Université de Montréal

Algorithmes pour la reconstruction de génomes ancestraux

par
Yves Gagnon

Département de biochimie
Faculté de médecine

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de Maître ès sciences (M.Sc.)
en bio-informatique

Mai, 2012

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé:

Algorithmes pour la reconstruction de génomes ancestraux

présenté par:

Yves Gagnon

a été évalué par un jury composé des personnes suivantes:

Sylvie Hamel,	président-rapporteur
Nadia El-Mabrouk,	directeur de recherche
Miklós Csűrös,	membre du jury

Mémoire accepté le: . . . . . .

RÉSUMÉ

L'inférence de génomes ancestraux est une étape essentielle pour l'étude de l'évolution des génomes. Connaissant les génomes d'espèces éteintes, on peut proposer des mécanismes biologiques expliquant les divergences entre les génomes des espèces modernes.

Diverses méthodes visant à résoudre ce problème existent, se classant parmis deux grandes catégories : les méthodes de distance et les méthodes de synténie. L'état de l'art des distances génomiques ne permettant qu'un certain répertoire de réarrangements pour le moment, les méthodes de synténie sont donc plus appropriées en pratique.

Nous proposons une méthode de synténie pour la reconstruction de génomes ancestraux basée sur une définition relaxée d'adjacences de gènes, permettant un contenu en gène inégal dans les génomes modernes causé par des pertes de gènes de même que des duplications de génomes entiers (DGE). Des simulations sont effectuées, démontrant une capacité de former une solution assemblée en un nombre réduit de régions ancestrales contigües par rapport à d'autres méthodes tout en gardant une bonne fiabilité. Des applications sur des données de levures et de plantes céréalières montrent des résultats en accord avec d'autres publications, notamment la présence de fusion imbriquée de chromosomes pendant l'évolution des céréales.

Mots clés: Algorithmes, génomes ancestraux, duplication de génomes, synténie, distances génomiques.

ABSTRACT

Ancestral genome inference is a decisive step for studying genome evolution. Knowing genomes from extinct species, one can propose biological mecanisms explaining divergences between extant species genomes.

Various methods classified in two categories have been developped : distance based methods and synteny based methods. The state of the art of distance based methods only permit a certain repertoire of genomic rearrangements, thus synteny based methods are more appropriate in practice for the time being.

We propose a synteny method for ancestral genome reconstruction based on a relaxed defenition of gene adjacencies, permitting unequal gene content in extant genomes caused by gene losses and whole genome duplications (WGD). Simulations results demonstrate our method's ability to form a more assembled solution rather than a collection of contiguous ancestral regions (CAR) with respect to other methods, while maintaining a good reliability. Applications on data sets from yeasts and cereal species show results agreeing with other publications, notably the existence of nested chromosome fusion during the evolution of cereals.

Keywords: Algorithms, ancestral genome, whole genome duplication, synteny, genomic distance.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES

ADN	Acide désoxyribonucléique
ARN	Acide ribonucléique
DGE	Duplication de génome entier
RAC	Région ancestrale contigüe

NOTATION

G	Génome
Σ	Alphabet représentant les familles de gènes
Σ_G	Familles de gènes présentent dans le génome G, $\Sigma_G \subseteq \Sigma$
Γ	Ensemble de génomes modernes
$\mathscr{F}_a$	Famille formée de toutes les occurences du gène a dans tout $G \in \Gamma$
$mult(a,G)$	le nombre de copies du gène a dans le génome G
S	phylogénie pour les espèces associées aux génomes de Γ
S^{DGE}	phylogénie augmentée de noeuds représenant des DGE
u_p	noeud pré-dupliqué de S^{DGE}
T_a	arbre de gènes pour $\mathscr{F}_a$
$\mathscr{I}_G$	ensemble de synténies pour le génome G

Je dédie ce travail à deux personnes disparues,

Jean, mon père et Steeve, mon grand frère.

REMERCIEMENTS

Je tiens tout d'abord à remercier mon mentor qui m'a enduré pendant ces trois ans de labeur, Dr. Nadia El-Mabrouk. Merci pour tes encouragements, ta patience et ton excellent encadrement ! Je ne pense pas que ça soit donné à tous les étudiants de maîtrise d'obtenir des publications, je te dois beaucoup pour cet accomplissement.

Merci également à Denis Bertrand et Mathieu Blanchette, les co-auteurs de deux articles qui ont permis d'aboutir à ce mémoire. Denis tu m'as beaucoup aidé à forger mon talent de programmeur.

Un merci particulier à Olivier. Nous nous sommes côtoyé pendant un total de six années, d'abord au baccalauréat et ensuite au laboratoire de Nadia. Toi aussi tes encouragements ont été d'un bon réconfort pendant les périodes plus difficiles. Nous sommes plus que de simple collègues de travail, nous sommes de bons amis.

Merci aux membre de mon jury, Dr. Sylvie Hamel et Dr. Miklós Csűrös, qui ont accepté de lire ce travail en plein été !

Merci également à tous les autres membres du laboratoire que j'ai eu la chance de connaître : Krister, Andréa, Billel, Manuel, Patrick.

Finalement un gros merci à ma famille qui m'ont encouragé dans la poursuite de mes études. Un gros merci également à toi Yves (pas moi, l'autre :P), je pense que c'est toi qui m'a le plus poussé à continuer à la maîtrise. Sans toi, peut-être que ce mémoire n'existerait pas.

CHAPITRE 1

INTRODUCTION

Avant la découverte par Watson et Crick de la structure en double hélice de l'ADN dans les années 50 [66], les différents caractères morphologiques étaient employés afin de comparer et classifier les espèces. De nos jours, quoique cette méhode de comparaison soit toujours utilisée, les données moléculaires sont plutôt celles priviligées pour remplir le même rôle. Depuis le séquençage des premiers génomes entiers, les avancées technologiques ont permis de démocratiser cet exercice autrefois coûteux. Ainsi, on retrouve maintenant les génomes de plus de 460 eucaryotes et plus de 1200 procaryotes dans les banques de données publiques tel que GenBank[3]. Cette manne de données est à l'origine de l'épanouissement de la génomique comparative. Non seulement la comparaison de génomes entiers permet de construire et de valider des arbres phylogénétiques, elle permet également de poser certaines hypothèses quant aux mécanismes d'évolution de ces génomes et d'étudier la co-évolution, la fonction ou l'inter-dépendance des gènes. L'annotation de génomes nouvellement séquencés est grandement facilitée par cette discipline.

Une étape essentielle à la génomique évolutive est l'inférence de génomes appartenant à des espèces ancestrales correspondant aux noeuds internes des arbres phylogénétiques. L'inférence de génomes ancestraux est, en particulier, une étape préliminaire au calcul de distances génomiques ainsi qu'à l'inférence d'un scénario évolutif. Ce mémoire est consacré au développement de méthodes algorithmiques pour l'inférence de génomes ancestraux.

La représentation de génomes sous forme de séquences nucléotidiques ne permet que difficilement la comparaison de génomes entiers. En effet, certaines régions de l'ADN sont mieux conservées que d'autres. Des mutations ponctuelles ou à petite échelle peuvent survenir (insertion, suppression et substitution de nucléotides) et comme les évènements de spéciation se produisent à des millions d'années d'intervalles, un nombre important de ces mutations peuvent brouiller un signal de similarité. De plus, l'alignement

de génomes entiers est compliqué par les évènements évolutifs à plus grande échelle décrits plus loin (inversion, translocation, fusion/fission, duplication, perte). Afin de tenir compte de ces macro-évènements, une alternative est de représenter les génomes sous forme de séquences de gènes ou autres éléments constitutifs du génome. Plus spécifiquement, cette représentation permet de visualiser l'organisation structurale des chromosomes et de comparer leurs architectures chez une même espèce ou chez différentes espèces.

L'ordre des gènes dans les génomes peut être modifié au cours de l'évolution par des évènements de réarrangements tels que les inversions, les transpositions, les translocations, les fusions et les fissions de chromosomes. D'autres évènements ont pour effet de changer plus ou moins radicalement le contenu en gènes : les duplications et pertes de gènes. Déjà dans les années 60, ces évènements étaient reconnus comme moteur de l'évolution [49]. On distingue plusieurs sources de duplication à petite et grande échelle, la plus impressionnante étant la duplication de génome entier (DGE). Dans tous les cas, les gènes supplémentaires provenant des duplications ne sont pas tous nécessaires à la survie des espèces et subissent différents sorts. Les copies non perdues, à l'intérieur d'un même génome ou dans différents génomes, évoluent indépendemment les unes des autres (sauf exception, par exemple dans des cas de co-évolution) et l'ensemble des copies «survivantes» constituent une famille de gènes. Le regroupement des gènes en familles est un problème complexe en soit, particulièrement dans le cas de duplications massives suivies de pertes importantes. Ce problème, plus précisément celui d'identifier les gènes provenant d'une DGE, est abordé au chapitre 2.

Le problème algorithmique étudié dans ce mémoire est le «small phylogeny problem». Comparativement au «large phylogeny problem», où on cherche à construire un arbre phylogénétique, l'arbre des espèce est déjà connu dans ce cas-ci. Le problème est alors d'identifier le contenu et l'ordre des gènes aux noeuds internes de l'arbre phylogénétique étant donné ceux aux feuilles. Une approche parcimonieuse à la résolution de ces problèmes est utilisée dans un contexte de reconstruction de génome ancestraux. De nombreuses contributions ont été apportées à ce problème dans le passé [4, 8, 11, 17, 44, 46, 57–59, 69], et les méthodes peuvent se classer dans deux catégories [16, 24] : les

méthodes de distances [11, 16, 46, 57, 59] et les méthodes de synténie [8, 17, 44, 47].

La distance génomique dont les méthodes de distances font usage est le nombre minimal de réarrangements requis pour transformer un génome en un autre. Un modèle biologique précis et restreint est utilisé par ces distances, car il faut choisir un seul type ou une certaine combinaison de réarrangements à minimiser. Le génome ancestral inféré avec les méthodes de distances est un génome médian : étant donné trois génomes A, B, C, on veut identifier un génome M tel que la somme $d(A,M) + d(B,M) + d(C,M)$ est minimale. Pour résoudre le «small phylogeny problem», on peut identifier les génomes ancestraux en résolvant itérativement le problème du génome médian aux noeuds internes de l'arbre, de manière à minimiser la distance totale de l'arbre. Par la suite, on peut reconstituer l'historique des réarrangements génomiques chez les espèces étudiées. Cependant, le problème a été prouvé être NP-difficile pour un nombre de distances classiques (breakpoint, inversion, transposition)[2, 13, 54].

Les méthodes de synténie quant à elles, aussi dites «locales» ou «sans modèle», sont appelées ainsi car à la différence des méthodes de distances, où un modèle biologique bien précis et souvent loin de la réalité est utilisé pour l'inférence, les méthodes «sans modèle» n'utilisent qu'une information observable directement dans les génomes contemporains : la synténie. Selon les méthodes, la définition de synténie diffère. Soit le terme possède son sens original, qui est la présence de gènes sur un même chromosome. Soit une spécification lui est apporté, indiquant que deux gènes sont non seulement sur un même chromosome, mais en plus sont adjacents [8, 44, 47], ou encore toujours en «équipe» avec les mêmes autres gènes [5, 17]. Selon la définition de synténie utilisée, on obtiendra un ordre de gènes précis à l'ancêtre, ou bien on permettra une incertitude à cet ordre. Parmis tous les algorithmes de cette catégorie, le karyotype ancestral est rarement complètement assemblé, mais des régions ancestrales contiguës (RAC) sont plutôt reconstruites en nombre plus ou moins grand. Ces algorithmes sont discutés au chapitre 3.

Les méthodes d'inférence de génomes ancestraux ont toutes leur lots d'inconvénients. Elles peuvent suivre un modèle de réarrangement limité, ou encore elles ne permettent pas de considérer des espèces dont le contenu en gène diffère, ce qui est très

restrictif, surtout dans le cas d'une évolution par DGE. L'essentiel des travaux présentés dans le chapitre 4 est une nouvelle méthode de synténie pour l'inférence de génome ancestraux pouvant s'appliquer à des génomes avec des contenus en gènes différents et des gènes ayant évolué suite à des évènements de DGE. Une évaluation de la méthode est effectuée à l'aide de simulations et des applications sur des données provenant du séquançage de levures et de plantes céréalières sont décrites. Ce dernier chapitre est constitué d'un article soumis à la conférence internationale RECOMB-CG 2012.

CHAPITRE 2

MODÈLE BIOLOGIQUE

Dans ce chapitre nous commençons par décrire brièvement les macro-molécules essentielles à la vie ainsi que le dogme central de biologie moléculaire, expliquant les relations entre ces trois molécules. Nous introduisons ensuite les évènements de réarrangements génomiques modifiant l'ordre des gènes dans les génomes ainsi que les évènements de duplications modifiant le contenu en gènes.

2.1 Hérédité

2.1.1 ADN

L'hérédité des êtres vivants est encodé dans une molécule organique, l'**acide désoxyribonucléique ou ADN**, présente dans chacune des cellules. Cette molécule est formée d'une chaîne de nucléotides, eux même composés d'un sucre (désoxyribose), d'un groupement phosphate et d'une base azotée. Les nucléotides sont séparés en deux familles, distinguées par les bases azotées présentes dans leur composition : les bases adénine (A) et guanine (G) forment les purines, les bases thymine (T) et cytosine (C) forment les pyrimidines. Une complémentarité existe entre ces deux familles, permettant les appariements des nucléotides A avec T et C avec G [1]. Ainsi, deux séquences de nucléotides, ou **brins**, s'assemblent pour former une structure en double hélice, l'ADN, chaque nucléotide étant apparié à son complément. Ces deux brins sont orientées en sens opposé : l'un dans le sens 3' vers 5', l'autre 5' vers 3'. 3' et 5' représentent les atomes de carbones du désoxyribose aux extrémités des brins. La lecture des brins se fait toujours de 5' vers 3'. Par convention, ces deux orientations sont notées par les signes $+$ et $-$, le brin de référence étant positif. Une molécule d'ADN à deux brins se nomme **chromosome**. Chez les procaryotes, la majorité des espèces ne contiennent qu'un seul chromosome circulaire, c'est à dire que les deux extrémités de la molécule d'ADN sont connectées ensemble.

[1] D'autres combinaisons sont possibles mais moins fréquentes.

Chez les eucaryotes par contre, on retrouve des chromosomes linéaires dont les deux extrémités sont libres, en nombre variable. L'ensemble de tout le matériel héréditaire, donc de tous les chromosomes d'une espèce forme le **génome**.

2.1.2 Réplication de l'ADN et mitose

Afin de permettre aux cellules de proliférer par division dans un organisme, l'ADN doit se répliquer afin que les cellules issues de la division (cellules filles) contiennent chacune une copie de l'information génétique. Ainsi, juste avant la division, des protéines reconnaissent le site d'initiation de la réplication sur les molécules d'ADN et recrutent un complexe permettant le déroulement de la structure en double hélice. Les deux brins ainsi séparés, l'enzyme ADN polymérase complémente chacun des brins à l'aide des nucléotides appropriés, corrigeant même des erreurs au besoin, permettant d'obtenir une copie quasi identique de la molécule originale. Une série d'évènements positionne ensuite les deux copies de chaque chromosome en paires sur un axe linéaire, les sépare puis les transporte vers les positions des futures cellules filles. Ce type de division cellulaire se nomme la **mitose**.

2.1.3 Méïose

Un autre type de division cellulaire intervient chez les espèce à reproduction sexuée, la **méïose**. Chez ces espèces, un type de cellule, les gamètes, doivent fusionner lors de la reproduction pour donner naissance à un nouvel individu. Les gamètes, à la différences des autres cellules sont **haploïdes** : elle ne contiennent qu'une seule copie de chaque chromosome. Toutes les autres cellules sont **diploïdes** [2] : elles contiennent deux copies, chacune provenant d'un des deux parents. La méïose est le processus par lequel une cellule diploïde est divisée pour obtenir quatres cellules haploïdes. Lors de ce processus, les chromosomes homologues (lire similaires) du père et de la mère peuvent se chevaucher en certains endroits. Les parties chevauchantes sont échangées, aboutissant à un mélange du patrimoine génétique des deux parents dans les gamètes haploïdes. Ce

[2] À l'exception des espèces polyploïdes, ou plus de deux copies sont présentes.

phénomène d'échange se nomme la **recombinaison**.

