrees.** *RECOMB 2009, Volume 5541 of LNCS* Springer; 2009, 46-58.
21. Doroftei A, El-Mabrouk N: **Removing Noise from Gene Trees.** *WABI 2011, Algorithms in Bioinformatics, Volume 6833 of LNCS/LNBI* 2011, 76-91.
22. Lafond M, Swenson K, El-Mabrouk N: **An Optimal Reconciliation Algorithm for Gene Trees with Polytomies.** *Algorithms in Bioinformatics, proceedings of WABI'12, Volume 7534 of LNCS/LNBI* 2012, 106-122.
23. Guidon S, Gascuel O: **A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood.** *Systematic Biology* 2003, **52**:696704.
24. Shimodaira H, Hasegawa M: **Consel: for assessing the confidence of phylogenetic tree selection.** *Bioinformatics* 2001, **17(12)**:1246-1247.

doi:10.1186/1471-2105-14-S15-S5
Cite this article as: Lafond *et al.*: Gene tree correction guided by orthology. *BMC Bioinformatics* 2013 **14**(Suppl 15):S5.

A.5 Efficient gene tree correction guided by species and synteny evolution

Auteurs : Emmanuel Noutahi, Magali Semeria, Manuel Lafond, Jonathan Seguin, Bastien Boussau, Laurent Gueguen, Nadia El-Mabrouk, Eric Tannier

Revue :

Statut : in prep

Efficient gene tree correction guided by species and synteny evolution

Emmanuel Noutahi[1,◕], Magali Semeria[2,◕], Manuel Lafond[1], Jonathan Seguin[1], Bastien Boussau[2], Laurent Guéguen[2], Nadia El-Mabrouk[1], Eric Tannier[2,3,*]

1 Département d'Informatique (DIRO), Université de Montréal, H3C3J7, Canada
2 LBBE, UMR CNRS 5558, Université de Lyon 1, F-69622 Villeurbanne, France
3 INRIA Grenoble Rhône-Alpes, F-38334 Montbonnot, France

◕These authors contributed equally to this work.
*** Eric.tannier@inria.fr**

Abstract

Gene trees are traditionally inferred from multiple alignments of homologous sequences according to a model of sequence evolution. They often contain unresolved parts or weakly supported resolutions. Information for their full resolution may lie in the poorly exploited dependency between gene families and their genomic context. Integrative methods use species tree information in addition to sequence information. They often rely on computationally intensive tree space searches which forecloses a large application to whole genomic databases. We propose a method called ProfileNJ that takes a gene tree with statistical supports on its branches and corrects its weakly supported parts by a using a combination of information from the species tree and a distance matrix. This reduction in tree space exploration allows for a significant gain in running time. ProfileNJ is part of RefineTree, a modular software package of gene tree correction techniques using several kinds of information: species tree, extant synteny through ortholog predictions, and ancestral synteny through a model of gene neighborhood evolution.

The low running time allows to propose an alternative set of gene trees for the whole Ensembl Comapara database, which allows a genome-wide analysis of duplication and loss patterns on the history of 65 eukaryote species, including ancestral genes and gene orders of all ancestors along this phylogeny. We show that according to several measures including running time, likelihood, stability of genome content and linearity of ancestral chromosomes, trees corrected by RefineTree are arguably more plausible than the ones stored by Ensembl. We finally discuss the quality criteria in the light of gene definition as a sequence or as a locus and extract some cases where a "true" gene tree should depend on this definition.
RefineTree web interface is available at:
`http://www-ens.iro.umontreal.ca/~adbit/polytomysolver.html`

Introduction

Several gene tree databases from whole genomes are available, including Ensembl Compara [1], Hogenom [2], Phog [3], MetaPHOrs [4], PhylomeDB [5], Panther [6]. However they are known to contain many errors and uncertainties, in particular for unstable families [7]. Their use for accurate ancestral genome inference, orthology detection, or the study of genome dynamics is still uncertain.

For example Ensembl Compara trees, when reconciled with a species tree to annotate gene duplication and loss, systematically and unrealistically overestimate the number of genes in ancestral genomes, and lead

to erroneous predictions of ancestral chromosome structures [8]. It is a known artifact and a significant number of nodes in the Ensembl gene trees are labeled as "dubious" [9].

Reasons for errors in gene trees are numerous. They are usually constructed from a DNA or protein sequence alignment with a model of substitution. Thus they are dependent on gene annotations, gene family clustering or alignment errors. In addition models make simplifying assumptions, inducing known systematic artifacts, and algorithms may not properly explore the solution space. But above all, gene sequences often do not contain enough substitutions to resolve all the branches of a phylogeny, or alternatively too many substitutions such that the substitution history is saturated. Therefore *sequence based methods*, computing gene trees from sequence information (e.g. PhyML [10], RAxML [11], MrBayes [12], PhyloBayes [13]), are usually accompanied with measures of statistical supports on their branches or *a posteriori* distributions of likely trees.

It is possible to choose among many statistically equivalent trees or sharpen the *a posteriori* distribution by modeling gene gains and losses inferred from the reconciliation of gene and species trees [14]. It is the goal of *integrative methods* (e.g. TreeBeST [15], TreeFix [16], BBCA [17], PhylDog [8], ALE [18], GSR [19, 20], SPIMAP [21], Giga [22], MowgliNNI [23]). They use species tree information in addition to sequence information. They all report gene trees with better accuracy compared with gene trees obtained solely from sequence alignment. But they leave a large space for improvement, both in terms of tree quality and in terms of computing time. Most methods use tree space exploration strategies based on small modifications or *local moves* on branches (typically NNI, SPR, TBR), usually proposed at random. Moves are accepted or rejected according to hill-climbing, Metropolis-like criteria, or other statistical or empirical arguments. For example, TreeFix [16] proposes a neighboring gene tree at random and retains it if it improves fitness with the species stree and is statistically equivalent to the input gene tree. Such exploration space methods are computationally intensive and do not scale well as databases grow in size. Consequently, database construction pipelines such as TreeBeST (contructing the Ensembl Compara gene trees [1]) have to adopt compromises, exploring limited subsets of tree spaces.

Gene tree reconstruction methods also fail to use other information from whole genome evolution such as synteny (gene relative positions on the chromosomes). Synteny is often seen as one of the best ways to predict orthology [24], a task that is theoretically contained in the phylogeny problem. But, probably due to the complexity of including this information in an evolutionary model, this has never been really exploited in a phylogenetic context (with the exception of a few pioneering studies applied on yeast genomes [25]).

In this paper, we present gene tree *correction* techniques. They consist in taking as input gene trees with supports on their branches, and proposing a different resolution for the weakly supported parts, based on information from a species tree or from synteny. Thus this information is used not only in the evaluation step, but directly in the proposition step [26–37].

We propose a correction method called ProfileNJ. It contracts all unsupported branches and reconstructs a binary tree according to the duplication and loss cost of reconciliation with the species tree. It extends an algorithm by our group [34] by integrating Neighbor-Joining principles to choose among the numerous optimal solutions, and by allowing unrooted trees as input.

We compare ProfileNJwith TreeFix [16], adopting the same evaluation strategy, that is, exploring statistically equivalent neighboring trees. On simulations both algorithms achieve results of comparable quality, ProfileNJ being several times faster. The gain in running time allowed us to run ProfileNJ, along with correction tools according to synteny, on the whole set of gene families from the Ensembl database, which is out of reach of alternative methods. For each family, starting with a PhyML maximum likelihood starting tree, we provide an alternative gene tree set as follow: (1) contract unsupported branches, and apply ProfileNJ; (2) construct a set of putative orthologs from synteny blocks between genomes (obtained with PhylDiag [38]), and apply ParalogyCorrector [35]; (3) construct ancestral chromosomes with DeCo [39], and correct some trees which are responsible for a non linear chromosome structure [36]. Thus we use a range of information in the construction of gene trees, taking into account gene sequence evolution, gene content evolution and chromosome structure evolution. We evaluate the results according to several criteria: (1)

likelihood ratio based on the Ensembl alignments between our results and the ones stored at Ensembl; (2) ancestral genome sizes based on a duplication-loss reconciliation; (3) linearity of ancestral chromosomal segments computed with DeCo.

The trees for the whole database are obtained in a few hours on a desktop computer (not including the starting tree construction) and compare very favorably with the trees stored in Ensembl. The set of trees and ancestral genomes are made accessible. We also use the reconstructed trees and ancestral genomes to study genome evolution across all the 69 eukaryotic species from the Ensembl database. In particular, a whole genome analysis of duplication patterns is provided, pointing at certain branches which seem to show acceleration of duplication or loss processes. We finally discuss the distance to "true" gene trees, in the light of incomplete lineage sorting and gene conversion.

1 ProfileNJ

Consider an initial tree with branch support, which can be retrieved from a gene tree database, or alternatively produced with a phylogenetic reconstruction tool. We introduce a new correction technique based on contracting weakly supported branches of the tree (under a certain threshold), and applying a new algorithm, ProfileNJ, that computes a binary refinement of the obtained multifurcated gene tree. ProfileNJ uses three kinds of information:

- The minimum duplication and loss cost of reconciliation with the species tree;
- The number of ancestral gene copies leading to the minimum cost (like in phyletic profiles methods such as [40]);
- The Neighbor-Joining [41] distance between gene sequences.

Previous algorithm: ProfileNJ is based on a previous algorithm, developed by our group [34], for finding a binary refinement of a multifurcated gene tree G, minimizing the duplication and loss cost of reconciliation with a given species tree S. As explained in [34], each polytomy (multifurcated node) of G can be considered independently. Therefore, in the following, we restrict the presentation to a single polytomy P.

