Université de Montréal

Approches algorithmiques pour l'inférence d'histoires de duplication en tandem avec inversions et délétions pour des familles multigéniques

par
Mathieu Lajoie

Département de biochimie
Faculté de médecine

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de Philosophiæ Doctor (Ph.D.)
en bio-informatique

août, 2009

Université de Montréal
Faculté des études supérieures

Cette thèse intitulée:

Approches algorithmiques pour l'inférence d'histoires de duplication en tandem avec inversions et délétions pour des familles multigéniques

présentée par:

Mathieu Lajoie

a été évaluée par un jury composé des personnes suivantes:

Muriel Aubry,	présidente-rapporteuse
Nadia El-Mabrouk,	directrice de recherche
Mathieu Blanchette,	membre du jury
Dannie Durand,	examinatrice externe
Gertraud Burger,	représentante du doyen de la FES

Thèse acceptée le: 29 janvier 2010

RÉSUMÉ

Une fraction importante des génomes eucaryotes est constituée de Gènes Répétés en Tandem (GRT). Un mécanisme fondamental dans l'évolution des GRT est la recombinaison inégale durant la méiose, entrainant la duplication locale (en tandem) de segments chromosomiques contenant un ou plusieurs gènes adjacents.

Différents algorithmes ont été proposés pour inférer une histoire de duplication en tandem pour un cluster de GRT. Cependant, leur utilisation est limitée dans la pratique, car ils ne tiennent pas compte d'autres évènements évolutifs pourtant fréquents, comme les inversions, les duplications inversées et les délétions.

Cette thèse propose différentes approches algorithmiques permettant d'intégrer ces événements dans le modèle de duplication en tandem classique. Nos contributions sont les suivantes:

- Intégrer les inversions dans un modèle de duplication en tandem simple (duplication d'un gène à la fois) et proposer un algorithme exact permettant de calculer le nombre minimal d'inversions s'étant produites dans l'évolution d'un cluster de GRT.

- Généraliser ce modèle pour l'étude d'un ensemble de clusters orthologues dans plusieurs espèces.

- Proposer un algorithme permettant d'inférer l'histoire évolutive d'un cluster de GRT en tenant compte des duplications en tandem, duplications inversées, inversions et délétions de segments chromosomiques contenant un ou plusieurs gènes adjacents.

Mots clés: arbre de duplication, arbre de gènes, duplication inversée, famille de gènes, médiane, perte de gène, réarrangement génomique, réconciliation.

ABSTRACT

Tandemly arrayed genes (TAGs) represent an important fraction of most genomes. A fundamental mechanism at the origin of TAG clusters is unequal crossing-over during meiosis, leading to the duplication of chromosomal segments containing one or many adjacent genes. Such duplications are called tandem duplications, as the duplicated segment is placed next to the original one on the chromosome.

Different algorithms have been proposed to infer the tandem duplication history of a TAG cluster. However, their applicability is limited in practice since they do not take into account other frequent evolutionary events such as inversion, inverted duplication and deletion.

In this thesis, we propose different algorithmic approaches allowing to integrate these evolutionary events in the original tandem duplication model of evolution. Our contributions are summarized as follows:

- We integrate inversion events in a tandem duplication model restricted to single gene duplications, and we propose an exact algorithm allowing to compute the minimum number of inversions explaining the evolution of a TAG cluster.
- We generalize this model to the study of orthologous TAG clusters in different species.
- We propose an algorithm allowing to infer the evolutionary history of a TAG cluster through tandem duplication, inverted duplication, inversion and deletion of chromosomal segments containing one or many adjacent genes.

Keywords: duplication tree, gene tree, inverted duplication, gene family, median, gene loss, genomic rearrangement, reconciliation.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES

ADN	Acide Désoxyribonucléique
AHR	Recombinaison Homologue Allélique
EGM	Éléments Génétiques Mobiles
GRT	Gènes Répétés en Tandem
LCR	*Low Copy Repeats*
NAHR	Recombinaison Homologue Non Allélique
NNI	*Nearest Neighbor Interchange*
DS	Duplication Segmentale
TAG	*Tandemly Arrayed Genes*

À Éléonore et Raphaël.

REMERCIEMENTS

Tout d'abord, je tiens à remercier les personnes avec qui j'ai eu le plaisir de travailler durant mes études doctorales. NADIA EL-MABROUK, ma directrice de recherche, pour son aide, sa patience, sa disponibilité et ses encouragements. DENIS BERTRAND, pour sa contribution aux travaux présentés dans cette thèse, les connaissances qu'il m'a transmises, ainsi que pour m'avoir gentiment rappelé à l'ordre lorsque mes digressions dépassaient l'entendement. OLIVIER GASCUEL, pour nous avoir proposé d'étendre le modèle de duplication aux inversions, et avec qui j'ai la chance de poursuivre des études post-doctorales à Montpellier. OLIVIER TREMBLAY-SAVARD, pour sa bonne humeur, sa motivation et la relecture de ma thèse.

Je remercie les membres de ma famille, qui m'ont appuyé tout au long de mes études. Mes parents, pour avoir alimenté ma curiosité et mon intérêt pour les sciences dès mon plus jeune âge : *L'électricité expliquée aux enfants* m'a permis de rester en vie, et les camps de jours au mont Saint-Bruno de m'initier à la biologie. Ma soeur JULIE et mon beau-frère FRÉDÉRIC, pour m'avoir accueilli dans leur maison durant les derniers mois de mon doctorat, ainsi que leurs enfants, RAPHAËL et ÉLÉONORE, pour m'avoir si souvent fait sourire.

Un merci tout spécial pour le personnel du Département de Biochimie et du DIRO. GERTRAUD BURGER, pour avoir mis sur pied les programmes de bio-informatique et de bourses biT. ÉLAINE MEUNIER, grâce à qui remplir un formulaire n'a jamais été aussi agréable. MARIE PAGEAU, qui m'a permis de découvrir l'enseignement et avec qui j'ai eu beaucoup de plaisir à travailler.

AMANDINE, ANNIE, DELPHINE, ÉVELYNE, FRED, GUILLAUME, JACQUES, LAURE, LOUIS, MATTHIEU, PIERRE, QUENTIN, SIMON et VICKI, je vous dis également merci. À un moment ou un autre, d'une manière ou d'une autre, vous avez contribué à la réalisation de cette thèse.

INTRODUCTION

Au début des années soixante, l'évolution était considérée essentiellement comme le résultat de mutations ponctuelles affectant les gènes. Les décennies suivantes ont vu s'amorcer un important changement de paradigme, en particulier grâce aux travaux d'Ohno [86], qui proposa la duplication génique comme l'un des principaux moteurs de l'évolution. Par la suite, les avancées technologiques en matière de séquençage ont permis la caractérisation de centaines de *familles multigéniques*, c'est-à-dire d'ensembles de gènes ayant évolué par duplication et spéciation à partir d'un gène ancestral commun. Plus récemment, des analyses comparatives ont révélé que plusieurs de ces familles ont connu des phases d'expansion et de contraction spécifiques à certaines espèces, ce qui suggère fortement que les duplications et les pertes de gènes jouent un rôle primordial dans l'adaptation des organismes à leur environnement, ainsi que dans le processus de spéciation [105, 47].

Bien que l'importance des duplications dans l'évolution des espèces ne soit plus à démontrer, beaucoup reste à faire afin de mieux comprendre les mécanismes de duplication, ainsi que le mode d'évolution des copies dupliquées. De nombreuses études ont donc porté sur l'inférence d'histoires évolutives de familles multigéniques, la plupart reposant sur l'utilisation de méthodes d'inférence phylogénétiques classiques. Celles-ci permettent d'obtenir, à partir des séquences d'ADN ou d'acides aminés, des arbres de gènes représentant les relations ancestrales à l'intérieur des familles multigéniques. Cependant, cette approche est incomplète et présente certaines limites. Premièrement, contrairement à l'inférence d'arbres d'espèces qui peut s'appuyer sur des génomes entiers (l'approche phylogénomique), l'inférence d'arbres de gènes repose sur une quantité d'information limitée par la taille des gènes dans la famille étudiée. En conséquence, les arbres obtenus ont souvent un faible support statistique. Deuxièmement, ces arbres ne représentent pas des histoires évolutives explicites. En effet, étant donné que plusieurs nœuds internes d'un arbre de gènes peuvent résulter d'un même évènement de duplica-

tion, il n'y a pas de correspondance bijective entre les nœuds de l'arbre et les évènements de duplication.

Une information qui n'est pas utilisée par les méthodes d'inférence phylogénétique classiques est l'ordre des gènes sur les chromosomes. Pourtant, certains mécanismes de duplication affectent l'ordre des gènes d'une façon particulière, et cet ordre constitue une information précieuse qui permet d'améliorer l'inférence de l'histoire évolutive des familles multigéniques. C'est sur cette idée que repose le modèle de duplication en tandem, introduit par Fitch [35] à la fin des années soixante-dix, utilisé pour inférer l'histoire évolutive des groupes (ou *clusters*) de gènes répétés en tandem (GRT). Ce modèle, que nous présenterons de façon formelle au chapitre suivant, est basé sur l'hypothèse que les clusters de GRT évoluent uniquement par des recombinaisons inégales menant à des duplications en tandem. Par conséquent, lorsqu'un segment chromosomique est dupliqué, la nouvelle copie et les gènes qu'elle contient se retrouve toujours adjacente à la copie originale, et dans la même orientation transcriptionnelle. Pour un cluster de GRT respectant ces contraintes, le modèle de duplication en tandem de Fitch permet, à partir d'un arbre de gènes et de l'ordre de ces derniers sur le chromosome, d'identifier chaque évènement de duplication sans ambiguïté. Autrement dit, il permet d'associer chaque nœud interne de l'arbre de gènes à un unique évènement de duplication pouvant impliquer plus d'un gène.

Cependant, du fait de la simplicité de ce modèle, qui ne considère que les duplications en tandem en ignorant les autres évènements évolutifs pouvant affecter le nombre, l'ordre et l'orientation transciptionelle des gènes (tels que les délétions, les inversions ou les duplications inversées), son usage est demeuré limité dans la pratique. C'est pour dépasser ces limitations et augmenter le réalisme du modèle que nous avons entrepris les travaux de recherche présentés dans cette thèse. À la suite du Chapitre 1, qui a pour but d'introduire de façon détaillée les modèles biologiques impliqués dans l'évolution des familles multigéniques, ainsi que la définition rigoureuse du modèle de duplication en tandem, nous présentons les trois chapitres principaux de cette thèse qui correspondent

chacun à des extensions différentes de ce modèle. Les Chapitres 2 et 3 correspondent à deux publications dans *Journal of Computational Biology* [62, 13], et le Chapitre 4 à une publication dans *Molecular Biology and Evolution* [61].

Dans le Chapitre 2, nous considérons l'ajout des inversions dans un modèle de duplication en tandem restreint aux duplications *simples* (qui impliquent un seul gène à la fois). Étant donné un cluster de GRT, un arbre de gènes pour ce cluster et un ordre sur ses feuilles (correspondant à l'ordre des gènes sur le chromosome), le problème consiste à inférer le nombre minimal d'inversions s'étant produites au cours de l'évolution de ce cluster. Nous présentons un algorithme de type *branch-and-bound* qui calcule la solution exacte, ainsi qu'une heuristique polynomiale basée sur la distance de points de cassure (*breakpoint*). Notre algorithme utilise le graphe des points de cassure de Hannenhalli et Pevzner, initialement introduit dans le but de calculer la distance d'inversion entre deux permutations [48]. Nous montrons ensuite, à l'aide de simulations, comment ces algorithmes peuvent être utilisés pour améliorer l'inférence phylogénétique pour des familles ayant évolué selon ce modèle. Une application à un cluster de gènes de type KRAB-ZNF est également présentée.

Le modèle de duplication en tandem classique ne permet pas d'étudier un ensemble de clusters simultanément dans plusieurs espèces. Dans le Chapitre 3, nous considérons donc un modèle d'évolution qui, en plus des duplications en tandem simples et des inversions, tient compte des pertes de gènes et des évènements de spéciation. Nous présentons ensuite une méthode générale permettant d'inférer l'ordre des gènes dans les génomes ancestraux qui minimise le nombre total d'inversions dans l'histoire évolutive de la famille étudiée. Au niveau méthodologique, ce chapitre intègre trois approches utilisées dans les études de génomique évolutive : la reconstruction d'un arbre de duplication [33], la réconciliation entre l'arbre des gènes et celui des espèces [91, 15], ainsi que le concept de médiane d'inversion utilisée dans l'inférence phylogénétique basée sur l'ordre des gènes [16, 18].

D'un point de vue algorithmique, les extensions présentées dans les Chapitres 2 et 3

constituent un premier pas vers la généralisation du modèle de duplication en tandem à d'autres types d'évènements évolutifs, mais leur utilité demeure limitée dans la pratique puisqu'elles ne considèrent que les duplications en tandem simples. Pour cette raison, nous proposons dans le Chapitre 4 une heuristique permettant d'inférer l'histoire évolutive d'un cluster de GRT en tenant compte d'un large spectre d'évènements évolutifs pouvant impliquer plusieurs gènes à la fois : les duplications en tandem, les duplications inversées, les inversions et les délétions menant à des pertes de gènes. Bien qu'ici les évènements de spéciation ne soient pas pris en compte, la richesse du modèle considéré fait de cette heuristique un outil très pratique pour étudier l'histoire évolutive des familles multigéniques à l'intérieur d'une seule espèce. L'intérêt de ce modèle est illustré par une application à deux familles multigéniques chez l'humain, à savoir les récepteurs olfactifs et les protocadherines.

Nous concluons cette thèse par une discussion sur les avantages et les limites de nos approches, ainsi que sur les directions de recherche futures envisagées.

CHAPITRE 1

MODÈLES BIOLOGIQUES ET INFORMATIQUES

1.1 ADN et gènes

Un brin d'ADN est constitué d'un squelette de sucres et de phosphates auquel sont fixés quatre types de bases azotées : l'adénine (A), la cytosine (C), la guanine (G) et la thymine (T). En raison de l'asymétrie de la liaison entre les sucres et les phosphates, un brin d'ADN possède une orientation $5' \rightarrow 3'$, ce qui permet de lui associer une unique séquence correspondant à l'enchainement de ses bases selon cette orientation. Certaines paires de bases, lorsqu'elles se font face en orientations inverses, ont la possibilité de s'apparier et sont dites complémentaires : (A-T), (T-A), (C-G) et (G-C). Dans un chromosome, l'ADN se trouve sous la forme de deux brins complémentaires d'orientations inverses (voir Figure 1.1) qui adoptent une structure évoquant une longue échelle torsadée. Selon cette image, les squelettes de sucres et de phosphates correspondent aux montants de l'échelle, et les paires de bases à ses barreaux.

L'ADN possède trois caractéristiques essentielles au maintien du vivant. La première est sa capacité à stocker de l'information, grâce à l'enchainement précis des quatre types de paires de bases qui la composent. Par exemple, le génome diploïde contenu dans chaque cellule d'un être humain est formé de six milliards de paires de bases. Deuxièmement, l'ADN est facilement réplicable, ce qui permet aux organismes de croitre et de se reproduire. En effet, les paires de bases sont maintenues par des liaisons hydrogènes de faible énergie, ce qui permet la séparation momentanée des deux brins complémentaires, et leur utilisation comme matrices pour en synthétiser des nouveaux. Finalement, en raison des mutations qui l'affectent, la réplication de l'ADN est imparfaite et assure aux organismes une descendance génétiquement variée, ce qui permet aux espèces de s'adapter à des environnements changeants.

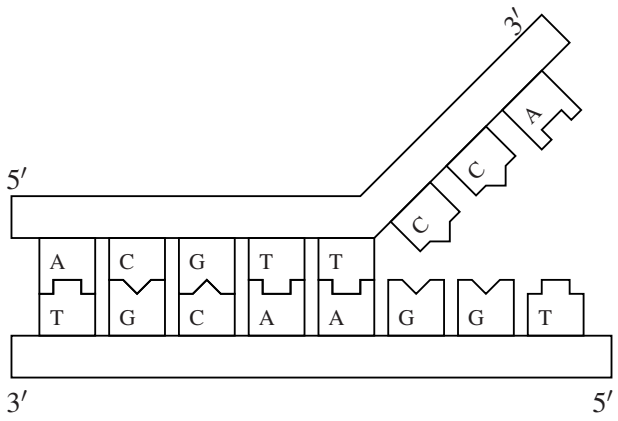

Figure 1.1: Schématisation d'une molécule d'ADN double brins illustrant la complémentarité entre les quatre types de bases azotées.

Dans cette thèse, nous considérons un gène comme une région de l'ADN codant pour un produit fonctionnel (protéine ou ARN). En considérant l'un des deux brins de l'ADN comme référence, un gène peut se trouver dans une seule des deux *orientations transcriptionnelles* possibles : soit il est encodé par ce brin, ce qui sera spécifié par le signe "+" (parfois omis par soucis de concision) ; soit il est encodé par le brin complémentaire, ce qui sera spécifié pas le signe "−". La séquence du brin complémentaire est le *complément inverse* de la séquence de référence. Par soucis de simplicité, nous considérerons que les *régions régulatrices* de l'ADN, qui déterminent dans quel(s) tissu(s) et dans quelle(s) circonstance(s) un gène est exprimé, ne font pas partie du gène lui-même.

1.2 Familles multigéniques

Une *famille multigénique* est un ensemble de gènes ayant évolué par duplications et spéciations à partir d'un gène ancêtral commun. Les familles multigéniques sont impliquées dans une grande variété de processus biologiques, comme la reconnaissance moléculaire (p. ex. les récepteurs olfactifs), le transport moléculaire (p. ex. les globines), ou la régulation de la transcription génique (p. ex. les KRAB-ZNF). Chez les mammifères, on estime à environ 12 000 le nombre de familles multigéniques, et environ

80% de ces familles auraient au moins un représentant chez l'humain [24, 106]. En supposant que le génome humain contienne 22 000 gènes, cela implique que plus d'un gène humain sur deux appartient à l'une de ces familles.

Dans la pratique, les gènes sont souvent regroupés en familles par comparaison de séquences : les paires de gènes ayant un score de similarité supérieur à un certain seuil sont considérées *homologues*. Ainsi, il n'est pas rare de rencontrer les termes "sous-famille" et "superfamille" pour désigner des ensembles de gènes correspondant à différentes valeurs de seuil de similarité. Dans cette thèse, nous utiliserons le terme "famille" de façon générale.

1.3 Organisation spatiale des gènes

Les membres d'une famille multigénique peuvent être dispersés aléatoirement dans le génome, ou regroupés en *clusters*[1], c'est-à-dire en suites de gènes adjacents sur le chromosome. Pour certains clusters, l'organisation spatiale des gènes joue un rôle biologique important et subit une pression de sélection négative, reflétée par un haut degré de conservation entre des espèces parfois très distantes. L'exemple le plus éloquent est celui des gènes partageant des exons (p. ex. UGT1 [131] ; PCDH α et γ [121] ; TRGV [66, 67] ; IGLC [49]). Ce type d'organisation se caractérise par un ensemble d'exons "variables", épissés de façon alternative à un ou plusieurs exons "constants". L'organisation en cluster pourrait aussi permettre la coordination de la transcription d'un groupe de gènes [52, 100]. C'est le cas par exemple des gènes de la famille Hox, impliquée dans le développement de l'axe antéro-postérieur des animaux à symétrie bilatérale [69]. Chez les mammifères, l'ordre de ces gènes correspond approximativement à l'ordre temporel et spatial dans lequel ils sont exprimés durant l'embryogenèse. Plus précisément, les gènes situés à l'extrémité 3' des clusters Hox sont exprimés en premier et participent au développement des structures antérieures, alors que ceux en 5' sont ex-

[1] L'équivalent français est "batterie de gènes", mais son usage est peu répandu. Nous utiliserons donc le terme anglais dans cette thèse.

primés en dernier et participent au développement des structures postérieures [25]. Le fait que cette organisation persiste depuis des centaines de millions d'années suggère qu'elle joue un rôle biologique très important. Cependant, nous devons noter que de nombreuses exceptions ont été observées depuis la découverte des gènes Hox. En particulier, plusieurs non-mammifères possèdent des *clusters* Hox plus ou moins fragmentés, utilisant les deux orientations transcriptionnelles [102, 68, 27]. Cela suggère que différentes combinaisons de processus régulatoires distincts peuvent produire des résultats similaires [58].

Cependant, dans plusieurs cas, le regroupement en cluster ne reflèterait pas des contraintes fonctionnelles entre les gènes, mais simplement un mode d'évolution par duplication locale. Les gènes récemment dupliqués sont alors physiquement rapprochés, mais ils auront tendance à se disperser au fil du temps, suite à divers réarrangements génomiques que nous décrirons plus loin. C'est probablement souvent le cas avec les récepteurs olfactifs et les KRAB-ZNF, qui forment deux des plus imposantes familles multigéniques chez les mammifères, avec plusieurs centaines de membres par espèce. En effet, plusieurs études ont révélé des différences importantes dans le nombre et la disposition de ces gènes dans les génomes, parfois même entre des espèces très rapprochées [42, 124, 83, 105, 47]. Il s'agit toutefois d'une hypothèse qui reste à démontrer.

1.4 Évolution des familles multigéniques

L'évolution des familles multigéniques découle de l'évolution des génomes, et cette évolution est la conséquence d'évènements ponctuels appelés mutations. Il existe une grande variété de mutations et celles-ci affectent l'ADN à différentes échelles, allant de la simple paire de bases (c-.à.-d. une mutation ponctuelle), jusqu'aux longs segments chromosomiques.

Le taux de mutation varie d'une région à l'autre d'un génome et dépend de plusieurs facteurs, comme la présence de séquences répétées ou de motifs sensibles à l'action de

certaines enzymes. Par exemple, un taux de substitution deux fois supérieur à la moyenne a été identifié dans une paire de gènes inversés chez *c. elegans*, et il a été suggéré que la formation d'une structure secondaire en forme de crucifix en était responsable [114].

Lorsqu'une mutation se produit et qu'elle est transmise à la descendance d'un organisme, il y a introduction d'un nouvel allèle (ou d'une variante structurale) au sein de la population. Sa fréquence fluctuera au fil des générations en fonction de son impact sur les taux de reproductivité, ainsi que des caractéristiques de la population. Souvent, le nouvel allèle disparaitra complètement, mais parfois il sera fixé dans la population puis éventuellement dans l'espèce. En l'absence de sélection (évolution neutre), le taux de fixation reflète le taux de mutation à une constante près. En présence de sélection positive (ou adaptative), il se trouve augmenté, tandis qu'en présence de sélection négative (ou purificatrice), il se trouve diminué.

Dans cette thèse, nous nous intéressons surtout aux mutations qui modifient le nombre et l'organisation spatiale des gènes dans les génomes. Ces mutations sont qualifiées de *réarrangements génomiques*. Elles impliquent souvent des échanges de matériel génétique entre deux régions similaires ou identiques de l'ADN, par un mécanisme appelé recombinaison homologue. Pour se produire efficacement chez les eucaryotes, celle-ci requiert deux segments d'identité parfaite d'environ 300 pb [95]. On peut différencier deux types de recombinaisons homologues, selon qu'elles impliquent une seule ou deux positions différentes du génome (c.-à-d. locus). Nous présentons ci-dessous ces deux types de recombinaisons homologues.

1.4.1 Recombinaison Homologue Allélique (AHR)

La Recombinaison Homologue Allélique (AHR) implique un seul locus (Figure 1.2). On la retrouve chez tous les organismes sexués, et elle n'engendre pas de réarrangements génomiques. Cependant, elle joue un rôle évolutif important en permettant de générer différentes combinaisons d'allèles à partir d'un génome diploïde lors de la méiose. L'AHR n'a pas lieu uniformément le long des chromosomes, celle-ci étant plus fréquente

dans les régions appelées *recombination hotspots* [56].

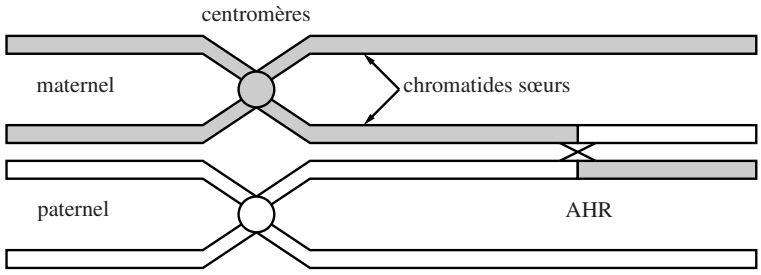

Figure 1.2: Recombinaison homologue allélique (AHR). Le mécanisme implique un échange réciproque d'ADN entre deux chromatides non sœurs d'une paire de chromosome homologues. L'AHR permet de générer des combinaisons d'allèles différentes de celles des parents.

1.4.2 Recombinaison Homologue Non Allélique (NAHR)

La Recombinaison Homologue Non-Allélique (NAHR), aussi appelée recombinaison ectopique ou inégale (*unequal crossingover* en anglais), est une recombinaison homologue impliquant deux locus différents. Elle peut impliquer une seule chromatide (intrachromatide), deux chromatides sœurs (intrachromosomale), ou encore deux chromosomes différents (interchromosomale).

Une grande variété de réarrangements génomiques peut résulter de la NAHR (voir [109] pour une revue). Dans cette thèse, nous nous intéressons aux réarrangements dits "locaux". Pour les illustrer, considérons une séquence d'ADN génomique partitionnée en trois sous-séquences quelconques, $S = ABC$, et notons par $\bar{B}$ le complément inverse de la sous-séquence B. Les réarrangements considérés dans cette thèse sont illustrés sur la sous-séquence B :

- **Duplication en tandem** : $ABC \rightarrow ABBC$.
- **Duplication inversée** : $ABC \rightarrow AB\bar{B}C$ ou $A\bar{B}BC$.
- **Inversion** : $ABC \rightarrow A\bar{B}C$.

- **Délétion** : $ABC \rightarrow AC$.

Chacun de ces évènements laisse une signature dans le génome, et celles-ci peuvent être visualisées à l'aide d'une représentation en *dotplot* (voir Figure 1.3). Certaines signatures sont visibles simplement en comparant la séquence de la région impliquée avec elle-même (p. ex. duplication en tandem, duplication inversée), alors que d'autres nécessitent une comparaison avec une séquence homologue qui n'a pas été affectée par l'évènement (p. ex. inversion, délétion). Ces signatures disparaitront progressivement avec le temps, suite à des mutations et réarrangements additionnels. Au bout de plusieurs millions d'années, seules les régions soumises à une sélection négative présenteront un degré de similarité encore identifiable par les méthodes d'alignement de séquences.

Les répétitions internes d'un génome causées par la NAHR sont appelées duplications segmentales (DS) ou *low copy repeat* (LCR) en anglais.

La NAHR est un phénomène courant qui contribue à la diversité génétique des populations humaines de façon importante [34]. À certains endroits du génome, le taux de NAHR est de 10 à 10 000 fois plus élevé que celui des mutations ponctuelles (substitution d'un nucléotide), causant des réarrangements sporadiques associés à diverses maladies appelées désordres génomiques [70]. Par exemple, le syndrome de Digeorge, causé par une délétion sporadique dans la région 22q11.2, affecte une naissance sur 4 000 [104]. Le principal facteur qui influence le taux de NAHR est la présence de répétitions locales nécessaires à la recombinaison homologue. Chez les humains, les polymorphismes d'inversions et de délétions sont 4 à 12 fois plus fréquents près des régions dupliquées [7]. Ainsi, en causant des répétitions internes, la NAHR se stimule elle-même.

Sur une plus longue période de temps, les réarrangements que nous venons de décrire contribuent également à l'évolution des familles multigéniques. En effet, lorsque les régions dupliquées contiennent des gènes, les nouvelles copies de ces derniers sont libres d'évoluer et peuvent acquérir de nouvelles fonctions. Chez les eucaryotes, le taux moyen

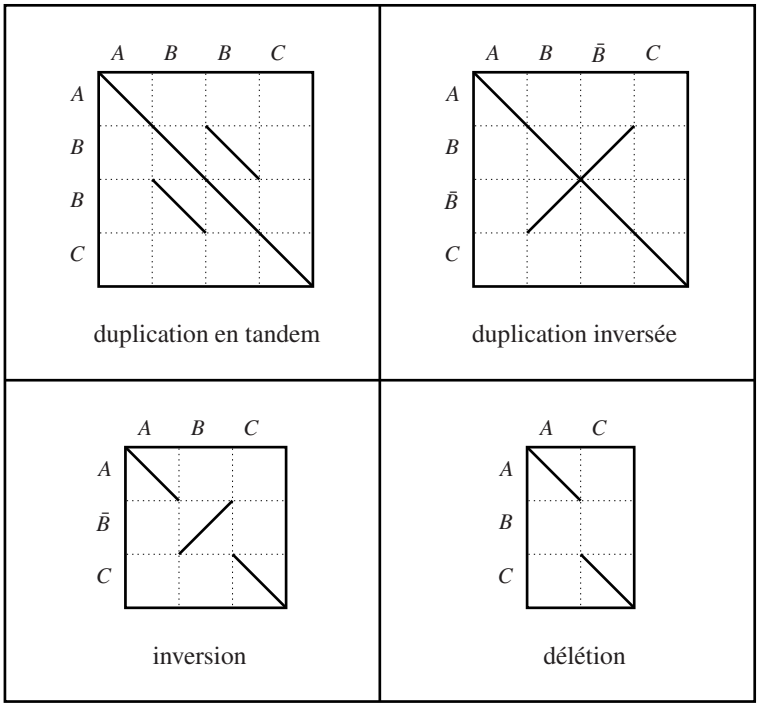

Figure 1.3: Schématisation des signatures laissées par différents réarrangements génomiques résultant de la NAHR. La séquence originale est *ABC*. Chaque trait illustre la similarité entre une région de l'axe horizontal et une autre de l'axe vertical. Les évènements du haut sont illustrés en comparant la région affectée avec elle-même, alors que ceux du bas nécessitent une comparaison avec la séquence d'origine.

de duplication génique est estimé à environ 0.01 par gène par million d'années, une valeur comparable au taux de mutation des nucléotides [71]. Cependant, la majorité des gènes dupliqués seront perdus, soit par inactivation suite à une mutation, soit par élimination d'une partie ou de la totalité de leur séquence. Les vestiges de gènes non fonctionnels sont appelés *pseudogènes*.

Les délétions, en éliminant certains gènes, peuvent également constituer un moyen d'adaptation pour les organismes et les espèces, par exemple en les rendant moins vulnérables à certaines maladies [89, 117]. Quant aux inversions, elles permettent d'associer les gènes à de nouvelles régions régulatrices, modifiant ainsi leurs conditions d'expression [22]. De plus, tous ces évènements peuvent mener à l'apparition de nouveaux gènes,

dits chimériques, par l'exploration de différentes combinaisons d'exons [123].

Bien que les traces laissées par ces évènements s'effacent progressivement avec le temps, certaines sont encore visibles dans plusieurs clusters de gènes. À titre d'exemples, les figures 1.4 et 1.5 présentent respectivement un *dotplot* des clusters Apobec3 et ZNF141 chez l'humain, dont les séquences ont été téléchargées à partir du *UCSC Genome Browser*[2] (hg18). La famille Apobec3 est impliquée dans la réponse immunitaire innée face aux rétrovirus, par le biais d'un mécanisme d'édition des ARN. Dans la lignée des primates, elle a connu une importante expansion par duplications en tandem et diversification [65]. Le cluster du clade ZNF141 contient quant à lui 6 gènes appartenant à la famille KRAB-ZNF [47], qui est impliquée dans la régulation de la transcription génique. En plus de contenir la signature de certaines duplications en tandem, le *dotplot* du clade ZNF141 contient la signature d'inversions et/ou de duplications inversées.

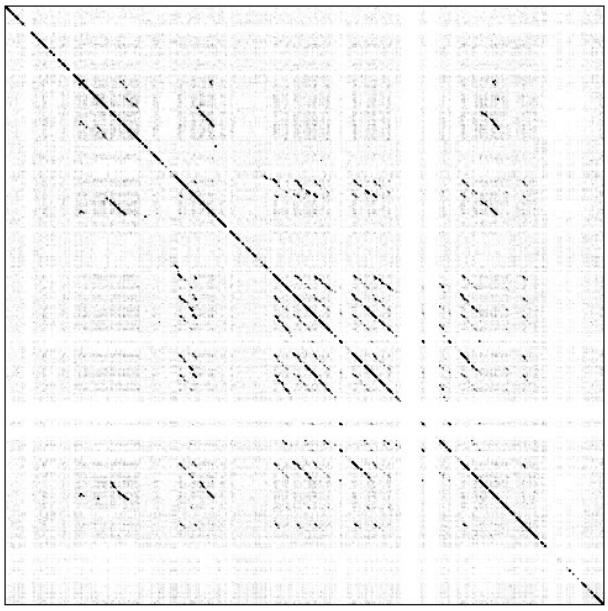

Figure 1.4: Représentation en dotplot des similarités locales dans la séquence du cluster Apobec3 humain (200kpb), obtenue avec le logiciel *Gepard* [59]. La trace des duplications en tandem et des délétions est reconnaissable par la disposition des lignes diagonales (hormis la diagonale principale).

[2] http://genome.ucsc.edu/

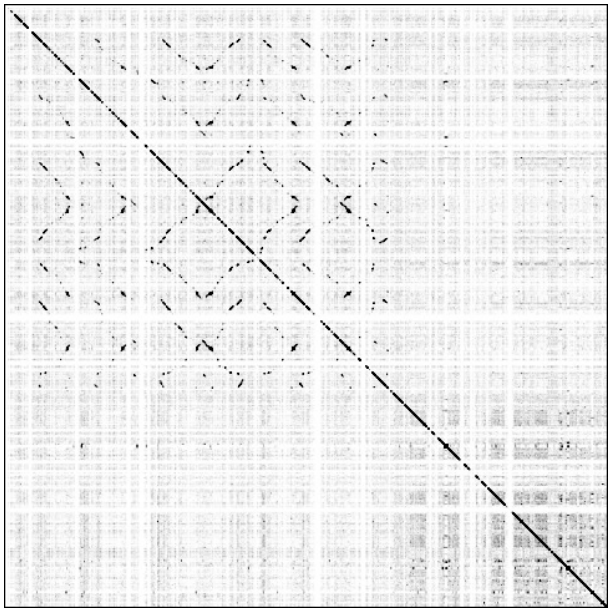

Figure 1.5: Représentation en dotplot des similarités locales dans la séquence du cluster ZNF141 humain (750kbp), obtenue avec le logiciel *Gepard* [59]. On voit clairement la signature de duplications inversées et/ou d'inversions.