2.1.4 Polyploïdie

L'haploïdie des gamètes est une règle générale mais pas absolue. Des accidents lors de la méïose peuvent empêcher les chromosomes homologues de se séparer, menant à des gamètes diploïdes. Lors de la fusion des gamètes mâles et femelles, on se retrouve avec un nouvel individu tétraploïde (ou triploïde si une des deux gamètes est diploïde et l'autre haploïde). Ce phénomène, nommé **polyploïdisation**, peut se produire à répétition et mener à un nombre élevé de copies de chromosomes, pair ou impair. Chez les animaux, la polyploïdie est un évènement plus rare que chez les plantes, où ce phénomène rend les spécimens plus adaptables à leur environnement à un point tel que la plupart des plantes sont polyploïdes. Cependant, des traces de polyploïdisation ancienne ont été observées chez un certain nombre d'espèces dans divers taxons eucaryotes tel que les levures, les mammifères et les poissons [50, 67].

2.2 Gènes et expression

Si le stockage de l'hérédité passe par l'ADN, l'expression des caractères héréditaires est plus complexe. Une deuxième molécule entre en jeu ici, l'**acide ribonucléique, ou ARN**. Sa structure est similaire à l'ADN, composée de nucléotides avec un sucre différent, (ribose, groupement phosphate, base azotée). Cependant, elle n'est en général formée que d'un seul brin. L'ARN est fabriquée par un processus nommé **transcription** : une enzyme catalysant la réaction de synthèse de l'ARN (ARN polymérase) se sert de l'ADN comme modèle pour former des molécules d'ARN. Les parties de l'ADN servant à la transcription sont appelés gènes.

Les **gènes** sont les sous-séquences de l'ADN contenant l'information nécessaire à la synthèse des ARN fonctionnels et des ARN codants [29]. Le brin de la double hélice d'ADN contenant la séquence d'un gène déterminera l'orientation + ou - du gène. Les ARNs fonctionnels, comme le suggère le nom, sont aptes à remplir une fonction. Les ARNs codant, ou ARNs messagers, sont un intermédiaire nécessaire à la synthèse

de **protéines**, la dernière macro-molécule essentielle à la vie, élément de base dans la construction des cellules. L'enzyme qui catalyse la réaction de synthèse des protéines, le **ribosome**, est composé d'ARNs fonctionnels. Le ribosome se fixe à un ARN messager et «lit» le brin d'ARN trois nucléotides à la fois. La **traduction** de l'information contenue dans la séquence nucléotidique de l'ARN en protéine consiste à faire correspondre un triplet de nucléotides à un acide aminé, élément constitutif des protéines. Le ribosome se déplace au prochain triplet et la traduction se poursuit, allongeant la séquence d'acides aminés constituant la protéine codée dans l'ARN messager. Le type de tissus auquel appartient la cellule détermine la portion du génome qui est transcrite et traduite en protéines dans cette cellule.

2.3 Réarrangements génomiques

La recombinaison, qui est une forme d'erreur de réplication de l'ADN, implique une coupure de la molécule d'ADN et des mécanismes de réparation. Cependant, ces mécanismes ne sont pas parfaits et sont à l'origine de modifications dans la structure des chromosomes. Par exemple, si un chromosome est brisé en deux points, le segment délimité par ces deux points peut être réinséré dans le chromosome dans un sens inversé par rapport à sa position originale. Ce type de réarrangement est appelé **inversion**. De plus, la recombinaison se fait souvent entre chromosomes homologues et de façon réciproque, c'est-à-dire que chaque chromosome «fournit» un segment d'ADN en échange d'un autre. On constate cependant des échanges d'ADN entre chromosome non homologues, phénomène nommé **translocation**. On peut considérer également un cas spécial de recombinaison non homologue où il y a un échange non réciproque qui de plus implique des chromosomes entiers : les **fusions et fissions** de chromosomes. Les mécanismes exacts de ces réarrangements ne sont pas évidents et l'existence de recombinaisons non réciproques est débattue. Récemment, Schubert et Lysak [60] ont proposé des mécanismes expliquant les recombinaisons non réciproque, les fusion et fissions de chromosomes, par le biais de translocations réciproques. Les conséquences des réarrangements chromosomiques sont multiples, car l'orientation, la position et l'ordre des

gènes dans le génome sont affectés.

2.4 Familles de gènes

En plus des erreurs lors de la réparation de l'ADN, les segments chromosomiques peuvent se chevaucher de manière inégale lors de la recombinaison, menant à des duplications d'un seul ou de plusieurs gènes en tandem (consécutifs dans le génome). D'autres mécanismes, tel la rétro-transposition de l'ARN vers l'ADN, donnent lieu à des duplicats dispersés. Un autre mécanisme plus radical, la polyploïdisation décrite à la section 2.1.4 entraîne la duplication non pas d'un seul ou d'un petit groupe de gènes, mais bien du génome dans sa totalité. Ces copies de gènes, après des milliers d'années d'évolution, accumulent des mutations. Intuitivement, on peut penser que la pression sélective sur les gènes produits par ces mécanismes de duplication doit être faible, étant donné la présence du gène original qui rempli toujours sa fonction biologique. Cependant, ce n'est pas nécessairement le cas et les duplicats subissent différents sorts [38]. Certains de ces duplicats perdent alors toute fonctionnalité, mais des vestiges du gène fonctionnel demeurent : ce processus se nomme pseudogénisation. D'autres subirons une sous-fonctionnalisation, c'est-à-dire qu'ils conserveront une partie de la même fonction, permettant à la cellule d'être plus efficace dans la production de la protéine encodée par le gène. Une autre possibilité est la néo-fonctionnalisation, soit l'acquisition d'une nouvelle fonction, ou encore une spécialisation. Par ces deux derniers scénarios, les évènements de duplications sont un facteur déterminant aidant à l'apparition de nouvelles espèces [49].

Tous les gènes ayant évolué à partir d'un même gène ancestral sont des **homologues** et forment une **famille de gène**. On distingue deux types de relation d'homologie : la **paralogie** et l'**orthologie**. Deux gènes sont orthologues lorsque la divergence des séquences nucléotidiques entre ces deux gènes est le résultat d'un évènement de spéciation. Dans le cas où la divergence a comme origine un évènement de duplication, ces deux gènes sont alors paralogues. Les paralogues d'intérêt pour ce mémoire sont ceux issus d'un évènement de DGE tel que la polyploïdie, nommés **ohnologues**. Étant donné que

la plupart des méthodes de reconstruction de génomes ancestraux ne permettent pas de prendre en compte des paralogues autres que des ohnologues, seulement ceux-ci seront considérés plus en profondeur dans la prochaine section. Pour une revue sur les autres types d'homologie, voir [40].

2.5 Identification d'ohnologues et détection de DGE

La méthode décrite dans ce paragraphe a été appliquée récemment pour inférerer une hexaploïdisation chez un ancêtre de la vigne [25]. La détection d'un évènement de DGE ancien se fait tout d'abord en identifiant les paralogues à l'intérieur d'une même espèce, par une comparaison de séquences de type tous contre tous. Par la suite, on cherche des paires de groupements (plusieurs gènes consécutifs) paralogues qui sont colinéaires, c'est-à-dire que les gènes dans chaque groupement conservent le même ordre. Si les paires de groupements colinéaires couvrent une portion significative du génome, on peut alors conclure qu'un évènement de DGE a eu lieu chez un ancêtre de l'espèce étudiée, étant donné que des copies de gènes issues d'une duplication massive ont conservé le même ordre. On peut également conclure que les paralogues formant les groupements colinéaires sont des ohnologues. La portion non couverte du génome s'explique par le fait que des évènements de réarrangements, ou encore des duplications provenant de différents mécanismes, brouillent les traces de colinéarité. En pratique, les traces d'un évènement de DGE peuvent être visualisées à l'aide d'un nuage de points, tel qu'illustré à la figure 2.1.

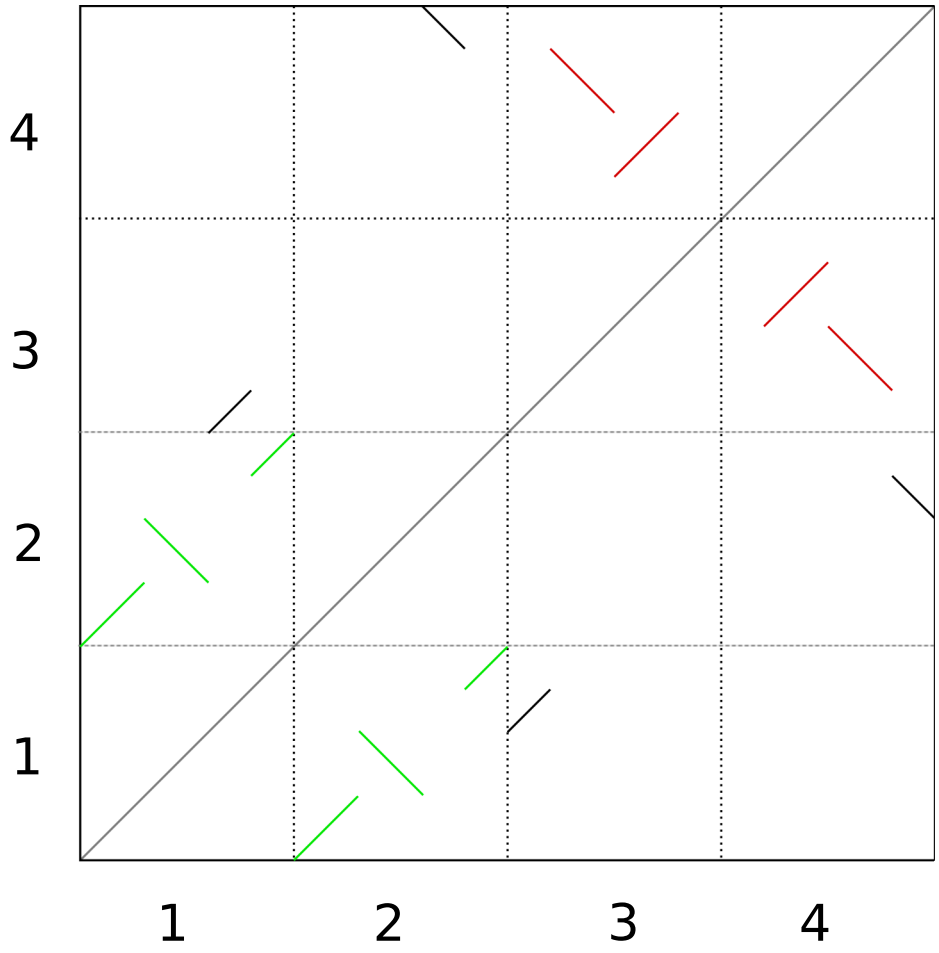


Figure 2.1 – Schématisation d'un nuage de points montrant les traces d'une DGE ancienne observable dans une espèce contenant quatres chromosomes. Sur les deux axes, on retrouve les quatres chromosomes du génome et chaque point représente des gènes paralogues. Les chromosomes 1 et 2 descendent d'un même chromosome ancestral, suggéré par la présence des groupements colinéaires en vert, ainsi que les chromosomes 3 et 4, suggéré par les groupements colinéaires en rouge.

CHAPITRE 3

MÉTHODES D'INFÉRENCE DE GÉNOMES ANCESTRAUX

Ce chapitre est dédié à une revue de littérature sur les méthodes de reconstruction de génomes ancestraux. La première section est consacrée à définir les concepts mathématiques et les notations nécessaires à la compréhension des méthodes. La seconde section porte sur les méthodes de distance et la troisième sur les méthodes de synténie pour l'inférence de génomes ancestraux, reflétant la classification déjà établie dans [16, 23].

3.1 Notations

Un génome G est un ensemble de chromosomes et un chromosome est composé d'une séquence de gènes, représentée par une chaîne de caractères sur un alphabet Σ. Chaque $a \in \Sigma$ est un identifiant pour tous les gènes d'une famille de gènes noté $\mathscr{F}_a$. Une famille $\mathscr{F}_a$ est composé de tous les gènes identifiés par a dans un ensemble Γ de génomes modernes. Σ_G représente l'ensemble des familles de gènes présentes dans le génome G, donc pour Γ contenant k génomes,

$$\Sigma = \Sigma_1 \cup \Sigma_2 \cup \Sigma_3 \cup ... \cup \Sigma_k$$

Un **génome simple** est un génome qui ne contient que des caractères uniques, donc un seul gène d'une même famille de gène. Un **génome avec duplicats** peut contenir des caractères dupliqués, donc plusieurs gènes d'une même famille. La multiplicité d'un gène a dans G, noté $mult(a,G)$, est le nombre de gènes de $\mathscr{F}_a$ dans G.

Deux génomes simples G,H ont un contenu égal en gène si $\Sigma_G = \Sigma_H$, autrement ils ont un contenu inégal en gène. De plus, dans le cas des génomes avec duplicats, le contenu en gène est égal si pour tout $a \in \Sigma_G$, $mult(a,G) = mult(a,H)$.

On note également $\pm\Sigma_G$ l'ensemble obtenu en considérant l'orientation des gènes, tel qu'illustré dans le tableau 3.I. Cependant, pour ne pas allourdir le texte, nous consi-

Tableau 3.I – Illustration des notations

G	A	B	C	D
Séquence	$+a+b+c+d$	$+a-d-c-b$	$+a+b+d$	$+a+b-d$, $+a+b+c$
Σ_G	$\{a,b,c,d\}$	$\{a,b,c,d\}$	$\{a,b,d\}$	$\{a,b,c,d\}$
$\pm\Sigma_G$	$\{+a,+b,+c,+d\}$	$\{+a,-b,-c,-d\}$	$\{+a,+b,+d\}$	$\{+a,+b,+c,-d\}$
$mult(a,G)$	1	1	1	2
$mult(b,G)$	1	1	1	2
$mult(c,G)$	1	1	0	1
$mult(d,G)$	1	1	1	1

dérerons les génomes comme non-signés dans les explications.

Un arbre d'espèce, ou phylogénie, noté S, est un arbre binaire[1] enraciné, représentant les relations de descendence d'un ensemble de k espèces modernes. Les noeuds internes représentent les espèces ancestrales éteintes et les k feuilles représentent les espèces modernes. Chacune des espèces ancestrales ou modernes de S est étiquetté par un génome. L'ensemble des génomes aux feuilles est Γ. Lorsqu'il n'y aura pas d'ambiguité, nous confondrons les noeuds et les feuilles avec leurs étiquettes.

Lorsque l'évolution des espèce implique un ou des évènements de DGE, on utilise alors un arbre S^{DGE}, qui est S augmenté de noeuds représentant les évènements de DGE. Plus précisément, ces noeuds ayant un unique fils sont ajoutés sur les branches où ont eu lieu ces évènements. Dans ce contexte, un noeud pré-dupliqué, noté u_p, est un noeud pour lequel il n'existe aucun noeud de DGE sur un chemin allant de la racine de S^{DGE} jusqu'à u_p.

Un arbre de gènes noté T_a pour une famille $\mathscr{F}_a$, est un arbre binaire enraciné représentant l'évolution des éléments de $\mathscr{F}_a$. Les feuilles sont étiquettés par les éléments de $\mathscr{F}_a$ et les noeuds internes représentent des gènes ancestraux.

Finalement, un ensemble de synténies (terme défini à la section 3.3.2) pour un génome G sera noté $\mathscr{S}_G$. Les notations étant introduites, nous pouvons maintenant entrer

[1] Un arbre d'espèce n'est pas nécessairement toujours binaire, mais dans ce mémoire nous nous intéressons seulement à ceux-ci.

dans le coeur du sujet et donner l'idée générale des méthodes existantes pour l'inférence de génomes ancestraux, en commençant par les méthodes de distance.

3.2 Méthodes de distance

Le problème que les méthodes de distance pour l'inférence de génomes ancestraux visent à résoudre est le suivant.