The algorithm, based on dynamic programming, computes a table M where, for each node (including leaves) u of S and each integer k (limits on k are discussed in [34]), $M(u, k)$ is the minimum duplication and loss cost required to have k genes in the branch of the species tree leading to u. The table is computed line by line begining with the leaves of S. The final cost of a minimum resolution of the polytomy is given by $M(r, 1)$, where r is the root of S.

When M values have been computed, a backtracking algorithm outputs a *count vector* containing the number of genes per branch of S. Such a vector, similar to a phyletic profile ([40]), is then associated with a scenario of gains and losses along the species tree.

As for the complexity of the algorithm, when the reconciliation cost is simply the number of operations (i.e. the same unit cost is attributed to each duplication or loss), table M can be constructed in time linear in the size of S [34], leading to a linear-time algorithm for finding one optimal binary resolution of the polytomy. Moreover, we showed recently [] that linearity can be extended to a whole gene tree involving multiple polytomies. The initial method developed in [34] has also been extended in [] to allow for weighted operations, i.e. different costs for duplications and losses. This generalization gives rise to a quadratic-time algorithm.

ProfileNJ: Several binary refinements of minimum reconciliation cost (optimal refinements) for a polytomy P can be output using the table M. ProfileNJ has been designed to explore the space of all optimal refinements. The main idea is to explore all possible count vectors. Each vector corresponds to a given duplication and loss scenario leading to a given topology of a binary tree. However, paralogous genes belonging to the same genome can be placed indistinguishably one from the other on the leaves. For each

topology, an appropriate placement of gene paralogs is then chosen according to a Neighbor joining (NJ) distance between gene sequences.

For example, the solution vector *count* output from Step 4 in Figure 1 accounts for two duplications. The first duplication, located on the branch leading to b, is deduced from the fact that extant genome b contains three gene copies, while *count* indicates two copies on the branch leading to b; the second one, located on the branch leading to d, is deduced from the fact that two genes are assigned respectively to the branches leading to a and b, while only one gene is assigned to the branch leading to d. The vector *count* does not however involve any information allowing to know which of the three genes b_1, b_2, b_3 are implicated in the first duplication.

We use an NJ criterion to choose among all possible leaf-labeling of a binary tree topology corresponding to a given count vector. The algorithm proceeds as follows. First, a gene tree reduced to a single leaf is associated to each leaf (gene) of P. Call D the pairwise distance matrix between genes. As in the NJ algorithm, a metric space, induced by D on the leaves of P, is progressively augmented with the genes created at the internal nodes of the binary tree in construction.

Let X be an internal node of the species tree with children X_1 and X_2. Suppose X, X_1, X_2 respectively have x, x_1, x_2 genes according to the count vector. Suppose by induction that the elements of X_1 and X_2 are in the metric space, but not the ones of X. If $x_1 > x$ or $x_2 > x$, the difference is explained by duplications. This duplication history is constructed in a first phase. So first suppose $x < x_i$. Choose, according to the NJ criterion, the couple of elements a, b from X_i which minimizes

$$Q(a,b) = (n-2)D(a,b) - \sum_{i \neq a} D(a,i) - \sum_{i \neq b} D(b,i). \quad (1)$$

Replace a and b by a new element ab in X_i, update the distances and the tree like in the NJ algorithm. This makes x_i decrease until we have $x \geq x_1$ and $x \geq x_2$.

So in a second phase we can suppose that the duplication history has been constructed and we suppose that $x \geq x_1$ and $x \geq x_2$. Suppose also without loss of generality that $x_1 < x_2$. This means that there are x_1 pairs of orthologs to couple. For this choose a from X_1 and b from X_2 which again minimize (1), and replace a and b by an element $x \in X$. When X_1 is empty replace every element from X_2 by one element in X.

At this step some elements of X might not be in the metric space. They correspond to an internal branch of the gene tree. Then it is easy to construct an element of the metric space by applying the NJ updating step on the fixed gene tree (for a fixed subtree there is not selection step)

At the end of these procedures all elements of X correspond to an element of the metric space, so an iteration is possible, to the next node of the species tree.

Note that if multifurcatef gene tree G is a star tree with genes from only one species, ProfileNJ reduces to NJ. If G is a star tree and no distance matrix is given, ProfileNJ gives a "profile tree", minimizing the weight of a duplication and loss scenario. If G is binary but not rooted (in that case d is useless and not required), ProfileNJ can be used to root the tree according to duplication and loss scenarios. ProfileNJ can also be used to reconcile a rooted binary gene tree with a species tree. So ProfileNJ is a phylogenetic tool that generalizes several usually unrelated standard methods.

Efficiency of the NJ criterion

In order to evaluate the relevance of the NJ criterion, we ran ProfileNJ twice on the same data sets, except that once the distance matrix was computed, using the Ensembl nucleotide alignments, with FastDist from the FastPhylo package [42], and once the distance matrix was random. The starting tree was computed for every family using PhyML on the nucleic alignments, and all branches with aLRT support < 0.95 were contracted. In average 55% of the branches were contracted. A histogram of the full distribution is shown on Figure S1. The species tree is taken from Ensembl.

Then we computed the likelihood of both trees for every family with PhyML. Among the trees for which the likelihood was different (55% of all tested trees), 76% were in favor of the trees built with the FastDist

distance matrix, and the log likelihood differences were much larger for those trees, contributing 95% of the total of log likelihood differences.

The comparisons are clearly in favour of the NJ criterion over no criterion at all, while quantitatively there remains a small but non negligible part of the trees for which no criterion (the random distance matrix) gives an unexplained slightly but significantly better likelihood.

Efficiency of the ProfileNJ tree space exploration strategy

As explained at the begining of this section, ProfileNJ can be used as a tree space exploration tool for the purpose of gene tree correction. Other tree space search strategies have been proposed for phylogenetic reconstruction, most of them based on random exploration of a tree neighborhood. The most common strategy is to select, in the space of trees obtained from the original one by performing some branch moves (NNI, SPR, TBR), the one best fitting the species tree in terms of reconciliation cost. In this class of algorithms, NOTUNG [27] and TreeFix [16] are the most closely related to ProfileNJ, with TreeFix being the most recent one. TreeFix generates a tree neighborhood from NNI and SPR moves and explores it randomly using a hill-climbing strategy. Instead we take a deterministic and more targeted approach by focusing on weakly supported branches of the tree, with a possibly deep modification of the tree. The comparison with TreeFix is intended to compare these two tree space exploration methods.

The authors of TreeFix [16] have compared it with SPIMAP [21], showing a similar accuracy and a higher speed for TreeFix. We perform a similar comparison on the same simulated dataset of 16 fungi. This dataset consists of simulated gene families generated under the SPIMAP model and their corresponding nucleotide alignments, for four different rates of duplication and loss (DL) events: ($1r_D$ - $1r_L$), ($2r_D$ - $2r_L$), ($4r_D$ - $4r_L$) and ($4r_D$ - $1r_L$); where r_D and r_L are respectively the estimated duplication and loss rates for fungi. Comparisons reported in this section are performed on 2575 simulated gene families randomly chosen from the four fungi datasets with different DL rates.

An initial maximum likelihood (ML) tree is constructed for each simulated gene family with RAxML v-8.1.2 [11], with the rapid bootstrap algorithm, under the GTR-Γ model and the majority rule consensus tree as bootstopping criterion. A randomly rooted tree is then provided as input to TreeFix (as TreeFix requires the input tree to be rooted), while a multifurcated unrooted tree obtained by contracting the branches with support lower than 95% is provided as input to ProfileNJ. We used default parameters for both programs. Among the set of resolutions output by ProfileNJ, the best supported tree was selected using consel [43] and site-wise likelihood values computed with RAxML (under the GTR-Γ model of nucleotide substitution).

For RAxML, TreeFix and ProfileNJ trees, we measured the Robinson-Foulds (RF) distance to true trees, compared the reconstructed tree with the true tree using site-wise likelihoods (Figure S7), measured the accuracy of the duplication and loss scenarios (Figure S5), the sensitivity of the accuracy to gene family size (Figure S6), the sensitivity to species tree errors (Figure S8), and the running time.

Figure 2 illustrates the results for the RF distance. It shows that sequence-only does not contain enough signal to lead to the true tree, and integrating information from the species tree is necessary. TreeFix and ProfileNJ reconstruct around 75% of true trees, compared with only 10% for RaxML (RF distances were computed for rooted trees with ProfileNJ and TreeFix, and unrooted trees for RaxML, so RF=0 means good topology and root for ProfileNJ and TreeFix). We investigated some cases where they were erroneous, and found that often, the true scenario was not parsimonious in terms of duplications and losses, while TreeFix and ProfileNJ chose too recent duplications in order to avoid losses. An example is given in supplementary material (Figure S4).

The performance of TreeFix and ProfileNJ are similar in terms of distance to the true tree. If we measure the likelihood of the reconstructed tree, RAxML of course gives the best likelihood. Its likelihood is even usually higher than the likelihood of the true tree, but not significantly according to an AU test. Treefix is designed to produce trees which are not significantly different than the ML tree, which we could check:

1.36% of the trees fail the AU test against the ML tree at $\alpha = 0.05$, while the proportion jumps to 9.17% for ProfileNJ. It is noticeable that this has no visible consequence on the distance to the true tree.

Figure 3 shows that ProfileNJ outperforms TreeFix in running-time. The gap in running-time between the two algorithms increases with tree size. This figure also shows that the most time-consuming step in ProfileNJ is the tree selection with consel. For a tree of size 30, ProfileNJ is about four to seven times faster than TreeFix and about 15 times faster without the tree selection step with consel. This includes the construction of the distance matrix. The construction of the initial RaxML tree is not included because it is common to both methods.

Other analyses, including the sensitivity to gene family size, number of duplications and losses, or errors in the species tree are reported in the supplementary material. They show similar tendencies, TreeFix and ProfileNJ have similar performances on all measures except running time, and RAxML has a lower performance except when there are errors in the species tree.