1.4.3 Contribution des Éléments Génétiques Mobiles (EGM)

Comme nous l'avons mentionné plus haut, la NAHR est favorisée par la présence de répétitions internes dans l'ADN. Or, chez plusieurs espèces, une fraction importante de ces répétitions provient des Éléments Génétiques Mobiles (EGM), qui sont des fragments d'ADN pouvant s'autorépliquer et s'insérer à différents endroits du génome de leur cellule d'origine [57, 23]. Il existe plusieurs familles d'EGM, et celles-ci ont connu des phases d'expansion et de diversification spécifiques aux différentes lignées d'espèces dont elles ont colonisé le génome. Les EGM constitueraient $\approx 46\%$ du génome humain et $\approx 39\%$ du génome de la souris [120, 64]. Alors que certains EGM encodent directement les protéines nécessaires à leur mobilité, d'autres dépendent entièrement des protéines produites par d'autres EGM. C'est le cas par exemple des séquences Alu, qui font partie de la famille SINE (pour *Short INterspersed Repetitive Element*). Les séquences Alu mesurent environ 300 pb et forment le plus grand groupe d'EGM chez l'humain,

avec plus d'un million de copies (environ 10% du génome) [9]. Une étude a d'ailleurs proposé que l'expansion spécifique de cette famille chez les primates, il y a 35 à 40 millions d'années, est à l'origine de nombreuses duplications dans des régions contenant plusieurs gènes [8].

1.5 Arbres de gènes

Lorsque les gènes sont dupliqués en entier, l'évolution d'une famille multigénique suit une structure arborescente. Les relations ancestrales entre n gènes $g_1, g_2, \ldots, g_n$ de la famille peuvent alors être représentées par un arbre binaire enraciné, que nous appellerons *arbre de gènes*, dont les feuilles sont bijectivement associées à ces n gènes. En particulier, les nœuds internes correspondent à des gènes ancestraux et le nœud racine à l'ancêtre commun de tous les gènes. Les arêtes sont orientées du passé vers le présent et peuvent être valuées en fonction du temps ou du nombre de mutations séparant les différents nœuds.

Il existe de nombreuses méthodes permettant d'inférer un arbre de gènes à partir de séquences nucléiques ou protéiques. Parmi les méthodes couramment utilisées, on retrouve les méthodes de distances [98, 36, 26]. Celles-ci visent à trouver l'arbre dont la distance entre chaque paire de feuilles correspond le mieux aux distances observées entre les gènes (par exemple le nombre de mutations dans un alignement). Elles sont très rapides et pour cette raison encore souvent employées pour étudier de très gros jeux de données. Cependant, ces méthodes présentent l'inconvénient de ne pas utiliser toute l'information disponible dans les séquences et d'autres méthodes sont à privilégier lorsque la taille des données le permet. En particulier, les méthodes probabilistes considèrent l'identité de chaque nucléotide (ou acide aminé) dans les séquences, en plus d'être basées sur des modèles d'évolution explicites. Dans le cas des méthodes du maximum de vraisemblance [46], seul l'arbre le plus vraisemblable est retourné, mais la probabilité qu'il soit exact n'est pas calculée. Les méthodes d'inférence bayésiennes quant à elles

retournent l'ensemble des arbres les plus probables, avec une estimation de leur probabilité postérieure respective [97]. Dans cette thèse, nous utiliserons une méthode d'inférence bayésienne pour la construction des arbres de gènes, afin de pouvoir exploiter la probabilité postérieure de chaque arbre inféré.

Lorsqu'un arbre de gènes est inféré à partir d'un ensemble de gènes appartenant à plusieurs espèces, ses nœuds internes correspondent *implicitement* à des évènements de duplication ou de spéciation. Cependant, lorsque la phylogénie des espèces considérées est connue, il est possible d'établir une correspondance *explicite* entre les nœuds de l'arbre de gènes et les événements évolutifs, en "réconciliant" l'arbre de gènes avec l'arbre des espèces à l'aide d'une méthode appropriée [90, 73, 15, 20]. Cela consiste à "emboîter" l'arbre de gènes dans l'arbre des espèces, et à en déduire une histoire de duplications et de pertes, la plus parcimonieuse possible, permettant d'expliquer la non-congruence éventuelle entre les deux arbres. Dans cette thèse, le thème de la réconciliation est abordé dans l'article constituant le Chapitre 3. Une explication plus détaillée du concept de la réconciliation d'arbres peut être trouvée dans ce chapitre.

1.6 Limites des arbres de gènes

Le modèle de duplication que nous allons introduire à la section suivante, ainsi que les extensions que nous proposons dans cette thèse, suppose que les relations ancestrales à l'intérieur d'une famille multigénique peuvent être représentées à l'aide d'un arbre de gènes. Bien que cela soit habituellement possible, il est important de mentionner qu'il existe des exceptions. En effet, lorsqu'une NAHR se produit à *l'intérieur* d'un gène (NAHR intragénique), les segments de part et d'autre du point de recombinaison sont associés à des arbres de gènes différents. Les conséquences d'un tel évènement sur les méthodes d'inférence phylogénétiques ont été étudiés dans [92] à l'aide de simulations. D'après les résultats, lorsque la NAHR se produit entre deux séquences très similaires (comme cela est généralement le cas), l'arbre inféré correspondra à l'arbre associé à

l'un des deux segments. Le modèle de duplication en tandem pourra donc être utilisé pour décrire correctement l'évolution *du segment correspondant*. Différentes méthodes permettent de détecter les NAHR intragéniques [76, 78, 79].

Même lorsque l'évolution d'une famille peut être représentée par un arbre de gènes, cet arbre ne peut pas toujours être inféré. C'est le cas par exemple des familles évoluant de façon concertée, c'est-à-dire sous l'action répétée des conversions géniques homogénéisant ses séquences. Plus précisément, une conversion génique se produit lorsqu'un brin d'ADN se substitue au brin complémentaire d'un autre segment similaire, et que la machinerie de réparation de l'ADN corrige les mésappariements en se servant d'un des deux brins comme matrice. Comme les conversions géniques ne se produisent qu'entre les gènes d'une même espèce, le nombre de mutations apparentes sera plus bas entre les gènes d'une même espèce qu'entre ceux d'espèces différentes. En conséquence, les méthodes d'inférence phylogénétique auront tendance à retourner des arbres inexacts pour ces familles (voir Figure 1.6). Heureusement, il semblerait que la majorité des familles multigéniques n'évoluent pas de façon concertée, et que les conversions géniques sont souvent trop peu nombreuses pour obscurcir complètement le signal phylogénétique (voir [81] pour une revue).

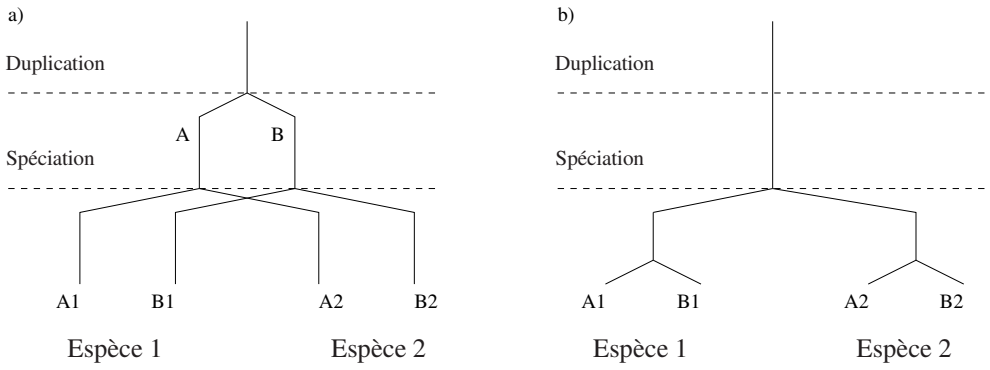

Figure 1.6: Conséquence de l'évolution concertée sur les arbres de gènes inférés. À gauche, l'arbre représentant les relations phylogénétiques de 4 gènes provenant d'une duplication suivie d'une spéciation. À droite, l'arbre qui pourrait être inféré à partir des séquences en cas d'évolution concertée.

1.7 Modèle de duplication

Dans cette section, nous présentons le modèle de duplication en tandem pour lequel nous proposons différentes extensions dans cette thèse. Ce modèle fut introduit par Fitch [35], en 1977, pour étudier l'évolution d'un ensemble de gènes (ou séquences) généré par une suite de duplications en tandem, à partir d'un unique gène ancestral.

Contrairement à l'inférence phylogénétique classique, basée uniquement sur les séquences nucléiques ou protéiques des gènes, le modèle de duplication en tandem a la particularité de considérer l'ordre de ces derniers sur le chromosome. Cette information, jumelée à la contrainte concernant le mécanisme de duplication, a deux conséquences importantes. La première est qu'elle restreint l'ensemble des arbres de gènes pouvant représenter l'évolution d'un cluster donné. En effet, si l'on tient compte de l'ordre des gènes sur le chromosome, seule une faible proportion des arbres de gènes est compatible avec le modèle. La deuxième est qu'elle permet d'identifier explicitement les évènements de duplication. Rappelons-nous qu'il n'y a pas de correspondance bijective entre les nœuds d'un arbre de gènes et les évènements de duplication.

À la base du modèle de duplication en tandem introduit par Fitch se trouvent donc les évènements de duplication, impliquant des segments chromosomiques contenant un gène (duplication simple) ou plusieurs gènes (duplication multiple). Une duplication en tandem a pour effet de placer le segment dupliqué adjacent au segment d'origine, dans la même orientation que celui-ci (voir Figure 1.7). L'évolution d'un cluster de GRT débute

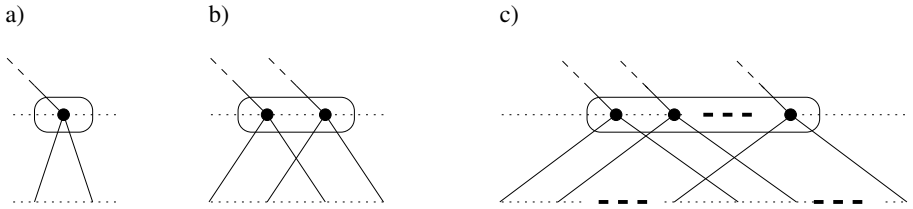

Figure 1.7: Évènements de duplication en tandem. a) Duplication simple. b) Duplication double. c) Duplication de $n \geq 3$ gènes.

avec un unique gène ancestral et se poursuit par une séquence de duplications en tandem appelée *histoire* de duplication. L'arbre de gènes résultant, avec l'ordre sur ses gènes, est appelé *arbre de duplication*. Même s'il peut exister différentes histoires de duplication menant à un même arbre de duplication, chaque arbre de duplication admet une unique partition de l'ensemble de ses nœuds internes en évènements de duplication (voir Figure 1.8). Autrement dit, les évènements de duplication sont *partiellement ordonnés*.

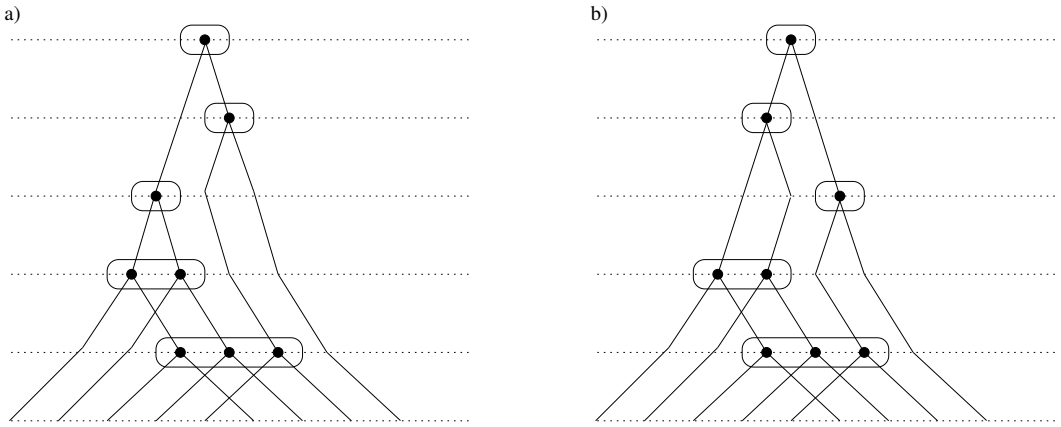

Figure 1.8: Deux histoires de duplication menant au même arbre de duplication. On remarque que la partition des nœuds internes en événements de duplication est identique pour les deux arbres de gènes ordonnés.

Nous allons maintenant présenter le modèle de façon formelle. Un cluster de GRT est représenté par un arbre de gènes ordonné, dénoté (T,O), où T est un arbre de gènes (binaire et enraciné) représentant les relations ancestrales entre les gènes, et $O=(v_1,\ldots,v_n)$ est une permutation des feuilles de T correspondant à l'ordre des gènes sur le chromosome. Une *cerise* de T est une paire de feuilles (g,d) séparée par un unique nœud appelé sa *racine*.

Définition 1.1. *Soit (T,O) un arbre de gènes ordonné. Une duplication en tandem a pour effet de remplacer une sous-séquence $(v_i,v_{i+1},\ldots,v_j)$ de O par une séquence de nouveaux éléments $(g_i,g_{i+1},\ldots,g_j,d_i,d_{i+1},\ldots,d_j)$. De plus, chaque feuille de T étiquettée par v_x, pour $i\leq x\leq j$, est substituée par la cerise (g_x,d_x).*

Définition 1.2. *Une histoire de duplication en tandem est une suite d'arbres de gènes ordonnés* $\mathscr{H} = ((T_1, O_1), (T_2, 0_2), \ldots, (T_n, O_n))$ *telle que :*

1. $T_1 = v$ *est un arbre constitué d'une unique feuille et* $O_1 = (v)$.
2. *Pour* $1 \leq k < n$, (T_{k+1}, O_{k+1}) *peut être obtenu en effectuant une duplication en tandem sur* (T_k, O_k)

Définition 1.3. *Un arbre de gènes ordonné* (T, O) *est un* arbre de duplication *si et seulement si il existe une histoire de duplication* $\mathscr{H} = ((T_1, O_1), (T_2, 0_2), \ldots, (T_n, O_n))$ *telle que* $(T_n, O_n) = (T, O)$.

1.7.1 Algorithmes de reconnaissance

Lorsqu'un arbre de gènes T est disponible pour un cluster de GRT avec un ordre O, différents algorithmes permettent de vérifier si celui-ci admet une histoire de duplication valide (c.-à-d. vérifier si (T, O) est un arbre de duplication). Le plus simple est l'algorithme récursif suivant [33] :

Algorithme *PossibleDuplicationHistory* (arbre T, ordre O)

1. Si T contient une unique feuille, retourner VRAI.
2. S'il existe une sous-séquence $(g_i, g_{i+1}, \ldots, g_j, d_i, d_{i+1}, \ldots, d_j)$ dans O, tel que (g_x, d_x) est une cerise de T pour tout $i \leq x \leq j$:
   (a) Remplacer cette séquence par $(v_i, v_{i+1}, \ldots, v_j)$ dans O, où v_x est la racine de (g_x, d_x) pour $i \leq x \leq j$.
   (b) Éliminer g_x et d_x dans T, pour $i \leq x \leq j$.
   (c) Retourner *PossibleDuplicationHistory*(T, O).
3. Sinon, retourner FAUX.

Pour un arbre ordonné (T, O) contenant n gènes, l'étape 2 de cet algorithme peut s'effectuer naïvement en parcourant l'ordre de gauche à droite, ce qui nécessite $O(n)$

opérations. Étant donné qu'au minimum une cerise est éliminée à chaque parcours, cette étape est répétée au plus n fois, pour une complexité globale en $O(n^2)$. Deux algorithmes de reconnaissance dont la complexité globale est linéaire ont également été proposés. L'un repose sur une approche descendante (de la racine vers les feuilles) [130], l'autre suit une approche d'agglomération ascendante semblable à l'algorithme décrit ci-haut, à la différence qu'il utilise une structure de données additionnelle pour ne pas répéter inutilement les mêmes comparaisons à l'étape 2 [38].

1.7.2 Propriétés des arbres de duplication

Différentes études ont porté sur les propriétés algorithmiques et combinatoires des arbres de duplication (voir [96] pour une revue). En particulier, des récurrences permettant de calculer le nombre exact d'arbres et d'histoires de duplication d'une certaine taille ont été proposées [38, 122, 32]. Puisque le nombre d'arbres de gènes à n feuilles peut facilement être obtenu, ces récurrences permettent de calculer la probabilité qu'un arbre de gènes ordonné quelconque soit un arbre de duplication. Cette probabilité devient rapidement très petite lorsque le nombre de gènes augmente. Une autre étude s'est intéressée au problème de l'exploration de l'espace des arbres de duplication en tandem, à l'aide de réarrangements topologiques [12]. En particulier, il a été démontré que les *nearest-neighbor-interchange* (NNI) ne sont pas suffisants pour explorer l'espace en entier, mais qu'un type restreint de *Subtree Pruning And Regrafting* le permet [12].

1.7.3 Algorithmes d'inférence

Pour un ensemble de GRT donné, une méthode classique d'inférence phylogénétique ne donne pas nécessairement lieu à un arbre de duplication. C'est la raison pour laquelle plusieurs méthodes d'inférence spécifiques aux GRT ont été développées dans le but d'inférer le meilleur arbre de duplication en tandem, selon différents critères (p. ex. distance, parcimonie) pour un cluster donné [10, 112, 31, 12]. Cependant, il a été démontré que l'inférence d'un arbre de duplication simple selon un critère de parci-

monie est un problème NP-Difficile [53]. De ce fait, la plupart des méthodes d'inférence existantes sont des heuristiques. Parmi celles-ci, on retrouve des méthodes d'agglomération gloutonnes, semblables à l'algorithme de Neighbor-Joining [98], utilisant un critère de parcimonie [10] ou une matrice de distance [112, 31]. On retrouve également des méthodes de recherche locale, qui débutent avec un arbre de duplication quelconque, puis explorent son voisinage en effectuant des réarrangements topologiques restreints à l'espace des arbres de duplication [12]. Des schémas d'approximation en temps polynomiaux ont également été proposés [111, 53]. Ceux-ci retournent une solution qui est dans le pire des cas $(1+\varepsilon)$ fois plus coûteuse que la solution optimale, mais leur temps d'exécution est généralement exponentiel en $1/\varepsilon$. Ces algorithmes sont intéressants d'un point de vue théorique, mais ils sont surpassés par les heuristiques en pratique. Finalement, il existe un algorithme polynomial exact (temps $O(n^3)$ et espace $O(n^2)$) permettant d'inférer un arbre de duplication simple selon le critère d'évolution minimum [32]. Ce critère utilise uniquement la matrice des distances entre les séquences et par conséquent est moins précis que les méthodes de parcimonie ou de maximum de vraisemblance. La solution retournée est l'arbre de duplication le plus court parmi l'ensemble des arbres de duplication existants.

1.8 Limites du modèle de duplication en tandem

Comme nous l'avons mentionné précédemment, suite à des inversions ou à des duplications inversées, la plupart des familles multigéniques contiennent des gènes dans les deux orientations transcriptionnelles. Le modèle de duplication en tandem ne peut être appliqué à de telles familles. De plus, les délétions entrainant des pertes de gènes sont relativement courantes, et le fait qu'elles ne soient pas prises en compte par le modèle constitue une limitation supplémentaire importante. Finalement, même lorsqu'un cluster a évolué selon les contraintes du modèle, les algorithmes de reconnaissance peuvent ne pas retourner d'histoire de duplication lorsque l'arbre de gènes utilisé diffère de l'arbre

véritable. En effet, la majorité des arbres de gènes ordonnés n'admet pas d'histoire de duplication. Afin d'augmenter le réalisme du modèle et de permettre de considérer un plus large éventail d'arbres, il est donc important d'étendre le modèle en considérant, en plus des duplications en tandem, d'autres évènements évolutifs, comme les inversions et les délétions. Pour des arbres qui diffèrent peu de l'arbre véritable, les évènements correctement inférés seront prédominants et pourraient permettre de vérifier certaines hypothèses concernant l'évolution du cluster considéré. L'importance d'étendre le modèle de duplication en tandem à d'autres évènements évolutifs comme les délétions et les inversions a d'ailleurs été souligné dans plusieurs publications, en particulier [37, 130].

CHAPITRE 2

DUPLICATION AND INVERSION HISTORY OF A TANDEMLY REPEATED GENE FAMILY

Mathieu Lajoie[1], Denis Bertrand[2], Nadia El-Mabrouk[3] et Olivier Gascuel[4]

Article publié dans *Journal of Computational Biology* [62]

[1]DIRO, Université de Montréal, Montreal (QC) Canada
[2]DIRO, Université de Montréal, Montreal (QC), Canada.
[3]DIRO, Université de Montréal, Montreal (QC), Canada.
[4]LIRMM, CNRS et Université Montpellier 2, Montpellier France.

Abstract

Given a phylogenetic tree for a family of tandemly repeated genes and their *signed* order on the chromosome, we aim to find the minimum number of inversions compatible with an evolutionary history of this family. This is the first attempt to account for inversions in an evolutionary model of tandemly repeated genes. We present a branch-and-bound algorithm that finds the exact solution, and a polynomial-time heuristic based on the breakpoint distance. We show, on simulated data, that those algorithms can be used to improve phylogenetic inference of tandemly repeated gene families. An application on a published phylogeny of KRAB zinc finger genes is presented.

2.1 Introduction

A large fraction of most genomes consists of repetitive DNA sequences. In mammals, up to 60% of the DNA is repetitive. A large proportion of such repetitive sequences is organized in tandem: copies of a same basic unit that are adjacent on the chromosome. The duplicated units can be small (from 10 to 200 bps) as it is the case of micro- and minisatellites, or very large (from 1 to 300 kb) and potentially contain several genes. The formation of those large duplicated sequences is widely assumed to be due to unequal recombination.

Many gene families are organized in tandem, including HOX genes [128], immunoglobulin and T-cell receptor genes [3], MHC genes [39] and olfactory receptor genes [41]. Reconstructing the duplication history of each gene family is important to understand the functional specificity of each copy, and to provide new insights into the mechanisms and determinants of gene duplication, often recognized as major generators of novelty at the genome level [86].

Both the linear order among tandemly repeated sequences, and the knowledge of the biological mechanisms responsible for their generation, suggest a simple model of evolution by duplication. This model, first described by Fitch [35], introduces tandem

duplication trees as phylogenies constrained by the unequal recombination mechanism. The main features of this model are illustrated by the examples given Figure 2.1.

Figure 2.1(a) shows the duplication tree of the 13 Antennapedia-class homeobox genes [128] which contains only simple duplication events (duplication of a segment containing only one gene). Starting from the unique ancestral gene, this series of events has produced the extant locus containing the 13 linearly ordered contemporary genes. As described by [32], trees that contain only simple duplication events are equivalent to binary search trees with labeled leaves. The Fitch model also allows for the simultaneous duplication of several gene copies, as observed in the duplication tree of the 9 variable genes of the human T cell receptor Gamma (TRGV) locus [33] (see Figure 2.1(b)). This duplication tree contains a double duplication where two adjacent genes have been simultaneously duplicated.

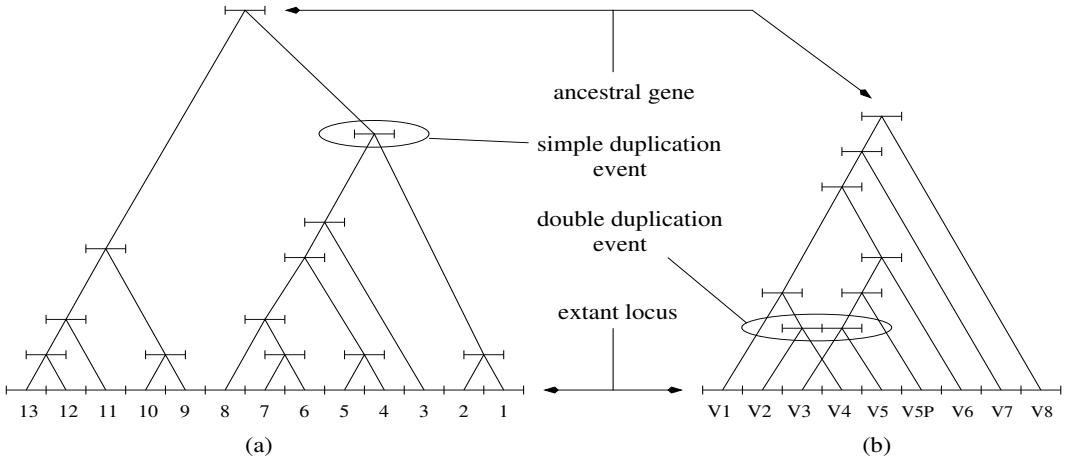

Figure 2.1: (a) Simple rooted duplication tree of the 13 Antennapedia-class homeobox genes from the cognate group [128]. (b) Rooted duplication tree of the 9 variable genes of the human T cell receptor Gamma (TRGV) locus [33]. In both examples, the contemporary genes are adjacent and linearly ordered along the extant locus.

Based on this model, a number of recent studies have considered the problem of reconstructing the tandem duplication tree of a gene family [10, 111, 33, 31, 53, 130, 12, 32]. These are essentially phylogenetic inference methods which compute the duplication tree that best explains the evolution of a gene family. When a phylogeny is

already available, a linear-time algorithm can be used to check whether it is a duplication tree [38, 130]. However, even for gene families that have evolved through tandem duplications, it is often impossible to reconstruct a duplication history [37]. This can be explained by the fact that the duplication model is oversimplified, and other evolutionary events have occurred, such as gene losses or genomic rearrangements.

Evidence of gene inversion is observed in many tandemly repeated gene families, such as zinc finger (ZNF) genes, where gene copies have different transcriptional orientations [105]. Although genome rearrangement with inversions has received great attention in the last decade [48, 30, 55, 107, 11], inversions have never been considered in the context of reconstructing a duplication history from a gene tree. In the case of general segmental duplications (not necessarily in tandem), potential gene losses have been considered to explain the incongruence between a gene tree and a species tree [45, 91, 72, 20]. Similarly, in the case of tandem duplication, the incongruence between a gene tree and an observed gene order can be naturally explained by introducing the possibility of segmental inversions.

In this paper, our goal is to infer an evolutionary history of a gene family accounting for both tandem duplications and inversions. As the number of such possible evolutionary histories can be very large, we restrict ourselves to finding the minimum number of inversions required to explain a given ordered phylogeny. The Fitch model allows for the simultaneous duplication of several gene copies, but there are now evidences that simple duplications are predominant over multiple duplications [128, 12]. As a first attempt, we only consider simple duplications.

After describing the evolutionary models in Section 2.2 and the optimization problem in Section 2.3, we present a branch-and-bound algorithm in Section 2.4. Then, in Section 2.5, we present a similar problem based on the breakpoint distance. This variant has a polynomial-time solution and can be used as an accurate heuristic to solve our original problem. Finally, in Section 2.6, we compare the time efficiencies of the two algorithms and show, using simulated data, their usefulness to improve phylogenetic inference. An

application on a KRAB zinc finger gene family is presented.

2.2 The evolutionary model

2.2.1 Duplication model

This model, first introduced by [35], is based on unequal recombination during meiosis. The later is assumed to be the sole evolutionary mechanism, with point mutations, acting on sequences. However, the model is robust to gene conversion, as long as the phylogenetic signal remains strong enough to reconstruct the correct tree, which seems a realistic assumption for many tandemly repeated gene families. Indeed, from a single sequence, the locus grows through a series of consecutive duplications, giving rise to a sequence of n adjacent copies of homologous genes *having the same transcriptional orientation*. We denote by $O = (l_1, \cdots l_n)$ the observed ordered sequence of extant gene copies.

A *tandem duplication history* (or just *duplication history* for brevity) is the sequence of tandem duplications that have generated O. It can be represented by a rooted tree with n ordered leaves corresponding to the n ordered genes, in which internal nodes correspond to duplication events (Figure 2.2(a)). Duplications may be *simple* (duplication of a single gene) or multiple (simultaneous duplication of neighboring genes). In our duplication/inversion model, we consider only simple duplications. As mentioned previously, simple duplications seems to be predominant over multiple duplications [128, 12], and there are examples of simple duplication trees in the literature, such as the one presented in Figure 2.1.

In a real duplication history, the time intervals between consecutive duplications are known, and the internal nodes are ordered from top to bottom according to the moment they occurred in the course of evolution. However, in the absence of a molecular clock mode of evolution, it is impossible to recover the order of duplication events. All we can infer from gene sequences is a phylogeny with ordered leaves (Figure 2.2(c)). Formally,

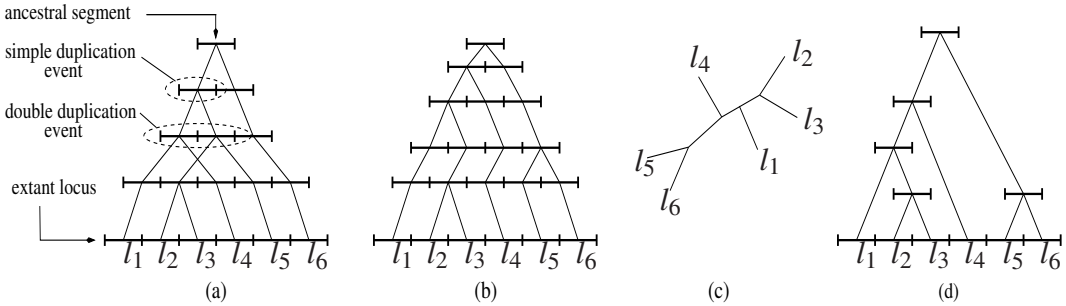

Figure 2.2: (a) Duplication history; each segment represents a copy. (b) Simple duplication history. (c) The unrooted simple duplication tree corresponding to history (b). (d) A simple rooted duplication tree corresponding to history (b).

an *ordered phylogeny* is a pair (T,O) where T is a phylogeny and O is the ordered sequence of its leaves. If an ordered phylogeny (T,O) can be explained by a duplication history $\mathscr{H}$, we say that (T,O) is *compatible* with $\mathscr{H}$, and that $\mathscr{H}$ *is a duplication history of* (T,O). If (T,O) is compatible with at least one duplication history, it is called a *duplication tree*. Choosing appropriate roots for unrooted duplication trees is discussed in [37].

In the rest of this paper, a *duplication tree* will refer to a *simple rooted duplication tree*, that is a rooted duplication tree compatible with at least one history involving only simple duplications (see Figure 2.2(d)). Unless otherwise stated, all the phylogenies are rooted.

2.2.2 A duplication/inversion model

Many tandemly repeated gene families contain members in both transcriptional orientations. The actual duplication model is thus inadequate to describe their evolution. To circumvent this limitation, we propose an extended model of duplication which includes inversions. Hereafter, the transcriptional orientations of the genes in a *signed* ordered phylogeny (T,O) are specified by signs $(+/-)$ in O. We denote by $d_{inv}(O_i,O_j)$ the inversion distance between the two signed orders O_i and O_j. Note that a signed ordered

phylogeny (T,O) cannot be a duplication tree unless all the genes in O have the same sign (although this is not a sufficient condition).

Definition 2.1. *A* simple duplication/inversion history *(or just* dup/inv history*) of length k is an ordered sequence $\mathscr{H}_k = ((T_1,O_1),...,(T_{k-1},O_{k-1}),(T_k,O_k))$ where :*

1. *Every (T_i,O_i) is a* signed *ordered phylogeny.*
2. *$T_1 = v$ is a single leaf phylogeny and $O_1 = (\pm v)$.*
3. *For $0 < i < k$,*
   - *if $T_{i+1} = T_i$, then $d_{inv}(O_i,O_{i+1}) = 1$. This corresponds to one inversion event.*
   - *if $T_{i+1} \neq T_i$, then T_{i+1} is obtained from T_i by adding two children u and w to one of its leaves v, and O_{i+1} is obtained from O_i by replacing v by (u,w), where u and w have the same sign as v. This corresponds to a simple duplication event.*

2.3 An inference problem

A signed ordered phylogeny is not necessarily compatible with a duplication history. The following lemma shows that additional inversions can always be used to infer a possible evolutionary history for the gene family.

Lemma 2.2. *A signed ordered phylogeny (T,O) is compatible with at least one simple duplication/inversion history.*

Proof. According to Definition 2.1, obtain a duplication tree (T,O') by successive duplication events. Then, transform O' into O by applying the required inversions. ■

As the number of possible dup/inv histories explaining (T,O) can be very large, we restrict ourselves to finding the minimum number of events involved in such evolutionary histories. More precisely, as the number of simple duplications is fixed by T, we are

interested in finding the minimum number of inversions involved in a dup/inv history. This parsimony approach relies on the biological assumption that the quantified events are scarce, which seems to be the case with genomic inversions. The next theorem shows that if i is the minimum number of inversions needed to transform O into O' such that (T,O') is a duplication tree, any dup/inv history of (T,O) contains at least i inversions.

Theorem 2.3. *Let (T,O) be a signed ordered phylogeny. For any dup/inv history $\mathscr{H}$ with i inversions leading to (T,O), there exists a duplication tree (T,O') such that $d_{inv}(O,O') \leq i$.*

Proof.

- Base case: Let $\mathscr{H}_1 = (T_1, O_1)$ be a dup/inv history with no duplication or inversion. Clearly $(T,O') = (T_1,O_1)$ is a duplication tree.