«SMALL PHYLOGENY PROBLEM» : Soit S une phylogénie pour un ensemble Γ de génomes modernes, le problème est de trouver un génome à chacun des noeuds internes de S.

Tel que la classification le suggère, ces méthodes utilisent les distances génomiques pour tenter de résoudre ce problème. Une distance génomique est le nombre minimal de réarrangements génomiques requis pour transformer un génome en un autre. Les réarrangements peuvent être un de ceux déjà abordés à la section 2.3, soit les inversions, les translocations, les fusions et fissions de chromosomes.

Par exemple, la formule d'Hannenhali et Pevzner (HP) [35, 63] permet de calculer le nombre minimum d'inversions, de translocations, ou d'inversions et de translocations requises pour transformer un génome en un autre. Publiée plus récemment, la distance DCJ [6, 7, 68] (de l'anglais double cut-and-join, double couper-et-coller) permet de simuler tous les types de réarrangements par une ou deux opérations DCJ consécutives.

L'idée générale de la reconstruction de génomes ancestraux à l'aide de distances génomiques, introduite par Sankoff *et al.* [59], est de choisir des ordres de gènes aux noeud internes de manière à minimiser la somme des distances génomiques sur S. Si S est un arbre binaire non enraciné avec $|\Gamma| = k$, alors S peut être décomposé en $k-2$ étoiles à 3 branches, chacune centrée sur un unique noeud interne de S. Le problème se ramène donc a résoudre $k-2$ fois le problème du génome médian : étant donné une distance d et trois génomes A, B, C aux pointes de l'étoile, on veut identifier un génome M au centre de l'étoile tel que la somme $d(A,M)+d(B,M)+d(C,M)$ soit minimale (voir figure 3.1). Au

fil des ans, plusieurs implémentations du calcul de distance et du problème du génome médian ont été proposées afin de généraliser la méthode originale en traitant des cas spécifiques. Ainsi, le calcul d'une distance d'inversion est maintenant possible entre deux génomes, même en présence d'insertions et/ou suppressions de gènes [21, 45]. Lorsque la phylogénie contient un évènement de DGE sur une branche menant à une feuille, l'algorithme «genome halving» [22] permet d'identifier le génome ancestral pré-dupliqué, c'est-à-dire le génome immédiatement après la duplication, lorsqu'il est formé par deux copies identiques de chaque chromosome. Une généralisation du problème du génome médian a aussi été proposé pour inclure les évènements de DGE, le «guided genome halving» [70].

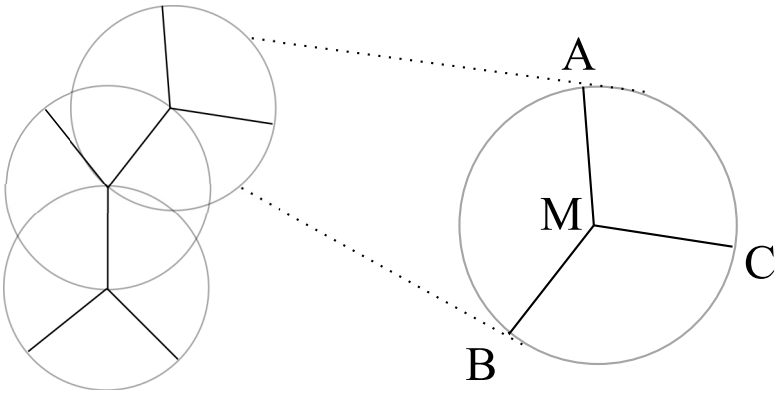


Figure 3.1 – Pour k espèces modernes, le «small phylogeny problem» (à gauche) peut se résoudre en itérant $k-2$ fois le problème du génome médian (à droite).

Les limitations de ces méthodes, recensées par Gordon *et al.* [31] et par Sankoff [56] sont de deux ordres. D'abord, un nombre important de solutions différentes peuvent être optimales selon la fonction objective. Deuxièmement, les distances sont restreintes à un type de réarrangement ou une combinaison de quelques uns seulement, ce qui peux mener à des solutions inexactes. Malgré les avancées faites dans cette classe de méthodes, tant que les modèles biologiques sous-jacents au calcul de distance ne pourront être plus complets, le plein potentiel de cette classe de méthodes ne pourra être exploitée.

3.3 Méthodes de synténie

3.3.1 Définition du problème

Le problème considéré par les méthodes de synténie diffère légèrement de celui des méthodes de distance. Redéfinissons-donc le problème, illustré à la figure 3.2.

PROBLÈME D'INFÉRENCE :
Soit S, une phylogénie pour un ensemble Γ de génomes modernes et soit u un noeud interne de S. Le problème est de trouver un génome A pour le noeud u.

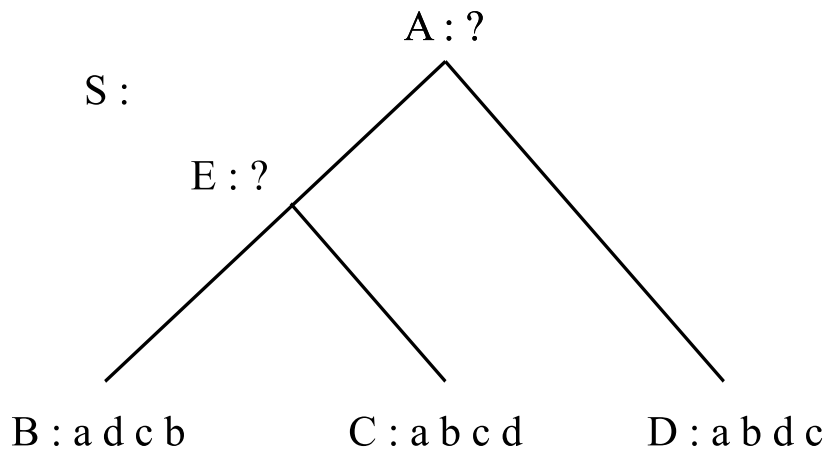


Figure 3.2 – Illustration du problème d'inférence avec un arbre phylogénétique S pour les espèces $\Gamma = \{B, C, D\}$. Ici, $\Sigma = \{a, b, c, d\}$.

Excluant toute autre source de duplication, s'il y a présence d'une ou de plusieurs DGE dans S, alors un sous-ensemble de génomes modernes $\gamma \subseteq \Gamma$ contiendra des ohnologues. C'est-à-dire que pour tout $G \in \gamma$, G sera un génome avec duplicats. De plus, le nombre de répétitions autorisés dans G ne peut dépasser deux fois le nombre de DGE présent sur un chemin évolutif partant de la racine de S et menant à G. Supposant qu'au moins un gène dans tout génome de γ reflétera ces duplications (c'est-à-dire qu'il n'y a pas eu de perte des autres membre de sa famille), un historique avec un minimum de DGE peut facilement être déduit de la multiplicité des gènes les plus fréquents dans

chaque génome. Cet historique permet la construction de S^{DGE}. Ceci nous mène à une variante du problème décrit plus haut, illustré à la figure 3.3.

PROBLÈME D'INFÉRENCE AVEC DGE :
Soit S^{DGE}, une phylogénie augmentée de noeuds de DGE pour un ensemble Γ de génomes modernes simples ou avec duplicats et u_p, un noeud de S^{DGE} pour lequel il y a absence de noeuds de DGE sur le chemin évolutif entre la racine et u_p. Le problème consiste alors à trouver un génome simple A_p pour le noeud u_p.

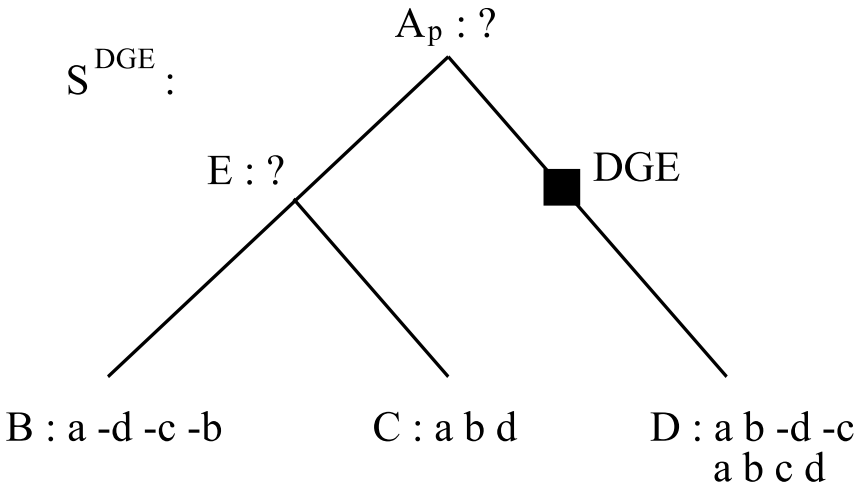


Figure 3.3 – Illustration du problème d'inférence avec DGE avec une phylogénie augmentée S^{DGE} pour les espèces $\Gamma = \{B, C, D\}$. Ici, $\Sigma = \{a, b, c, d\}$

La difficulté empêchant d'inférer tout génome à n'importe quel noeud de S^{DGE} est l'incapacité de distinguer les différents gènes d'une famille quelconque $\mathscr{F}$ pour un génome G avec duplicats. S'il était possible, pour chaque gène dans $\mathscr{F}$, d'étiquetter différemment les gènes originaux et les copies et de les identifier dans tous les génomes de Γ, à ce moment on pourrait inférer tout génome ancestral dans S^{DGE}. Un ensemble d'adjacences potentielles pourrait être attribué à une copie étiquetté de gène et non pas seulement à un identifiant d'une famille. Par exemple, considérons un génome avec duplicats $G = \{(abcd), (adcb)\}$ dont les copies de gènes ne sont pas identifées et un gé-

nome $H = \{(a_1 b_2 c_1 d_1), (a_2 d_2 c_2 b_1)\}$ où les copies de gènes sont identifiées. Dans G, on peut seulement dire qu'une copie de a est adjacent à b ou d dans un génome ancestral. Cependant, dans H, on peut dire que la copie a_1 est adjacente à b_2 et que la copie a_2 est adjacente à d_2. Dans le cas du génome G, cette situation ne permet que d'inférer des génomes ancestraux simple, ne contenant qu'une seule copie de chaque gène. Dans le cas du génome H, cela permet d'inférer des génomes ancestraux avec duplicats.

3.3.2 Concepts de synténie

Le problème étant posé, définissons maintenant le concept central de cette classe de méthodes : la synténie. Le terme synténie fit son apparition en génétique, pour caractériser la co-localisation de marqueurs génomiques (gène ou autre) sur un même chromosome. En génomique comparative, on ajoute une dimension d'organisation linéaire à ce terme. Ainsi, selon les travaux, le concept de synténie se défini en terme d'adjacence [8, 44], d'α-ajdacence (chapitre 4) ou d'intervalle [17].

Définition 3.3.1. *Soit G un génome et a, b deux gènes de G. Soit α un entier strictement positif. Alors, a et b sont dits* ***α-adjacent*** *dans G si a et b sont séparés par au plus $\alpha - 1$ gènes dans G.*

Plus précisément, le segment de G délimité par les gènes a et b contient au plus $\alpha + 1$ gènes. Une 1-adjacence sera tout simplement appelé adjacence ou adjacence immédiate. De plus, considérant une orientation arbitraire du chromosome contenant a et b, on dira que b est une α-adjacence droite de a si b est positionné à droite de a sur le chromosome. De façon similaire, on dira que b est une α-ajdacence gauche de a si b est à gauche de a sur le chromosome.

Exemple 3.3.1. *Dans le génome C de la figure 3.2, l'adjacence gauche de b est a et l'adjacence droite de b est c.*

Exemple 3.3.2. *Dans le génome C de la figure 3.2, {b} est une 1-adjacence droite, $\{b, c\}$ sont des 2-adjacences droites et $\{b, c, d\}$ sont des 3-adjacences droites de a.*

Définition 3.3.2. *Soit G un génome. Un **intervalle** de gènes est un sous-ensemble $\sigma \subseteq \Sigma_G$ tel que tous les éléments de σ sont consécutifs dans G.*

Exemple 3.3.3. *Dans le génome B de la figure 3.2, les intervalles de taille deux sont :$\{a,d\}$, $\{c,d\}$, $\{b,c\}$; ceux de taille trois sont : $\{a,c,d\}$, $\{b,c,d\}$; celui de taille quatre est $\{a,b,c,d\}$.*

Définition 3.3.3. *Un intervalle σ est un **intervalle commun** pour les génomes G,H si σ est un intervalle dans G et σ est un intervalle dans H.*

Exemple 3.3.4. *L'intervalle commun de taille deux pour les génomes B,D de la figure 3.2 est $\{c,d\}$, celui de taille trois est $\{b,c,d\}$ et celui de taille quatre, $\{a,b,c,d\}$.*

3.3.3 Cadre commun aux méthodes de synténies

Le concept de synténie étant défini, expliquons maintenant l'idée générale des méthodes de cette catégorie. Toutes les méthodes de synténie pour l'inférence de génomes ancestraux suivent le cadre décrit par Chauve et Tannier [17], qui consiste en quatre étapes.

Étape 1 : Identifier les gènes ancestraux aux noeuds internes de *S*.

Étape 2 : Identifier un ensemble de synténies ancestrales potentielles à partir des génomes actuels.

Étape 3 : Attribuer un poids à chacune des synténies ancestrales.

Étape 4 : Chaîner de manière optimale les synténies ancestrales pour obtenir des régions ancestrales contigües (RAC).

Les différences entres les méthodes de cette classe résident principalement dans la définition de synténie, le poids qui est attribué ainsi que dans la manière de chaîner les synténies ancestrales. Ce type de méthodes peut donner lieu, tout comme pour les méthodes de distance, à un nombre important de solutions, étant donné la phase d'optimisation à l'étape 4. Cependant, les méthodes de synténies ne sont pas, jusqu'à un certain

point, limitées par un modèle biologique contraignant. En effet, le concept de synténie tel que défini à la section 3.3.2 n'implique aucune notion de réarrangement génomique. Par contre, les méthodes de synténie ne sont pas totalement sans modèle biologique. À l'étape 1, l'ensemble des gènes présents à l'ancêtre dépendra du fait que l'on permette des contenus en gènes inégaux aux génomes aux feuilles, dû à des pertes ou duplications de gènes.

3.3.4 InferCar par Ma *et al.*

InferCar [44] est la première méthode de synténie implémentée. Cette méthode considère les adjacences immédiates de gènes signés et ne permet que des génomes simples dont le contenu en gènes est égal aux feuilles.

Étape 1 : $\Sigma_A = \Sigma$, comme le contenu en gène est égal dans tous les génomes de Γ.

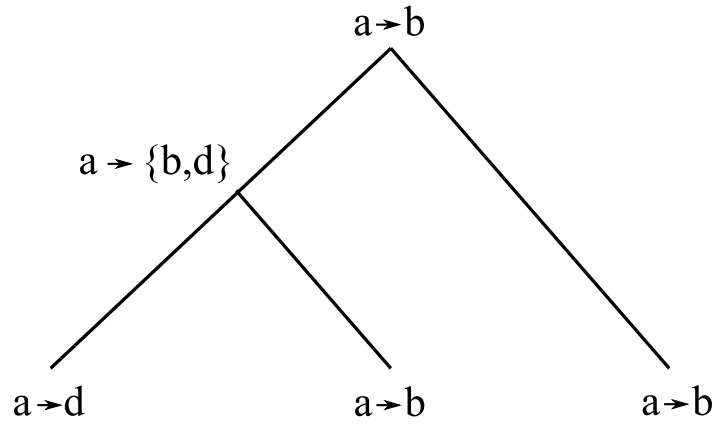


Figure 3.4 – Représentation de l'étape 2 de la méthode de Ma *et al.* pour déterminer les adjacences droites potentielles du gène a aux noeuds internes de S étant donné l'instance du problème de la figure 3.2. $a \rightarrow b$ indique que b est l'adjacence droite de a.