Indeed we also investigated the impact of the species tree used to reconstruct gene trees. We use an incorrect specie tree (Figure S9 (A)) as input for TreeFix and ProfileNJ. We found that the reconstructed gene trees became less accurate than RAxML gene trees. This impact however was limited to the branches that had been rearranged; the rest of the branches in the TreeFix and ProfileNJ gene trees remained more accurate than RAxML gene trees.

2 Results

RefineTree

We integrated ProfileNJ in a modular online software, called RefineTree, combining a number of correction techniques, with an easy-to-use interface (see Figure S2 in supplementary material).

Two additional correction techniques are included, that were previously published by our group. They use information from extant or ancient genome organization. The first one uses PhylDiag [38] (see Methods) to compute statistically supported synteny blocks of genes between every pair of genomes from the Ensembl database. We then assume that if several genes are found consecutive in one genome, and their homologs are also found consecutive in the other genome, the common linear arrangement was in the ancestor and the homologous genes are probably orthologous. This hypothesis is incorrect in at least three cases : (1) if the whole block of genes was duplicated, (2) if there is a tandem duplication of a gene followed by a differential loss in the two species, or (3) if a gene is converted by a paralog. To handle these cases, we require that (1) the majority of the homologous genes are indeed predicted as orthologs by phylogeny, (2) the common ancestor of two homologous genes does not lead to two paralog descendants placed in tandem in one species. In case (3), we are in a situation where the loci are orthologous but not the sequences. In that case we construct the "locus tree" [44] and trust syntenic information over gene sequence information. Details of these constrains are given in the Method section. Given couples of putative orthologous genes, we use ParalogyCorrector [35] with the output tree from ProfileNJ as input. This method, integrated in RefineTree, constructs the tree which is the closest to the input (in that case, ProfileNJ) tree according to a Robinson Foulds distance, with the constraint that couples of putative orthologs are found orthologs in a reconciled output phylogeny.

The other correction technique integrates information from the linearity of ancestral genomes computed with DeCo [39] (see Methods). Linearity means that genes are linearly ordered along chromosomes, which is true for extant genomes, but not guaranteed in ancestral genomes computed by DeCo. What could seem as a drawback is used here to detect errors in a gene tree: the "Unduplicator" correction [35] algorithm consists in fusing two ancestral copies of a gene when the two copies disrupt the linearity of an ancestral genome. Details can be found in the methods section or the associated publication.

A typical run of RefineTree, integrating all described correction techniques, is illustrated in Figure 4.

2.1 Results on Ensembl gene trees

On the whole Ensembl gene family database (version 73, sept 2013), we compared three sets of trees constructed by a modular use of RefineTree, as in Figure 4.

- **Ensembl trees**: Trees stored in the Ensembl database;
- **ProfileNJ trees**: Output trees from ProfileNJ, with as input PhyML trees (where branches with a < 0.95 aLRT support are contracted) and FastDist distance matrices;
- **Synteny trees**: Output trees of either ParalogyCorrector and Unduplicator (the two are computed and the most likely is chosen) with ProfileNJ trees as input, using PhylDiag and DeCo to construct synteny constraints.

We evaluate the resulting trees with sequence likelihood, ancestral genome content and ancestral chromosome linearity. The results are shown on Figure 5.

The distribution of ancestral gene content sizes is expected to be close to that of extant genomes. Incorrect trees are known to require additional duplications to be reconciled with the species tree, and thus tend to increase the number of genes in ancestral genomes. The linearity of ancestral genomes is expected to be as close as possible to that of the extant genomes as well, with each gene having zero, one or two neighbors, with a peak at two (the 0s and 1s are due to partially assembled genomes). ProfileNJ trees show a better behaviour than Ensembl trees according to the three measures: more than 2/3 of the trees have a better likelihood than Ensembl trees, the ancestral genome content distribution is much closer to the extant one, and the linearity of chromosomes is higher. So this set of trees, achieving better performance according to sequence evolution, gene content evolution and chromosome evolution, is arguably a better dataset than the one stored in the Ensembl database.

However the content and synteny signals are still distant from what we could expect from true trees. The behavior of synteny trees is interesting from this point of view. Their performance drops in terms of likelihood (Figure 5 (A)), but jumps in terms of the stability of gene content and the linearity of ancestral chromosomes (Figure 5 (B) and (C)). One interpretation is that the synteny corrections, while improving synteny signal, are not yet able to propose reliable gene trees. A reason is that they can break well supported branches to achieve orthology constraints. Branches might be highly misplaced while preserving the ancestral content according to the LCA. We however noticed a correlation between the size of the families (number of genes) and the loss of likelihood in the Synteny trees. Part of the likelihood drop could also be interpreted as an inadequacy of the phylogenetic models to appropriately account for gene families with a high rate of duplications. As observed in our simulations, the true tree is not necessarily the ML tree. Add that likelihood is computed with an alignment which results from a guiding tree which is different from the tested tree. Some synteny trees might therefore be considered as better trees even with these equivocal results.

However there is a third interpretation. Synteny information describes the history of loci [44], while phylogenetic models describe the evolution of sequences. Loci and sequences often have the same history, but they may differ following gene conversion or incomplete lineage sorting (ILS).

In case of ILS or gene conversion, two different true versions of the gene history are concurrent. In Figure 6 the gene as a locus has a history depicted by the right tree, while the gene as a sequence has a history depicted by the left tree. None of the two are wrong, but they are significantly different. They highlight the ambiguity of the definition of a gene, which yields an ambiguity in its history. Sequence trees will have a high likelihood and mediocre results for gene contents and synteny when constructed from duplication and loss scenarios, while it is the opposite for loci trees. ILS in sequence and duplications and losses in loci have been modeled [44], handling this difference in one case. However, no model is currently able to handle conversion.

Modes of evolution in eukaryotes

With the gene trees we reconstruct all gene contents of ancestral genomes and the way they are organized along ancestral chromosomes. Gene content is computed according to the LCA reconciliation (see Methods), and genome organizations consist in sets of links between consecutive genes. Ancestral genomes are not exactly linearly arranged, but sufficiently close to be often interpreted as chromosomes. We do not linearize them by removing links because the non linearity has diverse causes that we do not wish to mask. But this method also highlights genes or groups of genes evolving together in parts of the tree. For example there are 8488 blocks of co-duplicated genes according to DeCo. Most of them contain only a few number of genes (83% contain 2 genes). The largest blocks are found in the terminal branches leading to *Danio rerio* and *Caenorhabditis elegans*.

As seen in Figure 7, branches of the phylogeny which carry large numbers of duplications are also visible. Patterns of duplications in mammals have been studied by Boussau et al [8] with a subset of gene families or in vertebrates by Mahmudi et al [45] with a subset of species, but few methods are able to handle whole databases and provide a complete view on the duplication and loss pattern. Figure 7 shows the result for the full genomes of the full phylogeny of the 65 Ensembl species. Branches with a large number of duplications (hot branches) are those leading to vertebrates, which is in agreement with the two rounds of whole genome duplication hypothesis. Interestingly, the speciation event leading to *Petromyzon marinus*, which is usually thought to have diverged after these events [46], preceeds the hot branches. This may be in agreement with recent results based on the analysis of Hox clusters in the Japanese lamprey [47]. Another hot branch leads to eutherian mammals, which was also found by two other studies [8, 45] with partial data. These two hottest internal branches are exactly the ones found by Mahmudi et al [45] using a probabilistic technique, but using only 9 species due to computational cost. Other hot branches are terminal, the hottest being those leading to *Caenorhabditis elegans* and *Danio rerio*. This is possibly due to ongoing dynamics of polymorphic copy number variations. The same tree showing the number of losses is provided in the supplementary material (Figure S10)

Discussion

Possible uses of RefineTree

RefineTree is a gene tree correction toolbox that explores part of the tree space around a given gene tree, using information from a species tree and synteny. As such, it is modular and can be used with variations.

Various ways of contracting the branches of the starting tree can be considered, varying thresholds or choosing specific branches to contract. For example an exploration scheme contracting the branches one by one and applying ProfileNJ can be considered, which would be equivalent to local modifications [30]. A more radical modification would be to contract all branches. Other kinds of contraction schemes can be imagined, as contracting branches around "Non Apparent Duplications" [48], or "Dubious duplications" stored in the Ensembl trees.

Notice that moves considered here are not local reversible moves such as NNI, SPR, TBR, that can be used in a Monte Carlo exploration framework with a Metropolis algorithm. However our method could be used to produce a starting tree to speed-up a burn-in step and start a sampling from plausible trees. It might also be useful to guide proposals that would have a good probability to be accepted. These steps would speed-up the convergence, which could be useful as these techniques are known to be rather time consuming on large data.

An integrated model of genome evolution

The corrections and evaluations we propose are not integrated in a mathematical framework of genome evolution. They are fast and intuitive ways to construct, according to a range of different criteria and on a

whole genome scale, gene trees that are better rated than the current state of the art. In order to integrate these principles in a model of genome evolution, we would need to model the stability of genome content and the linearity of ancestral chromosomes. While a local version of the former is contained in gene content evolution and can be integrated [8], modeling linearity is more out of reach. A lot of models for genome evolution have linear structures to handle chromosomes [49], but none is able to include duplications and losses at a whole genome scale. For chromosomes defined as a list of local neighborhoods, like here using DeCo, a probability distribution of ancestral genomes according to their linearity, as well as a probabilistic version of DeCo, that could be used in an integrative model, still need to be developed.