- Induction step (on the number k of events):
  Let $\mathscr{H}_{k+1} = ((T_1,O_1),...,(T_k,O_k),(T_{k+1},O_{k+1}))$ be a dup/inv history involving $k+1$ events and i inversions, and $\mathscr{H}_k = ((T_1,O_1),...,(T_k,O_k))$. From Definition 2.1, there are two possibilities:

  - If $T_{k+1} = T_k$, then the last event is an inversion, and $\mathscr{H}_k$ is a dup/inv history involving $i-1$ inversions. By induction hypothesis, there exists a duplication tree (T_k,O'_k) such that $d_{inv}(O_k,O'_k) \leq i-1$. Let O_{k+1} be the order obtained from O_k by applying the last inversion. Then we have $d_{inv}(O_{k+1},O'_k) \leq d_{inv}(O_k,O'_k)+1 \leq i$.
  - If $T_{k+1} \neq T_k$, the last event is a duplication, that is a leaf v of (T_k,O_k) is replaced by two consecutive leaves (u,w) in (T_{k+1},O_{k+1}). Let (T_k,O'_k) be the duplication tree associated to $\mathscr{H}_k$ and suppose that all elements of O'_k are positive. If v has positive sign in O_k, we obtain O'_{k+1} by replacing v in O'_k by (u,w). Otherwise, v has negative sign in O_k and we obtain O'_{k+1} by replacing

v in O'_k by (w,u). Thus, $d_{inv}(O_{k+1},O'_{k+1}) = d_{inv}(O_k,O'_k) \leq i$ and (T_{k+1},O'_{k+1}) is a duplication tree. The case where the elements of O'_k have a negative sign is similar.

■

Corollary 2.4. *Let (T,O) be a signed ordered phylogeny and (T,O') a duplication tree such that $d_{inv}(O,O') = i$ is minimum. There exists a dup/inv history $\mathscr{H}$ for (T,O) with exactly i inversions, which is optimal.*

Proof. The existence of $\mathscr{H}$ for (T,O) with exactly i inversions follows directly from the proof of Lemma 2.2. The number i of inversions in $\mathscr{H}$ must be optimal, otherwise, from Theorem 2.3, it would contradict the hypothesis that $d_{inv}(O,O') = i$ is minimum. ■

Corollary 2.4 allows to reformulate our problem in the following way :

MINIMUM-INVERSION DUPLICATION PROBLEM
Input: A signed ordered phylogeny (T,O),
Output: An order O' such that (T,O') is a duplication tree and $d_{inv}(O,O')$ is minimal.

2.4 A Branch-and-Bound algorithm

We begin by briefly summarizing the Hannenhalli-Pevzner method [48], as it will be used in our approach.

2.4.1 Hannenhalli-Pevzner (HP) algorithm

Given two signed orders O,O' of size n on the same set of genes, the problem is to find the minimal number $d_{inv}(O,O')$ of inversions required to transform O to O' (or similarly O' to O). As the orders considered in this paper do not represent a whole chromosome, but rather a cluster of tandemly repeated genes, we can always consider them as linear (not circular), with a leftmost and rightmost gene. The algorithm is based on a

bicolored graph, called the *breakpoint graph*, constructed from the two signed orders as follows: if gene x of O has a positive sign, replace it by the pair $x_t x_h$, and if it is negative, by $x_h x_t$. Then the vertices of the graph are just the x_t and the x_h for all genes x plus two additional vertices, α and β, which represent the two extremities of the order. The graph contains two classes of edges: the real and desired edges [103]. Any two vertices which are adjacent in O, other than x_t and x_h deriving from the same x, are connected by a *real edge*, and any two adjacent in O', by a *desired edge* (see Figure 2.3(c)).

This graph decomposes naturally into a set of c disjoint color-alternating cycles. An important property of the graph is its decomposition into components, where a *component* is a maximal set of "crossing" cycles.

Based on this graph, the inversion distance can be computed according to the following formula [48]:

$$d_{inv}(O,O') = n + 1 - c + h + f,$$

where h and f are quantities related to the presence of "hurdles" (components of a particular type). As the probability for a component to be a hurdle is low, h and f are usually close to 0. Therefore, the number of cycles c is the dominant parameter in the formula. In other words, the more cycles there are, the less inversions we need to transform O into O'. For example in Figure 2.3(c), $n = 4$, $c = 3$, $h = 0$ and $f = 0$, which leads to $d_{inv}(O,O') = 2$.

2.4.2 Enumerating the compatible orders

We say that an order O' is *compatible* with a phylogeny T iff (T,O') is a duplication tree. As mentioned in the introduction, the considered duplication trees are equivalent to binary search trees. Therefore, to enumerate all the orders compatible with T, we associate a binary variable b_i to each internal node i of T. Each b_i defines an order relation between the left and right descendant leaves of i. By setting b_i to 0, we make

all the left descendants smaller than the right ones. Conversely, by setting b_i to 1, all left descendants are considered larger than the right ones (see Figure 2.3(a)(b)). Assigning a value to all internal nodes of T defines a total order O' on its leaves: the order between two leaves is determined by the b_i value of their closest common ancestor. Otherwise, the order is partial since some pairs of leaves are incomparable. We will denote such a partial order as O^*. Note that every order admits two transcriptional orientations according to our definition of a duplication tree. Therefore, if n is the number of leaves in T, there are 2^{n-1} possible assignments of the b_i variables, each with two possible transcriptional orientations. This leads to 2^n distinct orders O' compatible with T. Hereafter, for clarity of presentation and w.l.o.g., we will only consider the positive orientations for O'.

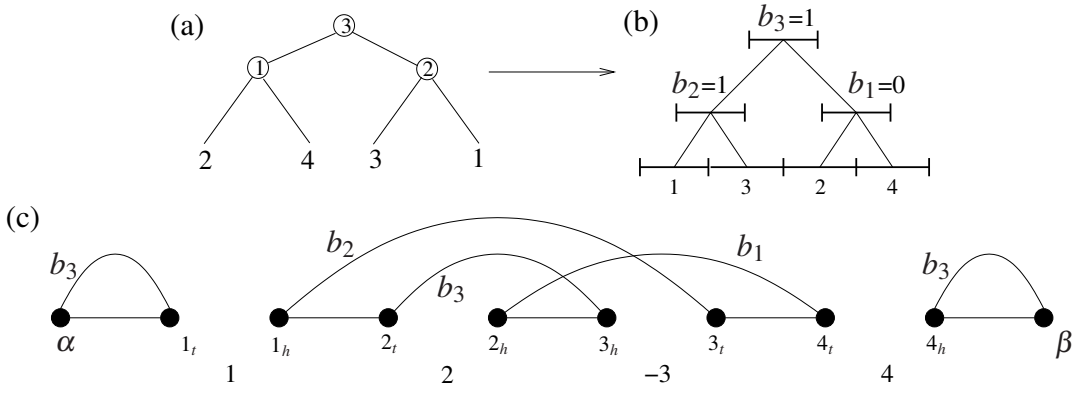

Figure 2.3: (a) A phylogeny with an appropriate post-order labeling of its internal nodes; (b) The duplication tree corresponding to an assignment of the b_i variables of (a); (c) The breakpoint graph illustrating the difference between the gene order $O' = (1,3,2,4)$ obtained from the duplication tree (b) and the gene order $O = (1,2,-3,4)$ observed in the genome. *Desired edges* (curved edges) are added in the same order as the corresponding b_i values (b_1 then b_2 then b_3).

2.4.3 A lower bound for the inversion distance

To avoid computing $d_{inv}(O,O')$ for each of the 2^{n-1} orders O' compatible with T, we consider a branch-and-bound strategy similar to the one used by [134]. The idea is to compute a lower bound on $d_{inv}(O,O')$ as we progressively define O^* by updating the partial breakpoint graph of (O,O^*). In order to progressively construct this graph, it is

essential to define the b_i values in a post-order traversal of T: the binary variables of all the descendant nodes of i should be defined before b_i. This insures that the two subtrees of i have a total order on their leaves.

Consequently, if we set b_i to 0, the greatest left descendant leaf l_{max} of node i will immediately precede its smallest right descendant leaf r_{min} in O'. Conversely, if b_i is set to 1, the greatest right descendant r_{max} will immediately precede the smallest left descendant l_{min}. Therefore, the assignment of a b_i value allows us to add a desired edge in the partial breakpoint graph between l_{max} and r_{min} (or r_{max} and l_{min}) (see Figure 2.3(c)).

Let O^* be the partial order obtained at a given stage of the procedure. Let e be the number of cycles and p the number of paths of the corresponding *partial breakpoint graph*. The remaining desired edges can create at most p cycles, ending with a breakpoint graph with at most $c = e + p$ cycles. Thus, any total order O' that can be obtained from the partial order O^* is such that:

$$d_{inv}(O,O') = n+1-c+h+f \geq n+1-c \geq n+1-p-e = d^*_{inv}(O,O^*).$$

2.4.4 Algorithm

The branch-and-bound algorithm proceeds as follows (see Algorithm 1). Denote O' the best order obtained so far at a given step and $\min_{inv} = d_{inv}(O,O')$ the corresponding inversion distance. Each following step assigns the values of the binary variables in a post-order traversal of T that progressively defines a partial order O^*. This procedure stops and backtracks when the current partial order O^* is such that $d^*_{inv}(O,O^*) > \min_{inv}$. This is justified by the fact that any total order that can be obtained from O^* cannot lead to a smaller inversion distance. If no bound were used, the assignment procedure would explore all the 2^{n-1} possible configurations of the binary variables. Finally, every time a total order is reached, the inversion distance is computed using the HP algorithm and $\min_{inv}$ and O' are updated, if necessary.

The efficiency of a branch-and-bound algorithm is usually correlated with its initial solution. Here, we use the initial order O' obtained with the polynomial-time algorithm described in the next section.

Algorithm 1: Branch-and-bound

Data: A signed ordered phylogeny (T, O) with n leaves.
Result: An order O' such that (T, O') is a duplication tree and $d_{inv}(O, O')$ is minimal.
begin
  O' is the initial order obtained with the polynomial-time algorithm (c.f. Section 2.5.2)
  $min_{inv} \leftarrow d_{inv}(O, O')$
  O^* is an empty partial order, and PBPG(O, O^*) the corresponding partial breakpoint graph
  Label the $n-1$ internal nodes of T according to a post-order traversal ($i < j$ if node i is a descendant of node j)
  Associate a binary variable b_i to each internal node i of T
  Add a positive sign to each leaf of T
  RECURSIVE_EXPLORE(1)
  Add a negative sign to each leaf of T
  Reset O^* and PBPG(O, O^*)
  RECURSIVE_EXPLORE(1)
  return O'
end

2.5 Minimizing the breakpoint distance

2.5.1 The minimum breakpoint duplication problem

Genome rearrangement mechanisms such as inversions cannot be observed directly from the data and can only be inferred from different theoretical probabilistic, algorithmic or phylogenetic methods. Evidence for the occurrence of such mechanisms during

Procedure RECURSIVE_EXPLORE(*integer i*)

```
begin
    if i = n − 1 then
        if d_inv(O, O*) < min_inv then
            O' ← O*
            min_inv ← d_inv(O, O*)
        end
    else
        b_i ← 0
        Add adjacency (l_max, r_min) in PBPG(O, O*)
        if d*_inv(O, O*) < min_inv then
            RECURSIVE_EXPLORE(i + 1)
        end
        Remove adjacency (l_max, r_min) in PBPG(O, O*)
        b_i ← 1
        Add adjacency (r_max, l_min) in PBPG(O, O*)
        if d*_inv(O, O*) < min_inv then
            RECURSIVE_EXPLORE(i + 1)
        end
        Remove adjacency (r_max, l_min) in PBPG(O, O*)
    end
end
```

Where l_{min} and l_{max} are respectively the smallest and greatest left descendant leaf of node i, and r_{min}, r_{max}, the smallest and greatest right descendant leaf of i.

evolution is reflected by the presence of breakpoints, that is inverted genes or genes that are adjacent in one genome but separated in another related genome. In contrast with rearrangement mechanisms, breakpoints can be directly observed from data. The breakpoint distance is the most widely used measure of gene order conservation, and usually considered as a first attempt to solve a given genome rearrangement problem. Moreover it provides an upper bound for the inversion distance.

As mentioned in Section 2.4.1, the orders considered in this paper represent clusters of tandemly duplicated genes, and as such, can always be considered linear. Let O and $\hat{O}$ be two signed orders, not necessarily on the same set of genes. A *breakpoint* of $\hat{O}$ with respect to O is a pair (j,k) of consecutive elements in $\hat{O}$ which is not present in O, neither in the form (j,k) nor in the form $(-k,-j)$. To account for breakpoints at cluster extremities, we add two "artificial genes" α and β so that O becomes (α,O,β) and $\hat{O}$ becomes $(\alpha,\hat{O},\beta)$ (see Figure 2.4).

$$\begin{array}{ll} (\alpha,O,\beta) & \alpha \;\; 2 \;\; 3 \;\; \text{-4} \;\; \text{-5} \;\; \text{-6} \;\; \beta \\ (\alpha,\hat{O},\beta) & \alpha \bullet 6 \;\; 5 \bullet 2 \;\; 3 \bullet 4 \bullet \beta \end{array}$$

Figure 2.4: Breakpoints (black dots) in $(\alpha,\hat{O},\beta)$ with respect to (α,O,β).

We denote by $d_{bp}(O,\hat{O})$ the number of breakpoints in $(\alpha,\hat{O},\beta)$ with respect to (α,O,β). When O and $\hat{O}$ are two permutations on the same set of genes, then $d_{bp}(O,\hat{O}) = d_{bp}(\hat{O},O)$, and d_{bp} is a distance.

The breakpoint distance is correlated to the inversion distance. Indeed, any sequence of inversions transforming $\hat{O}$ into O will eliminate all the breakpoints of $\hat{O}$ with respect to O. The following is a well known property.

Property 2.5. *Let O and $\hat{O}$ be two signed orders on the same set of genes. We have:*

$$\frac{d_{bp}(O,\hat{O})}{2} \leq d_{inv}(O,\hat{O}) \leq d_{bp}(O,\hat{O}).$$

In this section, we present an exact polynomial-time algorithm solving the following problem.

MINIMUM-BREAKPOINT DUPLICATION PROBLEM
Input: A signed ordered phylogeny (T, O),
Output: An order $\hat{O}$ such that $(T, \hat{O})$ is a duplication tree and $d_{bp}(O, \hat{O})$ is minimal.

A solution to this problem is an upper bound for the MINIMUM-INVERSION DUPLICATION PROBLEM. Indeed, let (T, O) be a signed ordered phylogeny, and O' and $\hat{O}$ be two orders such that (T, O') and $(T, \hat{O})$ are two duplication trees and $d_{inv}(O, O')$, $d_{bp}(O, \hat{O})$ are minimal. Then, from Property 2.5 we have:

$$d_{inv}(O, O') \leq d_{inv}(O, \hat{O}) \leq d_{bp}(O, \hat{O}).$$

The bound $d_{bp}(O, \hat{O})$ is not very tight as each inversion can create two breakpoints. A much better bound is $d_{inv}(O, \hat{O})$, which is obtained by using the HP algorithm with $\hat{O}$ generated by the polynomial-time algorithm we present in the next section.

2.5.2 A dynamic programming algorithm

For the purpose of our dynamic programming algorithm, orders extremities must be ignored while computing the number of breakpoints in intermediate sub-orders. Hence, extremities α and β will only be considered at the end of the procedure. We use the notation $d^*_{bp}(O, \hat{O})$ to refer to the number of breakpoints in $\hat{O}$ with respect to O. We denote by $\hat{O}[x, y]$ the sub-permutation of $\hat{O}$ beginning with element x and ending with element y. For example if $\hat{O} = (4, 2, 3, 5, 1)$, then $\hat{O}[2, 5] = (2, 3, 5)$.

Let (T, O) be a signed ordered phylogeny, and $\hat{O}$ be an alternative order on the leaves of T such that $(T, \hat{O})$ is a duplication tree. By definition, all genes in $\hat{O}$ must have the

same sign. For clarity of presentation and w.l.o.g, we suppose that they are positive.

Assume that $\hat{O} = \hat{O}[i,l]$, that is $\hat{O}$ begins with element i, ends with element l. Then, the duplication tree $(T, \hat{O}[i,l])$ can be defined recursively as the combination of two duplication trees $(T_1, \hat{O}[i,j])$ and $(T_2, \hat{O}[k,l])$ (see Figure 2.5), where j and k are two adjacent elements in $\hat{O}$ such that the least common ancestor of i, j and the least common ancestor of k, l are the two children of the root of T. Consequently, the breakpoint distance between $\hat{O}[i,l]$ and O can be expressed as follows:

$$d_{bp}(O, \hat{O}[i,l]) = d^*_{bp}(O, \hat{O}[i,j]) + d^*_{bp}(O, \hat{O}[k,l]) + bp(\alpha, i) + bp(j,k) + bp(l, \beta),$$

where $bp(x,y) = \begin{cases} 1 & \text{if the pair } (x,y) \text{ is a breakpoint with respect to } (\alpha, O, \beta) \\ 0 & \text{otherwise.} \end{cases}$

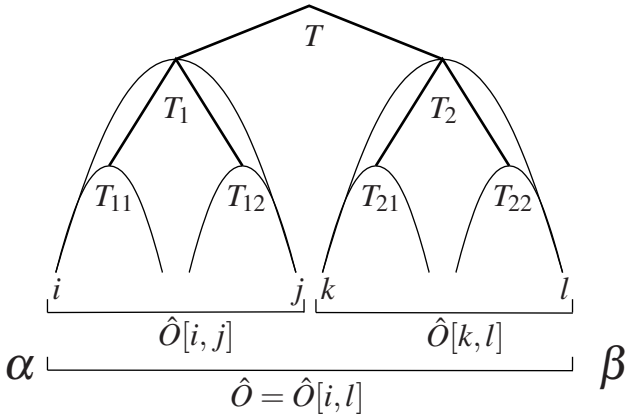

Figure 2.5: The duplication tree $(T, \hat{O}[i,l])$ can be obtained by combining two duplication trees $(T_1, \hat{O}[i,j])$ and $(T_2, \hat{O}[k,l])$. The "artificial genes" α and β allow to consider breakpoints at cluster's extremities.

Now we describe the central recursion of the algorithm. Let denote by $B[i,l]$ the minimum number of breakpoints (with respect to (α, O, β)) we can get among the set of orders compatible with T (or one of its subtrees) which start with i and end with l. Consider the subtree labeling of Figure 2.5 and assume that $i \in T_{11}$ and $l \in T_{22}$.

Then,

$$B[i,l] = \min_{(j \in T_{12}\,,\, k \in T_{21})} \Big(B[i,j] + B[k,l] + bp(j,k) \Big) \tag{2.1}$$

with the initial condition $B[i,i] = 0$ for every leaf i.

The $B[i,l]$ values can be computed recursively as follows. We consider every subtree T_x of T in a bottom-up approach (post-order traversal), beginning with the leaves of T and ending with T itself. For each T_x, using Recurrence 2.1, we compute $B[i,l]$ for every pair of leaves (i,l) whose least common ancestor is the root of T_x. It is easy to see from Figure 2.5 that this condition on (i,l) is necessary and sufficient for the existence of a duplication tree $(T_x, \hat{O}[i,l])$.

Finally, the breakpoint distance $d_{bp}(O,\hat{O})$ for an optimal order $\hat{O}$ (with positive signs) such that $(T,\hat{O})$ is a duplication tree is

$$d_{bp}(O,\hat{O}) = \min_{(i,l)} \Big(B[i,l] + bp(\alpha,i) + bp(l,\beta) \Big)$$

over the pairs (i,l) whose least common ancestor is the root of T. The order $\hat{O}$ is then simply constructed by backtracking in the dynamic programming table. The procedure above must be repeated with negative signs for the elements of $\hat{O}$ to get the global optimal.

Computing a given $B[i,l]$ value using Recurrence 2.1 takes $O(n^2)$ time in the worst case when the tree is balanced ($O(n)$ for a caterpillar tree). Since $B[i,l]$ is computed once for every pair (i,l), the worst-time complexity for the whole algorithm is $O(n^4)$.

2.6 Results with simulated and biological data

To simulate the evolution of a gene family and obtain the corresponding ordered phylogeny (T,O), we first generate T using the [125] model and define an order O' such that (T,O') is a duplication tree. Then, we obtain O by applying a fixed number of inversions to O'.

2.6.1 Execution time

To compare the execution time of the algorithms, we applied them on simulated ordered phylogenies with size varying from 10 to 40 leaves, that underwent 4, 8, 16 and 32 inversions. Results are averaged over 1,000 phylogenies and are given in Figure 2.6. We observe that the branch-and-bound performance depends significantly on the number of inversions. Nevertheless, it can be used on relatively important phylogenies within reasonable time (30 seconds on average for an ordered phylogeny with 40 leaves and 32 inversions). On the other hand, the execution time of the polynomial-time algorithm depends uniquely on the size of the phylogeny and requires less than a second for all the instances.

2.6.2 Using the polynomial-time algorithm as a heuristic

The polynomial-time algorithm finds a duplication tree $(T, \hat{O})$ such that the breakpoint distance between $\hat{O}$ and the order O observed on the chromosome is minimal. To see if $\hat{O}$ can be used as an approximation to the MINIMUM-INVERSION DUPLICATION PROBLEM, we applied the algorithm on simulated data and compared $d_{inv}(O, \hat{O})$ (computed using the HP algorithm) with the optimal value returned by the branch-and-bound. We used ordered phylogenies with 10 and 20 leaves, which underwent 1 to 16 inversions. The results are averaged over 1,000 phylogenies and are presented in Figure 2.7. We see that when the number of inversions is low, the inversion distance obtained with the polynomial-time algorithm is very close to the optimal one.

2.6.3 Improving phylogenetic inference

We applied our algorithms on simulated data to verify how they could be used to validate inferred phylogenies of tandemly repeated gene families. The idea is that a wrong phylogeny should require more inversions than the true one. We simulated ordered phylogenies with 10 and 20 leaves, which underwent 1, 2, 4 and 6 inversions.

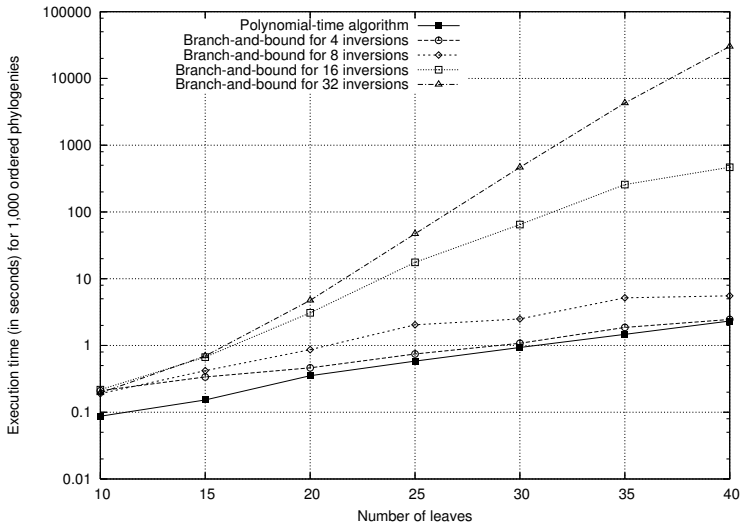

Figure 2.6: Execution times (in seconds) for 1,000 signed ordered phylogenies with 4, 8, 16 and 32 inversions. The execution time of the polynomial-time algorithm is not affected by the number of inversions.

These are the observable states (T_{true}, O) resulting from "true" duplication/inversion histories. For each T_{true}, we then generated four "wrong" (but close) phylogenies T_{wrong}, by applying one to four random Nearest Neighbor Interchange rearrangements (NNI) [e.g 110, chap. 7]. Those "wrong" phylogenies can be seen as the ones we would obtain from biological data when a few branches have weak statistical support. We then used the branch-and-bound algorithm to compute the minimum number of inversions inv() necessary to explain (T_{true}, O) and its associated (T_{wrong}, O). We did the same procedure with the polynomial-time algorithm which instead computes the minimum number of breakpoints bp() between the order of an inferred duplication tree and the order on the chromosome. The results are averaged over 1,000 phylogenies and are presented in Figure 2.8 and 2.9. Surprisingly, the results are very similar although the breakpoint distance is slightly less sensitive to wrong phylogenies.

Results can be interpreted as follows. For a wrong 10-leaf phylogeny that differs by one NNI from the true one, roughly 50% of the time on average our algorithms report an excess of inversions/breakpoints, otherwise they report the same number com-

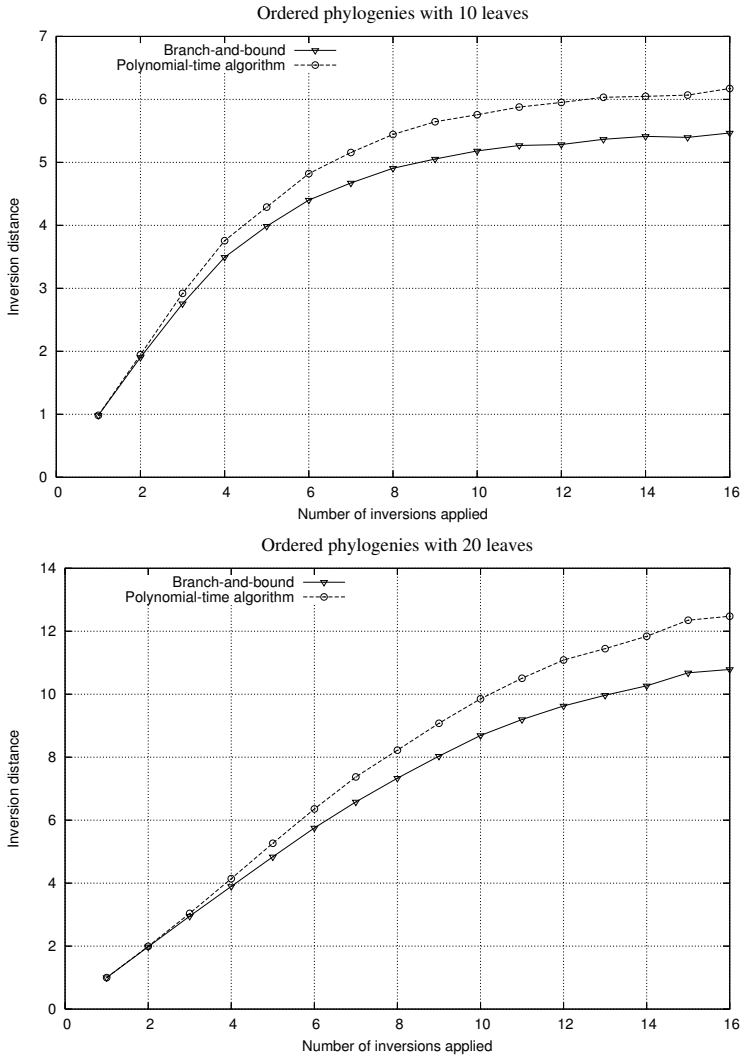

Figure 2.7: Number of inversions inferred by the polynomial-time algorithm compared to the optimal value obtained by the branch-and-bound. Results are averaged over 1,000 phylogenies.

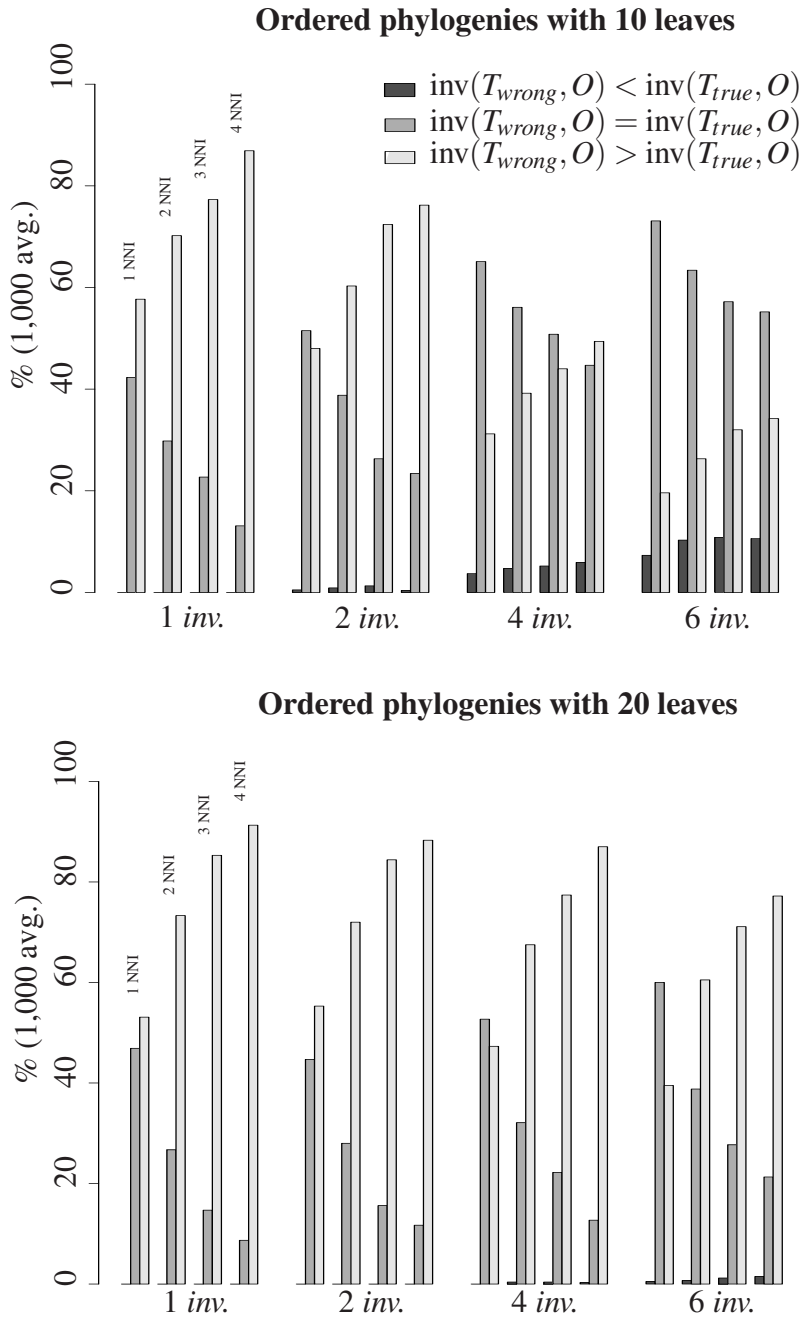

Figure 2.8: Fraction of times inv(T_{wrong}, O) is less, equal or greater than inv(T_{true}, O).

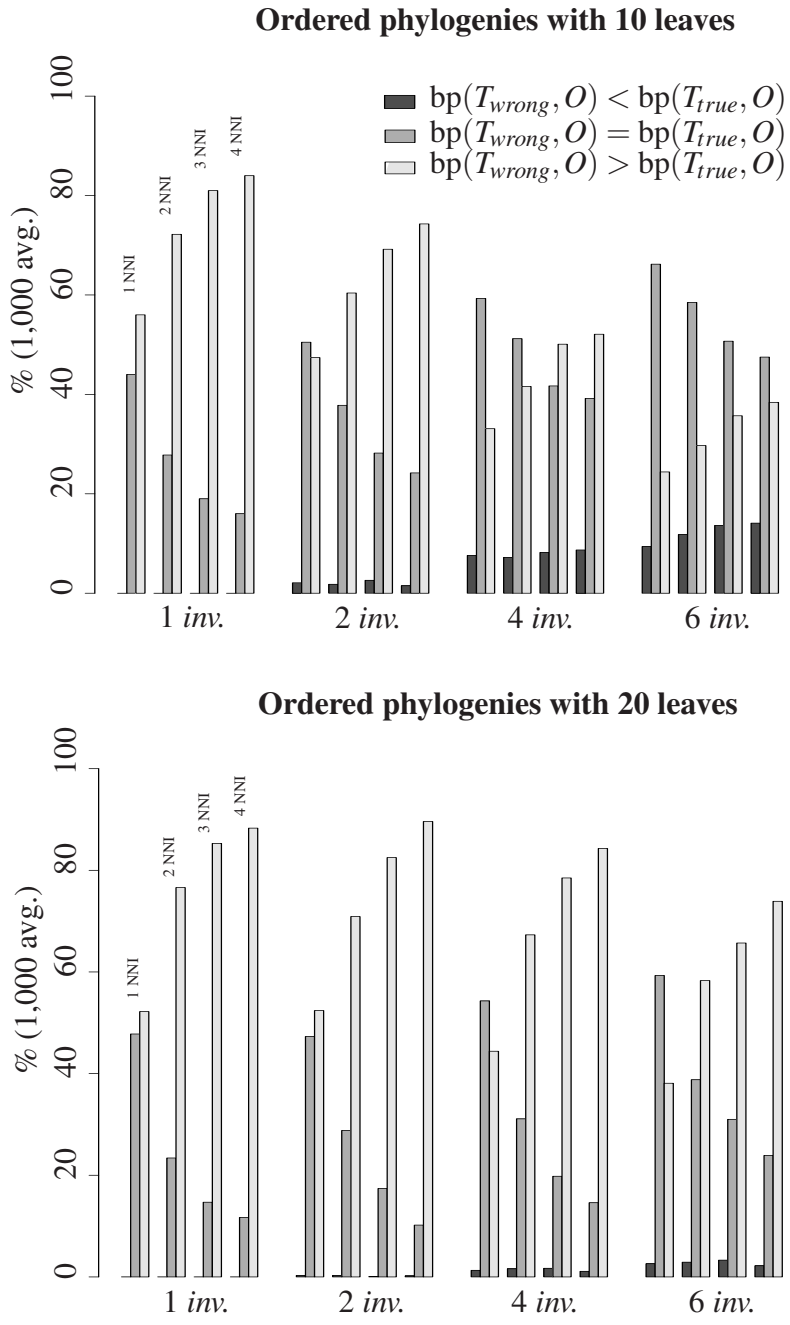

Figure 2.9: Fraction of time bp(T_{wrong}, O) is less, equal or greater than bp(T_{true}, O).

pared to the true phylogeny. Suppose we have a set of putative phylogenies for a given gene family, and one is correct while the others differ by a few NNI. According to Figure 2.8 and 2.9, for wrong trees, the algorithms almost always reports the same number of inversions/breakpoints or more as in the true tree. Thus, choosing the phylogeny with the lowest number of inversions/breakpoints is either a winning strategy, or not enough to select a single phylogeny as several ones require the same number of inversions/breakpoints, but is almost never misleading. Of course, this ability to discard wrong phylogenies decreases as the true number of inversions increases, but even with 6 inversions and 4 NNI the number of misleading cases remains low.

2.6.4 Application on biological data

The KRAB-zinc finger gene family encodes for transcription factors. It contains more than 400 active members physically grouped into clusters. In a recent study, [47] proposed a phylogeny of the primate specific ZNF91 sub-family based on their tether[5] and flanking sequences. This phylogeny (obtained by Neighbor-Joining [98]) contains a monophyletic group of 6 genes clustered at the telomere of HSA4p, which may have been derived from a single ancestor through successive tandem duplications.