Étape 2 : Pour chaque gène $a \in \Sigma$ les adjacences ancestrales potentielles aux noeuds internes de S sont inférées par une méthode semblable à la parcimonie de Fitch. Soit $AD_{(a,u)}$ l'ensemble des adjacences droites potentielles de a à un noeud interne u, et soit

v, w les fils gauche et droit respectivement de u. En faisant un parcours postfixe de S,

$$AD_{(a,u)} = \begin{cases} AD_{(a,u)} & \text{si } u \text{ est une feuille,} \\ AD_{(a,v)} \cup AD_{(a,w)} & \text{si } AD_{(a,v)} \text{ et } AD_{(a,w)} \text{ sont disjoints,} \\ AD_{(a,v)} \cap AD_{(a,w)} & \text{autrement.} \end{cases}$$

De façon similaire, l'ensemble des adjacences gauches est calculé. Cette étape est illustrée à la figure 3.4.

Étape 3 : Dans le cas où il y a ambiguité sur les adjacences de a, c'est-à-dire que a possède plus d'une adjacence droite (ou gauche) potentielle, un poids est attribué aux adjacences de a afin de les départager. Le poids $p_u(a,b)$ d'une adjacence entre deux gènes a et b à un noeud u est également calculé par un parcours postfixe de S. Plus précisément,

$$p_u(a,b) = \begin{cases} 1 & \text{si } u \text{ est une feuille et l'adjacence (a,b) existe à } u \\ 0 & \text{si } u \text{ est une feuille et l'adjacence (a,b) n'existe pas à } u \\ \frac{L_v p_v(a,b) + L_w p_w(a,b)}{L_v + L_w} & \text{autrement} \end{cases}$$

où L_v, L_w sont les longueurs de branches entre le noeud u et ses deux fils (la longueur de branche représente la distance évolutive ; InferCar ne la calcule pas et doit être fournie avec l'arbre d'espèce en entrée). Une schématisation et un exemple de la procédure sont présentées à la figure 3.5, à gauche et à droite respectivement.

Étape 4 : Enfin, une heuristique gloutonne permet de construire les RAC, en ajoutant dans l'ordre décroissant, les adjacences aux poids élevés et rejetant celles de poids plus faible causant des ambiguités. Plus spécifiquement, la solution ne doit contenir aucun des trois cas suivant : (1) une gène a possède plus d'une adjacence gauche, (2) un gène a possède plus d'une adjacence droite, (3) le RAC est circularisé par l'adjacence ajoutée

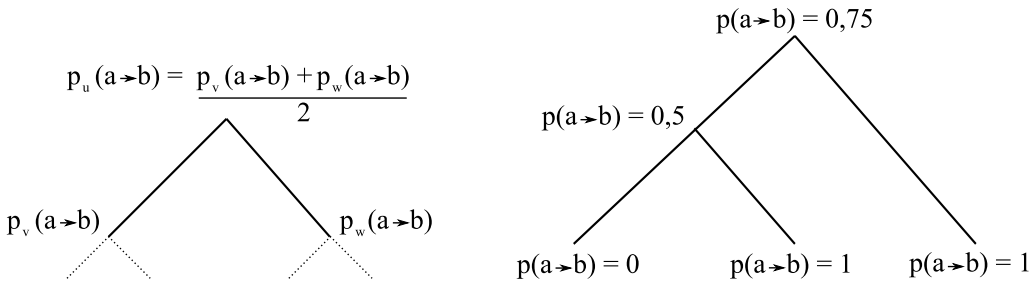


Figure 3.5 – Gauche : Schématisation de la récurrence permettant de calculer par programmation dynamique le score d'une adjacence selon la méthode InferCar, étant donné un sous-arbre de S enraciné au noeud u. Droite : Valeurs du poids de l'adjacence (a,b), étant donné l'ensemble des adjacences droites potentielles de a aux noeuds internes à la figure 3.4 et en suivant la procédure décrite dans le texte. Les longueurs de branches sont fixées à une valeur de 1.

(voir figure 3.6). Si l'adjacence ajoutée crée l'un de ces trois cas, elle est ignorée et la prochaine adjacence dans la liste des adjacences potentielle est considérée.

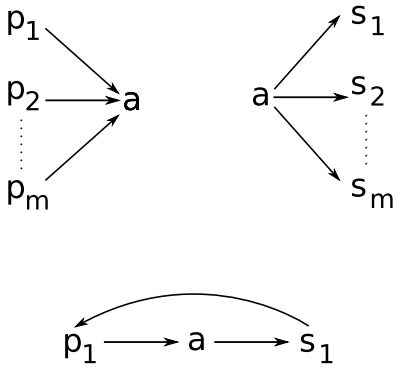


Figure 3.6 – Cas ambigus lors de l'assemblage des adjacences en RAC. Haut Gauche : un gène a possède plus d'une adjacence gauche. Haut Droite : un gène a possède plus d'une adjacence droite. Bas : circularisation du RAC.

3.3.5 DupCar

Une extension à la méthode de la dernière section, DupCar [43], permet des génomes avec duplicats. La présence de duplicats peut être le résultat de n'importe quel type de

duplication en plus des DGE, donc la restriction sur le nombre de duplicats ne tient plus. De plus, la méthode peut inférer tout génome à n'importe quel noeud interne de S. La méthode requiert pour tout $a \in \Sigma$ un arbre de gènes T_a représentant l'évolution de $\mathscr{F}_a$.

Pour une famille de gènes $\mathscr{F}_a$ ayant évolué sans duplications ni pertes, la topologie de T_a devrait reflèter en théorie celle de S. Cependant, lorsque la famille contient des paralogues, la topologie de T_a peut différer de celle de S. Pour expliquer la divergence entre les deux topologies, la méthode la plus utilisée est celle de la «réconciliation» d'arbres [9, 15, 20, 32, 33, 42, 52, 53]. Initialement introduite par Goodman en 1978 [30], elle consiste à emboîter l'arbre de gènes dans l'arbre d'espèce et à interpréter la divergence entre les deux.

Plus précisément, une réconciliation entre un arbre de gène T_a et un arbre d'espèce S est une cartographie des noeuds de T_a sur S posant des hypothèses de duplications et pertes de gènes pour expliquer la divergence entre les topologies de T_a et de S. Les feuilles de T sont associées aux feuilles de S correspondant à leur espèce de provenance. Les noeuds internes de T_a représentant des spéciations sont associés aux noeuds internes correspondant à la même spéciation dans S. Les noeuds internes de T_a représentant des duplications sont associés à la branche de S où a eu lieu la duplication.

Revenant à la description de DupCar, la première étape consiste en une réconciliation des arbres de gènes avec l'arbre des espèces. Le processus de réconciliation permet deux choses essentielles pour traiter les gènes dupliqués dans la reconstruction de génomes ancestraux. Premièrement, cela permet de distinguer clairement et sans ambiguité les différentes copies d'un gène chez toutes les espèces modernes considérées, car les relations d'orthologies et de paralogies entre les gènes d'une même famille sont révélées par la réconciliation. Cela résoud le problème soulevé à la fin de la section 3.3.1. Deuxièment, elle permet de connaître la multiplicité des gènes aux noeuds internes de l'arbre des espèces, puisque l'on sait par la réconciliation où les évènements de duplications ont eu lieu dans S.

Ces deux problèmes résolus, DupCar fonctionne essentiellement de manière identique à InferCar, nonobstant quelques technicalités.

Ce qui fait la force de la méthode DupCar fait également sa faiblesse. En utilisant les

arbres de gènes et la réconciliation, la méthode s'expose à des failles supplémentaires. En effet, la véracité des arbres de gènes est débattable car on ne connait pas l'histoire évolutive réelle de ces gènes. Cette incertitude peut avoir des conséquences importantes sur les résultats de la réconciliation. En effet, seulement quelques feuilles incorrectement positionnées dans l'arbre de gènes peut mener à un historique de duplications et pertes totalement différent [34, 55]. De plus, le nombre de famille de gènes peut se compter par milliers selon les génomes considérés, donc la méthode se base sur des milliers de résultats possiblement erronés. Pour ces raisons, cette façon de tenir compte des duplications n'est pas idéale dans un contexte de reconstruction de génomes ancestraux et c'est pourquoi les autres méthodes n'en font pas l'usage.

3.3.6 Méthode des uns consécutifs par Chauve et Tannier

Une méthode différente de synténie par Chauve et Tannier [17] utilise les intervalles de gènes comme définition de synténie. Originellement conçue pour résoudre le problème d'inférence formulé à la section 3.3.1, la méthode a ensuite été modifiée pour résoudre le problème d'inférence avec DGE [62].

Étape 1 : L'inférence du contenu de l'ensemble Σ_A est étroitement liée à l'étape 2.

Étape 2 : Les synténies ancestrales potentielles sont des intervalles communs dans un minimum de deux espèces dont le chemin évolutif passe par l'ancêtre d'intérêt. L'ensemble $\mathscr{I}_A$ des intervalles communs présents à l'ancêtre déterminera du même coup l'ensemble Σ_A. Reprenant l'exemple d'instance du problème de la figure 3.2, on considère les intervalles communs entre les génomes B et D, soit $\{\{c,d\},\{b,c,d\},\{a,b,c,d\}\}$ de même que ceux entre les génomes C et D, soit $\{\{a,b\},\{b,c,d\},\{a,b,c,d\}\}$. Alors, $\mathscr{I}_A = \{\{a,b\},\{c,d\},\{b,c,d\},\{a,b,c,d\}\}$ et $\Sigma_A = \{a,b,c,d\}$.

Une relaxation à la définition d'intervalle communs peut également être employée, permettant que les gènes ne soient pas nécessairement consécutifs dans les génomes des deux espèces, ce qui permet la présence de génomes aux contenus en gènes inégaux aux feuilles. Pour le problème d'inférence avec DGE, une synténie ancestrale potentielle

pourra alors en être une qui est conservée deux fois chez une espèce descendente d'une DGE. Cependant, l'ancêtre inféré doit être un ancêtre pré-dupliqué car la méthode de construction des RAC utilisé, qui est décrite à l'étape 4, ne permet pas la présence de gènes dupliqués. En effet, aucune distinction entre les copies de gènes ne peut être faite lorsque l'on compare plusieurs espèces (voir les explications à la fin de la section 3.3.1). Quoique les autres types de duplications doivent être ignorés, l'identification des ohnologues ne requière pas d'arbre de gènes ni de réconciliation (voir section 2.5), ce qui est un avantage sur la méthode de Ma.

Étape 3 : Le poids qui est attribué aux synténies ancestrales potentielles est calculé de la même manière que dans la méthode InferCar décrite à la section 3.3.4.

Étape 4 : Étant donné un ensemble Σ_A à l'ancêtre et un ensemble $\mathscr{I}_A$ d'intervalles conservés, l'information des synténies ancestrales potentielles est encodée dans une matrice de taille $|\Sigma_A| \times |\mathscr{I}_A|$. Chaque colonne représente un élément de Σ_A et chaque ligne un élément de $\mathscr{I}_A$. La présence d'un gène dans un intervalle est encodé par l'entrée 1 et l'absence par 0. Le problème devient alors le «problème des uns consécutifs» [26], qui consiste à ordonner les colonnes de manières à ce que tous les 1 soient consécutifs sur chacune des lignes. S'il est possible d'effectuer une telle réorganisation des colonnes, alors la matrice possède la propriété des 1 consécutifs (C1P en abrégé) et un tel ordre de colonnes représente également l'ordre des gènes à l'ancêtre. S'il est impossible de réordonner les colonnes de manière à ce que la matrice respecte la C1P, alors il y a de l'information conflictuelle dans les synténies ancestrales considérées. On cherche alors à retirer un minimum de lignes de la matrice afin d'éliminer les faux positifs dans l'ensemble de synténies ancestrales potentielles et résoudre les conflits. Les synténies de plus faible poids sont retirées en priorité. Tout comme InferCar à la section 3.3.4, une heuristique gloutonne permet d'effectuer cette opération.

Souvent, plusieurs réorganisations différentes des colonnes de la matrice respecteront la C1P, donnant lieu à plusieurs ordres de gènes ancestraux, ce qui est un problème

$$\begin{array}{cccc} a & b & c & d \\ \end{array} \qquad \begin{array}{cccc} c & d & b & a \\ \end{array}$$
$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \qquad \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \end{pmatrix}$$

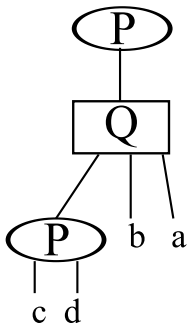


Figure 3.7 – Haut : Deux matrices encodant $\mathscr{I}_A = \{\{a,b\},\{c,d\},\{b,c,d\},\{a,b,c,d\}\}$, provenant de l'instance de la figure 3.2 et respectant la propriété des 1 consécutifs. Bas : Arbre PQ encodant les différentes solutions des deux matrices.

récurrent pour les méthodes décrites dans ce mémoire. De ce fait, la solution finale obtenue par la méthode de Chauve et Tannier est une représentation compacte de l'ensemble des solutions encodée dans une structure de données : les arbres PQ [10]. Ces arbres non-binaires sont composés de feuilles et de deux types de noeuds internes : les noeuds P et les noeuds Q. Les feuilles sont des éléments de Σ_A, les fils des noeuds P peuvent être réordonnés de quelque manière que ce soit et les fils des noeuds Q possèdent un ordre bien précis. L'ordre des fils des noeuds Q est connu à une inversion près, c'est-à-dire qu'il peut être lu de gauche à droite ou de droite à gauche. La racine de l'arbre est toujours un noeud P et chacun des sous-arbres enraciné à l'un de ses fils représente un RAC de l'ancêtre considéré. Par exemple, l'arbre PQ au bas de la figure 3.7 n'encode qu'un seul RAC étant donné que le noeud P à la racine ne possède qu'un seul fils. De plus, l'arbre encode autant les ordres de gènes $(abcd)$ que $(cdba)$.

3.3.7 Méthode des adjacences de Bertrand *et al.*

Notre laboratoire a également développé une méthode de synténie pour le problème d'inférence avec DGE [8]. Cette approche permet des génomes avec un contenu inégal en gènes.

Étape 1 : Pour chaque génome ancestral A de S, Σ_A de même que la multiplicité de chaque élément de Σ_A sont déterminés par une procédure en deux traversées de bas-en-haut de S^{DGE}.

La première permet de déterminer la multiplicité de chaque gène a à chaque noeud de DGE u. Pour ce faire, on considère $mult_{max}(a,u)$, la multiplicité maximale d'un gène à une feuille de S descendante de u (ou a un autre noeud de DGE si celui-ci descend de u). La multiplicité de a à u est alors $\lfloor \frac{mult_{max}(a,u)}{2} \rfloor$.

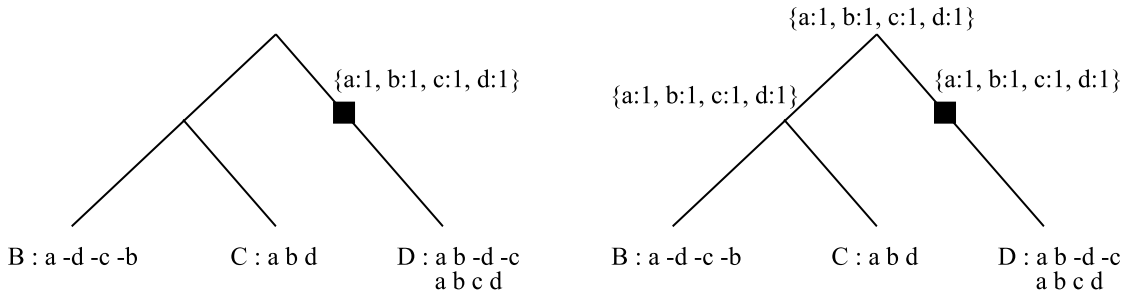


Figure 3.8 – Inférence de Σ_u pour chaque noeud interne u de l'instance de la figure 3.3 suivant la procédure décrite à l'étape 1 de la méthode de Bertrand *et al.* Gauche : État de Σ_u après la première passe de la procédure. Droite : État de Σ_u après la seconde passe.