Phylogeny and the quest for orthologs

As gene trees contain the most complete information about a gene family history, detecting orthologs or studying gene repertoire evolution should be achieved by interpreting trees. But due to the rate of errors in the current trees stored in databases, orthology is often assessed with a series of techniques including synteny [50] and Reciprocal Best Hits, while the evolution of gene repertoire is often studied with phyletic profile techniques [51]. What we present here is a way of integrating those diverse techniques into a phylogenetic framework. Full sets of orthology relations may de derived from our set of trees, while lists are more incomplete when derived from the Ensembl trees.

Not only the gene trees

Using genome evolution in the construction of the gene trees, we get ancestral genomes as a byproduct. They are made of genes and sets of gene adjacencies. They are still too big (in terms of gene number) and too non linear to be fully trusted. This is partly due to incorrect gene trees in our output, or incorrect inferences from DeCo, but also to problems in sequencing, assembling, annotating genomes, clustering families or inferring the species tree. Good methods for finding linear structures from a set of adjacencies exist [52]. Here we rather used non-linearity as a testimony of the flaws of the data and methods used to reconstruct genome evolution.

Although gene trees are "better" with our correction, they are still not good enough. The likelihood drop for synteny correction is indeed surprising, as these corrections lead to ancestral genomes that are closer to gene content and gene neighborhoods of extant genomes. We would need better exploration schemes with integrated models to really trust gene trees on a whole genome database within a deep phylogeny.

3 Methods

Families, alignments and trees are taken from Ensembl Compara release 73. They were computed with a pipeline called TreeBest, but we simply call them the "Ensembl trees". Trees are rooted and available with branch support and annotation. There are 20529 trees, each corresponding to a gene family, for a total of 1091891 genes taken from 67 species. Information on gene position on chromosomes, scaffolds or contigs is available. See `ftp://ftp.ensembl.org/pub/release-73/emf/ensembl-compara/homologies/`.

Use of ProfileNJ on Ensembl

PhyML was used with default parameters to compute maximum likelihood trees from the protein multiple alignments from Ensembl. An aLRT support was computed, and all branches with aLRT < 0.95 were contracted. FastDist was run on DNA alignments to provide a distance matrix. Then ProfileNJ was run with the command (an example is given for the first family).

```
ProfileNJ -s Compara.73.species_tree  \\
          -g data/famille_1.start_tree \\
```

```
-d data/famille_1.dist \\
-o data/famille_1.tree \\
-n -r best -c nj --slimit 1 --plimit 1 --firstbest --cost 1 0.99999
```

We tested the sensitivity of the method to the choice of the threshold parameter for contracting unsupported branches. The threshold is a trade-off between the amount of change in a tree and the probability that the resulting tree is rejected. Too small values would avoid exploring a large space around the starting tree while high values would lead to low likelihood trees. It has to be settled empirically. For example .80 was considered an acceptable threshold in some genomic studies [53].

Ancestral Genomes (gene content and order) from the LCA Reconciliation

LCA reconciliation is used to infer ancestral gene contents, one family at a time. It consists in labeling every node x of the gene tree with a node of the species tree corresponding to the last common ancestor of all extant species containing a gene which is a descendant of x (including x itself if x is a leaf). Then every internal node x is labeled with an event: a duplication if the species label of x is equal to the label of one of its children, and a speciation otherwise.

The LCA reconciliation induces sets of ancestral genes: for a species S (extant or ancestral), draw a graph in which every leaf of a gene tree which maps to S or one of its descendant is a node. Then draw an edge between two homologous genes x and y if their last common ancestor in the gene tree maps to a proper descendant of S, or to S but is a speciation node. Connected components of this graph are the ancestral genes in S: there is exactly one ancestral gene per component, and its descendants are the nodes of the component.

We organized the ancestral genes in the genomes using DeCo [39]. This algorithm aims at constructing the neighborhoods between ancestral genes, but starts by inferring ancestral gene contents.

This LCA method assumes the trees to be binary and does not take branch support and node annotation into account. In particular, the algorithm ignores the fact that a branch may be uncertain due to a weak support, or that a node may be labeled as dubious as in Ensembl. However, part of our goal is precisely to resolve unsupported and dubious parts of gene trees, by considering the validity of the obtained ancestral genomes.

Testing the linearity of ancestral genomes with DeCo

DeCo [39] computes ancestral gene neighborhoods that are highly dependant on the shape of the considered gene tree. Indeed, adjacencies in extant genomes, *i.e.* the immediate proximity of two consecutive genes, are taken as input and putative adjacencies in ancestral genomes are constructed by a parsimony principle minimizing the number of gains and losses of adjacencies. As two contemporaneous adjacencies are supposed to evolve independently one from the other, the linearity of extant genomes, *i.e.* the property that one gene never has more than two neighbors linked by an adjacency, does not guarantee the linearity of ancestral ones.

The apparent weakness of this feature is in fact a strength to evaluate the quality of gene trees. Indeed, a high part of the non linearity of ancestral genomes is not due to the inadequacy of the software itself, but to the quality of the input data. Indeed it has been remarked that a significant improvement in the linearity of ancestral genomes was obtained by constructing gene trees according to more complete models [8, 54].

Note that in extant genomes, no gene can have more than two neighbors, and most genes have two. But many genes have 1 or 0, because of the poor assembly of some genomes, many contigs contain one or a few genes.

Information from extant synteny

First we ran PhylDiag as follows, for each pair of genomes. Files genome_1, genome_2 and ancestral_genes respectively contain the ordered list of genes from each genome, and the list of families clustering the genes

as in the Ensembl database.

```
phylDiag.py genome_1 genome_2 ancestral_genes \\
    -gapMax=2 -pThreshold=0.00000005 \\
    -filterType=InBothSpecies -multiprocess \\
    -minChromLength=2 >syntenyblocks_1_2
```

The statistical threshold is calculated in order to minimize the number of false positives, taking into account the number (2211) of comparisons between pairs of species and the expected number (500) of synteny blocks for each comparison ($0.05/(2211 * 500) \approx 5e-8$).

For each synteny block found by PhylDiag, we kept only the genes that had one single exemplar in the two blocks from both species. We counted the number of such pairs of genes, and refered to an LCA reconciliation of the output trees of ProfileNJ to check that most pairs are orthologs (their common ancestor is labeled by a speciation). We discarded the blocks that did not fit this condition. This discards possible block duplications.

For the remaining blocks, and for each couple of uniquely represented genes a and b, we required that the LCA node X of a and b in the reconciled ProfileNJ tree is not a supported duplication: let X_1 and X_2 be the two children of the node X labeled as a duplication (so X_1 and X_2 are in the same species as X), the genes a and b are not kept as putative orthologs if one of the branches XX_1 and XX_2 has a high support (> 0.95), and there are two genes, x_1 and x_2, which respectively descend from X_1 and X_2, which are located on the same genome. This discards possible tandem duplications in the block, followed by differential losses of copies.

The output trees from ProfileNJ as well as the filtered pairs of putative orthologs were givent as input to ParalogyCorrector, which finds the tree that is as close as possible to the input tree in terms of RF distance, such that in an LCA reconciliation, all pairs of putative orthologs have an LCA node annotated as a speciation.

Information from ancestral synteny

From the results of DeCo on the output gene trees produced by ProfileNJ, we used an "unduplication" principle as in [36] everytime we found that an ancestral gene x had three neighbors a, b, c, two of them (say a, b) arising from a duplication node d in a single gene tree. In that case, we rearranged the four grand children of d so that the clade under d has an LCA which is annotated as a speciation in the LCA reconciliation. See an insight into its functionning in Figure 8.

Likelihood ratio tests

We computed the likelihood of all trees according to the HKY85 model with PhyML on nucleotide alignments. To test the significance of a likelihood difference, we computed the AU (Appoximately unbiased) tests with Concel. They consist in bootstraping the sites of an alignment, each site having a likelihood according to several trees. Then a probability is associated to each tree from this bootstrap, according to the number of replicates which place it above the others in terms of the bootstrapped likelihood. Unless otherwise stated, we use "significantly" better for a likelihood with a AU value > 0.95.

Data access

The 2575 simulated gene families used for our simulation represent a subset of the original SPIMAP simulated fungi datasets (see http://compbio.mit.edu/spimap/). Those data and the RAxML trees constructed from sequence alignment are available. We also provide the two sets of 20529 trees, as an output

from ProfileNJ and with the additional synteny-aware corrections. All softwares are freely accessible for academic purpose, under a GPL license.

Figure legends

Acknowledgments

MS, LG, BB and ET were supported by the French Agence Nationale de la Recherche (ANR) through Grant ANR- 10-BINF-01-01 "Ancestrome". EN, ML, JS and NEM were supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), and the "Fonds de recherche du Québec Nature et technologies" (FRQNT) of Quebec. Computations were made on the supercomputer "Briarée" from Université de Montréal, managed by Calcul Québec and Compute Canada.

References

1. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. EnsemblCompara gene trees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Research. 2009;19:327-335.

2. Penel S, Arigon AM, Dufayard JF, Sertier AS, Daubin V, Duret L, et al. Databases of homologous gene families for comparative genomics. BMC Bioinformatics. 2009;10 Suppl 6:S3. Available from: http://dx.doi.org/10.1186/1471-2105-10-S6-S3.

3. Datta RS, Meacham C, Samad B, Neyer C, Sjölander K. Berkeley PHOG: PhyloFacts orthology group prediction web server. Nucleic Acids Research. 2009;37:W84-W89.

4. Pryszcz LP, Huerta-Cepas J, Gabaldón T. MetaPhOrs: orthology nd paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. Nucleic Acids Research. 2011;39:e32.

5. Huerta-Cepas J, Capella-Gutierrez S, Pryszcz LP, Denisov I, Kormes D, Marcet-Houben M, et al. PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. Nucleic Acids Research. 2011;39:D556-D560.