We applied the branch-and-bound algorithm on this cluster using the proposed phylogeny, and found that a duplication/inversion history would require at least 4 inversions, which seems relatively high considering that only 6 genes are involved.

To test whether a "better" phylogeny could be proposed, we used the MrBayes software [97] to obtain a sample from the posterior probability distribution of all possible phylogenies. The tether (+100 flanking bp) sequences were downloaded from the Human KZNF Gene Catalog[6] [51] and aligned using ClustalW [115] with default settings. The ZNF160 tether sequence was used as an outgroup to obtain a rooted tree. We performed 500,000 MCMC generations with MrBayes under the GTR model [63, 113] and

[5] The region upstream from the first finger.

[6] http://znf.llnl.gov/catalog/

a gamma-shaped rate variation with a proportion of invariable sites. Convergence was easily attained and the experiment was repeated three times with similar results. Finally we applied the branch-and-bound on the sampled phylogenies and observed that the best one (p=0.4) is compatible with an optimal duplication/inversion history involving only two inversions. This provides strong support for the tandem duplication/inversion model and indicates that our phylogeny is probably the correct one. Results obtained with the polynomial-time algorithm are similar although less discriminative. Phylogenies are presented in Figure 2.10 with both their associated numbers of inversions / breakpoints.

2.7 Conclusion

This work represents the first attempt to account for inversions in an evolutionary model of tandemly repeated genes. We presented a time-efficient branch-and-bound algorithm for finding the minimal number of inversions in an evolutionary history of a gene family characterized by an ordered phylogeny. We have also developed a polynomial-time algorithm based on the breakpoint distance. We demonstrated, using simulations, that it is a good heuristic for the original problem. Though only simple duplications were considered here, the model has been shown useful to select an appropriate phylogeny among a set of possible ones. These are encouraging results that motivate further extensions.

One of the next steps of this work will be to account for multiple duplications in the evolutionary model, although generalizing the MINIMUM-INVERSION DUPLICATION PROBLEM to this model is far from being straightforward. From this perspective, it seems reasonable to begin with a simpler rearrangement distance such as the breakpoint distance. Another important generalization will be to consider a family of tandemly duplicated genes with orthologs in two or more genomes. For example, [105] identified homologous ZNF gene family regions in human and mouse. A phylogenetic tree involving such tandemly repeated genes in human and mouse clusters was established.

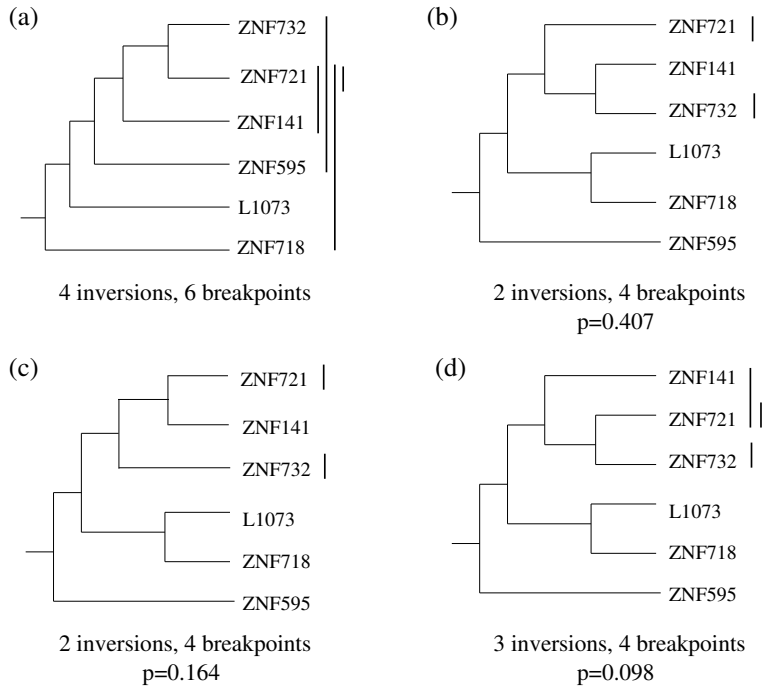

Figure 2.10: Different phylogenies for the ZNF141 clade on human chromosome 4, with the associated minimal number of inversions/breakpoints. The black vertical lines represent an optimal sequence of inversions leading to the *signed* gene order observed on the chromosome: (+ZNF595, +ZNF718, +L1073, −ZNF732, +ZNF141, −ZNF721). (a) The phylogeny published by [47] requires 4 inversions, which is relatively high for 6 genes; (b,c,d) The 3 best phylogenies we obtained with MrBayes, and their associated probabilities. The first two ones require only 2 inversions, which is optimal for this order. The position of the root was determined using ZNF160 as an outgroup.

It would be of major interest to develop an algorithm allowing one to explain such a phylogeny based on an evolutionary model involving tandem duplication, inversion and speciation events.

The biological assumption used in this paper is that inversion is the only rearrangement mechanism leading to the presence of different transcriptional orientations in a tandemly duplicated gene family. Another possible rearrangement mechanism that has been documented and that gives an alternative explanation for the presence of inverted genes is *inverted tandem duplication* as a single event [109]. However, "simple" inverted duplication (inverted duplication of a single gene) is not always sufficient to explain the fact that a tree representing a tandemly duplicated gene family is not a duplication tree, so other mechanisms (e.g. multiple duplication or gene deletion) have to be accounted for, which makes the computational problem even harder that the one we considered here. Future investigations should address such compound models, incorporating a wide range of duplication-deletion-rearrangement events.

Acknowledgments

The authors wish to thank M. Aubry and H. Tadepally for their help on zinc finger genes and the anonymous referees for their valuable comments. This work was supported by grants from the *Fonds québécois de la recherche sur la nature et les technologies (FQRNT)*, the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canadian Institutes of Health Research (CIHR) and the French ACI IMPBIO: REPEVOL project.

2.8 Contribution des auteurs

Mathieu Lajoie et Nadia El-Mabrouk ont écrit le manuscrit. Denis Bertrand et **Mathieu Lajoie** ont conçu et implémenté l'algorithme *branch-and-bound*. **Mathieu Lajoie** a conçu l'algorithme polynomial et Denis Bertrand l'a implémenté. Denis Bertrand et **Mathieu Lajoie** ont conçu et réalisé les expérimentations sur les données simulées. **Mathieu Lajoie** a conçu et réalisé les expérimentations sur les données biologiques. Olivier Gascuel a proposé le projet. Tous les auteurs ont participé à la révision du manuscrit. **Mathieu Lajoie** et Denis Bertrand sont considérés comme les co-premiers auteurs. Les résultats préliminaires ont été présentés par **Mathieu Lajoie** lors du *Fourth RECOMB International Workshop on Comparative Genomics* [14], le 25 septembre 2006 à Montréal au Canada.

CHAPITRE 3

INFERRING ANCESTRAL GENE ORDERS FOR A FAMILY OF TANDEMLY ARRAYED GENES

Denis Bertrand[1], **Mathieu Lajoie**[2] **et Nadia El-Mabrouk**[3]

Article publié dans *Journal of Computational Biology* [13]

[1]DIRO, Université de Montréal, Montreal (QC), Canada.
[2]DIRO, Université de Montréal, Montreal (QC), Canada.
[3]DIRO, Université de Montréal, Montreal (QC), Canada.

Abstract

Tandemly arrayed genes (TAG) constitute a large fraction of most genomes and play important biological roles. They evolve through unequal recombination, which places duplicated genes next to the original ones (tandem duplications). Many algorithms have been proposed to infer a tandem duplication history for a TAG cluster. However, the presence of different transcriptional orientations in many clusters highlights the fact that processes such as inversions also contribute to their evolution. Moreover, existing algorithms are restricted to the study of TAGs evolution in a single species (only paralogous genes are considered). To circumvent these limitations, we consider an evolutionary model for TAGs involving duplication, gene loss, inversion and speciation events. A general framework to infer ancestral gene orders that minimize the number of inversions in the whole evolutionary history is presented. At the methodological level, this paper integrates three approaches to genome evolution: the duplication tree reconstruction, the gene tree/species tree reconciliation theory, and the concept of inversion median used in order-based phylogeny reconstruction. An application on a cluster of olfactory receptor genes in 4 mammals is presented.

3.1 Introduction

A multigene family is a set of genes that have evolved by duplication and speciation from a common ancestral gene, and share a similar sequence and usually a similar function. Members of a gene family in a given genome may appear in clusters, or scattered on a single or many chromosomes. In this paper, we focus on clusters of tandemly arrayed genes (TAG): copies that are adjacent on the chromosome. TAGs have been shown to represent a large proportion of genes in mammalian genomes. In particular, they represent about 14-17% of all genes in human, mouse and rat [106]. Clusters of TAGs may vary in size from two to hundreds genes, though small clusters are largely predominant (an average of 3 to 4 genes in mouse, rat and human) [106]. They are involved

in many functions of binding or receptor activities. In particular, the olfactory receptor genes constitute the largest multigene family in vertebrate genomes, with several hundred genes per species [2]. Other families of TAGs include the HOX genes [128], the immunoglobulin and T-cell receptor genes [3], the MHC genes [39] and the Zinc Finger genes [105].

TAGs are widely viewed as resulting from unequal recombination during meiosis [35], generating clusters of similar genes with the same transcriptional orientation. When fixed in a genome, such duplicates increase the chance of giving rise to other mispairings, thus leading to other duplicates.

Several studies have considered the problem of inferring an evolutionary history for a TAG cluster [112, 33, 31, 53, 130, 12]. These are essentially phylogenetic inference methods using the additional constraint that the resulting tree should induce a duplication history according to the given gene order. Such trees are called *duplication trees*. When a gene tree is already available for a TAG cluster, a linear-time algorithm can be used to check whether it is a duplication tree [38, 130]. As the probability for an arbitrary gene tree to be a duplication tree is very low (2.10^{-5} for a random tree with 15 leaves [38]), the fact that a gene tree is a duplication tree is a strong argument in favor of the tandem duplication model of evolution for the associated gene family. However, it is often impossible to reconstruct a duplication history for a TAG cluster [37], even from well supported gene trees. This is due to the occurrence of other mechanisms, such as deletions and genomic rearrangements [29], during the evolution of the gene family. In particular, [106] have observed that more than 25% of all neighboring pairs of TAGs in human, mouse and rat have non-parallel orientations. This highlights the fact that other mechanisms, such as inversions, should be considered in an evolutionary model of TAGs. In a previous publication [62], we have presented an algorithm that finds the minimum number of inversions involved in the evolutionary history of a TAG cluster, assuming single gene duplications.

An important restriction of the above models of evolution is the fact that they are lim-

ited to the analysis of a TAG cluster located in a single species and on a single chromosome. However, the increasing availability of complete genomic sequences and of many different TAG databases [2, 51] makes it possible to study the evolution of gene families with members belonging to different species. Such a global evolutionary study may help deciphering the common origins of TAGs, highlighting the inter-species differences and identifying the genetic basis of species-specific features. Various phylogenetic studies have been conducted on different TAG families such as the Zinc-Finger genes in human and mouse [105], and the olfactory receptor genes in various mammalian species [2]. However, no rigorous approach has been developed so far to explain the non agreement between a gene tree of a TAG family and a duplication and speciation history.

This paper is the first attempt to account for tandem duplication, speciation, gene loss and inversion events in an evolutionary model of TAGs. Given the gene and species trees for a set of orthologous TAG clusters and their respective gene orders, we aim to infer the *ancestral* gene orders leading to a most parsimonious sequence of evolutionary events. Clearly, an important prerequisite is to have, as an input, a well supported gene tree. This is unrealistic in the framework of "concerted evolution", where all the members of a gene family are assumed to evolve in a concerted manner by repeated occurrences of gene conversions. Hopefully, evidences for many TAG families (e.g. MHC, immunoglobulin and olfactory receptor genes) is in favor of a "birth-and-death" model of evolution [81], in which gene conversion is much less important than previously believed.

At the methodological level, this paper integrates three approaches to genome evolution: the duplication tree reconstruction, the gene tree/species tree reconciliation, and the concept of inversion median used in order-based phylogeny reconstruction. We proceed in two steps. First, ignoring gene orders, a classical gene tree/species tree reconciliation method is used to infer a "minimal" duplication, speciation and loss history in agreement with a known species tree [90]. Second, we infer the ancestral gene orders that minimize the number of inversions required to obtain a valid duplication tree. This problem is related to the more classical one of inferring gene orders of the ancestral genomes in a

species tree [99, 16, 80, 75].

This paper is organized as follows. We describe the evolutionary model in Section 3.2 and our optimization problem in Section 3.3. The general iterative method used for minimizing the inversions in a whole species tree is then presented in Section 3.4. The detailed algorithm used for a single branch is then presented in Section 3.5. In Section 3.6 we present an exact branch-and-bound algorithm and a heuristic to solve the median problem. In Section 3.7, we compare the running times and the accuracy of our algorithms on different simulated data sets. Finally, an application on a set of orthologous TAG clusters in four mammalian species is presented.

3.2 The evolutionary model

The classical model of evolution considered for TAGs is based on tandem duplications resulting from unequal recombination during meiosis, which together with point mutations are assumed to be the sole evolutionary mechanisms acting on sequences. Formally, from a single ancestral gene at a given position in the chromosome, the locus grows through a series of consecutive duplications placing the created copy next to the original one. Such *tandem duplications* may be *simple* (duplication of a single gene) or multiple (simultaneous duplication of neighboring genes). In this paper, we only consider simple duplications. From now on, a *duplication* will refer to a simple tandem duplication.

Consider a set of m orthologous TAG clusters located on m different genomes. We denote by $\mathscr{O} = \{O_1, O_2, \ldots, O_m\}$ the set of gene orders, i.e. for $1 \leq i \leq m$, O_i is the signed order of the family members in genome i. The sign $(+/-)$ of a gene represents its transcriptional orientation. In addition to the observed gene orders, a gene tree can be inferred from the TAG sequences. In this paper, a *gene tree* T for a TAG family is a rooted binary tree with labeled leaves, where each label represents a gene copy. A leaf labeled by a gene copy in genome i is said to *belong to genome* i. For conciseness, we

make no difference between a leaf and its label. The pair $(T, \mathscr{O})$ is called the *ordered gene tree* for the gene family (see Figure 3.1(a)).

We denote by $d_{inv}(O_i, \widehat{O}_i)$ the inversion distance between two orders O_i and $\widehat{O}_i$ on the same set of genes. Such a distance can be computed using the original [48] algorithm, or any of the existing optimizations [55, 6, 11].

The following is a formal definition of a Duplication, gene Loss, Inversion and Speciation history (DLIS-history) leading to an ordered gene tree $(T, \mathscr{O})$ (see Figure 3.1(b) for an illustration).

Definition 3.1. *A DLIS-history of $(T, \mathscr{O})$ is a sequence of ordered gene trees $\mathscr{H} = ((T^1, \mathscr{O}^1), (T^2, \mathscr{O}^2), \ldots, (T^h, \mathscr{O}^h))$ where:*

1. *T^1 is a tree consisting of a single leaf v and $\mathscr{O}^1 = \{O_1^1\} = \{(\pm v)\}$ is one of the two trivial orders.*

2. *For $1 \leq k < h$, there is a unique i such that exactly one of the four following situations holds:*

   a. Duplication event: *T^{k+1} is obtained from T^k by adding two children u and w to a leaf v belonging to genome i. Moreover $\mathscr{O}^{k+1}$ is obtained from $\mathscr{O}^k$ by replacing v by (u, w) in O_i^k, where u and w have the same sign as v.*

   b. Gene loss event: *T^{k+1} is obtained from T^k by removing a leaf v belonging to genome i. If v was the only leaf in O_i^k then $\mathscr{O}^{k+1} = \mathscr{O}^k \setminus \{O_i^k\}$, otherwise $\mathscr{O}^{k+1}$ is obtained from $\mathscr{O}^k$ by deleting v from O_i^k.*

   c. Inversion event: *$T^{k+1} = T^k$ and $d_{inv}(O_i^k, O_i^{k+1}) = 1$.*

   d. Speciation event: *T^{k+1} is obtained from T^k by adding two children u_j and w_j to each leaf v_j belonging to genome i. Moreover, $\mathscr{O}^{k+1}$ is obtained from $\mathscr{O}^k$ by replacing $O_i^k = (v_1, \ldots, v_t)$, by $\{(u_1, \ldots, u_t), (w_1, \ldots, w_t)\}$, where u_j and w_j have the same sign as v_j.*

3. $(T, \mathscr{O}) = (T^h, \mathscr{O}^h)$.

Any DLIS-history $\mathscr{H}$ of $(T, \mathscr{O})$ induces a unique species tree S obtained from the speciation events of $\mathscr{H}$. We say that $\mathscr{H}$ is *consistent with S* (see Figure 3.1).

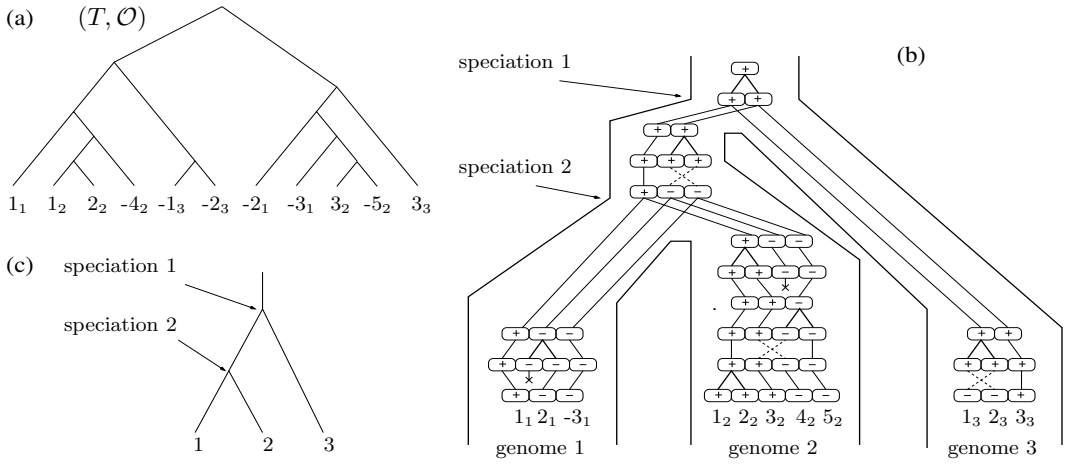

Figure 3.1: (a) An ordered gene tree $(T,\ \mathscr{O} = \{(1_1, -2_1, -3_1),\ (1_2, 2_2, 3_2, -4_2, -5_2),\ (-1_3,\ -2_3, 3_3)\})$. Genes are denoted as g_i meaning "the gth gene in genome i". (b) A DLIS-history for $(T, \mathscr{O})$. Duplications are indicated by bold lines, gene losses by 'X' and inversions by dashed lines. For clarity, we omitted successive identical configurations in each lineage. (c) The induced species tree for the three genomes.

3.3 An inference problem

Let $(T, \mathscr{O})$ be an ordered gene tree for a family of TAGs on m genomes. Suppose that a species tree S is already known for the m genomes. Then a natural problem is to find a DLIS-history of $(T, \mathscr{O})$ that is consistent with S. By Lemma 3.2, such a history exists. It follows from the existence of a duplication/speciation/loss history of T consistent with S in the general case of an unordered gene family. In this context, the *reconciliation* approach, first introduced by [44], and subsequently developed by many other authors [90, 45, 73, 15], can reconstruct such a history with a minimum number of duplication and/or loss events. The different reconciliation approaches are all based

on a particular mapping (Least Common Ancestor mapping) from the nodes of T to the nodes of S, allowing to "embed" the gene tree into the species tree.

Lemma 3.2. *Given an ordered gene tree $(T, \mathscr{O})$ on m genomes and a species tree S for the m genomes, there is at least one DLIS-history of $(T, \mathscr{O})$ consistent with S.*

Proof. Obtain a sequence of duplications, gene losses and speciations from the reconciliation of T and S. From that sequence, construct a DLIS-history $\mathscr{H}' = ((T^1, \mathscr{Q}^1), \ldots, (T^h = T, \mathscr{Q}^h))$ by applying the operations described in cases a, b and d of Definition 3.1. Then, obtain $\mathscr{H}$ from $\mathscr{H}'$ by performing any sequence of inversions transforming $\mathscr{Q}^h$ into $\mathscr{O}$ (case c in Definition 3.1). ■

As the number of possible DLIS-histories consistent with S is unlimited, reasonable criteria should be considered. Here, we restrict ourselves to the most parsimonious DLIS-histories that are in agreement with a given reconciled tree.
We proceed in two steps:

1. We obtain a reconciled tree G from T and S. In the present study, we used the parsimony method of [129], but any other method could be used (e.g. [4] and [118]).

2. We find the ancestral gene orders that minimize the total number of inversions involved in a DLIS-history of $(T, \mathscr{O})$. Formally, the problem considered in this step is the following:

MINIMUM-DLIS PROBLEM

Input: An ordered reconciled tree $(G, \mathscr{O})$.

Output: A gene order for each ancestral genome inducing a DLIS-history of minimum inversions.

In the rest of this paper, we focus on solving the MINIMUM-DLIS PROBLEM. We further introduce some additional information about the nodes of G and their implicit mapping to S (see Figure 3.2):

- A *duplication node* is an internal node which corresponds to a duplication event. It maps to a branch of S, i.e. the lineage in which the duplication occurred.
- A *speciation node* is an internal node which corresponds to an ancestral gene at the time of a speciation event. It maps to an internal node of S, i.e. the ancestral genome to which it belongs. It has either one child (in the case of a gene loss), or two children each belonging to a different lineage.
- A leaf is an extant gene and maps to a leaf of S, i.e. the extant genome to which it belongs.
- A maximal set of speciation nodes or leaves mapped to the same node A of S is defined as the *gene content* of A. When this set is ordered, we denote it by O_A.
- Let ρ be a direct descendant of a speciation node r. Then, the subtree rooted at ρ is said to be *externally rooted* at r.

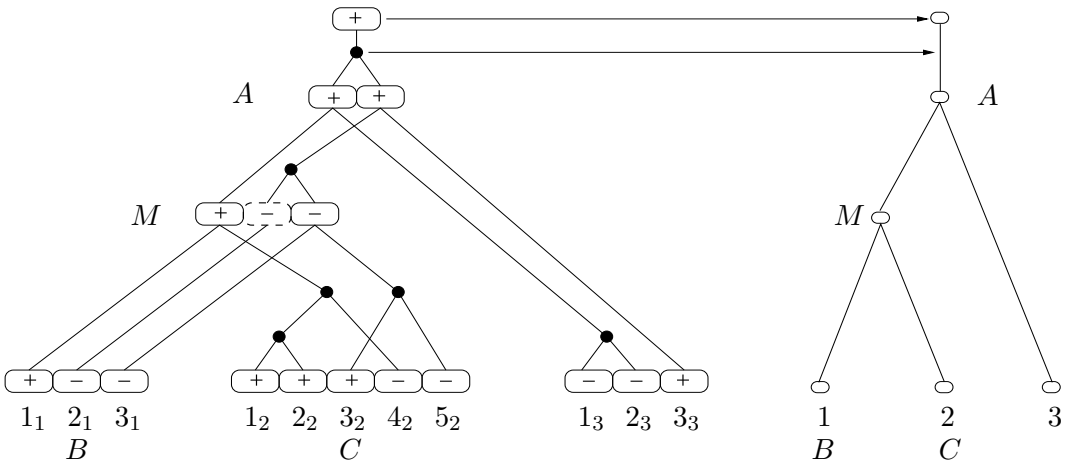

Figure 3.2: On the left, the ordered reconciled tree $(G, \mathscr{O})$ induced by the DLIS-history of Figure 3.1, with the corresponding ancestral gene orders. We see that each duplication node (black dot) in G implicitly maps to an *edge* of the species tree S (on the right), and each speciation node (box) to a *node* of S. The dashed gene in genome M has no descendants in lineage C, indicating a gene loss.

From now on, we consider the "embedded" representation of G in S. More precisely,

a *branch* (R,L) *of* S will denote the set of subtrees in G connecting the gene contents of R and L (see Figure 3.2).

Suppose that the gene content is ordered for each node of S. Then there exists a DI-history (a history restricted to duplication and inversion events) with a minimum number of inversions explaining each branch of S, and the minimum number of inversions in any DLIS-history of $(G,\mathscr{O})$ (and its corresponding ancestral gene orders) is the sum of the inversions involved in those minimal DI-histories. Thus, our problem reduces to the one of finding the ancestral gene orders minimizing the sum of inversions involved in a DI-history of each branch of S. The formal definitions of a branch of S and a DI-history are given in Section 3.5.

3.4 A general method based on the median problem

The Minimum-DLIS problem is related to the more classical one of inferring the gene orders at the internal nodes of a species tree, where each leaf is labeled by an ordered sequence of genes (see for example [99], [16] and [80]). After fixing the ancestral gene contents, which is an intricate problem in the general case of unequal gene content and gene paralogy, the problem is to find the ancestral gene orders minimizing a given genomic distance.

Although the case of an ordered reconciled tree G has the additional constraint of tandem duplications, the two problems are related, suggesting a similar global approach summarized below.

1. Begin with an initial order O_M for each internal node M of S.

2. Traverse S in a depth-first manner. For each subtree consisting of a branch (A,M), where A is the immediate ancestor of M, and two sister branches (M,B) and (M,C) (see Figure 3.2), ignore the assigned order for M, and reconstruct an order that

minimizes the value:

$$DI(O_A, O_M) + DI(O_M, O_B) + DI(O_M, O_C),$$

where $DI(O_R, O_L)$ is the minimum number of inversions in a DI-history explaining the branch (R, L). This step consists in solving the well known *median problem*.

3. Iterate Step 2. a given number of times, or until convergence to a local minimum.

In case no duplication and no gene loss have occurred during the evolution of the gene family along the branches from A to B and C, the DI value becomes the inversion distance, and the median problem formulated in Step 2 is just the *inversion median problem*, which has been proved to be NP-hard [18]. Therefore, the "generalized median problem" considered here is also NP-hard.

A rigorous definition and computation of $DI(O_R, O_L)$ for a branch (R, L) is given in the next section. We then present an exact algorithm and a heuristic to solve the median problem in Section 3.6. Finally, we present a heuristic allowing to begin with appropriate initial orders.

3.5 The generalized Minimum-DI problem

3.5.1 Definitions

We consider the problem of minimizing the number of inversions involved in a history explaining a given branch (R, L) of S when the gene contents are ordered. Formally, such a branch is called an *ordered forest* and is defined as follows:

Definition 3.3. *An* ordered forest (F, O_R, O_L) *is a forest of n gene trees* $F = \{T_1, T_2, \ldots, T_n\}$ *externally rooted at* $O_R = (r_1, r_2, \ldots, r_n)$*, with an order* O_L *on its leaves.*

We now formally define the notion of a DI-history explaining a given branch (R, L) of S. It is a generalization of the definition introduced in our previous paper [62] for a

single ordered gene tree.

Definition 3.4. *A* DI-history *of an ordered forest* (F, O_R, O_L) *is a sequence of ordered forests* $\mathscr{H} = ((F^1, O_R, O_L^1), (F^2, O_R, O_L^2), \ldots, (F^h, O_R, O_L^h))$ *such that:*

1. F^1 *is a set of single leaf gene trees externally rooted at* O_R *and ordered as* $O_L = O_R$.
2. *For* $1 \leq k < h$, *exactly one the following situations holds:*
   a. Duplication event: F^{k+1} *is obtained from* F^k *by adding two children* u *and* w *to one of its leaf* v, *and* O_L^{k+1} *is obtained from* O_L^k *by replacing* v *by* (u,w), *where* u *and* w *have the same sign as* v.
   b. Inversion event: $F^{k+1} = F^k$ *and* $d_{inv}(O_L^k, O_L^{k+1}) = 1$.
3. $(F, O_R, O_L) = (F^h, O_R, O_L^h)$.

From Definition 3.4, we also introduce the notion of a *duplication history*, which is simply a DI-history restricted to duplication events. A duplication history gives rise to a duplication forest, defined as follows.

Definition 3.5. *A* duplication forest *is an ordered forest* $(F = \{T_1, \ldots, T_n\}, O_R, O_L)$ *containing only* duplication trees, *and such that for every pair* (r_i, r_j) *in* O_R, *if* r_i *precedes* r_j, *then all the leaves of* T_i *precede all the leaves of* T_j *in* O_L. *Moreover, for any* $1 \leq i \leq n$, *the leaves of* T_i *have the same sign as* r_i.

The following theorem is a generalization of the result we obtained for a single ordered gene tree in one species [62].

Theorem 3.6. *For any DI-history of* (F, O_R, O_L) *with* i *inversions, there exists a duplication forest* $(F, O_R, \widehat{O}_L)$ *such that* $d_{inv}(O_L, \widehat{O}_L) \leq i$.

Proof. Let $\mathscr{H}^k = ((F^1, O_R, O_L^1), (F^2, O_R, O_L^2), \ldots, (F^k, O_R, O_L^k))$ be a DI-history of (F, O_R, O_L). We prove the theorem by induction on k:

- Base case: If $k = 1$, then $\mathscr{H}^1 = ((F^1, O_R, O_L^1))$ is a DI-history with no duplication and no inversion. Clearly $(F^1, O_R, \widehat{O}_L) = (F^1, O_R, O_L^1)$ is a duplication forest and $d_{inv}(O_L^1, \widehat{O}_L) = 0$.

- Induction step:
  Let $\mathscr{H}^{k+1} = ((F^1, O_R, O_L^1), \ldots, (F^k, O_R, O_L^k), (F^{k+1}, O_R, O_L^{k+1}))$ be a DI-history involving i inversions. From Definition 3.4, there are two possibilities:

  – If $F^{k+1} \neq F^k$, then the last event is a duplication, i.e. there is a leaf v of a tree of F^k that was replaced by two consecutive leaves u, w of the same sign in O_L^{k+1}. By the induction hypothesis, there exists a duplication forest $(F^k, O_R, \widehat{O}_L^k)$ such that $d_{inv}(O_L^k, \widehat{O}_L^k) \leq i$.
  Suppose v is positive in $\widehat{O}_L^k$. If v is also positive in O_L^k, we define $\widehat{O}_L^{k+1}$ as the order obtained by replacing $+v$ by $(+u, +w)$ in $\widehat{O}_L^k$. Otherwise, v is negative in O_L^k and we obtain $\widehat{O}_L^{k+1}$ by replacing $+v$ by $(+w, +u)$ in $\widehat{O}_L^k$. It follows that $d_{inv}(O_L^{k+1}, \widehat{O}_L^{k+1}) = d_{inv}(O_L^k, \widehat{O}_L^k) \leq i$ and $(F^{k+1}, O_R, \widehat{O}_L^{k+1})$ is a duplication forest. The case where v is negative in $\widehat{O}_L^k$ is treated similarly.

  – If $F^{k+1} = F^k$, then the last event is an inversion and $\mathscr{H}^k$ involves $i - 1$ inversions. By the induction hypothesis, there exists a duplication forest $(F^k, O_R, \widehat{O}_L^k)$ such that $d_{inv}(O_L^k, \widehat{O}_L^k) \leq i - 1$. Then we have $d_{inv}(O_L^{k+1}, \widehat{O}_L^{k+1}) \leq d_{inv}(O_L^k, \widehat{O}_L^k) + 1 \leq i$, where $\widehat{O}_L^{k+1} = \widehat{O}_L^k$.

■

The following result immediately follows from Theorem 3.6.

Corollary 3.7. *Let (F, O_R, O_L) be an ordered forest and $(F, O_R, \widehat{O}_L)$ be a duplication forest such that $d_{inv}(O_L, \widehat{O}_L) = i$ is minimal over all $\widehat{O}_L$. Then, there exists a DI-history*

of (F, O_R, O_L) *with exactly i inversions. Moreover, i is the minimum number of inversions in a DI-history of* (F, O_R, O_L).

Corollary 3.7 allows to reformulate the problem as follows:

GENERALIZED-MINIMUM-DI PROBLEM

Input: An ordered forest (F, O_R, O_L).

Output: An order $\widehat{O}_L$ on the leaves of F such that $(F, O_R, \widehat{O}_L)$ is a duplication forest and $d_{inv}(O_L, \widehat{O}_L)$ is minimal.

Given a branch (R, L) of S and the orders O_R and O_L, the minimum number of inversions involved in a DI-history of the branch (R, L) is denoted as $DI(O_R, O_L)$. In the following section, we present an algorithm for solving the GENERALIZED-MINIMUM-DI PROBLEM.

3.5.2 A Branch-and-Bound algorithm

The algorithm is a generalization of the one we presented in a previous paper [62]. Given an ordered gene tree (T, O), the goal was to find an order $\widehat{O}$ such that $(T, \widehat{O})$ is a duplication tree and $d_{inv}(O, \widehat{O})$ is minimal.

Ordered gene tree: As mentioned by [37], simple duplication trees are equivalent to binary search trees. Therefore, to enumerate all the orders $\widehat{O}$ such that $(T, \widehat{O})$ is a duplication tree, we associated a binary variable b_i to each internal node i of T. By setting $b_i = 0$, we make the left descendant leaves of i *smaller* than the right ones in $\widehat{O}$, whereas by setting $b_i = 1$ we makes them *larger*. If we assign these values by a post-order traversal of T, then each b_i value induces an adjacency between two of its descendant leaves in $\widehat{O}$.

Hence, $(T, \widehat{O})$ is a duplication tree iff $\widehat{O}$ is defined by an assignment of all the binary variables in T, and all its genes have the same sign ($+$ or $-$). If n is the number of leaves in T, this leads to 2^n distinct orders.