La deuxième passe permet de compléter Σ_A pour tous les noeuds internes. À Σ_A, tout gène a est ajouté si A est situé sur un chemin entre le dernier ancêtre commun de toutes les feuilles contenant a et une de ces feuilles. La multiplicité de a à A, si elle n'est pas déjà établie, est définie par la multiplicité maximale de a chez les deux fils de A. Un exemple de la procédure est présenté à la figure 3.8.

Étape 2 : Pour chaque génome ancestral A dans S et pour chaque gène $a \in \Sigma_A$, l'ensemble $\mathscr{I}_A(a)$ de toutes les adjacences à gauche (et à droite) de toutes les copies de a observables dans chacun des génomes de Γ sont considérées comme étant potentielle-

ment présentes dans le génome A si les gènes adjacents sont égalements présent dans Σ_A. Plus spécifiquement, $\mathscr{I}_A(a)$ est un multiensemble, car les différentes copies d'un gène peuvent être adjacent à des gènes de la même famille. Par exemple, les adjacences droite de a possible dans le génome D sont $\{b,b\}$ à la figure 3.8.

Étape 3 : La contribution principale de cette méthode par rapport à InferCar et celle des uns consécutifs est le fait que le poids attribué à chaque adjacence possède un sens rigoureux : il représente la proportion maximale de branches dans S où cette adjacence potentielle peut être conservée. Soit $b \in \mathscr{I}_A(a)$, une adjacence (gauche ou droite) d'un gène $a \in \Sigma_A$, le poids de l'adjacence (a,b) à l'ancêtre A, $AdjCons(a,b,A)$, est calculé par un algorithme de programmation dynamique en deux passes. À tout noeud u de S, la première passe calcule $AdjCons_{below}(a,b,u)$ qui est le nombre maximal d'adjacence conservée dans le sous-arbre enraciné en u, si dans le génome associé à u on infère l'adjacence b pour le gène a. À la deuxième passe, l'algorithme calcule $AdjCons_{above}(a,b,u)$ qui est le nombre maximal d'adjacences conservées dans toutes les branches de S hors du sous-arbre enraciné en u si on choisi l'adjacence b pour un gène a. Finalement,

$AdjCons(a,b,A) = AdjCons_{below}(a,b,A) + AdjCons_{above}(a,b,A)$.

La procédure est illustrée à la figure 3.9

Étape 4 : Plutôt que d'utiliser une heuristique gloutonne pour résoudre les ambigüités dans les adjacences potentielles et former les RAC, le problème est modélisé comme une instance du problème du commis voyageur (de l'anglais traveling salesman problem ou TSP en court). Plus précisément, un graphe non orienté est construit, avec chacune des deux extrémités d'un gène représentée par un sommet (cela permet de tenir compte de l'orietation des gènes). Une arête de poids élevé est ajoutée entre les deux extrémités d'un même gène. Les arêtes correspondants aux adjacences potentielles sont ajoutées avec leur poids respectif. Aussi, k sommets fictifs $(O_1, O_2, ..., O_k)$ représentants les extrémités de chromosomes sont ajoutés et reliés à chacun des autres sommets, k étant au minimum deux fois plus grand que le nombre de chromosomes ancestraux attendu. Un cycle hamiltonien de poids maximal construit à partir de ce graphe défini alors une col-

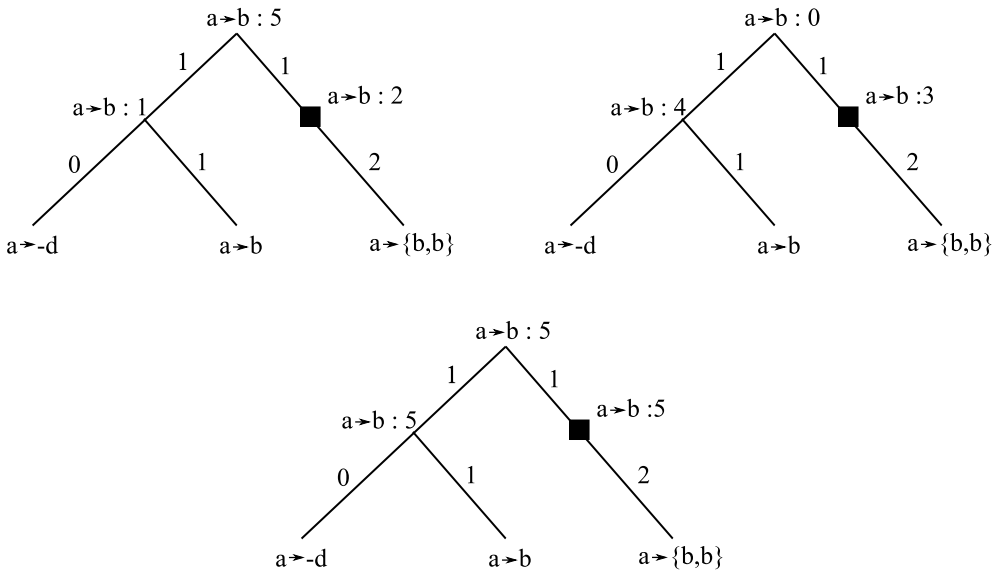


Figure 3.9 – Calcul du poids selon la méthode Bertrand *et al.* de l'adjacence $(a \rightarrow b)$ à tous les noeuds internes u de l'exemple de la figure 3.8. Haut Gauche : $AdjCons_{below}(a,b,u)$ Haut Droite : $AdjCons_{above}(a,b,u)$ Bas : $AdjCons(a,b,u)$

lection de chaînes de caractères (délimitées par les sommets O) représentant la solution désirée. Comme le TSP est un problème bien étudié, des heuristiques plus performantes qu'une méthode gloutonne existent. L'heuristique Lin-Khernigan [1] est celle utilisée par cette méthode. Enfin, il est toujours possible qu'une mauvaise adjacence soit retenue par l'heuristique. Pour cette raison, toutes les adjacences ayant un poids inférieur à un seuil τ (choisi par l'utilisateur) sont retirées de la solution, fragmentant l'ordre ancestral en un nombre plus important de RAC mais augmentant le niveau de confiance en cette solution.

Tout comme pour la méthode des uns consécutifs, la méthode est limitée à l'inférence d'un génome pré-dupliqué de S^{DGE}, mais avantagée par l'absence d'arbre de gènes ou de réconciliation. De plus, nous avons comparé cette méthode à InferCar décrite à la section 3.3.4. Des simulations montrent un gain sur la fiabilité de la solution au prix d'une fragmentation en un nombre plus important de RAC. Nous avons également appliqué la méthode à des espèces de levures et comparé à un ancêtre de ces espèces assemblé à la

main [31]. Nous obtenons 98% d'adjacences similaires, avec environ 120 RAC pour un seuil de 70% de conservation sur *S*.

Pour clore ce chapitre et faire le lien avec le chapitre suivant, introduisons la motivation à la base de la généralisation de l'approche de Bertrand *et al.* aux α-adjacences dont le but est de permettre une meilleure concaténation des RAC en un génome ancestral.

3.3.8 Généralisation aux α-adjacences

Supposant que les petites inversions forment la majorité des évènements de réarrangements expliquant l'évolution des génomes [39], les gènes impliqués dans les adjacences ancestrales aux extrémités de la séquence inversée avant l'inversion seront séparés dans les génomes aux feuilles par une petite quantité de gènes (la taille de la séquence inversée). Ces adjacences ne pourront facilement être retrouvées en considérant les adjacences immédiates, surtout si ce sont des inversions anciennes qui surviennent très tôt dans l'évolution. Dans ce cas particulier, les α-adjacences pourraient permettre de détecter, dans les génomes aux feuilles, les adjacences ancestrales exactes. Dans l'espoir de diminuer le nombre de RAC tout en conservant un bon niveau de confiance en la solution obtenue, nous avons généralisé notre méthode aux α-adjacences. Cette nouvelle approche est décrite dans le prochain chapitre qui est au format d'un article scientifique soumis à l'édition 2012 de la conférence internationale RECOMB - Comparative Genomics.

CHAPTER 4

ANCESTRAL GENOME RECONSTRUCTION BASED ON GAPPED ADJACENCIES

Ma contribution personnelle à ces travaux est d'avoir participé à l'élaboration de l'algorithme, à l'implémentation de l'algorithme, au design et à l'exécution des expériences de simulations de même que celles avec des données biologiques.

Yves Gagnon[1], **Mathieu Blanchette**[2] **and Nadia El-Mabrouk**[1]

Article soumis pour publication à la conférence internationale RECOMB - Comparative Genomics 2012

[1]Département d'Informatique (DIRO), Université de Montréal, H3C 3J7, Canada.
[2]McGill Centre for Bioinformatics, McGill University, H3C 2B4, Canada

Abstract

Motivation: The "small phylogeny" problem consists in inferring ancestral genomes associated with each internal node of a phylogenetic tree of a set of extant species. The existing methods can be grouped into two main categories: the distance based methods aiming at minimizing a total branch length, and the synteny-based (or mapping) methods that first predict a collection of relations between ancestral markers in term of "synteny", and then assemble this collection into a set of Contiguous Ancestral Regions (CARs). The predicted CARs are likely to be more reliable as they are more directly deduced from observed conservations in extant species. However the challenge is to end up with a completely assembled genome.

Results: We develop a new synteny-based method that is flexible enough to handle a model of evolution involving whole genome duplication events, in addition to rearrangements and gene insertions and losses. Ancestral relationships between markers are defined in term of *Gapped Adjacencies*, i.e. pairs of markers separated by up to a given number of markers. It improves on a previous, more conservative method, restricted to direct adjacencies, that revealed a high accuracy for adjacency prediction, but with the drawback being of generating a high number of CARs. Applying our algorithm on various simulated data sets reveal good performance as we usually end up with a completely assembled genome, while keeping a low error rate.

4.1 Introduction

One of the aims of comparative genomics is to reveal the evolutionary scenario that has led to an observed set of present-day genomes from hypothetical common ancestors. When a speciation history, represented as a phylogenetic tree, is already known, then the problem reduces to that of finding ancestral genomes, in terms of content and organization, for non-terminal nodes of the tree. The reconstruction of ancestral karyotypes and gene (or any markers) content and order has been largely considered by the computational biology community [4, 8, 11, 17, 44, 46, 58, 69]. For most formulations in terms of different kinds of genomes (circular, multichromosomal, single or multiple

gene copies, signed or unsigned genes) and different distance metrics, even the simplest restriction in term of the median of three genomes, has been shown NP-hard [54]. As reviewed in [16, 23], the considered methods can be grouped into two main classes. The distance-based methods aim at labeling ancestral nodes in a way minimizing total branch length over the phylogeny [11, 16, 46, 58, 69]. On the other hand, the synteny-based (or mapping) methods [8, 17, 44, 47] rely on three steps: (1) Infer a collection of ancestral genes; (2) Infer a collection of relations between ancestral genes in terms of "synteny"; (3) Assemble this collection into an ancestral genome. In contrast to a distance-based approach, the output of a synteny-based approach is a set of Contiguous Ancestral regions (CARs) that is not guaranteed to be completely assembled into a genome. However, the predicted CARs are likely to be more reliable as they are more directly deduced from observed conservations in extant species. The first formal method based on this approach was developed by Ma *et al.* [44]. In this algorithm, syntenies are adjacencies, sets of ancestral relations are computed by the Fitch parsimony algorithm and a greedy heuristic is used for the assembly. Another class of synteny-based methods [17, 27] define ancestral relations in term of common intervals, represent them in a 0-1 matrix, and then use an approach known as the *Consecutive Ones problem (C1P)* [26] to translate the matrix into sets of ancestral CARs. The translation is direct in case of a collection of ancestral relations being all compatible, but in general the problem of transforming the matrix into a C1P matrix in an "optimal" way is hard, and appropriate simplifications are considered. The result of such methods is not a unique ancestral gene order but rather a PQ-tree representing a collection of possible orders.

Most computational methods for comparative genomics account only for markers with exactly one copy in every considered extant genome. A few extensions to genomes with unequal gene content have also been considered [8, 12, 27]. The case of multiple gene copies is more challenging as the one-to-one correspondence between orthologs is missing. Recently, a number of ancestral genome inference studies have accounted for multiple gene copies in the very special case of an evolution by Whole Genome Duplication (WGD). WGD is a rare but spectacular evolutionary event that has the effect of simultaneously doubling all the chromosomes of a genome. Evidence of WGD has

shown up across the whole eukaryote spectrum. A distance-based approach for inferring a pre-duplicated genome has been developed in 2003 [24], and extended to the median problem [28, 70, 71]. However, the synteny-based approach is more naturally extendable to WGD events. Indeed, as the pre-duplicated genome has single gene copies, as long as an appropriate way for inferring "Double Conserved Synteny" (DCS) relations between ancestral markers is found, the assembly part can be taken without any modification. In [31], Gordon *et al.* used a "manual" approach to reconstruct the ancestral yeast genome. Formal extensions of the synteny-based approach to handle WGD have also been developed [8, 16, 51].

In this paper, we present a new synteny-based method for ancestral genome inference, allowing for evolutionary scenarios involving WGDs and gene losses, where relations between ancestral genes are defined as *Gapped Adjacencies*, i.e. pairs of genes separated by up to a fixed number of genes. It is an extension of a previous method [8] where relations between genes were defined in term of "direct" adjacencies. The assembling step is based on the computation of a rigorous score for each potential ancestral gapped adjacency (g,h), reflecting the maximum number of times g and h can be adjacent in the whole phylogeny, for any setting of ancestral genomes. To make the link with the "consecutive one" framework [17, 27], the syntenies that we consider in this paper can be related to gapped gene teams, while those considered in [17] are related to various types of common intervals [5]. However the assembling methods and the output of the algorithms (a set of CARs versus a PQ-tree) are very different. In the absence of WGD events and gene losses, the approach most comparable to ours is the one developed by *Ma et al.* [44]. In case of direct adjacencies, the algorithm in [8] revealed a higher accuracy for adjacency prediction than *Ma*'s algorithm, but with the counterpart being a higher number of CARs, preventing from recovering a completely assembled genome. In this paper, relaxing the constraint of adjacency to gapped adjacency allows to improve on these results. Indeed, applying our algorithm on simulated data sets reveals that we usually end up with a completely assembled genome, while keeping a low error rate.

4.2 Methods

4.2.1 Problem statement and preliminary concepts

PROBLEM STATEMENT:

Input: A set Γ of n modern genomes, a species tree S for Γ, and an internal node v of S representing a speciation event of interest;

Output: An ancestral genome at node v.

Formally, a species tree (or phylogeny) for Γ is a tree S with n leaves, where each genome of Γ is the label of exactly one leaf, and each internal node (called *speciation node*) has exactly two children and represents a speciation event. We say that S is *labeled* if each internal node u of S has a label $G(u)$ corresponding to a hypothetical ancestral genome just preceding the considered speciation event.

Considering a set Σ of genes, a genome is a set $\{C_1, C_2, \cdots C_N\}$ of chromosomes, where each chromosome is a sequence of signed elements from Σ. In case of linear chromosomes, their ends are represented by an artificial gene O at one extremity of each chromosome, and then considering the augmented chromosomes as circular. Given a genome G, we call the *gene set* of G and denote by $\Sigma_G \subseteq \Sigma$ the set of genes present in G (including O). For example, the gene set of the genome labeling the leftmost leaf of the tree in Figure 4.1 is $\{O, a, b, c\}$. We further denote by $\pm\Sigma_G$ the set obtained from Σ_G by considering each gene in its positive and negative directions. By convention, the gene O is always considered positive. A *multiset* of $\pm\Sigma_G$ is a subset of $\pm\Sigma_G$ with possibly repeated genes. Given a gene $g \in \Sigma_G$, we denote by $mult(g, G)$ the *multiplicity*, i.e. number of copies, of g in G. In particular, the multiplicity of O is the number of chromosomes of G. For example, the multiplicity of gene a in the genome labeling the leftmost leaf of the tree in Figure 4.1 is 4. We extend our notation to define, for node u of the tree, Σ_u and $mult(g, u)$ as the set of genes present in the genome at node u and the multiplicity of g in that genome.