6. Mi H, Muruganujan A, Thomas PD. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Research. 2012;41:D377-D386.

7. Boeckmann B, Robinson-Rechavi M, Xenarios I, Dessimoz C. Conceptual framework and pilot study to benchmark phylogenomic databases based on reference gene trees. Brief Bioinform. 2011 Sep;12(5):423–435. Available from: http://dx.doi.org/10.1093/bib/bbr034.

8. Boussau B, Szöllősi GJ, Duret L, Gouy M, Tannier E, Daubin V. Genome-scale coestimation of species and gene trees. Genome Research. 2013;23:323-330.

9. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2014. Nucleic Acids Research. 2014 Jan;42(Database issue):D749–D755. Available from: http://dx.doi.org/10.1093/nar/gkt1196.

10. Guindon S, Gascuel O. A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology. 2003;52:696- 704.

11. Stamatakis A. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analysis with thousands of taxa and mixed models. Bioinformatics. 2006;22:2688-2690.

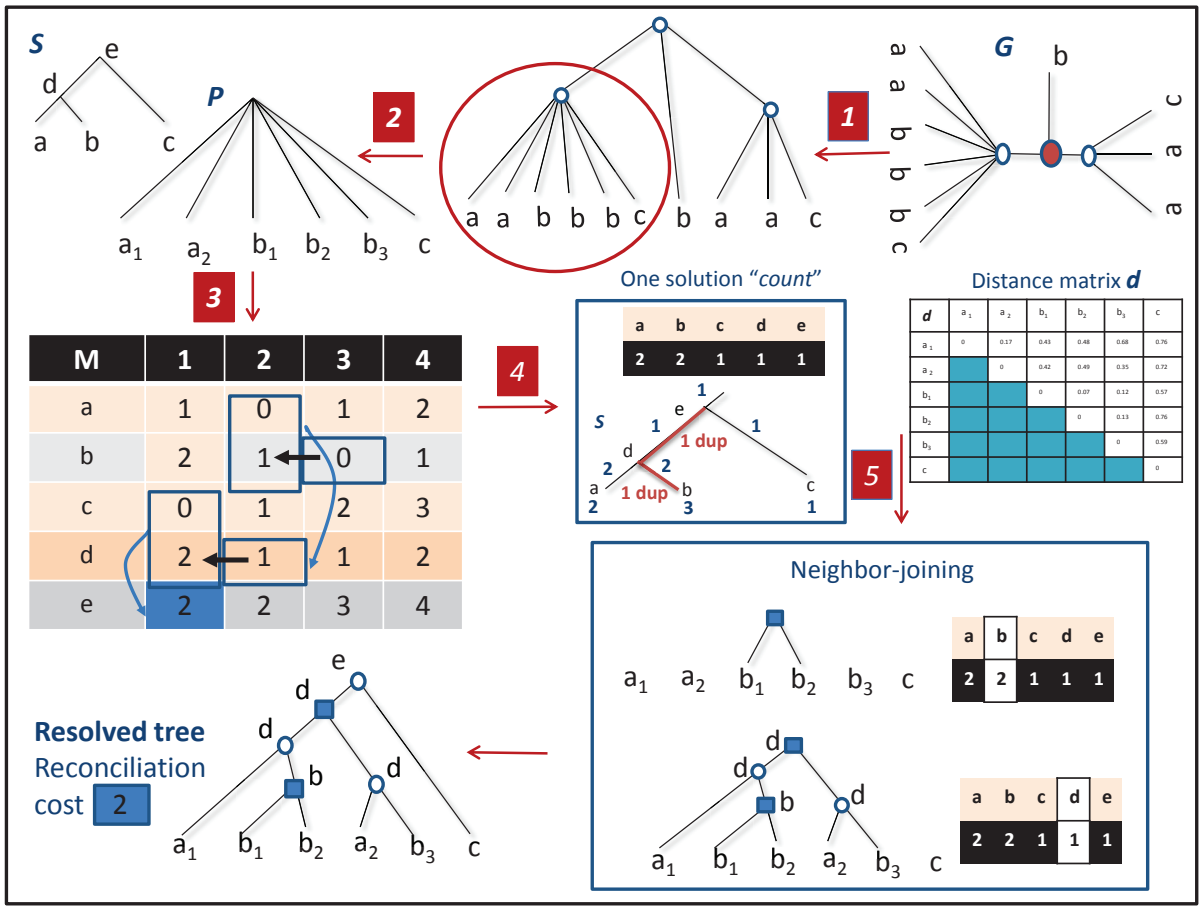


Figure 1. ProfileNJ at a glance. The input is the species tree S, an unrooted gene tree G with multifurcations and a distance matrix d. Step 1. Each internal node of the gene tree is considered in turn for the root. Here the considered root is highlighted in red in G. Step 2. In a bottom-up traversal of the tree, each polytomy P (non-binary node) is considered in turn. Step 3. With P and S as input, a dynamic programming table M is constructed. $M(u, k)$ denotes the minimum weight scenario of duplications and losses required to have k genes in the branch of the species tree leading to u. Each entry is constructed from neighboring entries as in [34]. Step 4. All minimum weight duplication and loss *count* solutions are obtained by backtracking in M. Step 5. One *count* solution might correspond to many binary trees. The Neighbor joining (NJ) procedure computes the one that best agrees with the distance matrix. The final completely resolved tree is given bottom left.

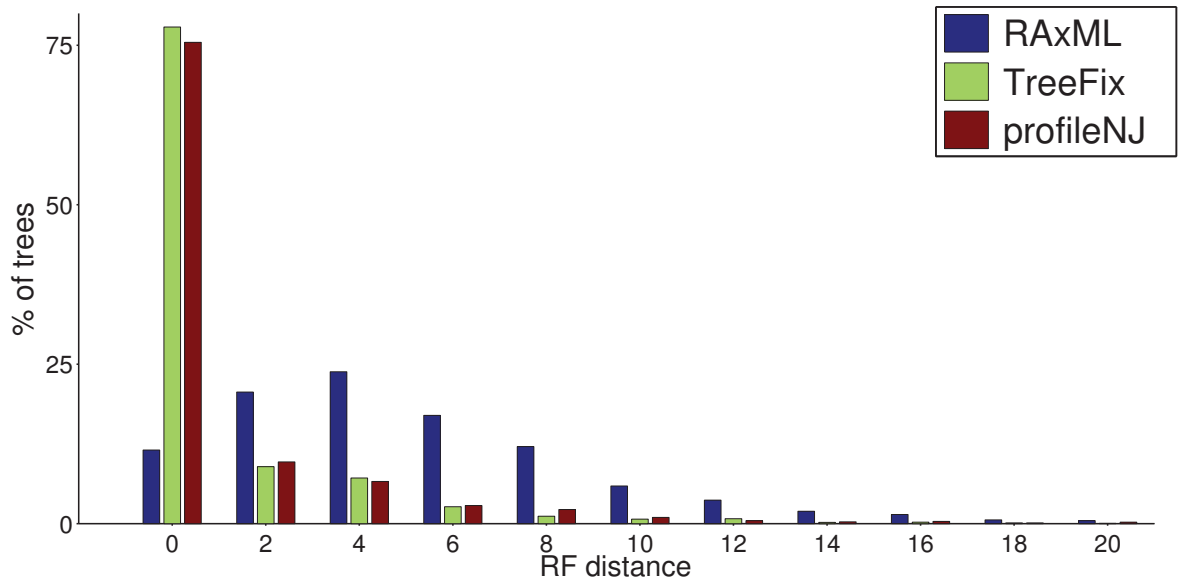


Figure 2. Topology accuracy of RAxML, TreeFix and ProfileNJ trees, mesured by RF distance with the true tree, on $\sim$ 2500 simulated trees from the fungal dataset. We use a sample of trees simulated under four different DL rate : ($1r_D$ - $1r_L$), ($2r_D$ - $2r_L$), ($4r_D$ - $4r_L$) and ($4r_D$ - $1r_L$). Percentage of reconstructed trees (y-axis) with a given RF distance (x-axis) to the true tree. TreeFix and ProfileNJ have a similar reconstruction accuracy (75% of trees match the true trees) while the input tree (RAxML) have the lowest accuracy. The graph is cut on the right, but contains more than 99% of the data.

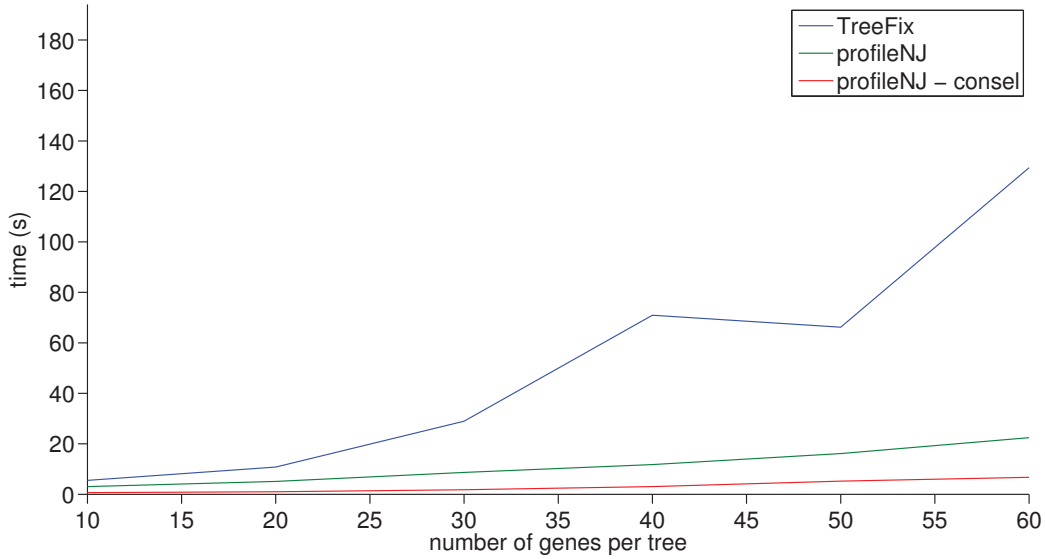


Figure 3. Runtime of TreeFix and ProfileNJ for increasing size of gene tree.