To avoid computing $d_{inv}(O,\widehat{O})$ for every possible order $\widehat{O}$, we considered a branch-and-bound strategy based on the following property: $d_{inv}(O,\widehat{O}) \geq n+1-c$, where n is the number of genes and c is the number of cycles in the *breakpoint graph* [48] of O and $\widehat{O}$. In this graph, each edge corresponds to an adjacency in one of the two orders (see Figure 3.3). The general idea is to bound c as we progressively add the edges induced by the assignment of a given b_i. More precisely, if at a given step we have e cycles and p remaining edges, we know that $c \leq e+p$ since each remaining edge can create at most one cycle. Therefore, we can use the following lower bound in a branch-and-bound strategy:

$$d_{inv}(O,\widehat{O}) \geq n+1-e-p.$$

Ordered forest: Generalization to an ordered forest (F,O_R,O_L) is straightforward. Indeed, let $(T_1,T_2,\ldots,T_t)$ be the trees in F ordered according to the order O_R of their external roots. From Definition 3.5 and the discussion above, it is clear that $(F,O_R,\widehat{O}_L)$ is a duplication forest iff $\widehat{O}_L$ is the concatenation of the t orders $(\widehat{o}_1,\widehat{o}_2,\ldots,\widehat{o}_t)$ respectively defined by an assignment of the binary variables in $T_1,T_2,\ldots,T_t$, and for each $1 \leq j \leq t$, all the genes belonging to $\widehat{o}_j$ have the same sign as r_j. Consequently, we can enumerate the orders $\widehat{O}_L$ as above and the same bound can be used (see Figure 3.3 for an example).

3.6 The median problem

To formally define the median problem, we need to extend the notion of an ordered forest (Definition 3.3) by allowing the orders to be defined only on the leaves or only on the external roots of the trees. A *leaf-ordered forest* will be denoted as (F_{RL},R,O_L) and a *root-ordered forest* as (F_{RL},O_R,L).

The median problem is formulated as follows. Given a root-ordered forest (F_{AM},O_A,M) and two leaf-ordered forests (F_{MB},M,O_B) and (F_{MC},M,O_C) (M is the set of ancestral genes generating both B and C), the goal is to find an order O_M minimizing

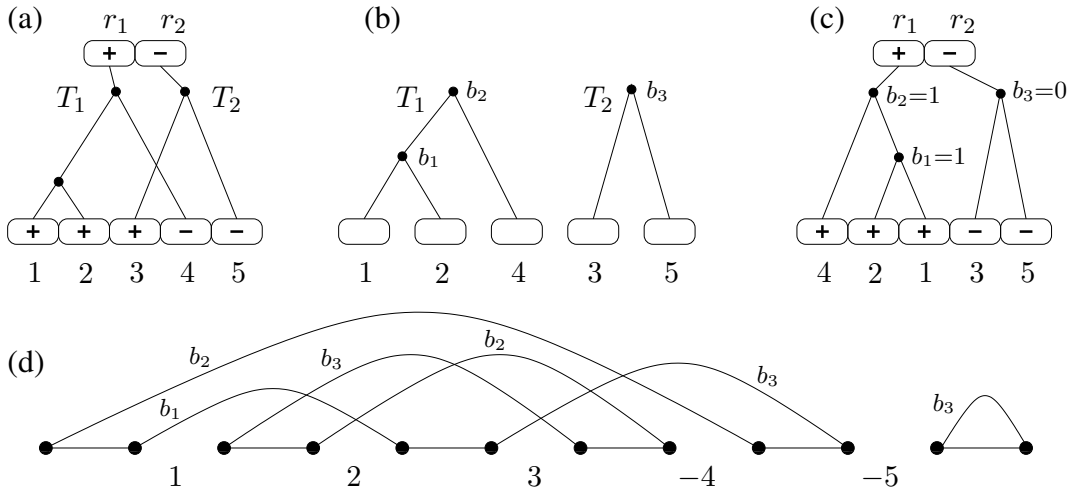

Figure 3.3: (a) The ordered forest corresponding to the branch (M,C) of the tree in Figure 3.2 ($F = \{T_1,T_2\}, O_R = (r_1,-r_2), O_L = (1,2,3,-4,-5)$). (b) The gene trees T_1 and T_2, with an arbitrary left/right orientation of the children at each internal node. (c) The duplication forest $(F, O_R, \widehat{O}_L = (4,2,1,-3,-5))$ induced by an assignment of the b_i variables. (d) The breakpoint graph of $\widehat{O}_L$ and O_L, with each curved edge labeled by the b_i inducing it, according to this assignment sequence: $(b_1 = 1, b_2 = 1, b_3 = 0)$.

the *median score*:

$$S(O_M) = DI(O_A,O_M) + DI(O_M,O_B) + DI(O_M,O_C)$$

3.6.1 A branch-and-bound algorithm

To avoid considering each of the $2^n n!$ possible orders O_M, where n is the number of genes in M, we consider a branch-and-bound strategy. The idea is to compute a lower bound on $S(O_M)$ as we progressively extend the prefix O_M^* of a candidate median O_M. This is justified by the following property.

Property 3.8. *Let (F_{RL}^*, O_R^*, O_L^*) be an ordered forest obtained from (F_{RL}, O_R, O_L) by removing a tree rooted at the last element of O_R, or the leaf corresponding to the last element of O_L. Then:*

$$DI(O_R^*, O_L^*) \leq DI(O_R, O_L)$$

From the property above, it follows that:

$$S(O_M) \geq S(O_M^*) = DI(O_A^*, O_M^*) + DI(O_M^*, O_B^*) + DI(O_M^*, O_C^*)$$

An exact branch-and-bound strategy for solving the median problem is sketched below. **Algorithm BBM-DI** (Branch-and-Bound for the Median with DI distance):

1. Consider an initial candidate O_M. Define the empty orders O_M^*, O_A^*, O_B^* and O_C^*.
2. Add a gene $\pm g_M \in M$ to the end of O_M^*. Then, insert the descendants of g_M in O_B^* and O_C^* according to their positions and signs in O_B and O_C. Moreover, if g_M is the descendant of a gene $g_A \in A$ that is not yet in O_A^*, insert g_A in O_A^* according to its position and sign in O_A.
3. If $S(O_M^*) < S(O_M)$:
   - If $S(O_M^*)$ contains less then n genes, then return to Step 2.
   - Else $O_M \leftarrow O_M^*$.
4. Backtrack to Step 2 and consider another gene g_M (or sign) for the last position of O_M^*. When all the genes have been considered, backtrack one position left. When all the positions have been tried, stop and return O_M.

This branch-and-bound approach can be used with medians containing up to a dozen of genes (see Execution time in Section 3.7.1). For larger instances, we next present a fast and simple heuristic which yields good approximations when the number of inversions is low.

3.6.2 A simple heuristic for the median problem

The idea is to consider an initial order and optimize the median score locally by successive applications of *transposition* or *transversion*[4] on that order. It is similar to

[4] A transposition followed by an inversion.

the exact algorithm of [108] except that our neighborhood is different, and we keep only the best candidates at each step. A local optimum is reached when no move can improve the median score. The algorithm is sketched bellow.

Algorithm LSM-DI (Local Search for the Median with the DI distance):

1. Consider an initial candidate median O_M. Set $S_{min} \leftarrow S(O_M)$.
2. For each of the $O(n^3)$ neighbors O_i of O_M:
   (a) Compute $S(O_i) = DI(O_A, O_i) + DI(O_i, O_B) + DI(O_i, O_C)$.
   (b) If $S(O_i) < S(O_M)$, then push O_i on the priority queue. Moreover, if $S(O_i) < S_{min}$, then set $S_{min} \leftarrow S(O_i)$.
3. As long as the priority queue contains an order O_i such that $S(O_i) = S_{min}$, set $O_M \leftarrow O_i$, remove O_i and return to Step 2.
4. Output O_M.

3.6.3 Getting the initial orders

The success of the above methods depends strongly on the choice of the initial candidates O_M. Our solution is to use a greedy version of the algorithm described in Section 3.5.2, but generalized to the *whole* reconciled tree G. More precisely, for each duplication node v of G, we set b_i to the value that maximizes the total number of cycles in the breakpoint graphs of the m extant genomes. Once all the b_i are defined, it is straightforward to obtain the orders in the ancestral genomes.

3.7 Results

3.7.1 Simulated data

Execution time

We measured the execution time of our general method for inferring ancestral orders (Section 3.4) using either the branch-and-bound (**BBM-DI**) or the heuristic (**LSM-DI**) for solving the median problem. Algorithms were implemented in C++ and run on a typical Linux workstation.

The ordered gene trees were obtained by simulating DLIS-histories consistent with balanced species trees with 2 or 4 leaves. The number of genes in the resulting genomes (extant or ancestral) depends uniquely on their depth in the species tree. Starting from the root which contains a unique ancestral gene, this number becomes respectively n, $\lfloor 3n/2 \rfloor$ and $\lfloor 9n/4 \rfloor$ as we reach depth 1, 2 and 3 (depth 3 applies only to species trees with 4 leaves). Inversion events are distributed evenly among the branches of the species trees and their cutting-points are chosen randomly.

Results are presented in Figure 3.4. We observe that the execution time of BBM-DI depends significantly on the number of inversions and rapidly becomes impractical. In contrast, LSM-DI can be used on relatively large datasets within reasonable time (100 seconds on average for a median of 30 genes with 12 inversions).

Algorithms Accuracy

We measured the accuracy of our general method for inferring ancestral orders on simulated data, using either the BBM-DI or LSM-DI for solving the median problem. Ordered gene trees were obtained as described above.

Accuracy is evaluated based on two criteria: the inferred number of inversions, and the inferred gene orders. Evaluation of the gene order is based on the percentage of adjacencies shared between the inferred order and the actual one. An inferred adjacency

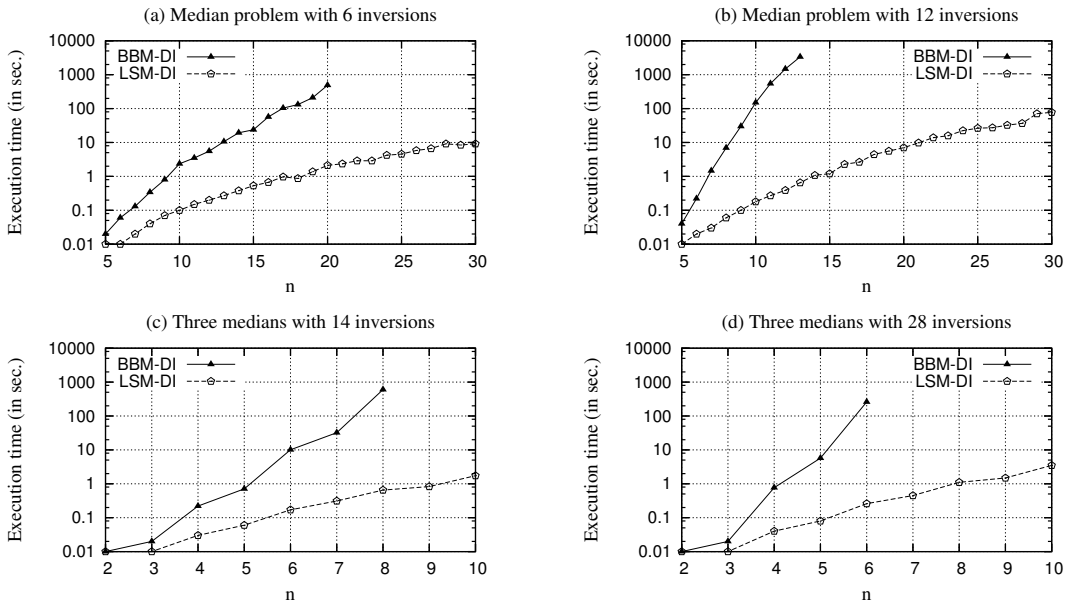

Figure 3.4: Average execution time in seconds on simulated data (50 replicates). (a and b) One ancestral genome with n genes and two extant genomes each containing $\lfloor 3n/2 \rfloor$ genes. (c and d) Three ancestral genomes containing respectively n, $\lfloor 3n/2 \rfloor$ and $\lfloor 3n/2 \rfloor$ genes, and four extant genomes each containing $\lfloor 9n/4 \rfloor$ genes.

(a,b) is shared iff (a,b) or $(-b,-a)$ is in the actual order.

We observe in Figure 3.5(a and b) that the number of inversions inferred by LSM-DI is very close to the global minimum found by BBM-DI. However, the probability that this global minimum corresponds to the reality decreases when the actual number of inversions increases in the DLIS-history. The same is observed for the percentage of correctly inferred adjacencies (see Figure 3.5(c and d)).

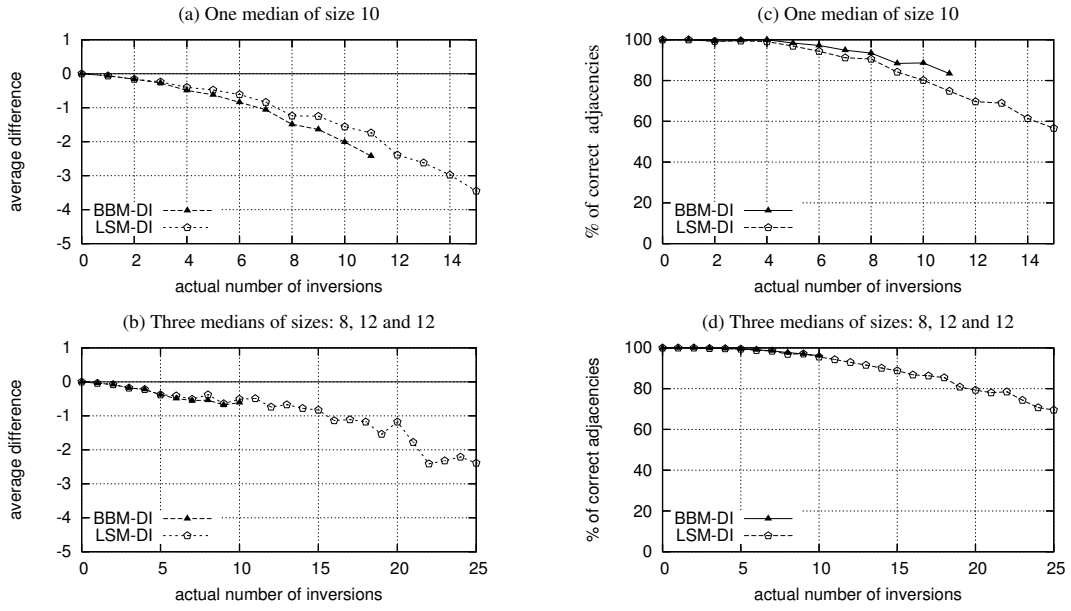

Figure 3.5: Comparison between BBM-DI and LSM-DI (100 replicates). (a and b) Average difference between the inferred number of inversions and the actual one (*inferred minus actual*). (c and d) Percentage of shared adjacencies between the inferred order and the actual one.

Effect of gene losses

To evaluate the effect of gene losses on the accuracy of LSM-DI, we generated appropriate DLIS-histories using a protocol similar to the one described above. Gene losses were distributed randomly among the branches of the species trees. Results are shown for correctly reconciled trees (80% and 60% respectively for 8 and 16 gene losses) in Figure 3.6. We see that gene losses have very little effect on the accuracy of our heuris-

tic when the reconciled gene tree is correct.

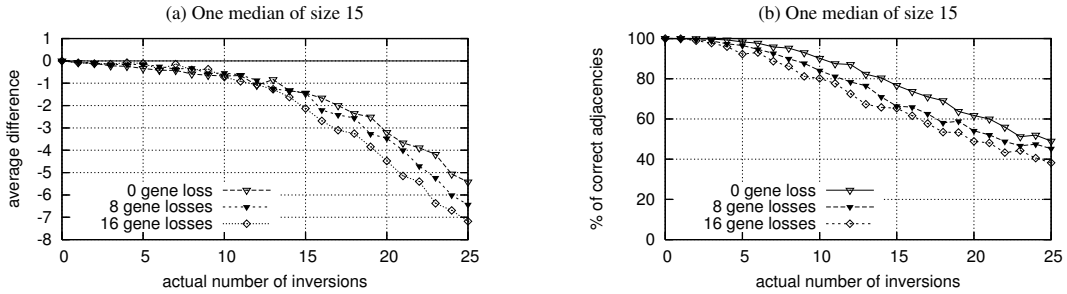

Figure 3.6: Accuracy of LSM-DI on ordered gene trees resulting from DLIS-histories with different numbers of gene losses (500 replicates). (a) Average difference between the inferred number of inversions and the actual one (*inferred minus actual*). (b) Percentage of shared adjacencies between the inferred order and the actual one.

Effect of double duplication (model deviation)

Recall that our DLIS model allows only simple tandem duplications. To measure the robustness of our inference method against model deviations, we simulated the evolution of orthologous clusters with a limited number of *double duplications*[5] (DD). As expected, we observe in Figure 3.7 that LSM-DI largely overestimates the number of inversions when DD are introduced, especially when few inversions really occurred (one DD can produce as much as 3 *false* inversions). However, the effect on the percentage of correctly inferred adjacencies is much smaller.

3.7.2 Application on biological data

The olfactory receptor (OR) gene family contains several hundred members in mammalian genomes, scattered in about 50 genomic clusters. We used our general method with LSM-DI to infer ancestral gene orders for one of these clusters, which is located on chr14@21.2 in the human genome. Four orthologous clusters were used in

[5] A double duplication simultaneously copies two adjacent genes as a single unit. For example, $O^k = (g_1, g_2, g_3, g_4)$ that becomes $O^{k+1} = (g_1, g_2, g_3, g'_2, g'_3, g_4)$.

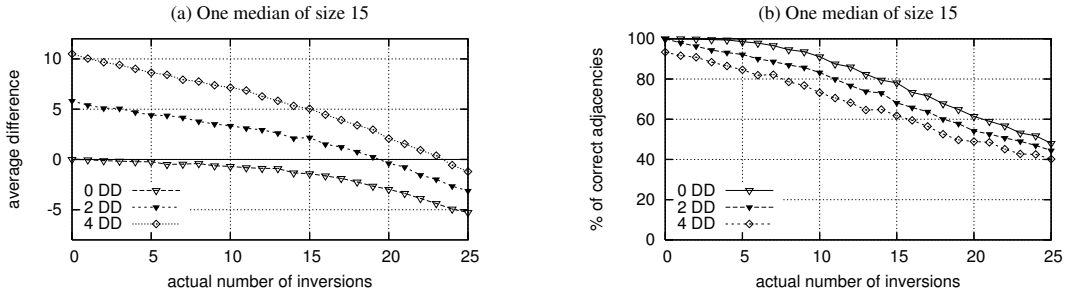

Figure 3.7: Accuracy of LSM-DI on ordered gene trees generated from DLIS-histories with different numbers of double duplications (500 replicates). (a) Average difference between the inferred number of inversions and the actual one (*inferred minus actual*). (b) Percentage of shared adjacencies between the inferred order and the actual one.

our study: human chr14@21.2; rat chr15@27.9; mouse chr14@47.5; opossum scaffold_19262@4.7. Protein sequences, gene orders and clusters orthology were all obtained from CLIC#35 in the HORDE database [2]. Human OR6Y1 gene was used as an outgroup. Sequences were aligned with ClustalW [115] and the gene trees with the largest posterior probability were obtained with MrBayes [97], using the Jones-Taylor-Thornton substitution matrix [54] and $500,000$ MCMC iterations.

The 16 most probable trees have a cumulative posterior probability of 0.8. For each of them, we obtained a reconciled gene tree with the RECONCILE software [101] and used our general algorithm to infer ancestral gene orders and the corresponding number of inversions. The most parsimonious DLIS-histories were obtained with the fourth $(p = 0.09)$ and the sixth $(p = 0.05)$ most probable trees returned by MrBayes. Both involve a single inversion and no gene loss. Other trees involve 4.7 gene losses and 1.8 inversions on average. The fourth tree is presented in Figure 3.8. According to this tree, a unique inversion event occurred before the divergence between eutheria and marsupialia. We point out that this scenario differs slightly with the one we obtained previously by considering only the human and rat clusters which involves an additional gene loss [60].

This simple application gives an example of a TAG cluster which is very likely to have evolved in agreement with our model of evolution.

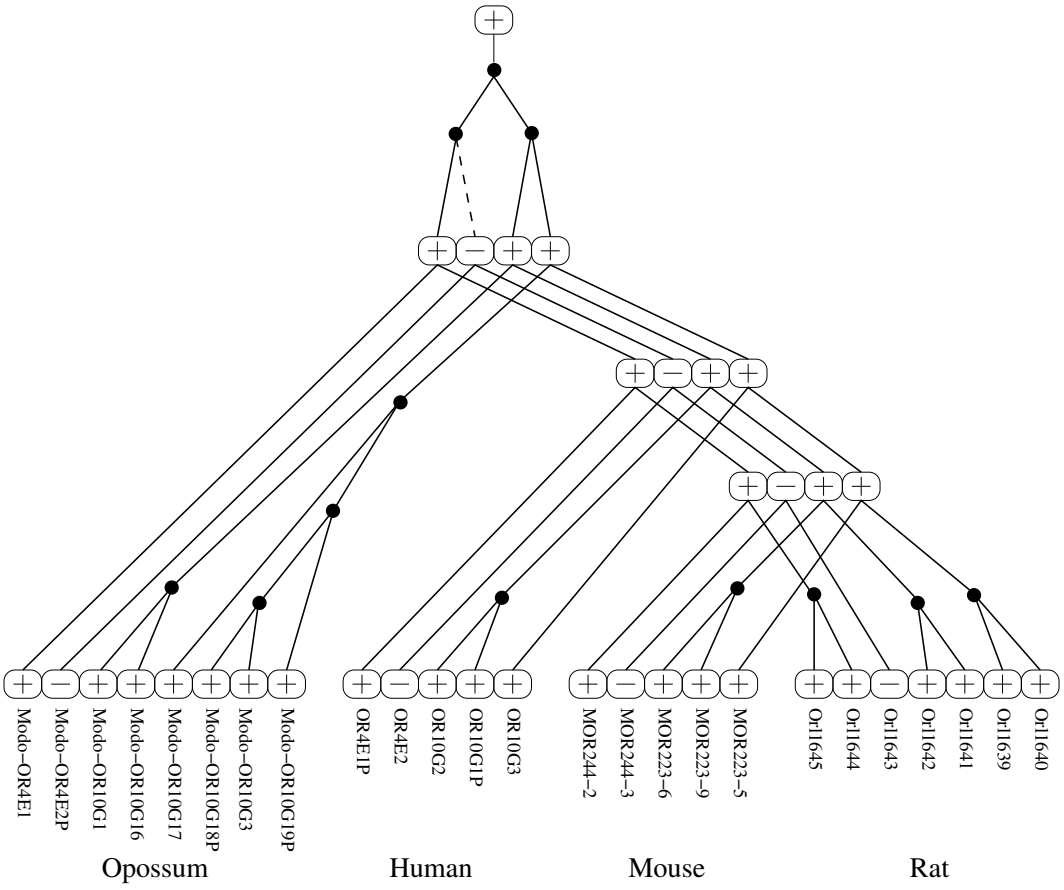

Figure 3.8: The ancestral gene orders inferred by our general method using LSM-DI on the orthologous clusters of CLIC#35. Transcriptional orientations are indicated by signs. The unique inversion event is indicated by the dashed edge. No gene loss was inferred by the reconciliation process.

3.8 Conclusion

We have presented a general framework for studying the evolution of tandemly arrayed gene families in multiple genomes. It is the first formal approach to integrate inversion and speciation events in a tandem duplication model of evolution.

Our study has been placed in the context of a known species tree. In the case of an unknown species tree, an alternative method for constructing the preliminary reconciled tree should be considered. Different methods have been developed in the literature based on different measures: the *duplication cost model* [73], the *mutation cost model* [73] and the *minimum loss model* [19].

The methods we presented make it possible to infer ancestral gene orders minimizing (locally) the number of inversions for *a given reconciled tree*. However, this tree is not guaranteed to provide the minimum number of inversions for *any* DLIS-history compatible with the species tree. Finding a DLIS-history of minimum inversions remains an open problem.

We point out the difficulty of measuring the accuracy of the phylogenetic methods used to infer the gene trees, especially for TAG families. Events such as gene conversions and unequal crossover can create "mosaic" genes that share more than one ancestor, and pseudogenization is a frequent process. Different strategies could be used to cope with these problems. For example, regions subject to gene conversions could be identified and excluded from the phylogenetic analysis, and the gene contexts could be considered. Pseudogenes could also be treated separately with more appropriate methods and models, or ultimately discarded from the analysis. Despite these efforts, it would remain difficult to infer the correct gene tree for several TAG families.

In this context, the minimum number of inversions for a TAG family could be used as an additional criteria for the comparison of different candidate gene trees [62]. Here again, results should be interpreted carefully since the actual model is limited to simple duplications. Although they are believed to be predominant, multiple duplications also

occur in TAG evolution.

An important improvement would thus be the extension of our model to multiple duplications. This poses many challenges since inferring a tandem duplication tree with multiple duplications and gene losses remains an open problem, even when inversions are not taken into account and only one species is considered.

Acknowledgments

This work was supported by grants from the *Fonds Québécois de la Recherche sur la Nature et les Technologies* (D.B. and N.E.M.), the Natural Sciences and Engineering Research Council of Canada (N.E.M.) and the Canadian Institutes of Health Research (M.L.).

3.9 Contribution des auteurs

Mathieu Lajoie et Nadia El-Mabrouk ont écrit le manuscrit. **Mathieu Lajoie** et Denis Bertrand ont conçu les algorithmes. Denis Bertrand a implémenté l'algorithme principal. **Mathieu Lajoie** a conçu et implémenté le simulateur de données. Denis Bertrand et **Mathieu Lajoie** ont conçu et réalisé les expérimentations sur les données simulées. **Mathieu Lajoie** a conçu et réalisé les expérimentations sur les données biologiques. Denis Bertand a réalisé les figures et participé à la révision du manuscrit. **Mathieu Lajoie** et Denis Bertrand sont considérés comme les co-premiers auteurs. Les résultats préliminaires de cet article ont été présentés par **Mathieu Lajoie** lors du *Fifth RECOMB International Workshop on Comparative Genomics* [60], le 16 septembre 2007 à San Diego aux États-Unis.

CHAPITRE 4

INFERRING THE EVOLUTIONARY HISTORY OF GENE CLUSTERS FROM PHYLOGENETIC AND GENE ORDER DATA

Mathieu Lajoie[1], **Denis Bertrand**[2] **et Nadia El-Mabrouk**[3]

Article publié dans *Molecular Biology and Evolution* [61]

[1]DIRO, Université de Montréal, Montreal (QC), Canada.
[2]DIRO, Université de Montréal, Montreal (QC), Canada.
[3]DIRO, Université de Montréal, Montreal (QC), Canada.

Abstract

Gene duplication is frequent within gene clusters and plays a fundamental role in evolution by providing a source of new genetic material upon which natural selection can act. While classical phylogenetic inference methods provide some insight into the evolutionary history of a gene cluster, they are not sufficient alone to differentiate single from multiple gene duplication events, and to answer other questions regarding the nature and size of evolutionary events. In this paper, we present an algorithm that infers a set of optimal evolutionary histories for a gene cluster in a single species, according to a general cost model involving variable length duplications (tandem or inverted), deletions and inversions. We applied our algorithm to the human olfactory receptor and protocadherin gene clusters, showing that the duplication size distribution differs significantly between the two gene families. The algorithm is available through a web interface at **http://www-lbit.iro.umontreal.ca/DILTAG/**

4.1 Introduction

As the genome sequences of many eukaryotes were sequenced and analyzed, it became clear that gene duplication plays a fundamental role in evolution, through the acquisition of new and complementary functions among gene families [127]. One of the most common mechanisms leading to gene duplication is unequal crossing-over during meiosis causing *tandem duplications*. As this phenomenon is favored by the presence of repetitive sequences, a single duplication can induce a chain reaction leading to further duplications, eventually creating large repetitive regions. When those regions contain genes, the result is a Tandemly Arrayed Gene (TAG) cluster: a group of paralogous genes that are adjacent on a chromosome. TAGs represent about 15% of all human genes [106] and are involved in a variety of functions such as binding and receptor activities. In particular, the olfactory receptor genes constitute the largest multigene family in vertebrate genomes, with several hundred genes per species [41]. Other examples of

TAG families include the APOBEC3 genes [65], the immunoglobulin and T-cell receptor genes [3] and the zinc finger genes [105].

In the absence of other evolutionary mechanisms, a TAG cluster containing n duplicated sequences would appear in a dotplot as a rectangle filled with $2n+1$ parallel stripes. Moreover, as the $5'-3'$ orientation is preserved through tandem duplication, all the genes in a TAG cluster would be on the same DNA strand. However, this is rarely observed in practice since most TAG clusters are affected by other events such as segmental deletion (the counterpart of a tandem duplication) and inversion. While the former can lead to gene loss, the latter can affect their order and transcriptional orientations. An alternative mechanism to explain the presence of genes on both DNA strands within a TAG cluster is *inverted duplication* [109], which has been observed in many cases.

Deciphering the evolutionary history of a TAG cluster is important, not only to understand the functional specificity of each gene inside the cluster, but also to provide new insights into the mechanisms of gene duplication, and to answer several questions regarding the nature and size of duplication and other evolutionary events. In most biology-oriented studies, evolutionary relationships between genes of a given cluster are deduced from a gene tree obtained by using a standard phylogenetic inference method. When data is available for several species, the obtained gene tree can be compared with the species tree, allowing to estimate the number of gene gains and losses in the different lineages. However, this number does not univocally correspond to the real number of evolutionary events, as multiple genes can be duplicated (or lost) in a single evolutionary event. Moreover, the gene tree alone provides no additional clues about the underlying evolutionary mechanisms. This motivates the need to develop novel algorithmic methods to infer histories in which evolutionary events are explicitly determined.

Recently, [133] considered self-alignment data and percent identity thresholds to represent a gene cluster as an ordered sequence of signed atomic segments, and developed a stochastic algorithm for reconstructing its evolutionary history using a model accounting for general segmental duplications (not necessarily in tandem) and deletions. Their

model was then extended in [5] for the study of orthologous TAG clusters in different species, and a Bayesian version has been implemented by [116].

While these methods are useful for inferring recent evolutionary events, they are less appropriate for longer time scales, as alignment of the non-functional regions becomes impossible due to mutations (such as indels and substitutions) continuously affecting each duplicated segment. An alternative and complementary approach is to focus on the genes present in the cluster. Indeed, as coding regions are usually characterized by lower evolutionary rates than surrounding non-coding regions, they provide a phylogenetic signal that can be used in combination with gene order data to infer evolutionary histories in which duplication events are explicitly determined, assuming they result solely from unequal crossing over. Based on this idea, a number of studies have considered the problem of reconstructing the *tandem duplication history* of a TAG cluster [*e.g.*, 35, 10, 111, 33, 130, 12]. However, none of the proposed algorithms account for deletion and inversion events, which strongly limits their applicability to biological data. This led us to propose an algorithm that finds the minimum number of inversions involved in the evolutionary history of a TAG cluster in a single species, assuming single-gene tandem duplications [62]. We then extended the model in [13] for the study of orthologous TAG clusters in different species. However, assuming single-gene duplication, while allowing for an exact algorithmic solution, presents an substantial limitation to its applicability.

In this paper, we describe a heuristic algorithm that seeks a set of optimal evolutionary histories for a TAG cluster in a single species, allowing for tandem duplications, inverted tandem duplications, inversions and deletions, each event involving one gene or a set of adjacent genes. To our knowledge, it is the first algorithmic study to explicitly account for this broad range of evolutionary events in a finite sites model of evolution (see [74] for an infinite sites model of genome evolution).

As in most related studies, we assume that the phylogenetic signal is not completely obscured by gene conversion and that a reliable gene tree can be obtained, using a classical phylogenetic inference method, as input to our algorithm. This assumption seems

realistic for most TAG clusters evolving according to a birth-and-death model of evolution [81].

We also assume that genes are duplicated entirely, *i.e.* that no unequal crossing-over occurs within the gene boundaries. Although this assumption can be violated in real data, the effect on our method is limited and could be worked out in future developments, as we discuss at the end of this paper.

We have applied our algorithm to various simulated datasets, showing that it can infer the number and size distribution of the duplication events with good accuracy. We have also shown that recent evolutionary events are predicted with higher accuracy than more ancient ones. An application to biological data suggests that the human protocadherin gene clusters evolved through successive tandem duplications and deletions, with as many as 45% of the duplications being multiple duplications (*i.e.* duplications involving more than a single gene). In contrast, a survey of 17 human olfactory gene clusters indicates that multiple duplications (in tandem or inverted) are less frequent in their evolution, representing approximately 20% of the duplication events.

4.2 The Evolutionary Model

4.2.1 The classical tandem duplication model

Our evolutionary model is an extension of the one introduced by [35] which considers only tandem duplications resulting from unequal crossing-over during meiosis. According to this model, the locus grows from a single gene through a series of consecutive duplications giving rise to a sequence of n adjacent copies of paralogous genes *having the same transcriptional orientation*. The main features of this model is illustrated by the example of Figure 4.1, which depicts a duplication history obtained by [33] for the 9 variable genes of the human T cell receptor Gamma (TRGV) locus. In particular, the model allows for the simultaneous duplication of several consecutive genes in a single evolutionary event (*e.g.* the last one in the TRGV history). When a gene tree is available

for a set of extant sequences that have evolved solely through tandem duplications, a simple greedy algorithm [33] can reconstruct a corresponding duplication history. However, even for a set of sequences that have evolved through unequal crossing-over, gene losses can prevent the existence of such a history. Moreover, many gene clusters contain genes in both transcriptional orientations, which reflects the occurrence of other mutational events during evolution, such as inversion or inverted duplication.

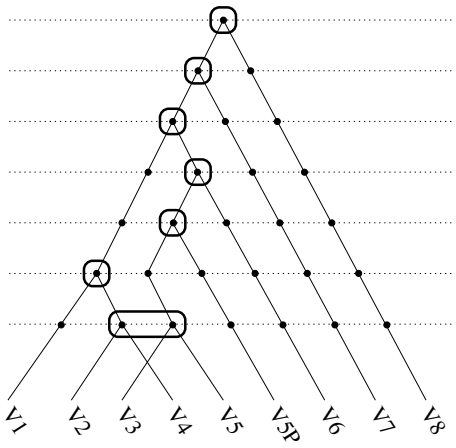

Figure 4.1: Duplication history of the 9 variable genes of the human T cell receptor Gamma (TRGV) locus [33]. White boxes depict duplication events. The last event is a double duplication.

4.2.2 An extended model

In this section we propose to extend the above model by also considering deletion, inversion, and inverted duplication events.

Formally, a gene cluster is represented as an *ordered gene tree* (T, O), where T is a rooted binary tree representing the phylogenetic relationship among the genes, and O is a *signed order* corresponding to the genes' arrangement and transcriptional orientation on the chromosome. Below is a formal definition of the evolutionary model considered in this paper. In this definition, a *cherry* of T is a pair of leaves (l, r) separated by a single vertex, called its *root*.