4.2.1.1 Evolutionary model

Our model involve rearrangements and content-modifying operations. As we adopt a synteny-based approach, rearrangements are only implicitly considered, as only traces of these rearrangements in term of disrupted gene adjacencies are considered. In other words, all kinds of rearrangement events can be present in the history. Our approach also allows for unequal gene content, resulting from gene losses or insertions.

As for the multiplicity of genes, the only operation leading to multiple gene copies (genes with multiplicity ≥ 2) considered is the *Whole Genome Duplication* (WGD). Tandemly duplicated genes can ben managed by concatenating them in one artificial gene. However, dispersed gene copies add ambiguity when considering potential ancestral adjacencies, as they are dispersed in the genome by their duplication mechanism. Thus, they are less likely to retain the same neighbours than genes arising from WGD, in which case they retain the same neighbours unless subjected to genome rearrangements. Formally, a WGD is an event transforming a genome $G = \{C_1, C_2 \cdots C_N\}$ of N chromosomes into a genome G^D containing $2N$ chromosomes, i.e. $G^D = \{C_1, C'_1, C_2, C'_2 \cdots C_N, C'_N\}$, where, for each $1 \leq i \leq N$, $C_i = C'_i$.

In addition to the assumption that WGDs are the only events responsible for gene multiplicity (in particular, single-gene duplications are not considered), we suppose that, in each genome, at least one gene reflects the doubling status of the genome, i.e. there exists a gene that has not lost any copy. As noticed by Zheng *et al.* [71], under these assumptions, a history with a minimum number of WGD events can be easily deduced from the multiplicity of the most frequent gene found in each genome. To account for such events, new internal nodes, called *WGD nodes*, are added appropriately on the edges of S (see Figure 4.1). Contrary to speciation nodes, each WGD node has only a single child. Moreover, if all extant genomes have a gene with multiplicity greater than 1, then a WGD node is inserted above the root of S.

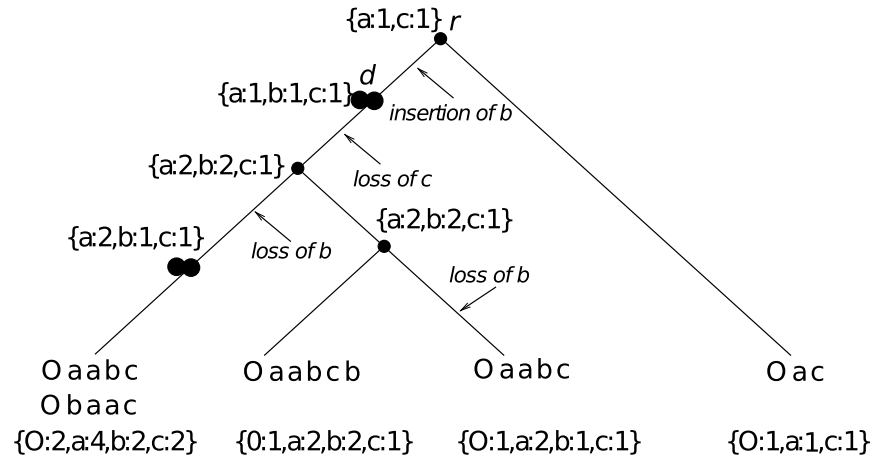


Figure 4.1: A species tree with each leaf labeled with its corresponding genome. For simplicity, we consider all the genes to be positively signed. The last line below each leaf is the gene set and multiplicity of each gene. Single circles indicate speciation nodes, while double-circles indicate WGD nodes. Applying the procedure described in the text leads to the gene set assignment and multiplicity given as labels of internal nodes. This assignment leads to the indicated insertion and losses.

4.2.1.2 Adjacencies

Given a genome G, let $g \in \Sigma_G$ and $h \in \pm\Sigma_G$. We say that h is a *1-adjacency*, a *direct adjacency* or simply an *adjacency* of g in G iff it is a left or right adjacency of g. h is a left-adjacency of g in G iff "$h\ +g$" or "$-g\ -h$" is a substring of G. Symmetrically h is a right-adjacency of g in G iff "$+g\ h$" or "$-h\ -g$" is a substring of G.

We now extend 1-adjacencies to *gapped adjacencies*, i.e. to α*-adjacencies* for an arbitrary value of α, by allowing for interleaving genes. Let $G = g_1 g_2 ... g_n$. As already defined, the set of 1-adjacencies of g_i is $\{g_{i-1}, g_{i+1}\}$. We can as well define the set of 2-adjacencies of g_i as $\{-g_{i-1}, g_{i-1}, -g_{i+1}, g_{i+1}\}$, etc. In general, for $\alpha \geq 1$, g_i is α-adjacent to $\{g_{i\pm k} \mid 1 \leq k \leq \lfloor(\alpha+1)/2\rfloor\} \cup \{-g_{i\pm k} \mid 1 \leq k \leq \lfloor\alpha/2\rfloor\}$.

We denote by $LA(g,\alpha,G)$ and $RA(g,\alpha,G)$, or just $LA(g,G)$ and $RA(g,G)$ if $\alpha = 1$, the *multisets* of left and right α-adjacencies of the one or more copies of g in G. For example, for the genome G labeling the leftmost leaf in the tree of Figure 4.1, we have $LA(a,1,G) = \{O,a,b,a\}$ and $RA(a,1,G) = \{a,b,a,c\}$.

4.2.1.3 Conserved Adjacencies

For genomes with single gene copies, it is easy to define the number of α-adjacencies preserved along a branch (u,v) of a labeled tree S as the number of substrings of size $\alpha+1$ between $G(u)$ and $G(v)$ bounded by the same genes. This definition is not directly transposable for genomes with multiple gene copies, as the one to one orthology between genes is not set. Instead, for each gene g, we compare its left and right α-adjacency multisets in $G(u)$ and $G(v)$. More precisely, we define $adjCons(g,\alpha,G(u),G(v)) = |LA(g,\alpha,G(u)) \cap LA(g,\alpha,G(v))| + |RA(g,\alpha,G(u)) \cap RA(g,\alpha,G(v))|$, as the number of left and right conserved α-adjacencies of g on (u,v), and

$$adjCons(\alpha,G(u),G(v)) = \sum_{g \in \Sigma_u \cap \Sigma_v} adjCons(g,\alpha,G(u),G(v))$$

as the number of conserved α-adjacencies on the branch (u,v). Finally, the number of conserved α-adjacencies over the whole tree S, denoted as $adjCons(\alpha,S)$ (or just $adjCons(S)$ for $\alpha=1$), is the sum of $adjCons(\alpha,G(u),G(v))$ for all branches (u,v) of S.

Remark: In $adjCons(\alpha,G(u),G(v))$ we account for each adjacency conservation twice. It may appear that right adjacencies alone (or, symmetrically, left adjacencies) are sufficient to reflect adjacency conservation between two genomes. But consider, for example, the sequence "$+1\ -2\ +3\ -4$". If we just consider right 1-adjacencies, then the subsequence "$+1\ -2$" will be considered twice (as -2 is the right adjacency of 1 and -1 is the right adjacency of 2) but the subsequence "$-2\ +3$" will not be considered (as -3 is the left adjacency of 2 and -2 is the left adjacency of 3).

4.2.2 Ancestral gene content

The first step of any ancestral inference method is to assign ancestral gene content and multiplicity at each ancestral node. We consider a natural procedure, inspired from [31], assuming a model with no convergent evolution and minimum losses. We say that a node v is a *direct descendant* of a WGD node u if and only if v is a WGD node

or a leaf and there is no other WGD node on the branch from u to v. To assign gene content Σ_u and gene multiplicity at each internal node u of S, we apply the two following operations in two bottom-up traversals of S: **(1)** For each WGD node u and each gene g, let v be the direct descendant of u with maximum multiplicity for g. If $mult(g,v) \geq 2$ then assign g to u and define $mult(g,u) = \lfloor \frac{mult(g,v)}{2} \rfloor$. For example after a traversal of the species tree S of Figure 4.1, the gene set of the WGD node d only contains a and b, as the maximum multiplicity of c in the direct descendants of d is 1; **(2)** Assign a gene g to any internal node u of S on a path from the node S representing the least common ancestor (LCA) of all the nodes containing a (leaves or WGD nodes), to any leaf containing g. Moreover, if not already defined, define $mult(g,u)$ as the maximum multiplicity of g in u's children.

In the rest of this paper, we will assume that gene content and multiplicity is set for all nodes of S. A *correct labeling*, or simply a *labeling* $G(u)$ of a node u of S will refer to a genome respecting the content and multiplicity constraints given by Σ_u. Notice that, by construction (taking the maximum multiplicity of each gene at each internal node), there is no increase of multiplicity (except in case of an insertion) from a node u to a child v, unless u is a WGD node, in which case the multiplicity of a gene is at most doubled. Such a construction guarantees that any labeling of S can be explained by an evolutionary scenario in agreement with the hypothesis of WGDs being the only events responsible for gene multiplicity.

4.2.3 A synteny-based method accounting for direct adjacencies

In [8], we have presented a synteny-based method that infers a pre-duplicated ancestral genome at a node v corresponding to a highest WGD node of S, or any node preceding a first WGD node. More precisely, the method infers a genome $G(v)$ such that $adjCons(S|G(v))$ is maximized, where $adjCons(S|G(v))$ is the maximum number of conserved adjacencies over the whole tree S, for any ancestral genome assignment, with the constraint that genome $G(v)$ is assigned at node v (see details in [8]).

For any node u of S, define $LeftAdj(g, S|_{LA(g,G(u))=X})$ (resp. $RightAdj(g, S|_{RA(g,G(u))=X})$) as the maximum number of left (resp. right) adjacencies that can be preserved over the

whole tree, for any ancestral genome assignment with the constraint that the genome $G(u)$ satisfies $LA(g,G(u)) = X$, where X is a multiset of $mult(g,u)$ potential adjacencies selected from $\pm\Sigma_u \setminus \{g\}$. The following upper bound on the objective function allows to treat each gene independently.

$$adjCons(S|G(u)) \leq \sum_g LeftAdj(g,S|_{LA(g,G(u))=X}) + RightAdj(g,S|_{RA(g,G(u))=X})$$

The method, that we call DirectAdj, proceeds in two steps summarized below.

STEP 1:

For each internal node u of the tree (speciation or WGD node), each gene $g \in \Sigma_u$, and each multiset X of possible adjacencies of g at node u, we compute $LeftAdj(g,S|_{LA(g,G(u))=X})$ and $RightAdj(g,S|_{LA(g,G(u))=X})$ using a Dynamic Programming Algorithm. The values at a node u are computed from the values at the two children and also at the parent of u. An illustration is given in Figure 4.2.

STEP 2:

For the node v for which an ancestral genome is sought, we obtain the desired pre-duplicated genome by chaining adjacencies. As v is a node in a tree with no WGD, or a first WGD node in a history, the multiplicity of genes can be ignored at v, as in the first case each gene g of Σ_v is present exactly once at v, and in the second case all copies of g have the same adjacency. At this node we use the notations $L(g,h) = LeftAdj(g,S|_{LA(g,G(v))=\{h\}})$ and $R(g,h) = RightAdj(g,S|_{RA(g,G(v))=\{h\}})$. We proceed by a reduction to the Traveling Salesman Problem (TSP) on a complete undirected graph Q where vertices correspond to genes, and an edge (g,h) is weighted according to a ratio $(L(g,h)+R(h,g))/MaxAdj(g,S)$, where $MaxAdj(g,S)$ is the number of nodes of S containing g. The division by $MaxAdj(g,S)$ allows to correct for genes that are lost in some parts of the tree, which avoids favoring genes with high multiplicity. Moreover, as noticed in [8], the result of the TSP is usually one long chromosome concatenating long

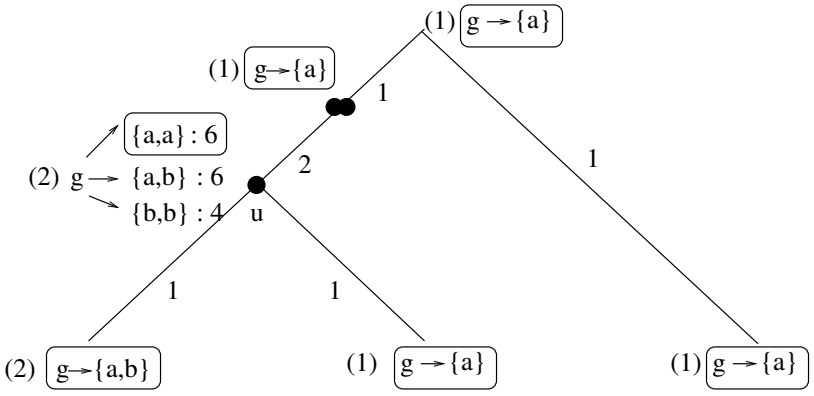


Figure 4.2: An illustration of STEP 1 for a gene g and an internal node u. Numbers in brackets indicate the multiplicity of gene g at each node of the tree. Multisets at leaves represent (say left) adjacencies of gene g in the corresponding genome. All multisets X of possible adjacencies of g at node u are shown, followed by the value of $LeftAdj(g, S|_{LA(g,G(u))=X})$. The rest of notation illustrates how the value 6 is obtained at u for the multiset $\{a,a\}$: the root and WGD node labels are the adjacencies that have to be set for g, and the label of an edge (v,w) is the number of conserved adjacencies for g on that branch.

CARs. To avoid this drawback, we define TSP-τ by augmenting the initial TSP heuristic with the procedure of cutting, from the inferred ancestor, all adjacencies with weight less than a certain threshold τ (see Section 4.2.4.2). All details on costs, the heuristic used to solve the TSP and how to handle chromosomal endpoints and gene signs, are given in [8]. In the following section, we generalize the approach described above to allow for a more flexible notion of synteny in term of gapped-adjacencies.

4.2.4 Generalization to gapped adjacencies

Before describing our new algorithm called GapAdj, which is a generalization of DirectAdj accounting for α-adjacencies for increasing values of α, we motivate our new approach in the following section.

Many adjacencies in an ancestral genome are likely to be no longer present in some present-day genomes due to rearrangements and content-modifying operations, preventing from reconstructing large CARs. However, assuming that small and local evolutionary events are more frequent than large and far-reaching operations, which has been

largely supported in the literature [39], we can expect to reconnect neighboring CARs by considering gapped adjacencies of increasing gap-size.

Consider for example the species tree (A) of Figure 4.3. As a and b are neighboring genes in all three genomes, we expect the inferred ancestral genome at the root of the tree to have a CAR with neighboring genes a and b. However, as all (right) direct adjacencies of a are different (b in 1, $-b$ in 2 and x in 3), none of these adjacencies would have a score attaining a reasonable minimum cost τ for the TSP, and a and b will end up in two different CARs with algorithm DirectAdj. However, as b (and also $-b$) is a $2-adjacency$ of a in two extent genomes, and a $3-adjacency$ of a in all three genomes, they will be in the same CAR after the second or third iteration of GapAdj algorithm.

As another example, consider a "true" evolutionary scenario depicted in Figure 4.3.(B). Consider a threshold τ for TSP-τ corresponding to an adjacency being present in two of the three extent species. Then, as the only direct adjacency present at least twice in extent genomes is bc, the result of DirectAdj is a set of CARs with a and bc being in two separate CARs. However, as b is a 3-adjacency of a in species 1 and 2 (it is actually the only adjacency reaching the threshold up to $\alpha = 3$), GapAdj would end up with a CAR containing the sequence abc after iteration $\alpha = 3$.

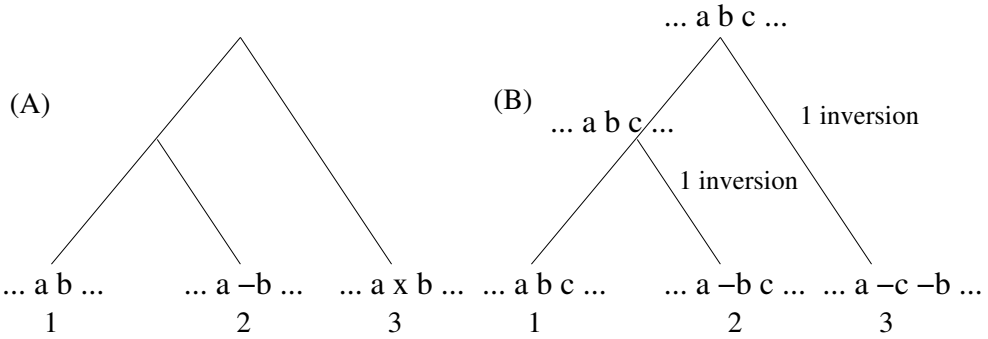


Figure 4.3: A species tree for the set of species $\Gamma = \{1,2,3\}$, with two different genome assignments at leaves. Example (B) depicts a most parsimonious inversion scenario leading to the observed genomes.