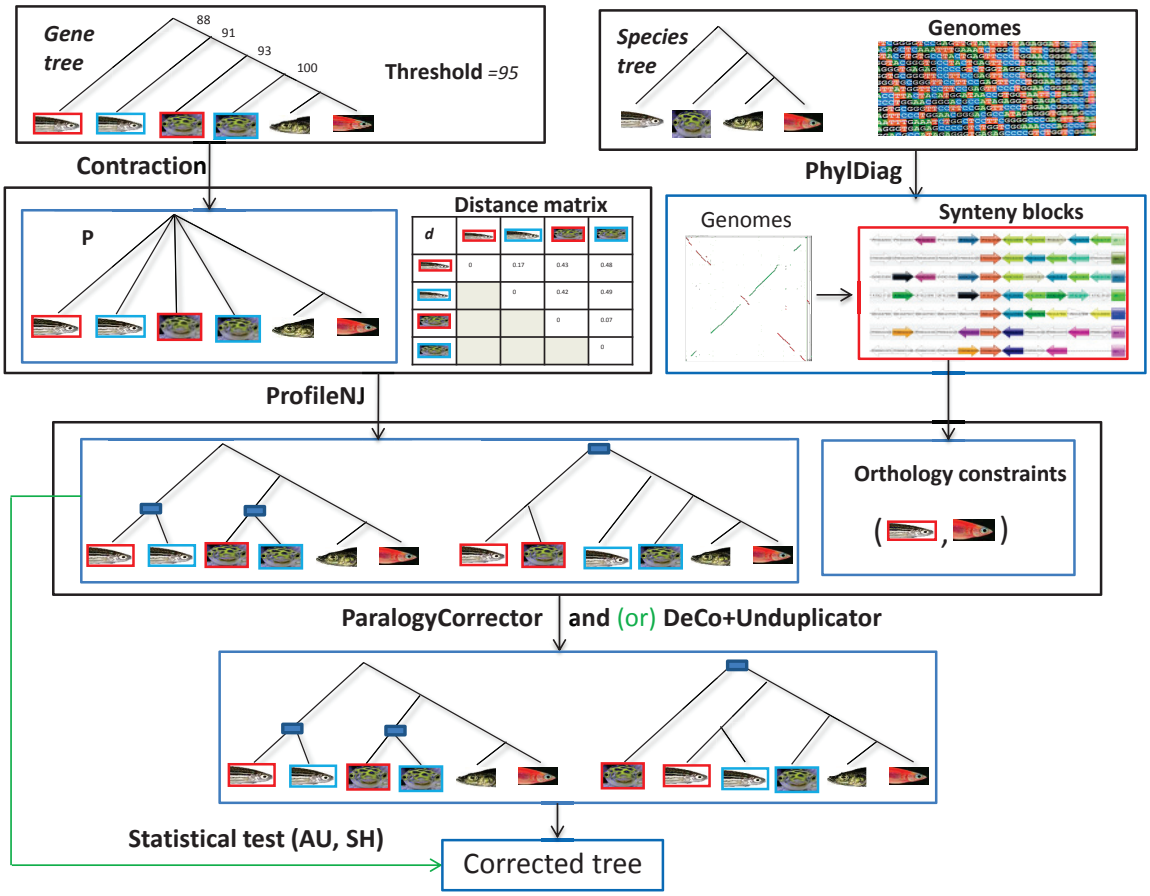


Figure 4. A general view on RefineTree when run on the Ensembl Compara gene families. An example is given for a species tree S of four fish species, a gene family of six genes (a gene is represented by the picture of the species it belongs to, and two paralogs belonging to the same species are distinguished by a different frame color), a rooted gene tree G (although it can be unrooted in general) with branch support, and a given threshold for branch contraction. Data framed in black are the input and those framed in blue are the output of the correction algorithm labeling the edge linking the considered frames. Black arrows depict the use we make of RefineTree on the Ensembl gene trees. The green arrow and the green "or" are alternative uses avoiding one or both of the correction tools ParalogyCorrector and Unduplicator. Any framed set of data can be alternatively provided to the pipeline as input. For example, orthology constraints obtained from various sources can be directly provided as input to ParalogyCorrector. The method for inferring orthology constraints from synteny blocks is described in the text.

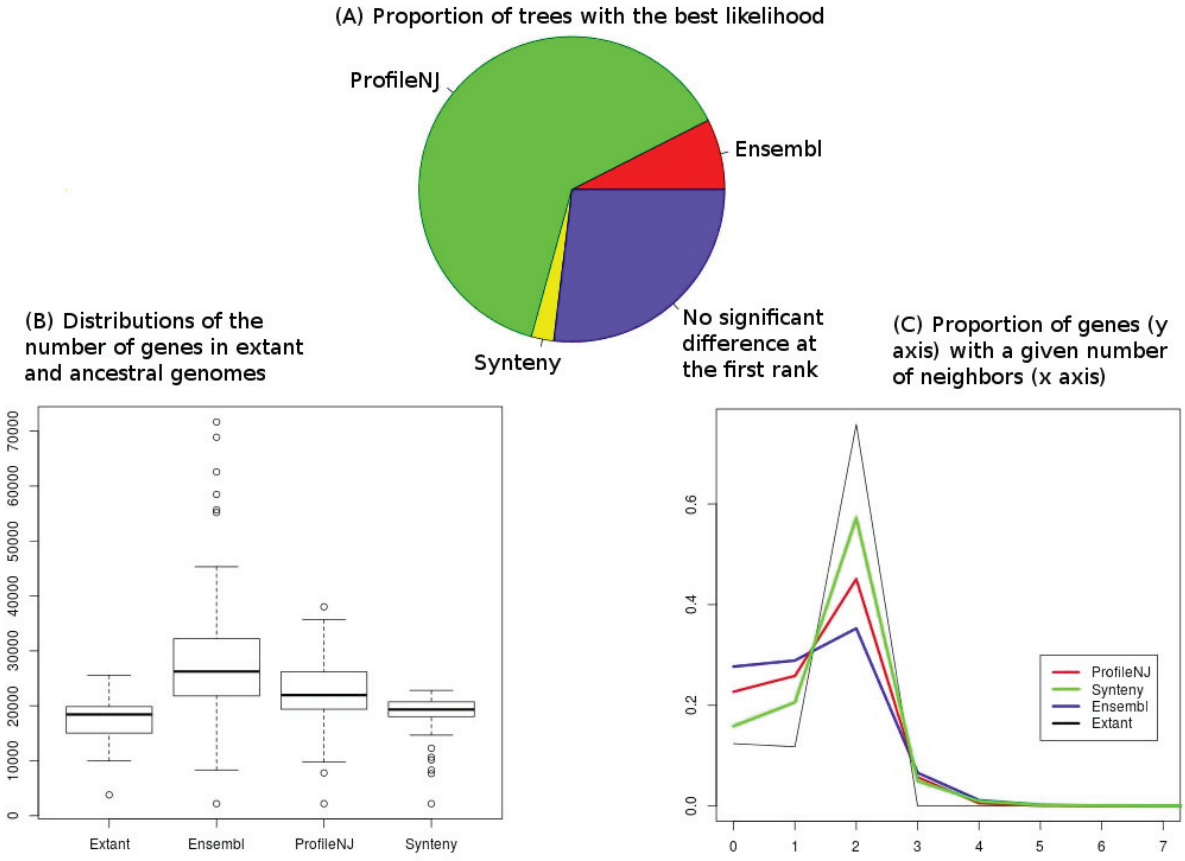


Figure 5. Sequence likelihood, ancestral genome content and ancestral chromosome linearity for ProfileNJ, Synteny and Ensembl trees: **(A)** Proportion of trees with a significantly better likelihood computed with PhyML. AU tests were computed for the three trees for each family, and if the tree at the first rank was significantly better than the second, it was stored as the best likelihood, and if not, it was stored as "no significant difference at the first rank". **(B)** Gene content computed with DeCo. Gene content has one value for each node of the phylogeny of 65 species, except for extant genomes, for which it has one value for each leaf. **(C)** Genome linearity computed with DeCo. Genome linearity is represented by a graph, whose x axis is the number of neighbors a gene can have, and the y axis shows the proportion of genes having this number of neighbors. Parameters from extant genomes are given as a reference in (B) and (C). Statistics for ancestral genomes are assumed better when close to the extant ones.

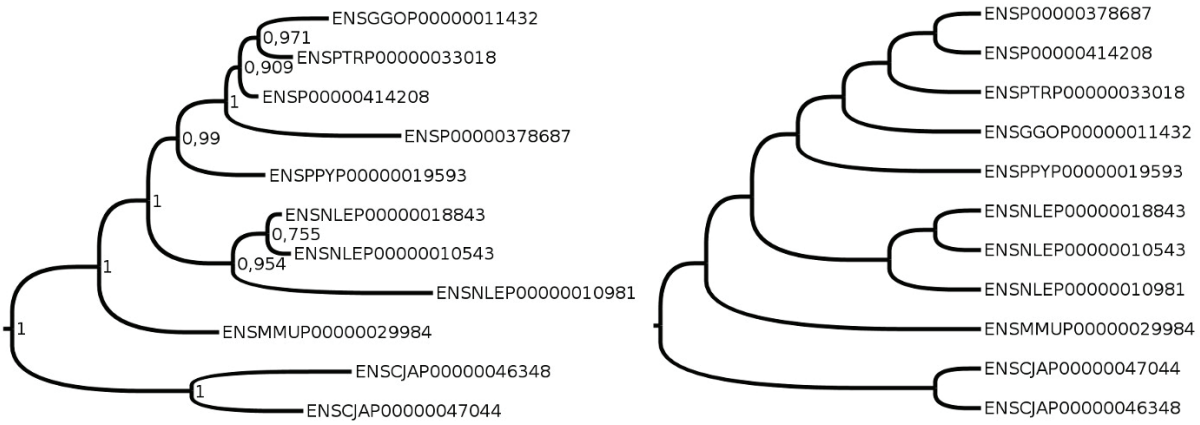


Figure 6. A probable example of ILS visible on a subtree of an ensembl gene family. The monophyly of the chimpanze and gorilla genes (ENSPTRP00000033018 and ENSGGOP00000011432) is well supported by the sequences (left tree, constructed by PhyML, with aLRT supports), while synteny argues for orthology of both with the human genes (ENSP00000414208 and ENSP00000378687) (right tree, constructed by ProfileNJ followed by ParalogyCorrector), so that a scenario of duplication and losses compatible with the left tree is unlikely.