Definition 4.1. *An* evolutionary history *of* $(\tilde{T},\tilde{O})$ *is a sequence of ordered gene trees* $((T_1,O_1),(T_2,O_2),\ldots,(T_h,O_h)=(\tilde{T},\tilde{O}))$*, where :*

1. T_1 *is a tree consisting of a single leaf* u*, and* $O_1=(+u)$ *or* $(-u)$*.*

2. *For* $1\leq k<h$*,* (T_{k+1},O_{k+1}) *can be obtained from* (T_k,O_k) *by applying one of the following evolutionary events:*

   (a) **Tandem-duplication:** *A sub-sequence* $(u_i,u_{i+1},\ldots,u_j)$ *of* O_k *is replaced by a sequence of new elements* $(l_i,l_{i+1},\ldots,l_j,r_i,r_{i+1},\ldots,r_j)$*, and each leaf* u_x *in* T_k*, for* $i\leq x\leq j$*, is replaced by a cherry* (l_x,r_x)*. Moreover,* l_x *and* r_x *have the* same *sign as* u_x *(see Figure 4.2 (left)).*

   (b) **Inverted-duplication:** *A sub-sequence* $(u_i,u_{i+1},\ldots,u_j)$ *of* O_k *is replaced by* $(-(l_i),-(l_{i+1}),\ldots,-(l_j),r_j,r_{j-1},\ldots,r_i)$ *or* $(l_i,l_{i+1},\ldots,l_j,-(r_j),-(r_{j-1}),\ldots,-(r_i))$*, where* l_x *and* r_x *have the* same *sign as* u_x*, with* $i\leq x\leq j$*. Moreover, each leaf* u_x *of* T_k *is replaced by a cherry* (l_x,r_x) *(see Figure 4.2 (right)).*

   (c) **Inversion:** *A sub-sequence* $(u_i,u_{i+1},\ldots,u_j)$ *of* O_k *is replaced by* $(-(u_j),-(u_{j-1}),\ldots,-(u_i))$ *and* T_k *remains unchanged.*

   (d) **Deletion:** *A sub-sequence* $(u_i,u_{i+1},\ldots,u_j)$ *of* O_k *is deleted, and the corresponding leaves (genes) are removed from* T_k *(each removed gene corresponds to a gene loss).*

To each event of a given type t acting on a subsequence $(u_i,u_{i+1},\ldots,u_j)$ of O, we associate a cost $C_t(n)=\alpha_t+(n\times\beta_t)$, where $n=j-i+1$ is the *size* of the event, and $\alpha_t>0$, $\beta_t>0$ are constants chosen to reflect the probability of each type of event. The cost of an evolutionary history is simply the sum of the costs associated with its events.

The set of evolutionary histories leading to a given ordered gene tree $(\tilde{T},\tilde{O})$ can be formally represented as the set of paths between (T_1,O_1) and $(\tilde{T},\tilde{O})$ in the *history graph* (see Figure 4.3), where vertices correspond to ordered gene trees and edges correspond

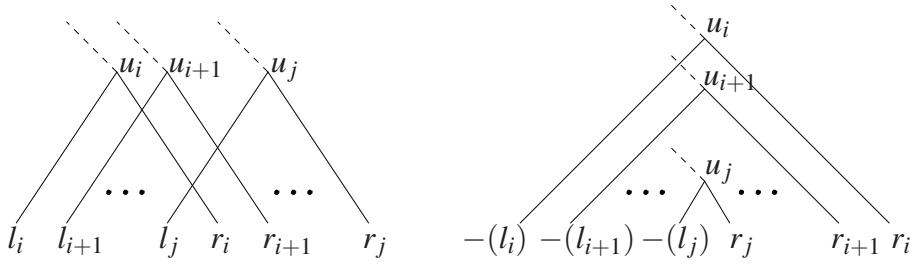

Figure 4.2: Two types of duplications acting on a subsequence $(u_i, u_{i+1}, \ldots, u_j)$ of an ordered gene tree. (Left) Tandem duplication. (Right) Inverted duplication.

to evolutionary events. More precisely, an edge from (T, O) to (T', O') is defined if and only if (T, O) can be transformed into (T', O') through one of the events defined above, and each edge is weighted by the cost of its corresponding event. Note that this graph is infinite due to the deletion edges (*e.g.* we can add an infinite number of duplications to (T, O) and delete them to re-obtain (T, O)). However, the set of *optimal* histories leading to $(\tilde{T}, \tilde{O})$ is finite and corresponds to the set of shortest paths (in term of cost) between (T_1, O_1) and $(\tilde{T}, \tilde{O})$.

4.3 Method

It is impractical to search the history graph leading to an ordered gene tree $(\tilde{T}, \tilde{O})$ in a *forward* manner. Indeed, for each vertex (T, O) of the history graph, an event can act over any subsequence of O. Thus, each vertex has $\Theta(n^2)$ *outgoing* edges, where n is the number of genes in O. An alternative approach is a *backward* search, *i.e.* starting at vertex $(\tilde{T}, \tilde{O})$ and following edges in their opposite direction. This represents an advantage since the number of *incoming duplication* edges (tandem or inverted) at a given vertex is linear in the worst case. However, as each vertex has an unlimited number of incoming deletion edges, an appropriate restriction of the search space should be considered.

In the following subsections, we successively describe: (1) the algorithms used to infer a restricted neighborhood for each ordered gene tree (T, O); (2) the heuristic used

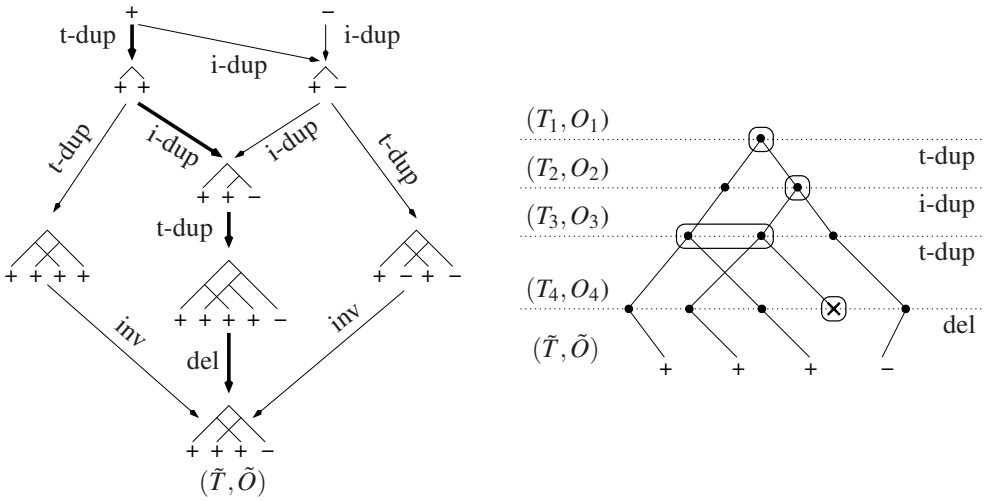

Figure 4.3: (Left) A subset of the history graph leading to $(\tilde{T},\tilde{O})$. Edges correspond to evolutionary events connecting different intermediate ordered gene trees. They are labeled "inv" for inversion, "t-dup" for tandem duplication, "i-dup" for inverted duplication and "del" for deletion. (Right) The evolutionary history corresponding to the path defined by the bold edges in the history graph.

to search the history graph and construct the sub-graph representing the set of optimal histories leading to $(\tilde{T},\tilde{O})$.

4.3.1 Computing the neighborhood of an ordered gene tree

Given an ordered gene tree (T,O) defined on n genes, the problem is to compute its (backward) neighborhood in the history graph, *i.e.* to find the set of ordered gene trees that can be reached from (T,O) through a given backward event. As noticed earlier, the main problem of the history graph is an unlimited number of incoming deletion edges at each vertex. However, as the problem is to find a most parsimonious sequence of events transforming a given ordered gene tree into a single ancestral gene, backward-deletion edges that do not allow for a subsequent backward-duplication (in tandem or inverted) can be removed from this graph, without loss of information. Therefore we only consider deletion in combination with duplication events. The algorithm used to compute the backward neighborhood of (T,O) is described below for each event:

Backward-tandem-duplication: An algorithm presented in [33] can be used to find in linear time the set of backward-tandem-duplications that can be applied to (T,O). This is done by traversing O from left to right, and for each gene, verifying whether or not it belongs to a cherry, and if so, whether it starts a new *candidate* backward-tandem-duplication, or if it can extend/complete the current one involving the gene just preceding it. The backward-tandem-duplication neighborhood is linear in space.

Backward-inverted-duplication: It is straightforward to adapt the algorithm described in the previous paragraph to find, in linear time, the backward-inverted-duplication neighborhood of (T,O). However, in contrast with backward-tandem-duplications, there are two possible orientations for the ancestral segment, and each one must be considered. The backward-inverted-duplication neighborhood is linear in space.

Inversion: The space complexity of the inversion neighborhood is $\Theta(n^2)$.

Backward-tandem-duplication-with-deletion: This operation consists of reinserting hypothetically deleted segments in (T,O) to allow a subsequent backward-tandem-duplication. More precisely, for any given cherry (w_i,w_j), with $i<j$ and $O=(w_1,w_2,\ldots,w_n)$, we seek the set of *optimal* reinsertion scenarios leading to an appropriate set of cherries according to Def.1.2.a. This can be done by considering two sets of global alignments: (1) between $(w_k,w_{k+1},\ldots,w_i)$ and $(w_{i+1},\ldots,w_{j-1},w_j)$, for $1\leq k\leq i$; (2) between $(w_i,w_{i+1},\ldots,w_{j-1})$ and $(w_j,w_{j+1},\ldots,w_k)$, for $j\leq k\leq n$. To reflect the cost of our evolutionary model, the following penalty scheme must be used:

- α_{del} is the cost of a gap opening;
- $\beta_{\text{del}}+\beta_{\text{dup}}$ is the cost of a gap extension;
- $M(w_x,w_y)=\beta_{\text{dup}}$ is the cost of a *match* between leaves w_x and w_y if they have the *same sign* in O and they form a cherry of T. Otherwise, $M(w_x,w_y)=\infty$.

The rationale is that each such alignment corresponds to a tandem duplication followed by some deletion events (see Figure 4.4 for an example). For any optimal

alignment of score s, the cost of the corresponding backward-tandem-duplication-with-deletion is simply $s + \alpha_{\text{dup}}$. We point out that this approach does not guarantee that the overall algorithm will find the true set of optimal evolutionary histories for any ordered gene tree. In particular, a single deletion spanning two adjacent tandem duplications can be considered as two independent deletions. It may also be possible that an optimal evolutionary history involves a suboptimal backward-tandem-duplication-with-deletion, though this seems unlikely when the size of the duplications and the number of deletions are expected to be small, as appears to be the case with biological data.

For any pair of sequences, the set of optimal alignments with affine gap cost can be found in $\Theta(n^2)$ time using the classical dynamic programming algorithm [119]. Since we initiate the alignment procedure in both directions for each cherry of the cluster, the overall time complexity is $O(n^3)$.

In general, the number of optimal alignments can be exponential, and thus so is the space complexity of the backward-tandem-duplication-with-deletion neighborhood. However, it is usually small in practice. As a precaution to guarantee tractability, we nevertheless limit the maximum number of deletions to six in any single duplication with deletion event.

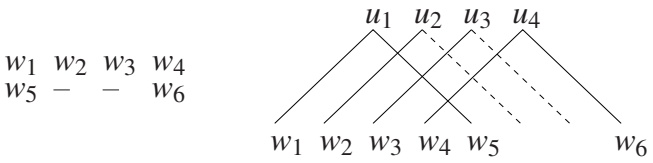

Figure 4.4: (Left) An alignment between (w_1, w_2, w_3, w_4) and (w_5, w_6). (Right) The corresponding tandem-duplication-with-deletion, involving a single deletion of size two.

Backward-inverted-duplication-with-deletion : This consists of reinserting hypothetically deleted segments in (T, O) to allow a subsequent backward-inverted-duplication. That is, for any given cherry (i, j), with $i < j$, we seek a set of optimal reinsertion scenar-

ios allowing for the creation of an appropriate set of cherries (according to Def.1.2.b). In a manner similar to above, this can be done by considering the set of global alignments between $(w_i, w_{i+1}, \ldots, w_k)$ and $(w_j, w_{j-1}, \ldots, w_{k+1})$, for $i \leq k < j$, with the difference that a match (w_x, w_y) is possible only if w_x and w_y have *opposite signs* in O, and they form a cherry of T (see Figure 4.5 for an example). In contrast with the backward-tandem-duplication-with-deletion neighborhood, the alignment procedure can be performed once for each (T, O), which leads to a time complexity of $\Theta(n^2)$. The same restriction concerning the maximum number of deletions in the alignments is applied.

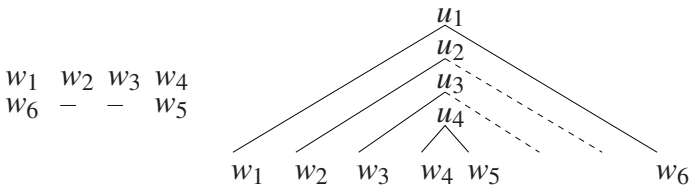

Figure 4.5: (Left) An alignment between (w_1, w_2, w_3, w_4) and (w_6, w_5). (Right) The corresponding inverted-duplication-with-deletion, involving a single deletion of size two.

4.3.2 A heuristic for the shortest path in the history graph

The set of optimal evolutionary histories leading to a gene tree $(\tilde{T}, \tilde{O})$ can be constructed by searching our search space with the well known Dijkstra's algorithm [21]. That is, we start with a min-priority queue Q_{cost} containing the neighborhood of $(\tilde{T}, \tilde{O})$, and each further step consists of (1) dequeuing the vertex with the minimum shortest-path estimate and (2) enqueueing its neighborhood. The algorithm stops when the vertex (T_1, O_1) defined on a single gene is dequeued. In practice, except for clusters restricted to a very small number of genes, Q_{cost} can grow to a point that exceeds time and memory resources. Our solution is a greedy heuristic that conserves only the most promising candidate solutions in Q_{cost}. To do so, each vertex (T, O) is additionally keyed by its number of genes in a max-priority queue Q_{genes}. If the size of Q_{cost} exceeds a predefined limit by a certain number d, we simply remove the d vertices with the highest number

of genes in Q_{genes}, as these are less likely to be on a shortest path leading to (T_1, O_1). Another restriction we use is to not consider inversion edges that reduce the number of available backward-duplications (in tandem and inverted) for a fixed T. For most size limits on Q_{cost}, this increases both the speed and the accuracy of the algorithm (data not shown).

4.4 Results

We implemented our algorithm in C++ and applied it to different simulated datasets in order to identify appropriate cost parameters (data not shown). To avoid introducing a bias, both types of duplications (tandem and inverted) must have equal costs. Moreover, deletions and inversions must be given higher costs to prevent their over representation in the inferred histories. The following cost configuration (where *t-dup* stands for tandem duplication, *i-dup* for inverted duplication, *del* for deletion and *inv* for inversion) works well for a broad range of simulated datasets:

- $\alpha_{\text{t-dup}} = 100$; $\beta_{\text{t-dup}} = 1$,
- $\alpha_{\text{i-dup}} = 100$; $\beta_{\text{i-dup}} = 1$,
- $\alpha_{\text{del}} = 500$; $\beta_{\text{del}} = 1$,
- $\alpha_{\text{inv}} = 500$; $\beta_{\text{inv}} = 1$.

Unless otherwise stated, we will use these costs.

4.4.1 Experiments on simulated datasets

In the following sections, we present various experiments devised to measure the accuracy of our algorithm. Rooted ordered gene trees were obtained by simulating evolutionary histories with different types and numbers of events. The size of each event was sampled according to a geometric distribution with parameter $p = 0.5$, truncated by

the number of genes in the cluster before the event. Given an infinite number of genes, this distribution would lead an expected event size of 2, but the actual value is smaller because the size of the first events is restricted by the size of the cluster. A simulated history containing 10 duplications leads to an average cluster size of 19 genes. For simulated histories containing deletion events, one duplication was added for each deletion in order to maintain the average cluster size over all the datasets. Clusters containing less than 10 genes were not included in the datasets. Additional results are provided as supplementary material (Figures 4.14-4.17) for parameter values of 0.8 and 0.3, corresponding respectively to an expected event size of 1.25 (average cluster size of 13) and 3 (average cluster of size of 24, upper limit fixed to 35).

The maximum size of the priority queue was set to 10,000. For each ordered gene tree, the values of interest are averaged over the set of all optimal histories inferred by our algorithm. Results are averaged over 500 replicates. For a cluster containing 19 genes, the average execution time of the algorithm is less than a second on a desktop computer running Linux.

Note that in the case of biological datasets, the ordered gene trees are obtained using classical phylogenetic inference methods, which are prone to errors. To measure the impact of such errors on our algorithm, we repeated each experiment after applying a fixed number of random nearest-neighbor-interchanges (NNIs) [110] to the simulated trees (data shown for 2 NNIs).

Estimating the number and type of the duplications events

In this section, we measure the ability of our algorithm to infer the correct number of duplication events, as well as the ratio between tandem and inverted duplication in the simulated histories. For histories containing only duplication events (50% in tandem, 50% inverted), we can see from Figure 4.6 (left) that our algorithm almost infers the exact number of duplications when the true trees are used, whereas two NNIs induce a slight overestimation. When four inversions are introduced, the overestimation is more

pronounced and increases with the total number of duplications (Figure 4.6 (right)). This can be partly explained by the insufficient size of the priority queue.

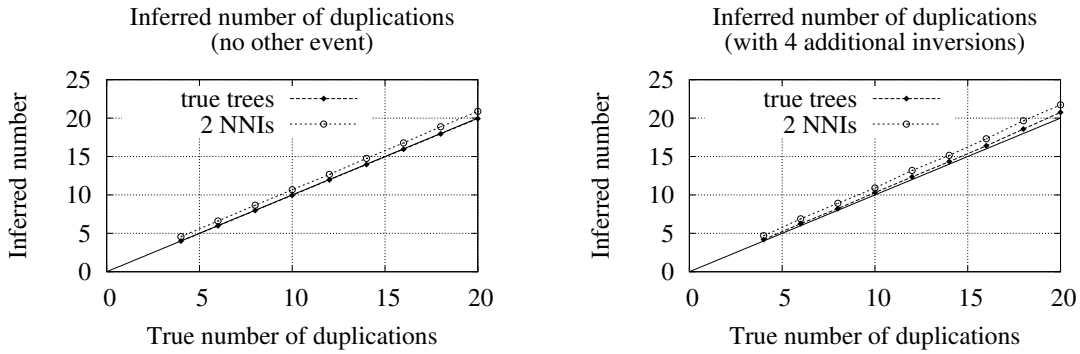

Figure 4.6: Inferred number of duplications (tandem + inverted). On each chart, an additional series shows the result obtained after performing two NNIs on the original trees. (Left) Histories with duplications only (50% tandem + 50% inverted). (Right) Histories with four additional inversions.

The ratio between inverted and tandem duplications is slightly more difficult to infer, particularly on trees perturbed by two NNIs, as we can see from Figure 4.7 (left). When two inversions and two deletions are applied to the histories, the bias can reach as much as 20% for the true trees, and 30% for two NNIs (Figure 4.7 (right)).

We argue that this is mostly a consequence of the considered evolutionary model and the parsimony assumption, rather than an inability of our algorithm to be close to obtain near-optimal solutions. In particular, a tandem duplication of size one followed by an inversion of the duplicated gene will be inferred as a single inverted duplication in an optimal history (and vice versa). Moreover, in some cases, an ordered gene tree resulting from x inverted duplications of size one can be explained by an optimal history of $x-1$ tandem duplications and one inverted duplication.

Accuracy of the inferred histories and duplication size distribution

As duplications are the major evolutionary events shaping a TAG cluster, we focus on the correctness of the inferred duplication events and their size distribution. For this

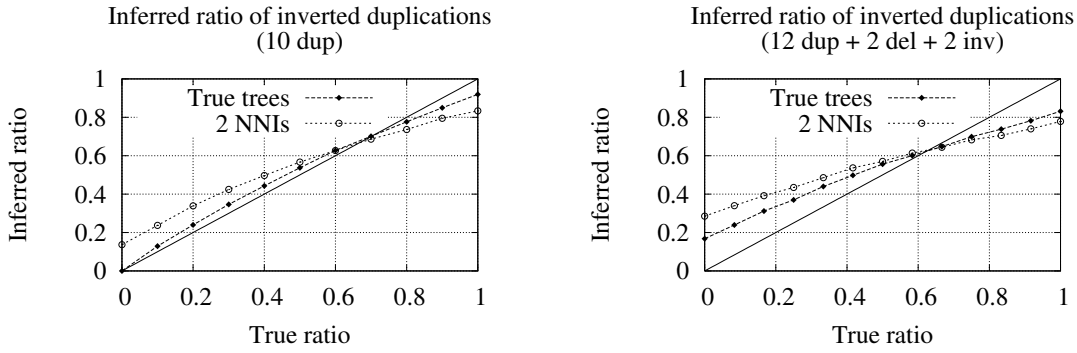

Figure 4.7: Inferred ratio of inverted duplications (inverted / (tandem + inverted)) on simulated histories. (Left) Histories with 10 duplications only. (Right) Histories with two inversions, two deletions and 12 duplications (two duplications were added to keep the same average tree size). On each chart, an additional series shows the results obtained after performing two NNIs on the original trees.

purpose, we introduce two types of error described below.

Consider a simulated history H_{true} and the corresponding inferred history $H_{inferred}$. For each predicted duplication E in $H_{inferred}$, three possibilities can be encountered:

- E is in H_{true}. In this case, the prediction is correct.
- There is a duplication in H_{true} involving exactly the same set of genes as E but in a different order, or the duplication is of a different type (tandem versus inverted). In this case, we say that we made a *partial* error.
- There is no duplication in H_{true} involving the same set of genes as E. In this case, we say that we made a *complete* error.

Note that the duplication events H_{true} that were totally obscured by subsequent deletion events are not considered. For a given history H_{true}, the error rate is simply the ratio between the number of errors and the number of predicted duplications, averaged over the set of all inferred optimal histories.

To evaluate how the error rate depends on the position of the events in the simulated history, we further introduce the notion of *depth* as follows: for a given simulated history,

the depth of an event E is the number of subsequent events acting on the descendants of E. For example, in the history depicted in the right part of Figure 4.3, the tandem duplication of size two has a depth of one since its descendants were affected by a single deletion, whereas the first tandem duplication of the history has a depth of three.

The left column of Figure 4.8 shows the error rates at different depths for simulated histories containing different numbers of events. The first two rows correspond to histories inferred from the true trees, whereas the last row corresponds to histories inferred after applying two NNIs. As expected, both types of error rates increase with the depth of the predicted events. In other words, our algorithm predicts recent evolutionary events with higher confidence than more ancient events. Hopefully, events at low depths are more numerous than events at high depths, which tends to lower the average error rates, which are respectively 0.15, 0.29 and 0.49 for the three experiments of Figure 4.8 (ordered from the top to the bottom).

Interestingly, despite some high error rates, our algorithm is very robust in inferring the duplication size distribution, as suggested by the right column histograms of Figure 4.8. Indeed, for the three simulated datasets, there is only a small difference between the true and inferred numbers of duplications of a given size. When the proportion of tandem duplications is low and 2 NNIs are applied to the gene trees (Figure 4.8 (third row)), the difference is mostly a slight overestimation $(+0.031)$ in the number of duplications of size one. However, we point out that for $p = 0.8$ (Figure 4.14 (third row) in Supplementary data), our algorithm tends to underestimate the proportion of duplications of size one in such histories (-0.030).

The results presented in Figure 4.9 are obtained with simulated histories restricted to tandem duplications and deletions. This is appropriate for TAG clusters where multiple variable exons share a set of common constant exons, such as the Pcdh cluster considered in the next section. We can see that error rates are lower when inversions and inverted duplication are not allowed. In particular, the two error types are almost equal since all the genes keep the same orientation, and that tandem duplications can no longer

be interpreted as inverted duplications. Moreover, in contrast with the case were all types of evolutionary events are considered, our algorithm tends to underestimate the proportion of duplications of size one (-0.020) for the hardest dataset when $p = 0.5$ (Figure 4.9 (third row)).

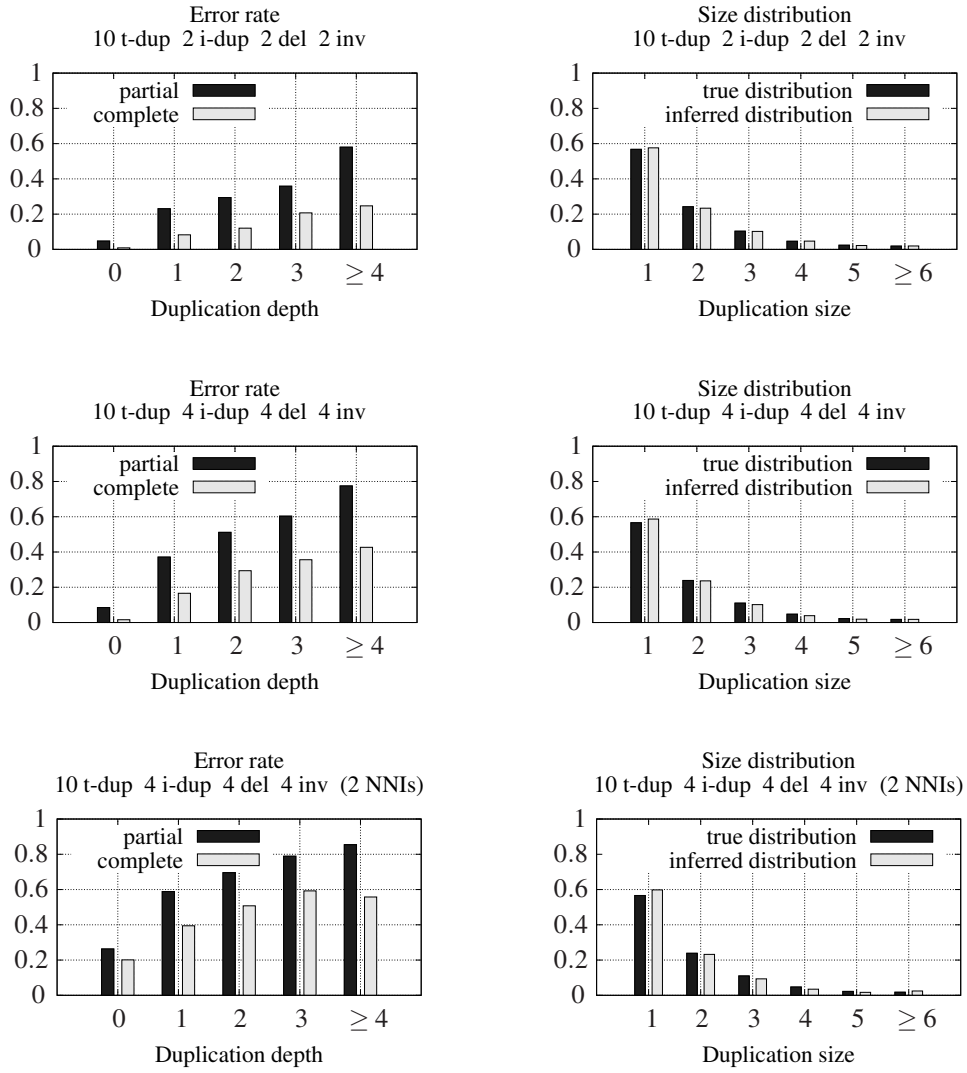

Figure 4.8: (Left) Error rates in predicting duplication event according to their depth. (Right) Comparison between the true and the inferred duplication size distribution.

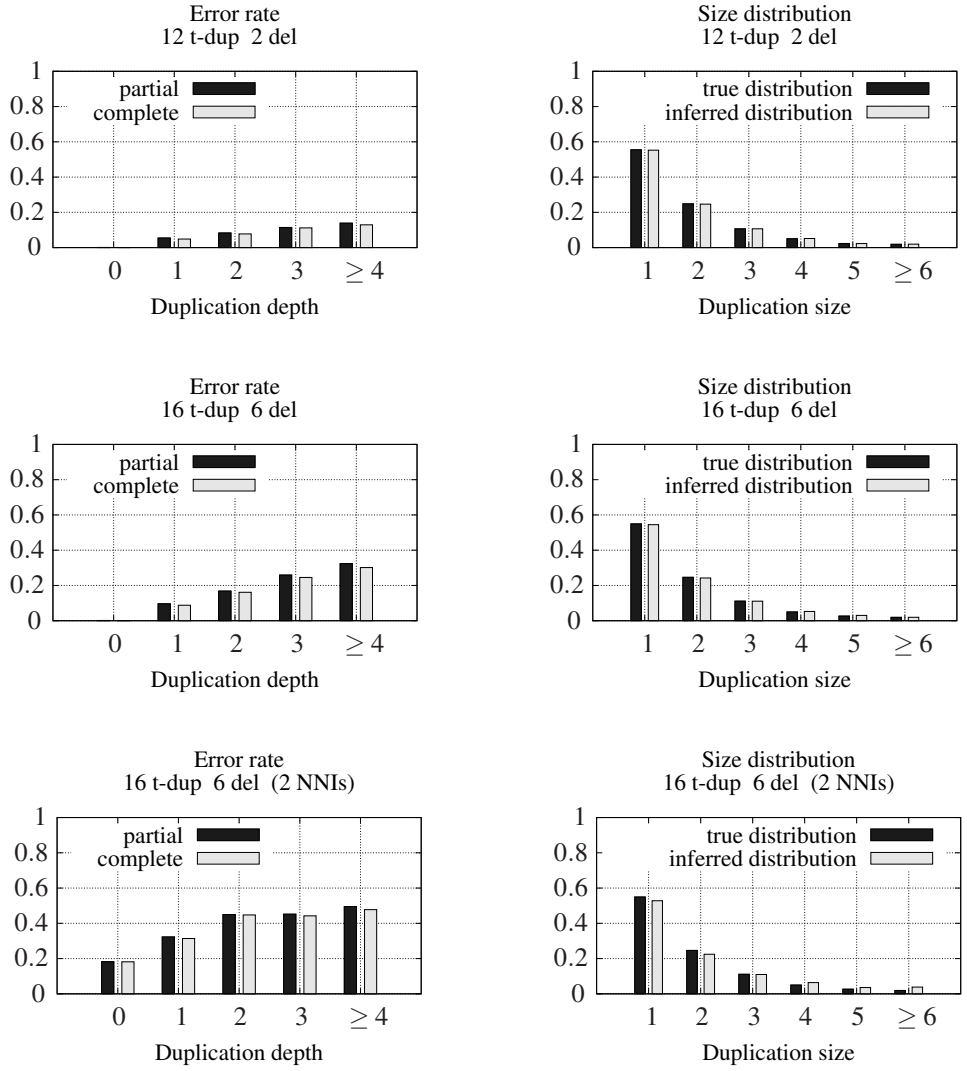

Figure 4.9: (Left) Error rates in predicting duplication event according to their depth. (Right) Comparison between the true and the inferred duplication size distribution.

4.4.2 Experiments on biological data

Many biological studies investigating the evolutionary history of a specific multigene family (or gene cluster) have invoked tandem duplication as an important evolutionary mechanism to explain their data. However, these conclusions are rarely based on a formal analysis that fully exploit phylogenetic and gene order data. Indeed, the usual argument is that functional genes that belong to the same phylogenetic clade are generally located nearby on the chromosome, and have in many cases the same transcriptional orientation [82, 1, 50]. Precise information about the individual duplication events rarely emerges from such a vague description.

In the following sections, we applied our algorithm to two gene families with the purpose of: (1) Assessing the relevance of the evolutionary model considered in the present study; (2) Inferring the duplication size distribution in these families.

The human protocadherin gene cluster

It has been hypothesized that the protocadherin gene cluster (Pcdh) is involved in the generation of synaptic complexity during brain development [121]. In human, the Pcdh cluster contains 53 tandemly arrayed genes located on chromosome 5, organized into three subclusters denoted α, β and γ [121, 85]. Each gene of the β subcluster consists of a single *variable* exon, while the α and γ subclusters each have three additional *constant* exons at their 3' end that are alternatively cis-spliced to each variable exon. This kind of genomic organization suggests a mode of evolution through tandem duplications and deletions of the variable exons in each subcluster (inversions and inverted duplications are not allowed here as they would be deleterious). Interestingly, [121] have noted that the successive alternation between Pcdhγ-a and Pcdhγ-b subfamily members on the chromosome strongly suggests that the Pcdhγ cluster evolved by duplications involving pairs of genes. However, they provided no further evidence supporting that hypothesis.

We downloaded the protein sequences for the whole Pcdh cluster from the UCSC

Genome Browser (March 2006, hg18), and aligned them with Muscle v3.6 [28]. It has been shown that the 3' ends of the variable exons were homogenized by gene conversion, while the regions encoding ectodomains 2 and 3 were subject to divergent evolution, retaining much of the phylogenetic signal [85]. Consequently, we restricted our analysis to ectodomains 2 and 3 to obtain the gene trees, using the delimitation presented in [121]. For each subcluster, a set of *unrooted* gene trees was obtained with MrBayes v3.1.2 [97], using the JTT model of evolution and performing 500,000 MCMC generations, from which 40,000 trees were sampled.

We then applied our algorithm to the first hundred most probable trees of each subcluster, considering the root positions leading to the lowest cost histories as the correct ones. To ensure that the results do not significantly depend on the choice of the cost parameters, the three following configurations were considered: $(\alpha_{\text{del}} = 500\ ;\ \beta_{\text{del}} = 1)$, $(\alpha_{\text{del}} = 250\ ;\ \beta_{\text{del}} = 250)$ and $(\alpha_{\text{del}} = 1\ ;\ \beta_{\text{del}} = 500)$.

The size distribution of the duplication events, inferred for the whole Pcdh cluster and weighted by the posterior probability of each tree, is presented in Figure 4.10. We can see that with all three configurations, there is a significant fraction of multiple gene duplications (up to 45%), most of them involving two genes. However, since the posterior probability distribution of the trees returned by Mr.Bayes is flat, our analysis relies on trees that obviously differ from the true ones. According to our results on simulated datasets with a significant fraction of multiple gene duplications (Figure 4.17 (third row)), we expect that 45% may represent a slight overestimation of the number of duplications involving multiple genes.