4.2.4.1 Algorithm

The full GapAdj algorithm is given in Supplementary Algorithm 1 (appendix). The output of GapAdj is the set of CARs C representing the ancestral genome at node v of S. This set is first initialized to the set Σ_v of genes at v (each gene being assigned to its own CAR). The algorithm proceeds by iterating the two-step procedure described in Section 4.2.3 on increasing values of α, from 1 to a constant MAX_α. Step 1 consists in computing α-adjacency scores. The dynamic programming algorithms detailed in [8] for computing the scores $LeftAdj(g, S|_{LA(g,G(u))=X})$ and $RightAdj(g, S|_{RA(g,G(u))=X})$ of left and right adjacencies of a gene g with a multiset X at a node u of S are directly generalizable to account for α-adjacencies, i.e. to compute the scores $LeftAdj(g, S|_{LA(g,\alpha,G(u))=X})$ and $RightAdj(g, S|_{RA(g,\alpha,G(u))=X})$.

As for Step 2, we proceed by constructing a complete undirected graph Q where vertices are the two extremities of each CAR, and edges are weighted according to α-adjacencies scores, computed at Step 1, of the two genes at the extremities of each CAR. A heaviest Hamiltonian cycle through Q, where edges under a threshold τ are excluded, corresponds to an hypothetical ancestral genome characterized by a set of CARs C_α with $|C_\alpha| \leq |C_{\alpha-1}|$. This instance of the TSP is solved using the Chained Lin-Khernigan heuristic implemented in the Concord package [41].

4.2.4.2 Choice of parameters

An important parameter of our algorithm is the cut-off value τ used to filter out less reliable adjacencies from the solution produced by the TSP algorithm. Based on the simulations that we have performed in [8], we choose a fixed threshold allowing for the best balance between error rate and number of CARs produced. The chosen threshold τ corresponds roughly to keeping an adjacency if and only if it is conserved in at least 70% of the tree branches. Another important parameter of our algorithm is the constant MAX_α, corresponding to the maximum value of α to be considered, which affects both the running time, the final number of CARs and their accuracy. Clearly MAX_α does not need to be more than the size of the longest chromosome of Γ, as no improvement can

be achieved for larger values. Unless explicitly indicated, we use $MAX_\alpha = 50$.

4.3 Results

To evaluate the accuracy and running time of our approach, we first used data obtained from simulated genome evolution. This allows us to dissect the impact of each aspect of the method and of the data on the accuracy of the reconstructed ancestor. Our simulations are based on the phylogenetic tree of yeast species shown in Figure 4.4 (A), which is ideal for this type of study as it contains a phylum affected by a whole-genome duplication and another that remains non-duplicated. Each of the simulation-based results reported in this section are averaged over 50 repetitions.

4.3.1 Simulations with no WGD

In the absence of WGD events, the method that is most comparable to ours is the one of Ma *et al.* [44], implemented in a program called InferCAR. As this method does not support gene losses, we first restrict our simulations to a model with rearrangements only. In addition, as a first validation, we consider single chromosomal genomes, and inversions as the only rearrangement events.

We simulated data sets based on the yeast phylogenetic tree but excluding the portion affected by the WGD. The tree contains six non-duplicated species. The node of interest is the root σ of the monophyletic group of five species (indicated by a simple circle in Figure 4.4 (A)). A simulated genome of two hundred genes is placed at the root ρ of the tree, and a number r of inversions are randomly performed on each branch of the tree, where r is chosen randomly in the interval $[\frac{rmax}{2}, rmax]$, for a given constant value $rmax$. Notice that the maximum value $rmax = 25$ considered in our simulations leads to some of the leaf genomes being almost completely shuffled, as four or five branches separate them from the root, which lead to the creation of about 160 to 200 breakpoints. The length of inverted segments follows a geometric distribution with $p = 0.5$, resulting in the majority of inversion involving a small number of genes, as previously suggested [39].

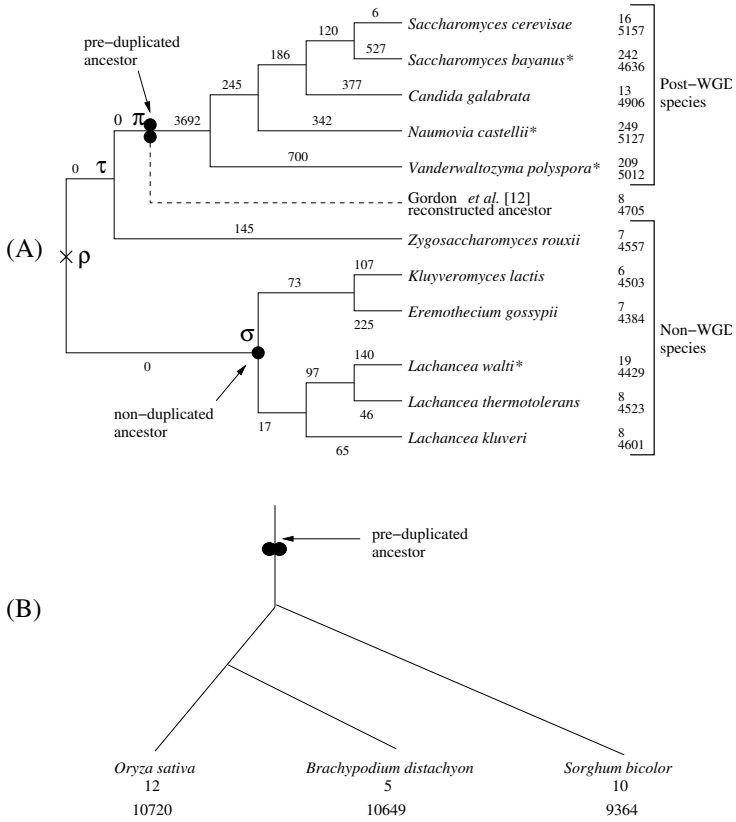


Figure 4.4: (A) Evolution of the 11 yeast species recorded in the Yeast Gene Order Browser, as given by [36]. The * indicates partially sequenced organisms. At leaves, the top number is the number of chromosomes, contigs or scaffolds. The bottom number is the number of genes, as reported in [31]. On each branch, the label is the number of gene losses, which is directly inferred from the gene content at leaves. The simple circle is the root of the monophyletic group of non-duplicated species, referred in the text by σ. (B) The phylogenetic tree for *Oryza sativa* (rice), *Brachypodium distachyon* (brachypodium) and *Sorghum bicolor* (sorghum). At leaves, the top number is the number of chromosomes. The bottom number is the number of markers used in the study of Section 4.3.4.

Figure 4.5 (left) illustrates the two algorithms' error rates, computed as the fraction of inferred α-adjacencies (for $1 \leq \alpha \leq MAX_\alpha$) that are not present as α-adjacencies in the true simulated ancestor at σ, while the right diagram illustrates the number of CARs obtained (on average) for that ancestor. Both algorithms show a high accuracy for adjacency prediction, as the error rate is always lower that 10%. Our GapAdj algorithm

almost always recovers a complete genome (i.e. a single CAR), which is very rarely the case of InferCAR, which yields an average of 6 CARs for $rmax = 25$. However, this increase in CAR concatenation is obtained at the cost of a small loss of precision.

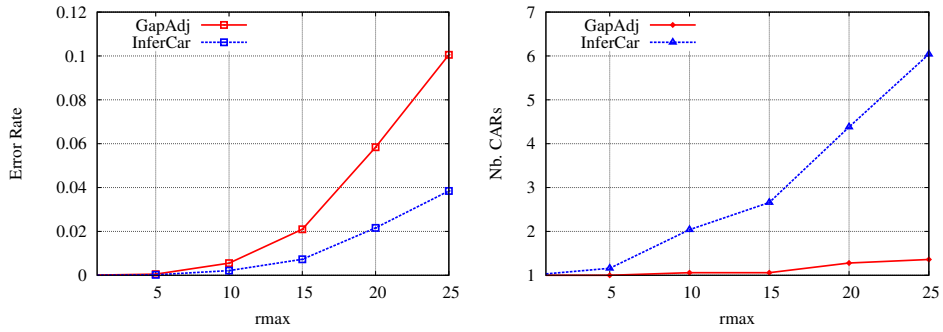


Figure 4.5: Simulations for a tree without WGD, and a maximum of *rmax* inversions (x-axis) on each branch. Red curves are the results of GapAdj and the blue ones those of InferCAR. (Left) Error rate for the inferred ancestral genome; (Right) Number of inferred CARs.

Figure 4.6 illustrates the progression of the error rate and CAR number for increasing values of α. It provides a comparison with the initial algorithm DirectAdj [8] that only considers direct adjacencies ($\alpha = 1$). From $\alpha = 1$ to $\alpha = 50$, the number of CARs drops from 20 to a single chromosome, while the error rate is increased by less than 4%. These preliminary results are promising as the initial goal of obtaining a completely assembled genome while keeping a low error rate is attained in this case.

We then consider an extended model of evolution for multichromosomal genomes that evolve through inversions, inter-chromosomal rearrangements (translocations, fusions, fissions) and gene losses. Based on the same six-leaf species tree described above, we simulate data sets starting with a 2-chromosome, 200-gene genome at the root ρ of the tree. The number of gene losses on each branch is proportional to that observed in actual yeast genomes, while the proportion of each type of rearrangement operation is chosen to be similar to that reported for *S. cerevisiae* in [31]: (Inv : Trans : Fus+Fiss) = (5 : 4 : 1). The results given in Figure 4.7 reflect the difference in gapped-adjacencies and number of chromosomes between the real and predicted genome at node σ. Notice that chromosomal fusions and fissions may occur on the branch from ρ to σ, so the true

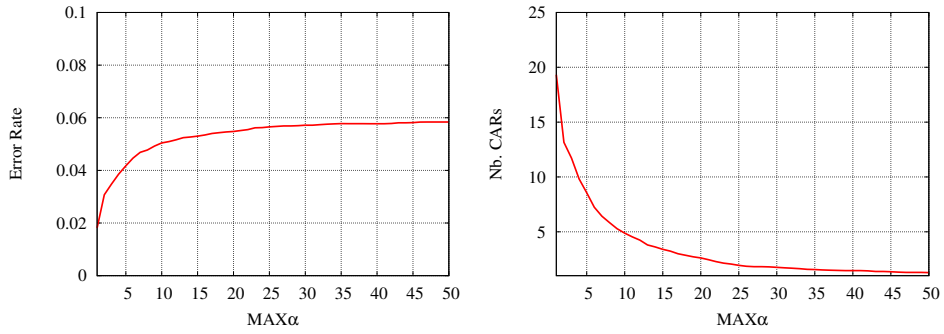


Figure 4.6: (Left) Error rate and (Right) Number of CARs obtained by our algorithm *GapAdj*. Values on the *x*-axis correspond to the parameter MAX_α, i.e. the maximum value considered for α. Simulations are performed with $rmax = 20$.

number of chromosomes depicted in the right diagram of Figure 4.7 is not always 2. Interestingly, the curve for inferred CARs roughly follows the curve for true CARs. In addition, the error rate remains lower than 10% in all cases.

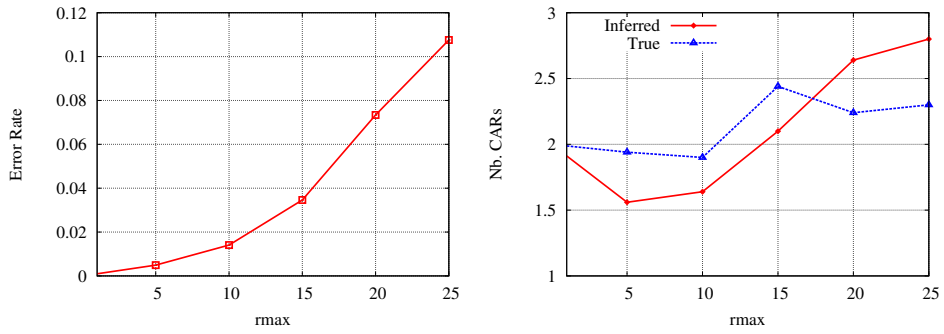


Figure 4.7: (Left) Error rate and (Right) Number of CARs obtained by *GapAdj* on simulations following a model accounting for multichromosomal genomes evolving through gene losses, and a maximum of *rmax* (x-axis) inversions and inter-chromosomal rearrangements per branch of the tree.

4.3.2 Simulations with WGD

For simulations with WGD, we used three trees: two being the subtrees of yeast (Figure 4.4 (A)) rooted at τ and π, and another (Figure 4.4 (B)) corresponding to the

evolution of three cereals (rice, brachypodium and sorghum), that we will study in Section 4.3.4. We simulate data sets starting with a pre-duplication 2-chromosome, 200-gene genome at the root of the tree and performing a number of gene losses and a maximum *rmax* of rearrangements on each branch. As WGD events are usually followed by extensive losses, we perform 50 or 100 random losses between the duplication and first speciation event, followed by 5 random losses on each branch of the tree. As for the rate of various rearrangements, we use the same as before. Error rates are given in Figure 4.8 (left). The number of CARs produced by the algorithm typically slightly overshoots the correct number, varying from 2 to 4. Note that the losses that occurred immediately after the duplication event result in many false adjacencies inferred, as depicted by the difference in error rate between simulations with only 50 post-duplication losses and those with 100. Since those are ancient events, their effects are seen on many or all of the leaf gene orders, preventing us from inferring the right order in areas surrounding the lost genes in the ancestor. Interestingly, the fact that an outgroup (a genome that is not descendant of the WGD) is available for yeast allows to circumvent this problem as adjacencies can be grasped from this genome not affected by losses, which explains the better results obtained for the yeast subtree rooted at τ.

Figure 4.8 (right) shows the running time of our algorithm for $rmax = 20$, as a function of MAX_α. Although the running time increases cubically with MAX_α, it remains quite manageable. In the absence of the WGD, the running time is significantly smaller, as it remains under 2 seconds even for $MAX_\alpha = 50$ (results not shown).

4.3.3 Study of yeast genome evolution

We applied our method to the full yeast species tree (Figure 4.4 (A)) with the gene data sets of the *Yeast Gene Order Browser* [31], to infer the pre-duplicated ancestral genome of *Sccharomyces cerevisiae*. We then compared our predicted ancestor with the 8-chromosome genome manually inferred by Gordon *et al.* [31]. Figure 4.9 (left) gives the fraction of α-adjacencies that we infer but are in contradiction with the genome inferred by Gordon *et al.* For all tested values of α, this fraction remains below 2%. Importantly, considering gapped adjacencies in addition to direct adjacencies allows to

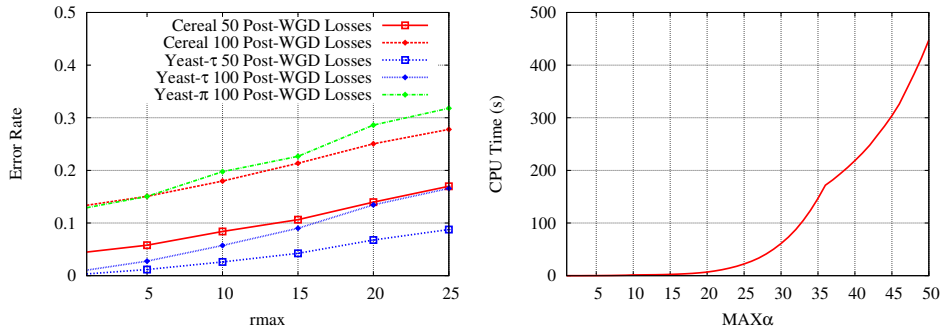


Figure 4.8: (Left) Error rate obtained by *GapAdj* on simulations performed according to the cereal tree (Figure 4.4(B)) and the subtrees of yeast rooted at τ and π (Figure 4.4(B)). The model accounts for inversions, inter-chromosomal rearrangements, gene losses and one WGD. The two red (resp. blue) curves correspond to the results for cereal (resp. yeast) by performing 50 and 100 losses just following the WGD. (Right) Running time of *GapAdj* for one data set following the "cereal 50" model, and with *rmax*=20.

reduce the number of CARs from 23 to 12, which is significantly closer to the number of ancestral chromosomes predicted by Gordon *et al.* Among the 11 additional inferred 1-adjacencies, 7 are shared with the ancestor of Gordon *et al*.