12. Ronquist F, Huelsenbeck JP. MrBayes3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 2003;19:1572- 1574.

13. Lartillot N, Philippe H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Molecular Biology and Evolution. 2004 Jun;21(6):1095–1109. Available from: `http://dx.doi.org/10.1093/molbev/msh112`.

14. Szöllősi GJ, Tannier E, Daubin V, Boussau B. The inference of gene trees with species trees. Systematic Biology. 2015 Jan;64(1):e42–e62. Available from: `http://dx.doi.org/10.1093/sysbio/syu048`.

15. Schreiber F, Patricio M, Muffato M, Pignatelli M, Bateman A. TreeFam v9: a new website, more species and orthology-on-the-fly. Nucleic Acids Research. 2013;Doi: 10.1093/nar/gkt1055.

16. Wu YC, Rasmussen MD, Bansal MS, Kellis M. TreeFix: Statistically informed gene tree error correction using species trees. Systematic Biology. 2013;62(1):110- 120.

17. Zimmermann T, S M, Warnow T. BBCA: Improving the scalability of BEAST using random binning. BMC Genomics. 2014;(15(Suppl 6)):S11. Proceedings of RECOMB-CG.

18. Szöllősi GJ, Rosikiewicz W, Boussau B, Tannier E, Daubin V. Efficient exploration of the space of reconciled gene trees. Systematic Biology. 2013 Nov;62(6):901–912. Available from: `http://dx.doi.org/10.1093/sysbio/syt054`.

19. Akerborg O, Sennblad B, Arvestad L, Lagergren J. Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. Proceedings of the National Academy of Sciences USA. 2009;106(14):5714–5719.

20. Arvestad L, Berglund AC, Lagergren J, Sennblad B. Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. In: RECOMB; 2004. p. 326-335.

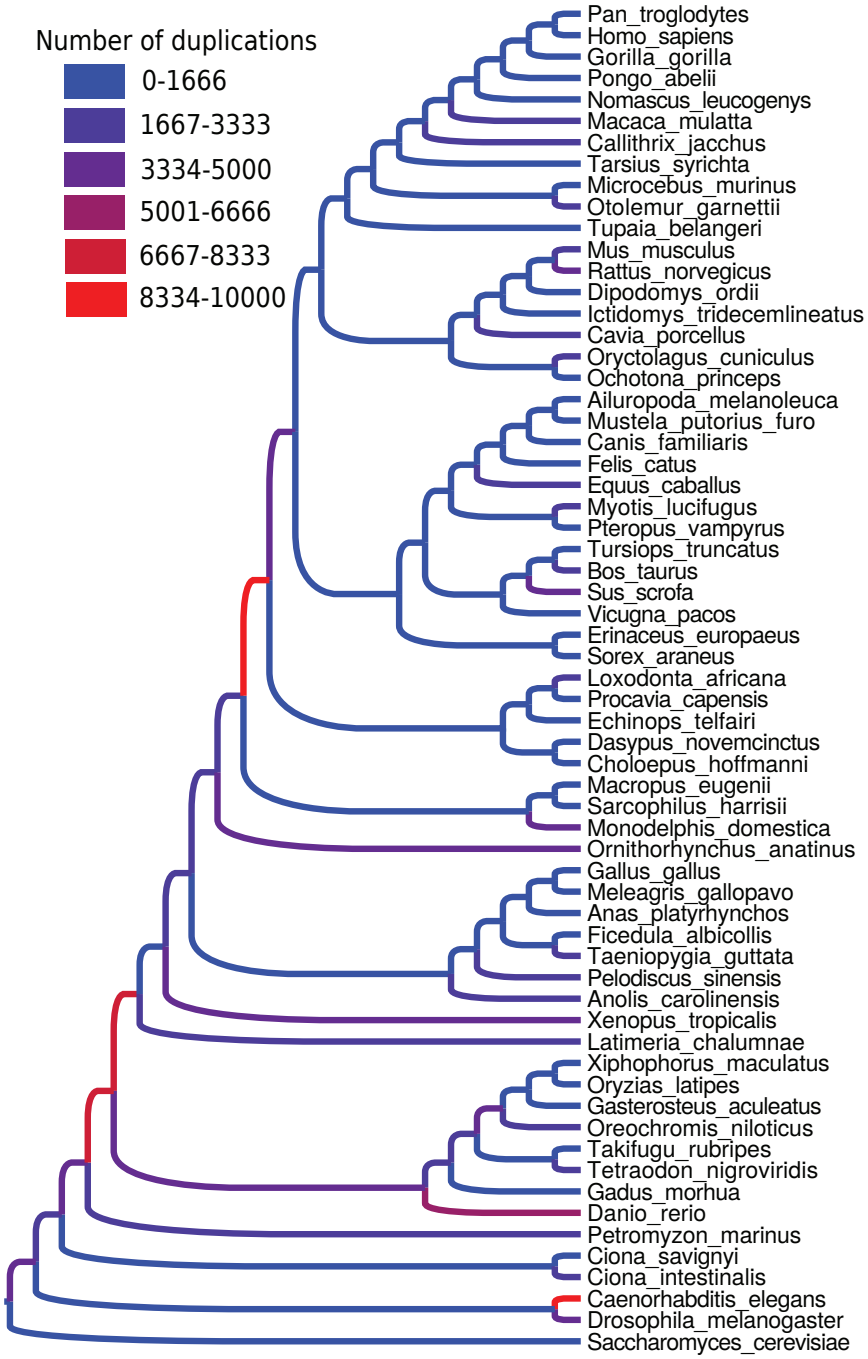


Figure 7. Numbers of duplications in the eukaryote phylogeny, estimated with reconciled ProfileNJ trees from PhyML starting trees on the whole Ensembl Compara database, version 73. Drawn with Figtree [55].

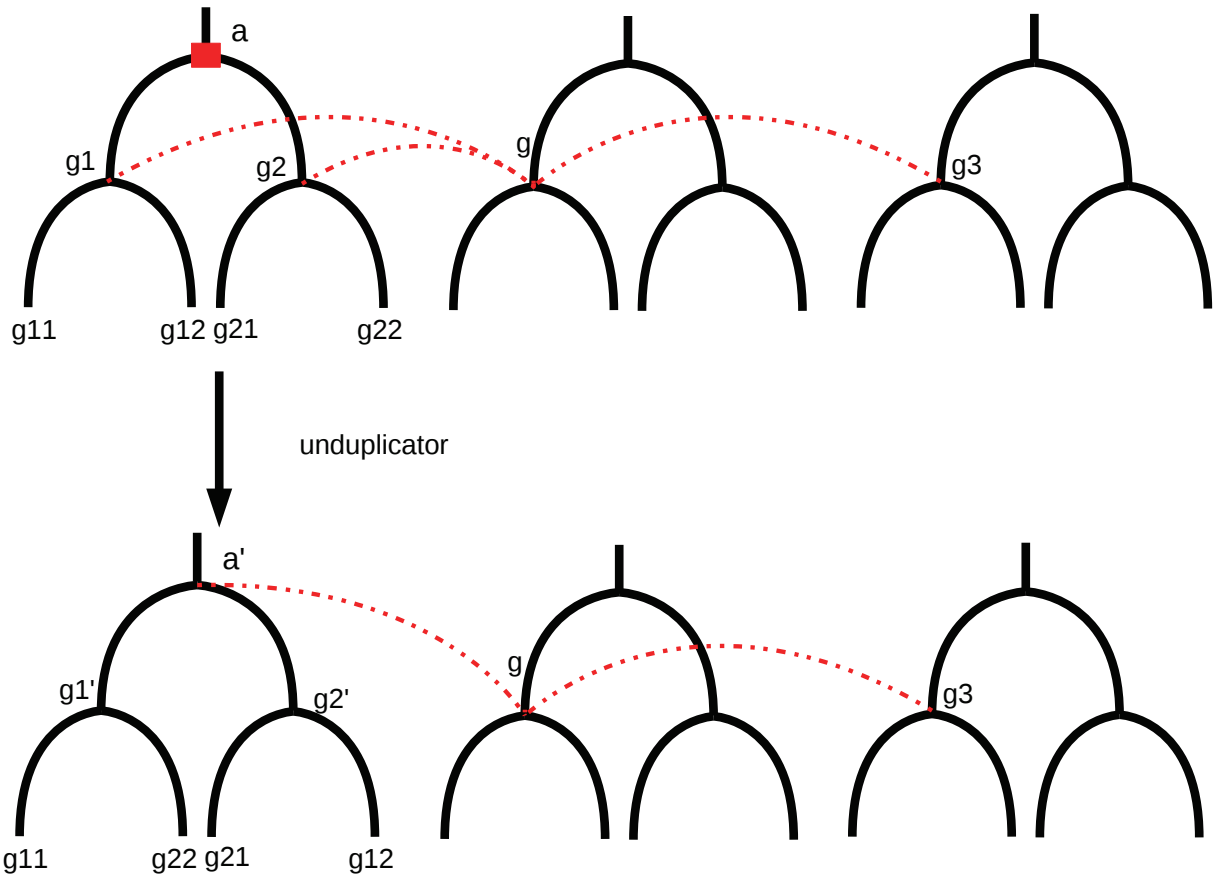


Figure 8. The unduplication principle (figure from [36]). A non linearity is detected in an ancestral genome (gene g has three neighbors). Two of its neighbors g_1 and g_2 are issued from a possibly dubious duplication labeled node. The tree is rearranged so that its root is labeled with a speciation instead of a duplication. In the resulting configuration g_1' and g_2' are in two different species, so that g can have only one neighbor in this family, and linearity is recovered.