From a detailed inspection of the inferred histories for the γ subcluster, we noticed two specific duplications of size two that are present in all the optimal histories, for every gene tree (see one of these optimal histories in Figure 4.11). Overall, these results tend to confirm the hypothesis of [121].

Table 4.I summarizes the average number of events involved in the sets of optimal histories we obtained for each subcluster (weighted by the posterior probability of each

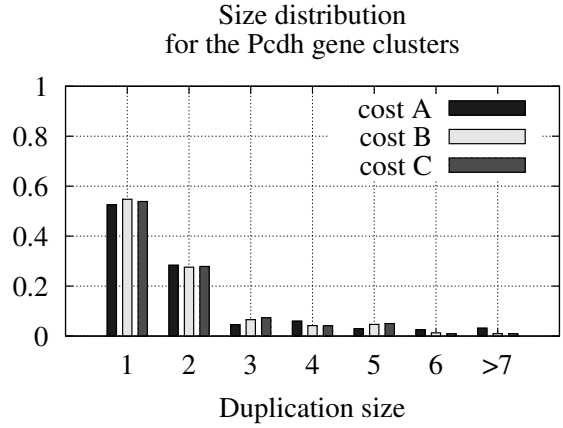

Figure 4.10: Size distributions of inferred tandem duplications for the three Pcdh gene clusters α, β and γ. Three different cost configurations were considered: A($\alpha_{\text{del}} = 500$; $\beta_{\text{del}} = 1$), B($\alpha_{\text{del}} = \beta_{\text{del}} = 250$) and C($\alpha_{\text{del}} = 1$; $\beta_{\text{del}} = 500$).

tree), using $\alpha_{\text{del}} = 500$ and $\beta_{\text{del}} = 1$. For each subcluster, we also estimated a p-value by comparing the cost of the optimal histories with the cost distribution obtained for 1,000 random gene trees of the same size (sampled from the uniform distribution). These values appears to be extremely low for the Pcdhα and Pcdhγ, which is a strong argument in favor of the considered model of evolution. The other cost configurations are omitted since they lead to nearly identical results.

Table 4.I: Estimated number of events for the three human Pcdh subclusters obtained with the following costs: $\alpha_{\text{del}} = 500$ and $\beta_{\text{del}} = 1$.

name	size	t-dup	del (loss)	p-value	cumul. prob.
Pcdhα	15	10.1	1.7 (4.3)	0.0124	0.918
Pcdhβ	18	13.5	3.6 (9.9)	0.0424	0.704
Pcdhγ	22	14.3	3.9 (9.8)	0.0002	0.731

Note.– The numbers in parentheses in the "del (loss)" column correspond to the total number of gene losses. The p-values represent the estimated probability of obtaining a history of score less or equal from a random gene tree. The numbers in the "cumul. prob." column indicate the posterior cumulative probability (according to MrBayes) of the considered genes trees (up to the 100 most probable trees).

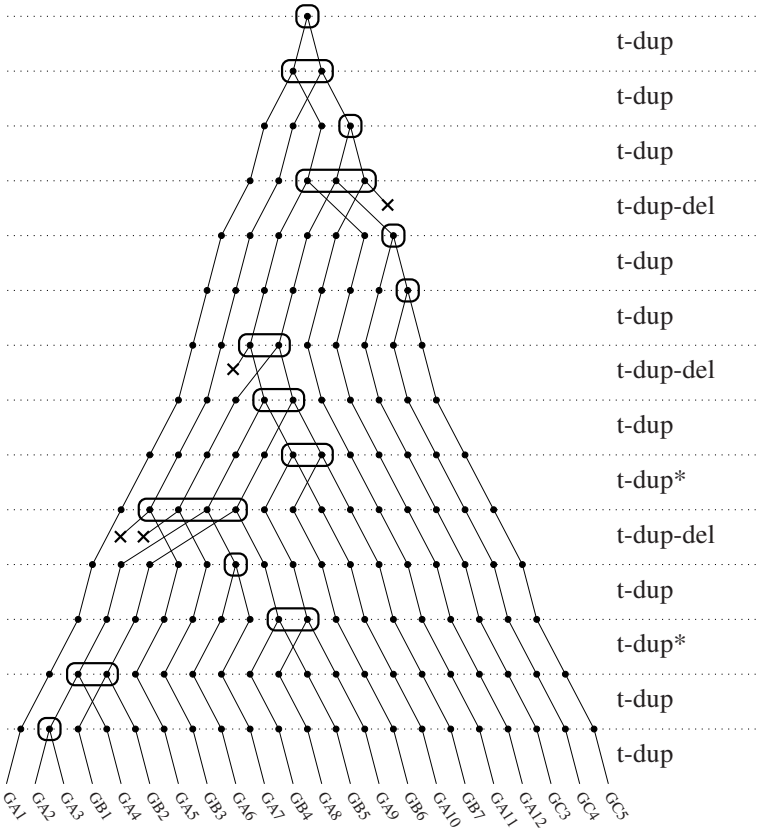

Figure 4.11: An optimal evolutionary history for the Pcdhγ gene cluster, obtained with the third tree returned by MrBayes (posterior probability of 0.036). Circled boxes correspond to tandem duplication events. Crosses correspond to gene losses resulting from deletion events. The two duplications that are marked with an asterisk are present in all the optimal histories, for every gene tree. The order of independent duplications is chosen arbitrarily.

The olfactory receptor genes

The olfactory receptor (OR) genes form the largest multigene family in mammals [124]. It comprises several hundred members per species, organized in clusters scattered throughout the genome. By allowing the recognition of a wide spectrum of odorant molecules, these G-protein-coupled receptors play a central role in food acquisition and environmental interactions [17]. With the increase in genome sequence data, many studies have focused on the identification and classification of OR genes in specific species (*e.g.* human [41, 87, 82]; dog [88] and mouse [43]). Because of their large number and their

simple and well conserved structure (each OR has a single coding exon of ≈310 aa encoding seven transmembrane domains), the OR family constitutes an ideal model for studying the dynamics of genome evolution through inter-species comparison such as human-chimp [40, 42], dog-rat [94] and human-mouse [124, 83]. These studies, based on the reconciliation between the species and gene trees, and in some case the identification of pseudogenes, have shown that the OR family has experienced extensive gains and losses in different mammalian lineages, representing an extreme form of birth and death evolution [81, 84]. However, as we mentioned previously, gains and losses do not necessarily correspond univocally to evolutionary events.

To see how the number of gene gains reflects the number of duplication events, we applied our algorithm to a set of human olfactory receptor gene clusters using a protocol similar to that of the previous section. A cluster was defined as a contiguous sequence of at least seven OR genes without interleaving non-OR genes. A perl script was used to search the *knownGene* table of the UCSC Genome Browser Database (March 2008, hg18), which led to the identification of 17 clusters containing a total of 216 genes (≈55% of the functional OR genes). For each cluster, a set of gene trees was obtained as in the previous section, with the difference that two melanocortin receptors (MC4R and MC5R) were used as an outgroup to root each tree [135, 132]. We then reconstructed the sets of optimal histories for the first hundred most probable trees of each cluster, using the three following configurations for the cost parameters: $A(\alpha_{\text{inv}} = 300\ ;\ \alpha_{\text{del}} = 700)$, $B(\alpha_{\text{inv}} = \alpha_{\text{del}} = 500)$ and $C(\alpha_{\text{inv}} = 700\ ;\ \alpha_{\text{del}} = 300)$.

As expected, the most probable gene trees tend to lead to the lowest cost histories. However, this is not always the case, as it has been observed, for example, with the OR13F1 cluster (data not shown). Indeed, its most probable gene tree ($p = 0.995$) leads to optimal histories with a relatively high cost (p-value $= 0.647$), whereas the third most probable tree leads to the lowest cost histories for this cluster (p-value $= 0.008$). This could indicate that this particular gene tree is wrong, perhaps because gene conversion or intragenic recombination events have affected the cluster.

The duplication size distribution for the 17 clusters is presented in Figure 4.12. In contrast with the Pcdh cluster, parameter values have a stronger effect on the shape of the distribution. Based on the two extreme configurations A and C, we estimate that the proportion of duplications involving a single gene lies between 75% and 90%. These results confirm and refine previous hypothesis that most tandem duplications are of size one in this particular family [82].

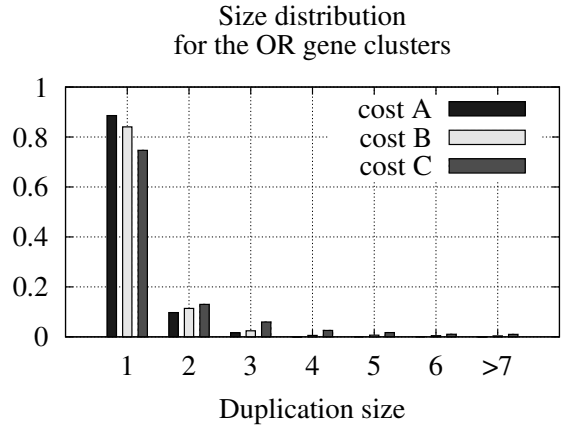

Figure 4.12: Size distributions of inferred duplication events (tandem + inverted) for the 17 olfactory gene clusters. Three cost configurations were considered: A($\alpha_{\text{inv}} = 300$; $\alpha_{\text{del}} = 700$), B($\alpha_{\text{inv}} = \alpha_{\text{del}} = 500$) and C($\alpha_{\text{inv}} = 700$; $\alpha_{\text{del}} = 300$).

In addition, we point out that inverted duplications, rather than tandem duplications followed by inversions, could explain why OR genes are often found in different transcriptional orientations among a single cluster. Indeed, the three considered cost configurations lead to optimal histories with a significant proportion of inverted duplications.

Tables 4.II-4.IV in Supplementary data summarize the average number of events involved in the set of optimal histories for each cluster, with p-values computed the same way as above. In contrast with the Pcdh subclusters, many of these p-values are high. This can be partly explained by the fact that the four types of evolutionary events are allowed here, which lowers the cost of the optimal histories obtained for the random gene trees. Moreover, there are many clusters with very few genes, which naturally leads

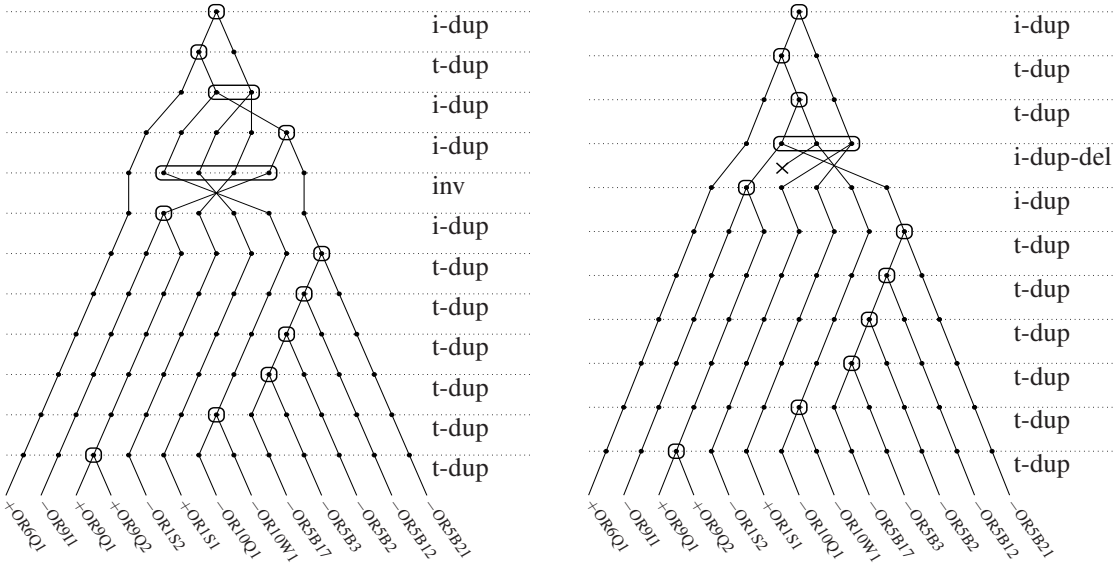

Figure 4.13: Two possible optimal histories for the OR6Q1 cluster, using the third tree returned by MrBayes (posterior probability of 0.068). (Left) History obtained with $\alpha_{\text{inv}} = 300$ and $\alpha_{\text{del}} = 700$. (Right) History obtained with $\alpha_{\text{inv}} = 700$ and $\alpha_{\text{del}} = 300$. Circled boxes illustrate the genes involved in each evolutionary event. A gene loss caused by a unique deletion is represented by a cross. The order of independent duplications is chosen arbitrarily.

to higher p-values. By comparing the tables, we can see that inversions are gradually replaced by deletions as the relative cost of the latter decreases in the inferred histories. However, the total number of duplications and their size remains similar over the three configurations. Figure 4.13 depicts two possible optimal histories (each involving 12 events) for the OR6Q1 cluster, illustrating how different types of evolutionary events can be substituted as the cost parameters change.

4.5 Conclusion

We have presented a heuristic algorithm to reconstruct a set of optimal evolutionary histories explaining the order and phylogenetic relationships among the genes of a TAG cluster. Experiments on simulated data showed that this algorithm can be used to infer the duplication size distribution with high accuracy, even for gene trees that are slightly inaccurate (*i.e.* those that differ from the true one by a few NNIs). Moreover, the last

duplication events of an evolutionary history can be inferred with good accuracy when a reliable gene tree is available for a particular family.

We used our algorithm to study the duplication size distribution among the Pcdh and olfactory receptor gene clusters, revealing a significant difference between the two gene families ($\approx$55% of single-gene duplications for Pcdh versus $\approx$80% for the olfactory receptor genes).

Another possible application of our algorithm is to use it as an additional tool to improve phylogenetic inference, by choosing among a set of gene trees having similar posterior probabilities the one leading to a most parsimonious evolutionary history. However, additional experiments would be required to measure how much weight should be given to this information, and how to choose the most appropriate cost parameter values for particular biological datasets.

An interesting extension of our algorithm would be to introduce branch length information in the considered gene trees. However, the molecular clock often does not hold for families of duplicated genes, and it could be difficult to determine how much weight to give to this type of information.

Among the four types of events considered here, duplications are the most informative as they leave a clear signature on an ordered gene tree. In contrast, inversions affect only gene orders and are difficult to infer since they can often be replaced by inverted duplications in an optimal evolutionary history, while deletions can only be detected at certain positions inside a multi-gene duplication. As comparative genomics is a more appropriate approach to infer these events, an extension of our algorithm to consider the evolution of a cluster in multiple species would be an interesting and challenging direction for future research.

In this study, we made the basic (and popular) assumption that no unequal crossing-over occurred inside the genes, as otherwise their phylogenetic relationships could no longer be represented by a unique gene tree. More precisely, each of the two segments defined by the recombination point would correspond to a different gene tree. Fortu-

nately, [93] have shown, using simulations, that phylogenetic methods tend to infer one of the two corresponding gene trees when the recombination occurs between two very similar genes, as appears to be the case most of the time in nature (≈300 bp of perfect identity is required for non-allelic homologuous recombination to occurs efficiently [95]). In such a case, our algorithm can still be applied to one of the gene trees, and the returned set of optimal histories will apply to the corresponding segment, still reflecting the evolution of the whole cluster (but somehow incompletely). Moreover, assuming the recombination points can be identified (*e.g.* by using the software presented in [77]) and the corresponding set of gene trees obtained, our algorithm could be generalized to handle a forest of gene trees and return more detailed histories.

Acknowledgments

This work was supported by grants from the *Fonds Québécois de la Recherche sur la Nature et les Technologies* (D.B. and N.E.M.), the Natural Sciences and Engineering Research Council of Canada (N.E.M.) and the Canadian Institutes of Health Research (M.L.). We thank Olivier Tremblay-Savard for the Java program allowing to draw the evolutionary histories produced by our algorithm on the web interface, and two anonymous reviewers for comments and suggestions that helped to improved our initial manuscript.

4.6 Supplementary data

Table 4.II: Estimated number of events for the 17 human olfactory gene clusters obtained with the following cost: $\alpha_{\text{inv}} = 300$ and $\alpha_{\text{del}} = 700$.

name	position	size	t-dup	i-dup	inv	del (loss)	p-value	cumul. prob.
OR2G2	chr1q44	7	2.4	3.5	1.9	–	0.613	1.000
OR2L12	chr1q44	26	14.1	7.0	6.4	–	0.001	0.896
OR5AC2	chr3q11.2	10	8.0	–	–	–	0.002	1.000
OR2F1	chr7q35	7	3.2	2.8	1.8	–	0.552	1.000
OR13F1	chr9q31.1	8	3.9	3.1	2.0	–	0.573	1.000
OR1J1	chr9q33.2	14	6.5	5.5	2.5	–	0.061	0.832
OR51D1	chr11p15.4	13	6.2	4.6	3.1	–	0.185	0.981
OR56B1	chr11p15.4	15	5.9	4.6	3.1	–	0.026	0.998
OR2AG2	chr11p15.4	8	2.7	4.2	0.3	–	0.056	1.000
OR4B1	chr11p11.2	7	3.2	2.8	0.9	–	0.226	1.000
OR4A16	chr11q11	14	4.7	4.6	2.0	–	0.002	1.000
OR5W2	chr11q11	30	15.5	10.2	6.4	0.2 (1.0)	0.003	0.208
OR6Q1	chr11q12.1	13	5.8	5.2	1.3	–	0.009	1.000
OR5AN1	chr11q12.1	7	3.1	2.8	0.3	–	0.063	1.000
OR8D4	chr11q24.1	8	2.2	4.7	0.5	–	0.057	1.000
OR9K2	chr12q13.2	14	5.7	5.1	3.3	–	0.101	0.983
OR4Q3	chr14q11.2	15	5.8	6.5	4.1	–	0.324	0.757

Note.– Each cluster is named after its leftmost gene. The numbers in parentheses in the "del (loss)" column correspond to the total number of gene losses. The p-values represent the estimated probability of obtaining a history of score less or equal from a random gene tree. The numbers in the "cumul. prob." column indicate the posterior cumulative probability (according to MrBayes) of considered genes trees (up to the 100 most probable trees).

Table 4.III: Estimated number of events for the 17 human olfactory gene clusters obtained with the following cost: $\alpha_{\text{inv}} = 500$ and $\alpha_{\text{del}} = 500$.

name	position	size	t-dup	i-dup	inv	del (loss)	p-value	cumul. prob.
OR2G2	chr1q44	7	2.2	3.6	1.6	0.4 (0.4)	0.663	1.000
OR2L12	chr1q44	26	16.2	6.0	3.4	2.1 (5.2)	< 0.001	0.896
OR5AC2	chr3q11.2	10	8.0	–	–	–	0.002	1.000
OR2F1	chr7q35	7	3.4	2.6	1.2	0.5 (0.5)	0.609	1.000
OR13F1	chr9q31.1	8	3.9	3.1	1.6	0.4 (0.4)	0.646	1.000
OR1J1	chr9q33.2	14	6.5	5.5	2.4	0.1 (0.2)	0.071	0.832
OR51D1	chr11p15.4	13	5.2	5.6	2.0	1.0 (1.1)	0.225	0.981
OR56B1	chr11p15.4	15	5.3	5.1	2.2	0.9 (0.9)	0.041	0.998
OR2AG2	chr11p15.4	8	2.7	4.2	0.3	–	0.062	1.000
OR4B1	chr11p11.2	7	3.2	2.8	0.6	0.3 (0.3)	0.243	1.000
OR4A16	chr11q11	14	4.7	4.6	2.0	–	0.002	1.000
OR5W2	chr11q11	30	13.9	11.3	2.3	3.2 (10.0)	< 0.001	0.208
OR6Q1	chr11q12.1	13	5.7	5.1	0.7	0.6 (0.7)	0.010	1.000
OR5AN1	chr11q12.1	7	3.1	2.8	0.3	–	0.069	1.000
OR8D4	chr11q24.1	8	2.1	4.8	0.3	0.2 (0.2)	0.071	1.000
OR9K2	chr12q13.2	14	5.9	4.9	2.0	1.0 (1.2)	0.084	0.983
OR4Q3	chr14q11.2	15	6.3	5.7	3.0	0.9 (2.6)	0.284	0.757

Note.– Same as Table 4.II.

Table 4.IV: Estimated number of events for the 17 human olfactory gene clusters obtained with the following cost: $\alpha_{\text{inv}} = 700$ and $\alpha_{\text{del}} = 300$.

name	position	size	t-dup	i-dup	inv	del (loss)	p-value	cumul. prob.
OR2G2	chr1q44	7	3.9	2.0	–	2.1 (3.9)	0.731	1.000
OR2L12	chr1q44	26	14.7	7.8	–	5.5 (14.2)	0.001	0.896
OR5AC2	chr3q11.2	10	8.0	–	–	–	0.002	1.000
OR2F1	chr7q35	7	6.0	–	–	1.8 (1.9)	0.451	1.000
OR13F1	chr9q31.1	8	4.3	2.7	–	2.0 (5.0)	0.579	1.000
OR1J1	chr9q33.2	14	5.9	6.5	–	2.8 (4.8)	0.065	0.832
OR51D1	chr11p15.4	13	4.4	6.6	–	3.7 (4.2)	0.347	0.981
OR56B1	chr11p15.4	15	4.3	7.0	–	3.1 (6.3)	0.032	0.998
OR2AG2	chr11p15.4	8	2.6	4.3	–	0.3 (0.6)	0.052	1.000
OR4B1	chr11p11.2	7	3.9	2.1	–	0.9 (0.9)	0.136	1.000
OR4A16	chr11q11	14	8.9	2.9	0.2	1.9 (2.4)	0.012	1.000
OR5W2	chr11q11	30	13.7	12.0	–	5.6 (18.7)	< 0.001	0.208
OR6Q1	chr11q12.1	13	7.5	3.4	–	1.3 (1.4)	0.007	1.000
OR5AN1	chr11q12.1	7	3.8	2.2	–	0.3 (0.5)	0.053	1.000
OR8D4	chr11q24.1	8	2.1	4.9	–	0.5 (0.7)	0.054	1.000
OR9K2	chr12q13.2	14	7.5	4.0	–	3.5 (5.9)	0.099	0.983
OR4Q3	chr14q11.2	15	9.5	3.5	–	4.2 (10.5)	0.281	0.757

Note.– Same as Table 4.II.

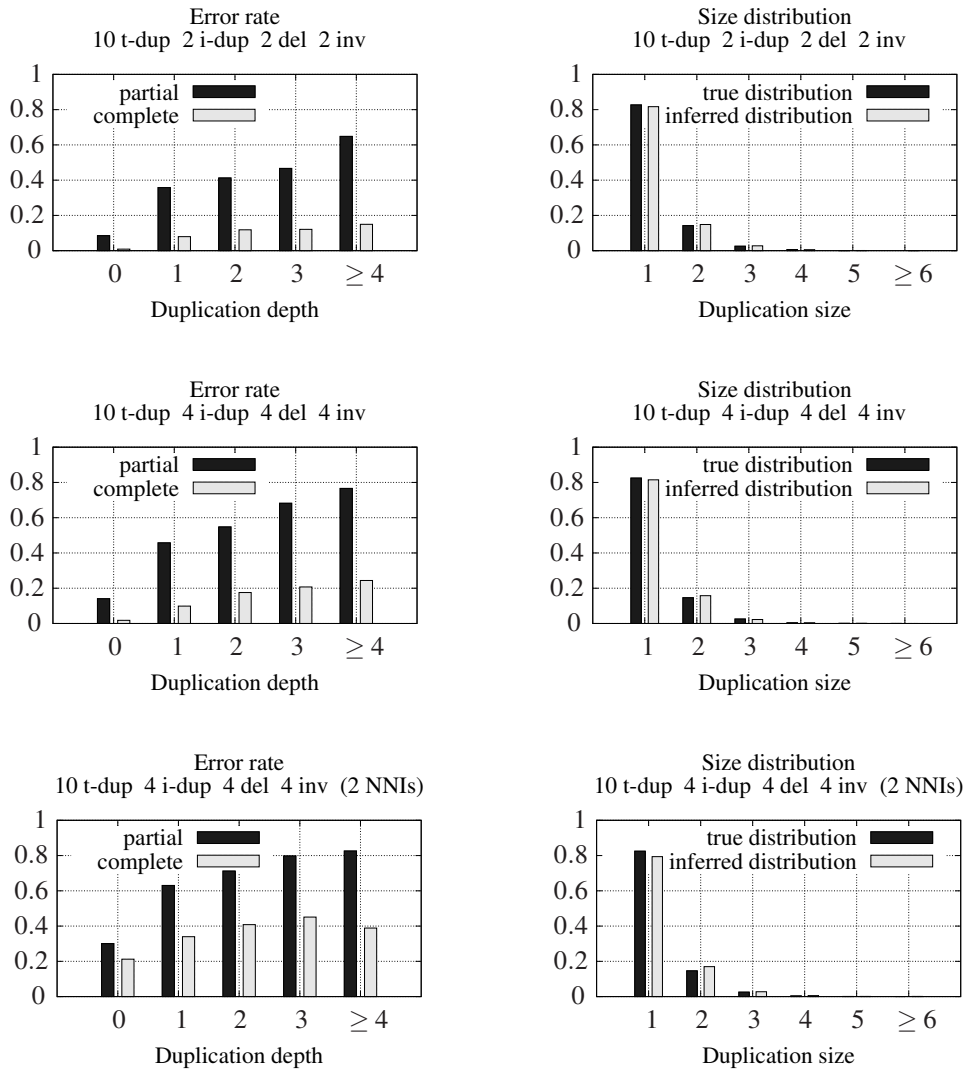

Figure 4.14: Result for simulated histories where the event size is sampled from a geometric distributions of parameter $p = 0.8$. (Left) Error rates in predicting duplication event according to their depth. (Right) Comparison between the true and the inferred duplication size distribution.

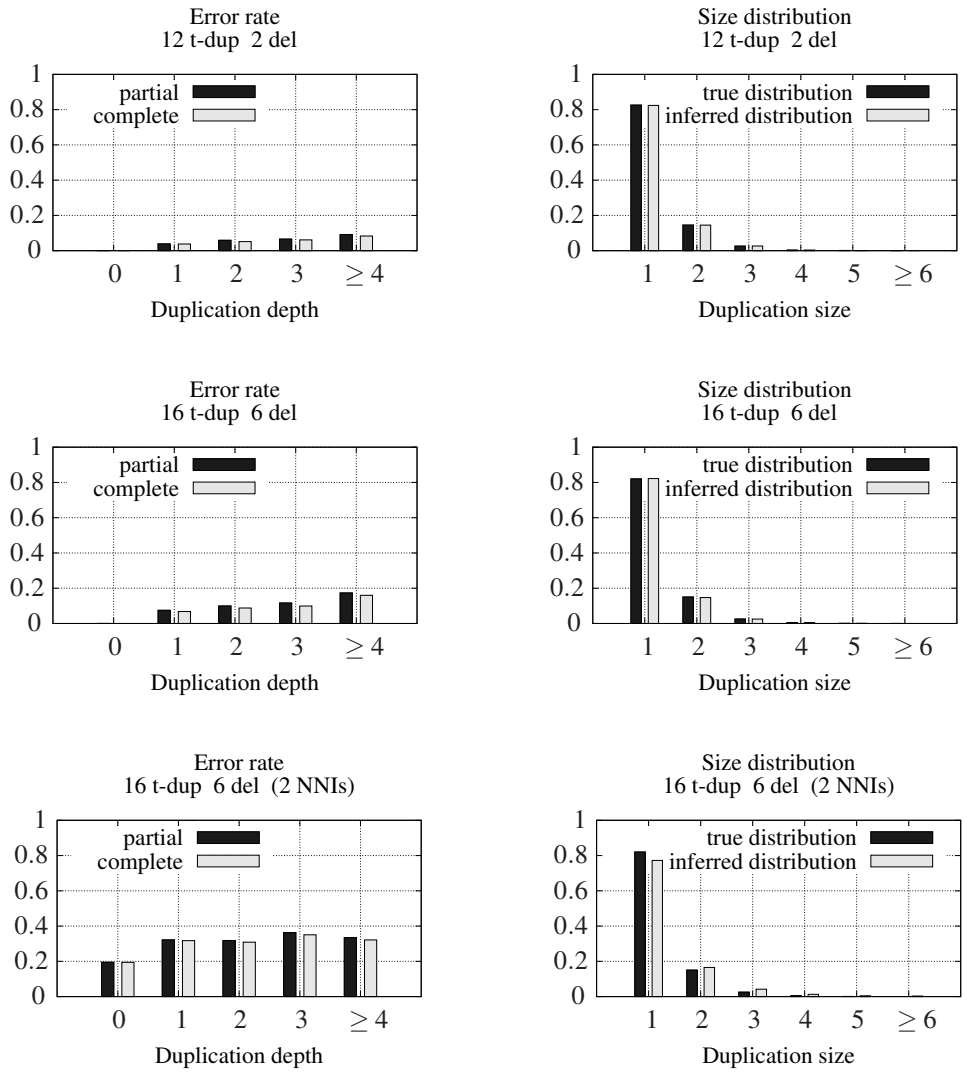

Figure 4.15: Result for simulated histories where the event size is sampled from a geometric distributions of parameter $p = 0.8$. (Left) Error rates in predicting duplication event according to their depth. (Right) Comparison between the true and the inferred duplication size distribution.

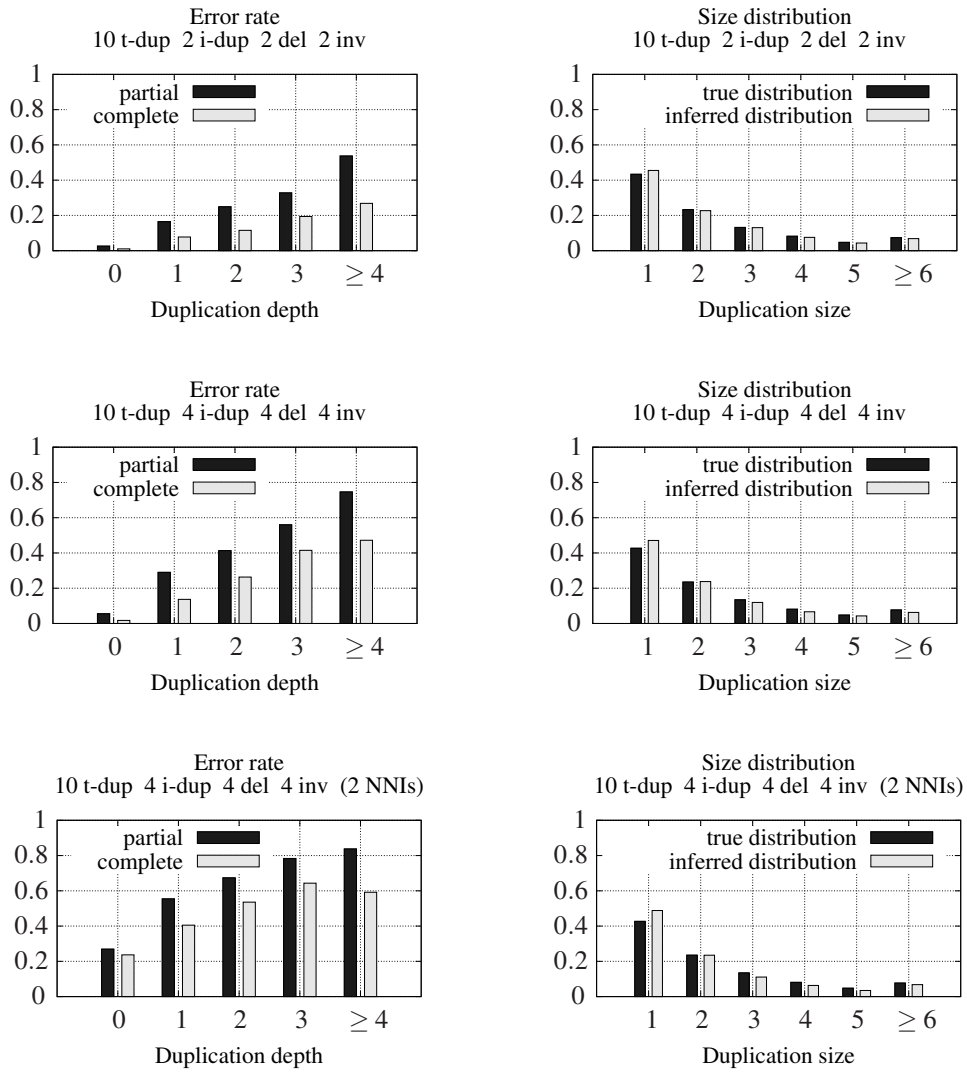

Figure 4.16: Result for simulated histories where the event size is sampled from a geometric distributions of parameter $p = 0.3$. (Left) Error rates in predicting duplication event according to their depth. (Right) Comparison between the true and the inferred duplication size distribution.