21. Rasmussen MD, Kellis M. A bayesian approach for fast and accurate gene tree reconstruction. Molecular Biology and Evolution. 2011;28(1):273- 290.

22. Thomas PD. GIGA: a simple, efficient algorithm for gene tree inference in the genomic age. BMC Bioinformatics. 2010;11:312.

23. Nguyen TH, Ranwez V, Pointet S, Chifolleau AMA, Doyon JP, Berry V. Reconciliation and local gene tree rearrangement can be of mutual profit. Algorithms for Molecular Biology. 2013;8(1):12. Available from: http://dx.doi.org/10.1186/1748-7188-8-12.

24. Jun J, Mandoiu II, Nelson CE. Identification of mammalian orthologs using local synteny. BMC Genomics. 2009;10:630. Available from: http://dx.doi.org/10.1186/1471-2164-10-630.

25. Wapinski I, Pfeffer A, Friedman N, Regev A. Automatic genome-wide reconstruction of phylogenetic gene trees. Bioinformatics. 2007;23(13):i549–i558. Available from: http://bioinformatics.oxfordjournals.org/content/23/13/i549.abstract.

26. Durand D, Haldórsson BV, Vernot B. A hybrid micro-macroevolutionary approach to gene tree reconstruction. Journal of Computational Biology. 2006;13:320–335.

27. Chen K, Durand D, Farach-Colton M. Notung: Dating Gene Duplications using Gene Family Trees. Journal of Computational Biology. 2000;7:429–447.

28. Gorecki P, Eulenstein O. A linear-time algorithm for error-corrected reconciliation of unrooted gene trees. In: ISBRA. vol. 6674 of LNBI. Springer-Verlag; 2011. p. 148-159.

29. Gorecki P, Eulenstein O. Algorithms: simultaneous error-correction and rooting for gene tree reconciliation and the gene duplication problem. BMC Bioinformatics. 2011;13(Supp 10):S14.

30. Chaudhary R, Burleigh JG, Eulenstein O. Efficient error correction algorithms for gene tree reconciliation based on duplication, duplication and loss, and deep coalescence. BMC Bioinformatics. 2011;13(Supp.10):S11.

31. Berglund-Sonnhammer AC, Steffansson P, Betts MJ, Liberles DA. Optimal gene trees from sequences and species trees using a soft interpretation of parsimony. Journal of Molecular Evolution. 2006;63:240-250.

32. Doroftei A, El-Mabrouk N. Removing Noise from Gene Trees. In: WABI. vol. 6833 of LNBI/LNBI; 2011. p. 76-91.

33. Swenson KM, Doroftei A, El-Mabrouk N. Gene Tree Correction for Reconciliation and Species Tree Inference. Algorithms for Molecular Biology. 2012;7(1):31.

34. Lafond M, Swenson KM, El-Mabrouk N. An Optimal Reconciliation Algorithm for Gene Trees with Polytomies. In: LNCS. vol. 7534 of WABI; 2012. p. 106-122.

35. Lafond M, Semeria M, Swenson KM, Tannier E, El-Mabrouk N. Gene tree correction guided by orthology. BMC Bioinformatics. 2013;14 (supp 15)(S5).

36. Chauve C, El-Mabrouk N, Guéguen L, Semeria M, Tannier E. Duplication, Rearrangement and Reconciliation: A Follow-Up 13 Years Later. In: Chauve C, El-Mabrouk N, Tannier E, editors. Models and Algorithms for Genome Evolution. London: Springer; 2013. p. 47–62.

37. Bansal MS, Wu YC, Alm EJ, Kellis M. Improved Gene Tree Error-Correction in the Presence of Horizontal Gene Transfer. Bioinformatics. 2014 Dec;Available from: http://dx.doi.org/10.1093/bioinformatics/btu806.

38. Lucas JM, Muffato M, Roest Crollius H. PhylDiag: identifying complex synteny blocks that include tandem duplications using phylogenetic gene trees. BMC Bioinformatics. 2014;15(1):268. Available from: http://dx.doi.org/10.1186/1471-2105-15-268.

39. Bérard S, Gallien C, Boussau B, Szöllősi GJ, Daubin V, Tannier E. Evolution of gene neighborhoods within reconciled phylogenies. Bioinformatics. 2012 Sep;28(18):i382–i388. Available from: http://dx.doi.org/10.1093/bioinformatics/bts374.

40. Csurös M. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics. 2010 Aug;26(15):1910–1912. Available from: http://dx.doi.org/10.1093/bioinformatics/btq315.

41. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution. 1987;4:406-425.

42. Khan MA, Elias I, Sjölund E, Nylander K, Guimera RV, Schobesberger R, et al. Fastphylo: fast tools for phylogenetics. BMC Bioinformatics. 2013;14:334. Available from: http://dx.doi.org/10.1186/1471-2105-14-334.

43. Shimodaira H, Hasegawa M. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics. 2001;17(12):1246–1247. Available from: http://bioinformatics.oxfordjournals.org/content/17/12/1246.abstract.

44. Rasmussen MD, Kellis M. Unified modeling of gene duplication, loss, and coalescence using a locus tree. Genome Research. 2012 Apr;22(4):755–765. Available from: http://dx.doi.org/10.1101/gr.123901.111.

45. Mahmudi O, Sjöstrand J, Sennblad B, Lagergren J. Genome-wide probabilistic reconciliation analysis across vertebrates. BMC Bioinformatics. 2013;14 Suppl 15:S10. Available from: http://dx.doi.org/10.1186/1471-2105-14-S15-S10.

46. Smith JJ, Kuraku S, Holt C, Sauka-Spengler T, Jiang N, Campbell MS, et al. Sequencing of the sea lamprey (Petromyzon marinus) genome provides insights into vertebrate evolution. Nature genetics. 2013;45(4):415–421.

47. Mehta TK, Ravi V, Yamasaki S, Lee AP, Lian MM, Tay BH, et al. Evidence for at least six Hox clusters in the Japanese lamprey (Lethenteron japonicum). Proceedings of the National Academy of Sciences. 2013;110(40):16044–16049. Available from: http://www.pnas.org/content/110/40/16044.abstract.

48. Lafond M, Chauve C, Dondi R, El-Mabrouk N. Polytomy refinement for the correction of dubious duplications in gene trees. Bioinformatics. 2014 Sep;30(17):i519–i526. Available from: http://dx.doi.org/10.1093/bioinformatics/btu463.

49. Fertin G, Labarre A, Rusu I, Vialette ETT. Combinatorics of Genome Rearrangements. MIT press; 2009.

50. Sonnhammer ELL, Gabaldón T, Sousa da Silva AW, Martin M, Robinson-Rechavi M, Boeckmann B, et al. Big data and other challenges in the quest for orthologs. Bioinformatics. 2014 Jul;p. btu492.

51. Cohen O, Ashkenazy H, Burstein D, Pupko T. Uncovering the co-evolutionary network among prokaryotic genes. Bioinformatics. 2012 Sep;28(18):i389–i394. Available from: http://dx.doi.org/10.1093/bioinformatics/bts396.

52. Maňuch J, Patterson M, Wittler R, Chauve C, Tannier E. Linearization of ancestral multichromosomal genomes. BMC Bioinformatics. 2012;13 Suppl 19:S11. Available from: http://dx.doi.org/10.1186/1471-2105-13-S19-S11.

53. Abby SS, Tannier E, Gouy M, Daubin V. Lateral gene transfer as a support for the tree of life. Proceedings of the National Academy of Sciences USA. 2012 Mar;109(13):4962–4967. Available from: http://dx.doi.org/10.1073/pnas.1116871109.

54. Patterson M, Szöllősi G, Daubin V, Tannier E. Lateral gene transfer, rearrangement, reconciliation. BMC Bioinformatics. 2013;14 Suppl 15:S4. Available from: http://dx.doi.org/10.1186/1471-2105-14-S15-S4.

55. ;. Available from: http://tree.bio.ed.ac.uk/software/figtree/.

A.6 Effects of successive predator attacks on prey aggregations

Auteurs : Christophe Lett, Magali Semeria, Andréa Thiebault, Yann Tremblay

Revue : Theoretical Ecology

Statut : publié

URL : http://link.springer.com/article/10.1007/s12080-014-0213-0

Référence Bibliographique : [80]

Le contrat d'édition de *Theoretical Ecology* ne permet pas la libre diffusion de cet article.

A.7 How to capture fish in a school ? Effect of successive predator attacks on seabird feeding success

Auteurs : Andréa Thiebault, Magali Semeria, Christophe Lett, Yann Tremblay

Revue : Journal of Animal Ecology

Statut : publié

URL : `http://onlinelibrary.wiley.com/doi/10.1111/1365-2656.12455/abstract`

Référence Bibliographique : [135]

Le contrat d'édition de *Journal of Animal Ecology* ne permet pas la libre diffusion de cet article.