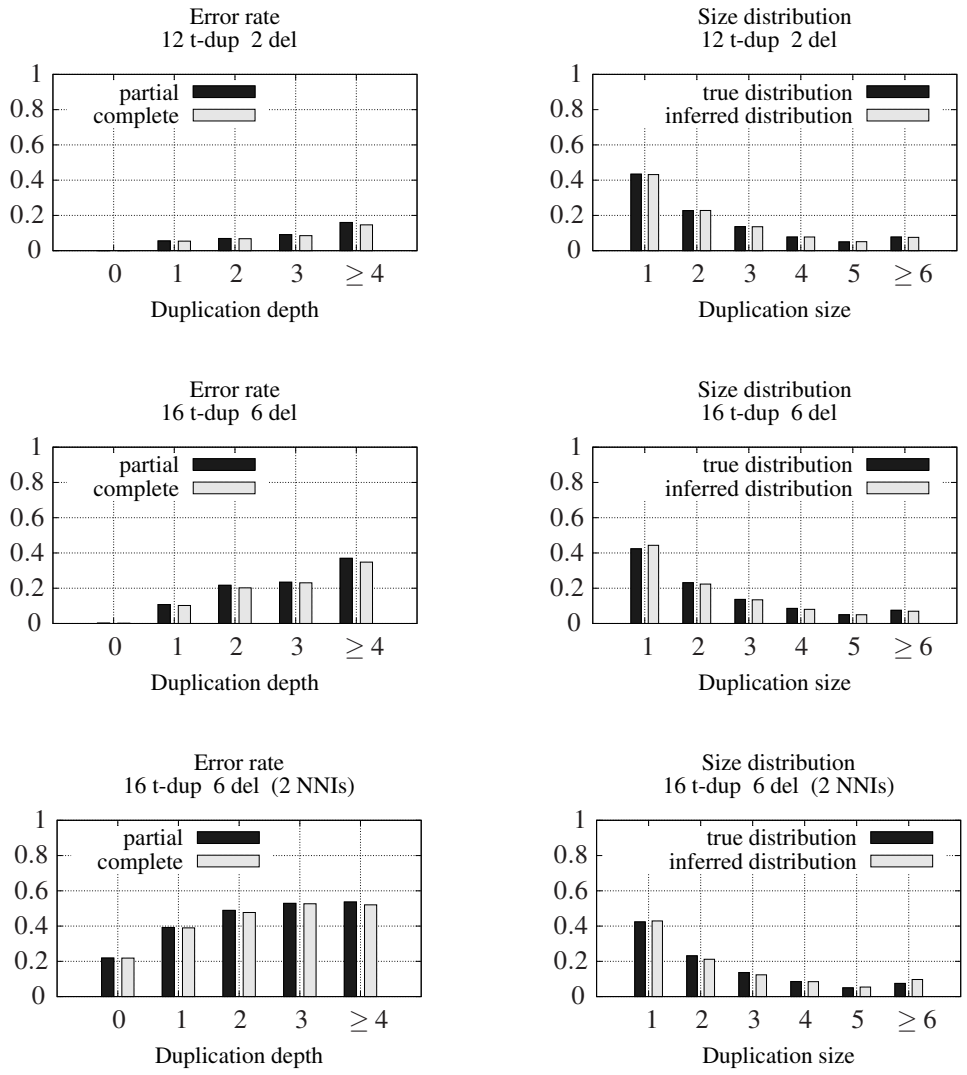

Figure 4.17: Result for simulated histories where the event size is sampled from a geometric distributions of parameter $p = 0.3$. (Left) Error rates in predicting duplication event according to their depth. (Right) Comparison between the true and the inferred duplication size distribution.

4.7 Contribution des auteurs

Mathieu Lajoie et Nadia El-Mabrouk ont écrit le manuscrit. **Mathieu Lajoie** et Denis Bertrand ont conçu et implémenté l'algorithme. **Mathieu Lajoie** a conçu et implémenté le simulateur de données. **Mathieu Lajoie** et Denis Bertrand ont conçu et réalisé les expérimentations sur les données simulées. **Mathieu Lajoie** a réalisé les expérimentations sur les données biologiques. Denis Bertand a réalisé les figures et participé à la révision du manuscrit. Denis Bertrand et **Mathieu Lajoie** ont réalisé l'interface Web.

CONCLUSION

Nous avons proposé dans cette thèse différentes extensions au modèle de duplication en tandem classique, permettant ainsi d'élargir ses champs d'applications. En effet, le modèle proposé par Fitch [35] ne s'applique qu'à l'ensemble restreint des clusters dont les gènes partagent tous la même orientation transcriptionnelle. En particulier, il n'est pas approprié à l'étude des récepteurs olfactifs et des gènes de type KRAB-ZNF, qui forment pourtant deux des plus importantes familles multigéniques chez l'humain, avec plusieurs centaines de membres par famille. De plus, même pour des clusters ayant évolué par duplications en tandem, la présence de délétions ayant mené à des pertes de gènes, comme c'est le cas par exemple dans l'évolution des protocadherines, peut rendre son utilisation inadéquate.

Dans un premier temps, nous avons donc intégré les inversions au modèle de duplications en tandem simple, et nous avons proposé un algorithme exact pouvant être appliqué à des familles multigéniques contenant des gènes dans les deux orientations transcriptionnelles. Nous avons ensuite généralisé ce modèle pour permettre l'étude d'un ensemble de clusters orthologues dans plusieurs espèces. Bien que ces deux extensions se limitent aux duplications simples (duplication d'un seul gène), elles ont démontré que le modèle classique pouvait être amélioré, ce qui a contribué à relancer l'intérêt pour le développement d'algorithmes visant à inférer l'histoire évolutive des clusters de gènes regroupés en tandem (voir, par exemple, les travaux de [133, 126] sur l'inférence d'histoires évolutives de clusters humains).

Finalement, nous avons proposé un algorithme qui n'est pas restreint aux évènements de duplications simples et qui tient compte d'un vaste ensemble d'évènements évolutifs pouvant impliquer plusieurs gènes à la fois : les duplications en tandem, les duplications inversées, les inversions et les délétions. Nous avons démontré, à l'aide de simulations, la capacité de cet algorithme à inférer la distribution de la taille des duplications avec une bonne précision, même en considérant des arbres de gènes "érronés", c.-à-d. qui

diffèrent des arbres véritables par un certain nombre de NNI. Les évènements évolutifs récents (n'ayant pas été obscurcis par des évènements subséquents), peuvent également être inférés avec une bonne précision. Nous avons utilisé cet algorithme pour étudier la distribution de la taille des duplications chez les récepteurs olfactifs et les protocadherines. Cette étude nous a permis d'observer une différence importante dans la proportion des duplications multiples entre les deux familles.

Un autre intérêt de notre algorithme est qu'il permet de comparer différents arbres de gènes (pour un même cluster) selon un nouveau critère objectif, à savoir le coût des histoires évolutives qu'ils induisent. Bien que des études supplémentaires soient requises pour préciser le poids à accorder à ce critère, il semble pour le moment raisonnable de privilégier, parmi l'ensemble des arbres ayant des probabilités postérieures similaires, ceux induisant des histoires évolutives ayant les coûts les plus bas. Notre algorithme est disponible via une interface Web[4] permettant de visualiser les histoires inférées grâce à une représentation graphique adéquate, ce qui, nous l'espérons, facilitera son utilisation.

Comme nous l'avons mentionné dans l'article constituant le Chapitre 4 de cette thèse, de nombreuses améliorations peuvent encore être apportées à cet algorithme. La première et la plus importante est sa généralisation à plusieurs espèces. En effet, considérer plusieurs clusters orthologues dans différentes espèces voisines permet de restreindre l'ensemble des histoires optimales pour chaque cluster et ainsi d'augmenter la précision des histoires inférées. Pour ce faire, l'approche utilisée dans l'article du Chapitre 3 doit cependant être modifiée, car la méthode de réconciliation utilisée minimise les pertes de gènes et ne reflète pas notre modèle évolutif, basé sur des délétions pouvant entrainer la perte simultanée de plusieurs gènes adjacents. Il faudrait donc intégrer les évènements de spéciations directement à notre modèle, élargissant ainsi notre espace de recherche. Heureusement, pour des clusters d'une vingtaine de gènes (comme ceux que nous avons considérés dans nos études), il semble possible d'envisager des algorithmes plus coûteux en temps et en espace que ceux développés dans cette thèse, tout en les maintenant

[4] http://www-lbit.iro.umontreal.ca/DILTAG/

raisonnables d'un point de vue pratique.

Une autre amélioration serait de considérer la longueur des branches dans les arbres de gènes utilisés pour l'inférence d'histoires évolutives. Cependant, étant donné que les taux d'évolution diffèrent souvent entre gènes dupliqués, le poids à donner à cette information est difficile à déterminer et nécessiterait l'ajout de paramètres additionnels. De plus, les évènements de délétion entraînent des difficultés techniques pour l'évaluation de la longueur des branches dans les arbres de gènes.

Finalement, il serait intéressant de généraliser notre algorithme pour permettre l'inférence d'histoires évolutives à partir d'une *forêt d'arbres ordonnés* $(\tilde{F}, \tilde{O})$. En effet, les relations phylogénétiques à l'intérieur d'une famille ne peuvent plus être représentées par un unique arbre lorsqu'une recombinaison non-allélique se produit à l'intérieur des limites d'un gène, car les segments de part et d'autre du point de recombinaison sont alors associés à des arbres différents. Pour qu'une telle généralisation soit acceptable au niveau biologique, il faudrait cependant restreindre notre espace de recherche aux forêts ordonnées intermédiaires (F, O) représentant des clusters "valides", c.-à-d. dans lesquels tous les segments d'un même gène sont correctement ordonnés et partagent la même orientation transcriptionnelle. Une telle généralisation permettrait aussi l'étude des clusters dont les gènes sont constitués de différentes combinaisons d'exons. Dans ce cas, la restriction précédente devrait être modifiée pour refléter les contraintes biologiques concernant l'ordre et le contenu en exons des gènes dans la famille étudiée.

BIBLIOGRAPHIE

[1] T. ALIOTO et J. NGAI : The repertoire of olfactory C family G protein-coupled receptors in zebrafish : candidate chemosensory receptors for amino acids. *BMC Genomics*, 7(1):309, 2006.

[2] R. ALONI, T. OLENDER et D. LANCET : Ancient genomic architecture for mammalian olfactory receptor clusters. *Genome Biology*, 7(10):R88, 2006.

[3] B. ARDEN, S. CLARK, D. KABELITZ et T. MAK : Human T-cell receptor variable gene segment families. *Immunogenetics*, 42(6):455-500, 1995.

[4] 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 Proceedings of the eighth annual international conference on Resaerch in computational molecular biology (RECOMB04)*, p. 326-335. ACM, 2004. ISBN 1-58113-755-9.

[5] L. ARVESTAD, A.-C. BERGLUND, J. LAGERGREN et B. SENNBLAD : Simultaneous history reconstruction for complex gene clusters in multiple species. *In Pacific Symposium on Biocomputing*, p. 162-173. World Scientific Publishing Co. Pte. Ltd., 2009.

[6] D. BADER, B. MORET et M. YAN : A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. *Journal of Computational Biology*, 8(5):483-491, 2001.

[7] J. BAILEY et E. EICHLER : Primate segmental duplications : crucibles of evolution, diversity and disease. *Nature Reviews Genetics*, 7(7):552-564, 2006.

[8] J. BAILEY, G. LIU et E. EICHLER : An Alu transposition model for the origin

and expansion of human segmental duplications. *The American Journal of Human Genetics*, 73(4):823-834, 2003.

[9] M. BATZER et P. DEININGER : Alu repeats and human genomic diversity. *Nature Reviews Genetics*, 3(5):370-379, 2002.

[10] G. BENSON et L. DONG : Reconstructing the duplication history of a tandem repeat. *In Proceedings of Intelligent Systems in Molecular Biology (ISMB1999), Heidelberg, Germany*, p. 44-53. AAAI, 1999.

[11] A. BERGERON, J. MIXTACKI et J. STOYE : Reversal distance without hurdles and fortresses. vol. 3109 de *Lecture Notes in Computer Science*, p. 388 - 399. Springer-Verlag, 2004.

[12] D. BERTRAND et O. GASCUEL : Topological rearrangements and local search method for tandem duplication trees. *IEEE Transactions on Computational Biology and Bioinformatics*, p. 15-28, 2005.

[13] D. BERTRAND, M. LAJOIE et N. EL-MABROUK : Inferring ancestral gene orders for a family of tandemly arrayed genes. *Journal of Computational Biology*, 15 (8):1063-1077, 2008.

[14] D. BERTRAND, M. LAJOIE, N. EL-MABROUK et O. GASCUEL : Evolution of tandemly repeated sequences through duplication and inversion. *In Fourth RECOMB International Workshop on Comparative Genomics*, vol. 4205 de *LNBI*, p. 129-140. Springer, 2006.

[15] P. BONIZZONI, G. VEDOVA et R. DONDI : Reconciling gene trees to a species tree. *Theoretical Computer Science*, 347(1-2):36-53, 2005.

[16] G. BOURQUE et P. PEVZNER : Genome-scale evolution : Reconstructing gene orders in the ancestral species. *Genome Research*, 12(1):26-36, 2002.

[17] L. BUCK et R. AXEL : A novel multigene family may encode odorant receptors : a molecular basis for odor recognition. *Cell*, 65(1):175-187, 1991.

[18] A. CAPRARA : The reversal median problem. *Journal on Computing*, 15(1):93-113, 2003.

[19] C. CHAUVE, J. DOYON et N. EL-MABROUK : Inferring a duplication, speciation and loss history from a gene tree. *In Fifth RECOMB International Workshop on Comparative Genomics*, p. 45-57. Springer-Verlag, 2007.

[20] K. CHEN, D. DURAND et M. FARACH-COLTON : Notung : Dating gene duplications using gene family trees. *In Proceedings of the Fourth Annual International Conference on Computational Molecular Biology (RECOMB 2000)*, New York, 2000. ACM.

[21] T. CORMEN, C. E. LEISERSON, R. L. RIVEST et C. STEIN : *Introduction to Algorithms*. MIT Press, 2001.

[22] S. DE, S. TEICHMANN et M. BABU : The impact of genomic neighborhood on the evolution of human and chimpanzee transcriptome. *Genome Research*, 19 (5):785, 2009.

[23] P. DEININGER, J. MORAN, M. BATZER et H. KAZAZIAN : Mobile elements and mammalian genome evolution. *Current Opinion in Genetics & Development*, 13 (6):651-658, 2003.

[24] J. DEMUTH, T. BIE, J. STAJICH, N. CRISTIANINI et M. HAHN : The evolution of mammalian gene families. *PLoS ONE*, 1(1):e85, 2006.

[25] J. DESCHAMPS et J. van NES : Developmental regulation of the Hox genes during axial morphogenesis in the mouse. *Development*, 132(13):2931-2942, 2005.

[26] R. DESPER et O. GASCUEL : Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *Journal of Computational Biology*, 9(5):687-705, 2002.

[27] D. DUBOULE : The rise and fall of Hox gene clusters. *Development*, 134(14): 2549-2560, 2007.

[28] R. EDGAR : MUSCLE : multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792-1797, 2004.

[29] E. EICHLER et D. SANKOFF : Structural dynamics of eukaryotic chromosome evolution. *Science*, 301:793-797, 2003.

[30] N. EL-MABROUK : Genome rearrangement by reversals and insertions/deletions of contiguous segments. *In* R. GIANCARLO et D. SANKOFF, éds : *Combinatorial Pattern Matching*, vol. 1848 de *Lecture Notes in Computer Science*, p. 222- 234, 2000.

[31] O. ELEMENTO et O. GASCUEL : A fast and accurate distance-based algorithm to reconstruct tandem duplication trees. *Bioinformatics*, 18:92-99, 2002.

[32] O. ELEMENTO et O. GASCUEL : An exact and polynomial distance-based algorithm to reconstruct single copy tandem duplication trees. *Journal of Discrete Algorithms*, 3(2-4):362-374, 2005.

[33] O. ELEMENTO, O. GASCUEL et M.-P. LEFRANC : Reconstructing the duplication history of tandemly repeated genes. *Molecular Biology and Evolution*, 19:278-288, 2002.

[34] L. FEUK, A. CARSON et S. SCHERER : Structural variation in the human genome. *Nature Reviews Genetics*, 7(2):85-97, 2006.

[35] W. FITCH : Phylogenies constrained by cross-over process as illustrated by human hemoglobins and a thirteen-cycle, eleven amino-acid repeat in human apolipoprotein A-I. *Genetics*, 86:623-644, 1977.

[36] O. GASCUEL : BIONJ : an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, 14(7):685-695, 1997.

[37] O. GASCUEL, D. BERTRAND et O. ELEMENTO : Reconstructing the duplication history of tandemly repeated sequences. *In* O. GASCUEL, éd. : *Mathematics of Evolution and Phylogeny*, p. 205-235. Oxford University Press, 2005.

[38] O. GASCUEL, M. HENDY, A. JEAN-MARIE et S. MCLACHLAN : The combinatorics of tandem duplication trees. *Systematic Biology*, 52:110-118, 2003.

[39] D. GERAGHTY, B. KOLLER, J. HANSEN et H. ORR : The HLA class I gene family includes at least six genes and twelve pseudogenes and gene fragments. *Journal of Immunology*, 149(6):1934-1946, 1992.

[40] Y. GILAD, O. MAN et G. GLUSMAN : A comparison of the human and chimpanzee olfactory receptor gene repertoires. *Genome Research*, 15(6):224-230, 2005.

[41] G. GLUSMAN, I. YANAI, I. RUBIN et D. LANCET : The complete human olfactory subgenome. *Genome Research*, 11(5):685-702, 2001.

[42] Y. GO et Y. NIIMURA : Similar numbers but different repertoires of olfactory receptor genes in humans and chimpanzees. *Molecular Biology and Evolution*, 25(9):1897, 2008.

[43] P. GODFREY, B. MALNIC et L. BUCK : The mouse olfactory receptor gene family. *Proceedings of the National Academy of Sciences*, 101(7):2156-2161, 2004.

[44] M. GOODMAN, J. CZELUSNIAK, G. MOORE, A. ROMERO-HERRERA et G. MATSUDA : Fitting the gene lineage into its species lineage : A parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Zoology*, 28:132-168, 1979.

[45] R. GUIGÓ, I. MUCHNIK et T. SMITH : Reconstruction of ancient molecular phylogeny. *Molecular Phylogenetics and Evolution*, 6:189-213, 1996.

[46] S. GUINDON et O. GASCUEL : A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology*, 52(5):696-704, 2003.

[47] A. HAMILTON, S. HUNTLEY, M. TRAN-GYAMFI, D. BAGGOTT, L.GORDON et L. STUBBS : Evolutionary expansion and divergence in the ZNF91 subfamily of primate-specific zinc finger genes. *Genome Research*, 16(5):584-594, 2006.

[48] S. HANNENHALLI et P. PEVZNER : Transforming cabbage into turnip : polynomial algorithm for sorting signed permutations by reversals. *Journal of the ACM*, 46(1):1-27, 1999.

[49] P. HIETER, G. HOLLIS, S. KORSMEYER, T. WALDMANN et P. LEDER : Clustered arrangement of immunoglobulin λ constant region genes in man. *Nature*, 294 (5841):536-540, 1981.

[50] I. HORNE et V. HARITOS : Multiple tandem gene duplications in a neutral lipase gene cluster in Drosophila. *Gene*, 411(1-2):27-37, 2008.

[51] S. HUNTLEY, D. BAGGOT, A. HAMILTON, S. Y. M. TRANGYAMFI, J. KIM, L. GORDON, E. BRANSCOMB et L. STUBBS : A comprehensive catalogue of human krab-associated zinc finger genes : Insights into the evolutionary history of a large family of transcriptional repressors. *Genome Research*, 16:669-677, 2006.

[52] L. HURST, C. PÁL et M. LERCHER : The evolutionary dynamics of eukaryotic gene order. *Nature Reviews Genetics*, 5(4):299-310, 2004.

[53] D. JAITLY, P. KEARNEY, G. LIN et B. MA : Methods for reconstructing the history of tandem repeats and their application to the human genome. *Journal of Computer and System Sciences*, 65:494-507, 2002.

[54] D. JONES, W. TAYLOR et J. THORNTON : The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences*, 8 (3):275-282, 1992.

[55] H. KAPLAN, R. SHAMIR et R. E. TARJAN : A faster and simpler algorithm for sorting signed permutations by reversals. *SIAM Journal on Computing*, 29:880-892, 2000.

[56] L. KAUPPI, A. JEFFREYS et S. KEENEY : Where the crossovers are : recombination distributions in mammals. *Nature Reviews Genetics*, 5(6):413-424, 2004.

[57] H. KAZAZIAN : Mobile elements : drivers of genome evolution. *Science*, 303 (5664):1626-1632, 2004.

[58] M. KMITA et D. DUBOULE : Organizing axes in time and space ; 25 years of colinear tinkering. *Science*, 301(5631):331-333, 2003.

[59] J. KRUMSIEK, R. ARNOLD et T. RATTEI : Gepard : a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics*, 23(8):1026, 2007.

[60] M. LAJOIE, D. BERTRAND et N. EL-MABROUK : Evolution of tandemly arrayed genes in multiple species. *In Fifth RECOMB International Workshop on Comparative Genomics*, p. 98-109. Springer-Verlag, 2007.

[61] M. LAJOIE, D. BERTRAND et N. EL-MABROUK : Inferring the evolutionary history of gene clusters from phylogenetic and gene order data. *Molecular Biology and Evolution*, p. 761-772, 2010.

[62] M. LAJOIE, D. BERTRAND, N. EL-MABROUK et O. GASCUEL : Duplication and inversion history of a tandemly repeated gene family. *Journal of Computational Biology*, p. 462-478, 2007.

[63] C. LANAVE, G. PREPARATA, C. SACCONE et G. SERIO : A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution*, 20 (1):86-93, 1984.

[64] E. LANDER, L. LINTON, B. BIRREN, C. NUSBAUM, M. ZODY, J. BALDWIN, K. DEVON, K. DEWAR, M. DOYLE, W. FITZHUGH *et al.* : Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860-921, 2001.

[65] R. LARUE, S. JONSSON, K. SILVERSTEIN, M. LAJOIE, D. BERTRAND, N. EL-MABROUK, I. HÖTZEL, V. ANDRESDOTTIR, T. SMITH et R. HARRIS : The artiodactyl APOBEC3 innate immune repertoire shows evidence for a multi-functional domain organization that existed in the ancestor of placental mammals. *BMC Molecular Biology*, 9(1):104, 2008.

[66] M. LEFRANC, A. FORSTER, R. BAER, M. STINSON et T. RABBITTS : Diversity and rearrangement of the human T cell rearranging gamma genes : nine germ-line variable genes belonging to two subgroups. *Cell*, 45(2):237, 1986.

[67] M. LEFRANC, A. FORSTER et T. RABBITTS : Rearrangement of two distinct T-cell γ-chain variable-region genes in human DNA. *Nature*, 319(6052):420-422, 1986.

[68] D. LEMONS et W. MCGINNIS : Genomic evolution of Hox gene clusters. *Science*, 313(5795):1918-1922, 2006.

[69] E. LEWIS : A gene complex controlling segmentation in Drosophila. *Nature*, 276 (5688):565-570, 1978.

[70] J. LUPSKI : Genomic rearrangements and sporadic disease. *Nature Genetics*, 39:S43-S47, 2007.

[71] M. LYNCH et J. CONERY : The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494):1151-1155, 2000.

[72] B. MA, M. LI et L. ZHANG : On reconstructing species trees from gene trees in term of duplications and losses. *In* S. ISTRAIL, P. PEVZNER et M. WATERMAN, éds : *Proceedings of the Second Annual International Conference on Computational Biology (RECOMB 98)*, p. 182-191, New York, 1998. ACM.

[73] B. MA, M. LI et L. ZHANG : From gene trees to species trees. *SIAM Journal on Computing*, 30(3):729-752, 2000.

[74] J. MA, A. RATAN, B. RANEY, B. SUH, W. MILLER et D. HAUSSLER : The infinite sites model of genome evolution. *Proceedings of the National Academy of Sciences, USA*, 105(38):14254, 2008.

[75] J. MA, L. ZHANG, B. B. SUH, B. J. RANEY, R. C. BURHANS, W. J. KENT, M. BLANCHETTE, D. HAUSSLER et W. MILLER : Reconstructing contiguous regions of an ancestral genome. *Genome Research*, 16:1557-1565, 2006.

[76] D. MARTIN, C. WILLIAMSON et D. POSADA : RDP2 : recombination detection and analysis from sequence alignments. *Bioinformatics*, 21(2):260-262, 2005.

[77] D. MARTIN, C. WILLIAMSON et D. POSADA : RDP2 : recombination detection and analysis from sequence alignments. *Bioinformatics*, 21(2):260-262, 2005.

[78] J. MAYDT et T. LENGAUER : Recco : recombination analysis using cost optimization. *Bioinformatics*, 22(9):1064-1071, 2006.

[79] I. MILNE, F. WRIGHT, G. ROWE, D. MARSHALL, D. HUSMEIER et G. MCGUIRE : TOPALi : software for automatic identification of recombinant

sequences within DNA multiple alignments. *Bioinformatics*, 20(11):1806-1807, 2004.

[80] B. MORET, J. TANG, L. WANG et T. WARNOW : Steps toward accurate reconstructions of phylogenies from gene-order data. *Journal of Computer and System Sciences*, 65(3):508-525, 2002.

[81] M. NEI et A. ROONEY : Concerted and birth-and-death evolution of multigene families. *Annual Review of Genetic*, 39:121-152, 2005.

[82] Y. NIIMURA et M. NEI : Evolution of Olfactory Receptor Genes in the Human Genome. *Proceedings of the National Academy of Sciences*, 100(21):12235-12240, 2003.

[83] Y. NIIMURA et M. NEI : Evolutionary changes of the number of olfactory receptor genes in the human and mouse lineages. *Gene*, 346(14):23-28, 2005.

[84] Y. NIIMURA et M. NEI : Evolutionary dynamics of olfactory and other chemosensory receptor genes in vertebrates. *Journal of Human Genetics*, 51(6):505-517, 2006.

[85] J. NOONAN, J. GRIMWOOD, J. SCHMUTZ, M. DICKSON et R. MYERS : Gene conversion and the evolution of protocadherin gene cluster diversity. *Genome Research*, 14(3):354-366, 2004.

[86] S. OHNO : *Evolution by Gene Duplication*. Springer Verlag, New York.

[87] T. OLENDER, E. FELDMESSER, T. ATAROT, M. EISENSTEIN et D. LANCET : The olfactory receptor universe-from whole genome analysis to structure and evolution. *Genetics and Molecular Research*, 3(4):545-553, 2004.

[88] T. OLENDERAND, T. FUCHS, C. LINHART, R. SHAMIR, M. ADAMS, F. KALUSH, M. KHEN et D. LANCET : The canine olfactory subgenome. *Genomics*, 83:361-372, 2004.

[89] M. OLSON : When less is more : gene loss as an engine of evolutionary change. *The American Journal of Human Genetics*, 64(1):18-23, 1999.

[90] R. PAGE : Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Systematic Biology*, 43:58-77, 1994.

[91] R. PAGE et M. CHARLESTON : Reconciled trees and incongruent gene and species trees. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 37:57-70, 1997.

[92] D. POSADA et K. CRANDALL : The effect of recombination on the accuracy of phylogeny estimation. *Journal of Molecular Evolution*, 54(3):396-402, 2002.

[93] D. POSADA et K. CRANDALL : The effect of recombination on the accuracy of phylogeny estimation. *Journal of Molecular Evolution*, 54(3):396-402, 2002.

[94] P. QUIGNON, M. GIRAUD, M. RIMBAULT, P. LAVIGNE, S. TACHER, E. MORIN, E. RETOUT, A. VALIN, K. LINDBLAD-TOH, J. NICOLAS et F. GALIBERT : The dog and rat olfactory receptor repertoires. *Genome Biology*, 6:R83, 2005.

[95] L. REITER, P. HASTINGS, E. NELIS, P. DE JONGHE, C. VAN BROECKHOVEN et J. LUPSKI : Human meiotic recombination products revealed by sequencing a hotspot for homologous strand exchange in multiple HNPP deletion patients. *The American Journal of Human Genetics*, 62(5):1023-1033, 1998.

[96] E. RIVALS : A Survey on Algorithmic Aspects of Tandem Repeats Evolution. *International Journal of Foundations of Computer Science*, 15(2):225-257, 2004.

[97] F. RONQUIST et J. HUELSENBECK : MrBayes 3 : Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572-4, 2003.

[98] N. SAITOU et M. NEI : The neighbor-joining method : a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4:406-425, 1987.

[99] D. SANKOFF et M. BLANCHETTE : Multiple genome rearrangement and breakpoint phylogeny. *Journal of Computational Biology*, 5:555-570, 1998.

[100] M. SEMON et L. DURET : Evolutionary origin and maintenance of coexpressed gene clusters in mammals. *Molecular Biology and Evolution*, 23(9):1715-1723, 2006.

[101] B. SENNBLAD, E. SCHREIL, A. B. SONNHAMMER, J. LAGERGREN et L. ARVESTAD : Primetv : a viewer for reconciled trees. *BMC Bioinformatics*, (8):148, 2007.

[102] H. SEO, R. EDVARDSEN, A. MAELAND, M. BJORDAL, M. JENSEN, A. HANSEN, M. FLAAT, J. WEISSENBACH, H. LEHRACH, P. WINCKER *et al.* : Hox cluster disintegration with persistent anteroposterior order of expression in Oikopleura dioica. *Nature*, 431(7004):67-71, 2004.

[103] J. SETUBAL et J. MEIDANIS : *Introduction to Computational Molecular Biology*. PWS Pub. Co., 1997.

[104] L. SHAFFER et J. LUPSKI : Molecular mechanisms for constitutional chromosomal rearrangements in humans. *Annual Review of Genetics*, 34:297, 2000.

[105] M. SHANNON, A. HAMILTON, L. GORDON, E. BRANSCOMB et L. STUBBS : Differential expansion of Zinc- Finger transcription factor loci in homologous human and mouse gene clusters. *Genome Research*, 13:1097-1110, 2003.

[106] V. SHOJA et L. ZHANG : A roadmap of tandemly arrayed genes in the genomes of human, mouse, and rat. *Molecular Biology and Evolution*, 23(11):2134-2141, 2006.

[107] A. SIEPEL : Algorithm to find all sorting reversals. *In Proceedings of the second conference on computational molecular biology (RECOMB'02)*, p. 281 - 290. ACM Press, 2002.

[108] A. SIEPEL et B. MORET : Finding an optimal inversion median : Experimental results. *In Proceedings of the 2nd International Workshop on Algorithms in Bioinformatics(WABI2003)*, Lecture Notes in Computer Science, p. 189-203. Springer, 2001.

[109] P. STANKIEWICZ et J. LUPSKI : Genome architecture, rearrangements and genomic disorders. *TRENDS in Genetic*, 18(2):74-82, 2002.

[110] D. SWOFFORD, P. OLSEN, P. WADDELL et D. HILLIS : *Molecular Systematics*. Sinauer Associates, Sunderland, Massachusetts, 1996.

[111] M. TANG, M. WATERMAN et S. YOOSEPH : Zinc finger gene clusters and tandem gene duplication. *In Proceedings of International Conference on Research in Molecular Biology (RECOMB2001)*, p. 297-304, 2001.

[112] M. TANG, M. WATERMAN et S. YOOSEPH : Zinc finger gene clusters and tandem gene duplication. *Journal of Computational Biology*, p. 429-446, 2002.

[113] S. TAVARE : Some probabilistic and statistical problems on the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, 17, 1986.

[114] J. THOMAS : Concerted evolution of two novel protein families in Caenorhabditis species. *Genetics*, 172(4):2269-2281, 2006.

[115] J. THOMPSON, D. HIGGINS et T. GIBSON : CLUSTAL W : improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22):4673 - 4680, 1994.

[116] T. VINAŘ, B. BREJOVÁ, G. SONG et A. SIEPEL : Reconstructing histories of complex gene clusters on a phylogeny. *In* F. D. CICCARELLI et I. MIKLOS, éds : *Comparative Genomics, International Workshop (RECOMB-CG)*, vol. 5817 de *Lecture Notes in Computer Science*, p. 150-163, Budapest, 2009. Springer.

[117] X. WANG, W. GRUS et J. ZHANG : Gene losses during human origins. *PLoS Biology*, 4(3):366, 2006.

[118] I. WAPINSKI, A. PFEFFER, N. FRIEDMAN et A. REGEV : Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics*, 23:i549-i558, 2007.

[119] M. WATERMAN : *Introduction to Computational Biology*. Chapman & Hall New York, NY, 1995.

[120] R. WATERSTON, K. LINDBLAD-TOH, E. BIRNEY, J. ROGERS, J. ABRIL, P. AGARWAL, R. AGARWALA, R. AINSCOUGH, M. ALEXANDERSSON, P. AN *et al.* : Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520-562, 2002.

[121] Q. WU et T. MANIATIS : A Striking Organization of a Large Family of Human Neural Cadherin-like Cell Adhesion Genes. *Cell*, 97:779-780, 1999.

[122] J. YANG et L. ZHANG : On counting tandem duplication trees. *Molecular Biology and Evolution*, 21(6):1160-1163, 2004.

[123] S. YANG, J. ARGUELLO, X. LI, Y. DING, Q. ZHOU, Y. CHEN, Y. ZHANG, R. ZHAO, F. BRUNET, L. PENG *et al.* : Repetitive element-mediated recombination as a mechanism for new gene origination in Drosophila. *PLoS Genetics*, 4 (1):e3, 2008.

[124] J. YOUNG, C. FRIEDMAN, J. ROSS, L. TONNES-PRIDDY et B. TRASK : Different evolutionary processes shaped the mouse and human olfactory receptor gene families. *Human Molecular Genetics*, 11(5):535-546, 2002.

[125] G. YULE : A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis. *Philosophical Transactions of the Royal Society of London. Series*

B, Containing Papers of a Biological Character Immunology Review, 213:21-87, 1924.

[126] F. ZHANG, C. CARVALHO et J. LUPSKI : Complex human chromosomal and genomic rearrangements. *Trends in Genetics*, 25(7):298-307, 2009.

[127] J. ZHANG : Evolution by gene duplication : an update. *Trends in Ecology and Evolution*, 18(6):292-298, 2003.

[128] J. ZHANG et M. NEI : Evolution of antennapedia-class homeobox genes. *Genetics*, 142(1):295-303, 1996.

[129] L. ZHANG : On a Mirkin-Muchnik-Smith conjecture for comparing molecular phylogenies. *Journal of Computational Biology*, 4:177-187, 1997.

[130] L. ZHANG, B. MA, L. WANG et Y. XU : Greedy method for inferring tandem duplication history. *Bioinformatics*, 19:1497-1504, 2003.

[131] T. ZHANG, P. HAWS et Q. WU : Multiple variable first exons : a mechanism for cell-and tissue-specific gene regulation. *Genome Research*, 14(1):79-89, 2004.

[132] X. ZHANG et S. FIRESTEIN : The olfactory receptor gene superfamily of the mouse. *Nature Neuroscience*, 5(2):124-133, 2002.

[133] Y. ZHANG, G. SONG, T. VINAR, E. GREEN, A. SIEPEL et W. MILLER : Reconstructing the evolutionary history of complex human gene clusters. vol. 4955, p. 29-49. Springer, 2008.

[134] C. ZHENG, A. LENERT et D. SANKOFF : Reversal distance for partially ordered genomes. *Bioinformatics*, 21:i502 - i508, 2003.

[135] S. ZOZULYA, F. ECHEVERRI et T. NGUYEN : The human olfactory receptor repertoire. *Genome Biology*, 2(6):1-18, 2001.