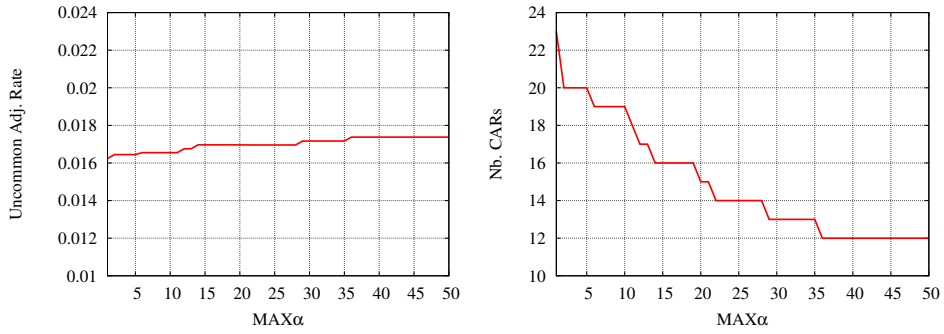


Figure 4.9: (Left) Fraction of adjacencies in disagreement between the pre-duplicated yeast ancestor inferred by *GapAdj* and that inferred by Gordon et al. in [31]. (Right) Number of CARs inferred with *GapAdj* algorithm.

4.3.4 Study of cereal genome evolution

We now focus on three of the four completely sequenced cereal crop genomes studied by Murat *et al.* [48], namely rice (*Oryza sativa*), sorghum (*Sorghum bicolor*) and brachypodium (*Brachypodium distachyon*). As demonstrated by various studies, these species have evolved following a whole genome duplication that has occurred about 60 million years ago (see Figure 4.4.(B)). Maize, the fourth species considered in [48] was excluded here to avoid noise due to an additional maize-specific WGD and ensuing massive gene loss. We used the sets of markers (10,720 from rice, 10,649 from brachypodium, and 9,364 from sorghum) and the homology relationships provided by Murat *et al.*, and the orders for these markers from the annotations in [37, 64, 65].

Supplementary Figure 1 shows the predicted pre-duplication genome and its extant descendants. Syntenic regions (homologous sets of genes with conserved order) are painted using the Cinteny web server [61]. Running GapAdj with a maximum value of α (up to the size of the largest chromosome which is about 3500), we end up with a set of 6 CARs (plain bars in Supplementary Figure 1), which is one more chromosome than that inferred by Murat *et al.* [48]. Looking carefully at the obtained results, we can see that the ancestral CARs 5 and 6 are clustered (and shuffled) into a single chromosome in Brachypodium (chromosome 2), and in two chromosomes in rice and sorghum (chromosomes 1 and 5 in the rice, and 3 and 9 in sorghum). Moreover there is no other segment of the CARs 5 and 6 in any other extant chromosome. This observation suggests that these two CARs should be concatenated into a single and complete chromosome. This would be consistent with the results reported by Murat *et al.* [48], who infer that a single pre-duplicated chromosome *C* is the ancestor of the same chromosome in Brachypodium (2) and the same two chromosomes in rice (1 and 5) and sorghum (3 and 9). The reason our algorithm did not concatenate them is probably that the genes at both extremities of the ancestral CAR 5 are in two different chromosomes in rice and sorghum. This suggests a future extension of our algorithm that would consider the α-extremities of each current CAR for subsequent concatenations.

Comparing our observations with Murat *et al.*, we notice a number of striking sim-

ilarities. In particular, one of the main discovery of the paper [48] is that some chromosomes have evolved following a particular evolutionary event, called nested fusion, resulting in the insertion of one chromosome inside another (non-telomeric fusion). Indeed, chromosome 2 of Brachypodium is explained in [48] as resulting from a nested chromosome fusion of the two copies of the chromosome C (introduced in the previous paragraph), that has occurred after the speciation leading to the Brachypodium lineage. Interestingly this nested fusion is clear in our results, as our chromosome painting is in agreement with chromosome 2 of Brachypodium being the result of an insertion of the ancestors of rice chromosome 5 in the middle of the ancestor of rice chromosome 1.

4.4 Conclusion

Any method for ancestral genome inference is debatable by nature, as it should be based on a model of evolution that is set *a priori*, even though we have no direct access to the history of genomes. Moreover, as real ancestors are not known, any validation method is open to criticism, and there is no direct way of evaluating one solution compared to another.

Based on the first observation, we opted for a synteny-based method that is based as much as possible on the observed data sets, without the need for explicitly defining the rearrangement events acting on these genomes. It is the first synteny-based method that fully capitalizes on the observed adjacencies in present day genomes in relation with their phylogenetic organization. It is flexible enough to apply to genomes that have evolved through whole genome duplication events, in addition to rearrangements and gene insertions and losses.

Based on the second observation, we first opted in [8] for a conservative approach concatenating two ancestral genes g and h only if the direct adjacency (g,h) is observed in a large fraction of extant genomes and sufficiently supported by the phylogeny. The result was an algorithm with high accuracy for adjacency prediction, but with the counterpart being a high number of CARs. Our generalization to gapped adjacencies while maintaining a conservative strategy for each gap size has led to a reasonable compromise

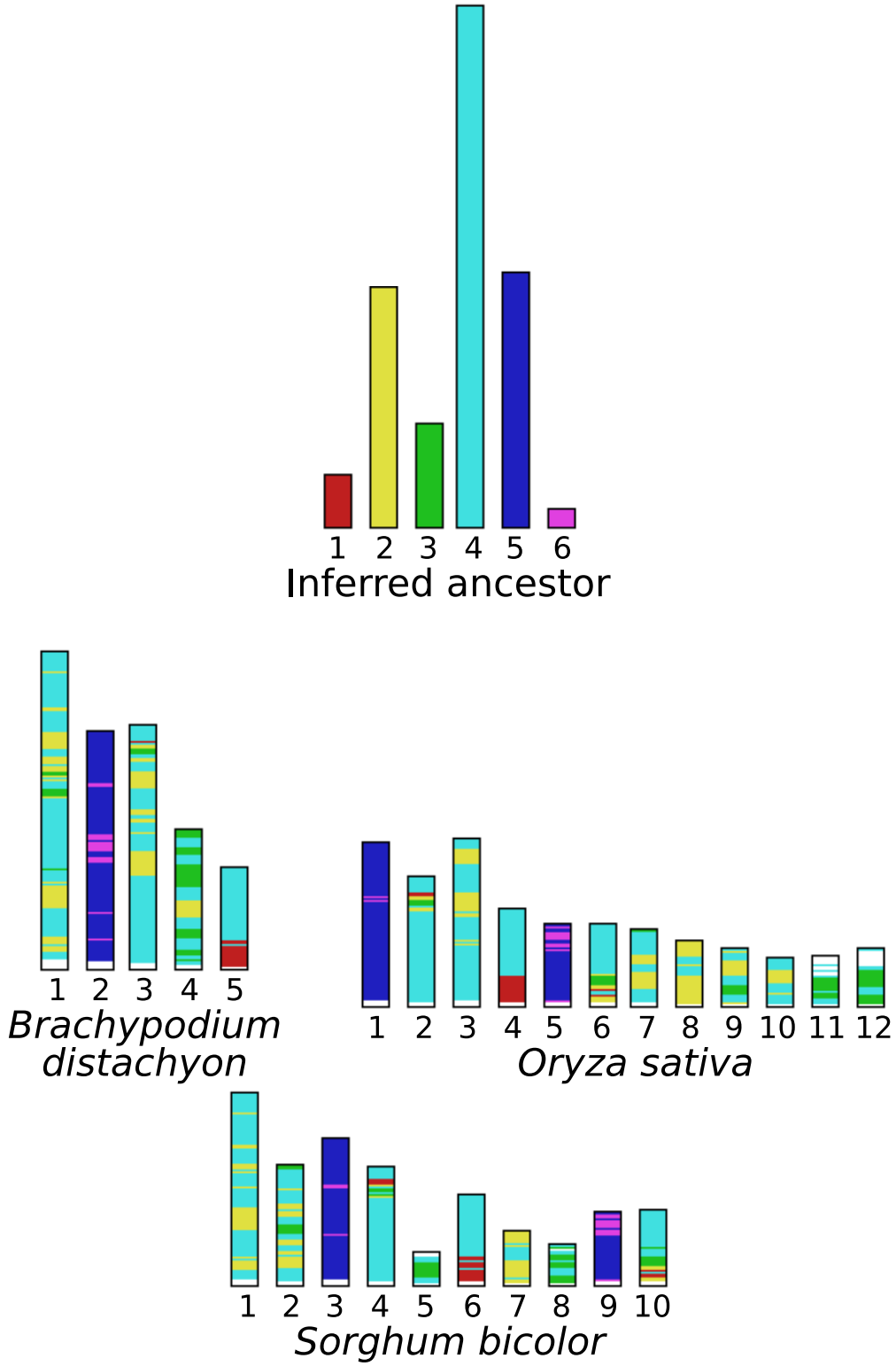
between accuracy in adjacency and karyotype reconstruction.

Supplementary algorithm 1

Algorithm Gapped-Adjacencies (GapAdj): (Σ, S, v, τ, MAX_α)

Initialize the set C of CARs to Σ_v;
For $\alpha = 1$ to MAX_α **Do**
STEP 1:
  For each internal node u of S (bottom-up traversal) **Do**
    For each $g \in \Sigma_u$ **Do**
      For each multiset X of possible adjacencies of g at u **Do**
        Compute $LeftAdj(g, \alpha, S|_{LA(g,\alpha,G(u))=X})$;
        Compute $RightAdj(g, \alpha, S|_{RA(g,\alpha,G(u))=X})$;
      End For
    End For
  End For
STEP 2:
  Construct the graph Q with vertices being the genes of Σ, and edges weighted according to computed α-adjacencies;
  By applying TSP-τ on Q, update the set C of CARs;
  Restrict Σ to the α-extremities of each CAR of C;
End For
Return (C);

Supplementary Figure



Supplementary Figure 1. Syntenic regions of three cereal species karyotype with respect to their ancestor inferred using our *GapAdj* algorithm.

CHAPITRE 5

CONCLUSION

L'apport principal de ce mémoire est une nouvelle méthode flexible pour la reconstruction de génomes ancestraux basée sur les α-adjacences et permettant des génomes aux contenus en gènes inégaux de même qu'une évolution par DGE. Un recensement des méthodes existantes a d'abord été fait, discutant des forces et des limites de chacune. Rappelons que les méthodes de reconstruction basées sur les distances génomiques sont limitées par les modèles biologiques incomplets utilisés : les distances ne peuvent tenir compte que d'un seul ou d'une certaine combinaison de type de réarrangements génomiques. Quant aux méthodes locales, la seule qui permet de gérer les paralogues (tous types confondus), DupCar, utilise la réconciliation d'arbres de gènes avec un arbre d'espèce, procédure qui est connue être problématique à cause de l'incertitude sur les arbres de gènes. Enfin, un problème majeur et récurrent pour toutes les méthodes est la quantité importante de solutions différentes mais équivalentes selon l'objectif de la méthode.

Rappelons également que la multiplicité des solutions provient de la présence d'information conflictuelle dans les synténies observables chez les espèces modernes. Sachant cela, notre laboratoire a d'abord développé une méthode basée sur les adjacences immédiates permettant des DGE et des insertions/suppressions de gènes. De la solution obtenue, on peut retrancher des adjacences moins bien supportées selon leur poid. Cela a pour effet d'augementer la confiance attribuée a cette solution, démontré par une comparaison avec la méthode InferCar sur des simulations, au prix de la fragmenter en un nombre plus important de RAC. Tentant de limiter l'effet du retranchement d'adjacences de faible poids, nous avons ensuite généralisé notre méthode aux α-adjacences. Des simulations ont démontré que cette généralisation permettait d'obtenir un nombre de RAC grandement réduit, sans pour autant diminuer de manière importante la confiance en la solution obtenue. Nous avons appliqué cette méthode chez trois espèces de céréales descendant d'une DGE, menant à certaines conclusions similaires à d'autres études sur l'évolution de ces trois espèces, notamment l'existence de fusions imbriquées de chro-

mosomes.

Avant de pouvoir inférer des génomes ancestraux à tous les noeuds d'un arbre d'espèce, un problème ouvert reste à résoudre : celui de tenir compte de tous les types de paralogues et non seulement des ohnologues. La solution peut se trouver dans les méthodes de réconciliation. En effet, certains efforts ont été portés dernièrement de manière à combler les lacunes de la réconciliation avec la correction des arbres de gènes [14, 18, 19]. La solution peut aussi se trouver dans une astuce lors des étapes du chaînage de synténies en RAC, permettant les paralogues mêmes si on ne peut clairement distinguer les différentes copies des gènes. La résolution de ce problème ouvert met en perspective la possibilité d'étudier une histoire évolutive complète de génomes d'espèces provenant de tous les clades de l'arbre de la vie.

BIBLIOGRAPHIE

[1] D. Applegate, W. Cook et A. Rohe. Chained lin-kernighan for large traveling salesman problems. *INFORMS Journal on Computing*, 15:82 - 92, 2003.

[2] M. Bader. The transposition median problem is np-complete. *Theoretical Computer Science*, 42(12-14):1099 - 1110, 2010.

[3] D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell et E.W. Sayers. Genbank. *Nucleic Acids Research*, 39:D32- D37, 2011.

[4] A. Bergeron, M. Blanchette, A. Chateau et C. Chauve. Reconstructing ancestral gene order using conserved intervals. Dans *Lecture Notes in Computer Science*, volume 3240 de *WABI*, pages 14- 25. Springer, 2004.

[5] A. Bergeron, C. Chauve et Y. Gingras. Formal models of gene clusters. Dans I. Mandoiu et A. Zelikovsky, éditeurs, *Bioinformatics algorithms : techniques and applications*, chaptitre 8. Wiley, 2008.

[6] A. Bergeron, J. Mixtacki et J. Stoye. Reversal distance without hurdles and fortresses. Dans S.C. Sahinalp, S. Muthukrishnan et U.l Dogrusoz, éditeurs, *Lecture Notes in Computer Science*, volume 3109 de *Combinatorial Pattern Matching*, pages 388- 399. Springer, 2004.

[7] A. Bergeron, J. Mixtacki et J. Stoye. A unifying view of genome rearrangements. Dans *Lecture Notes in Computer Science*, volume 4175 de *WABI*, pages 163- 173. Springer, 2006.

[8] D. Bertrand, Y. Gagnon, M. Blanchette et N. El-Mabrouk. Reconstruction of ancestral genome subject to whole genome duplication, speciation, rearrangement and loss. Dans V. Moulton et M. Singh, éditeurs, *Algorithms in Bioinformatics, WABI '10*, Lecture Notes in Computer Science, pages 78–89, 2010.

[9] P. Bonizzoni, G. Della Vedova et R. Dondi. Reconciling a gene tree to a species tree under the duplication cost model. *Theoretical Computer Science*, 347:36 - 53, 2005.

[10] K.S. Booth et G.S. Lueker. Testing for the consecutive ones property, interval graphs, and graph planarity using pq-tree algorithms. *Journal of Computer and System Sciences*, 13:335 - 379, 1976.

[11] G. Bourque et P.A. Pevzner. Genome-scale evolution : Reconstructing gene orders in the ancestral species. *Genome Research*, 12:26 – 36, 2002.

[12] David Bryant. A lower bound for the breakpoint phylogeny problem. Dans *CPM'00*, pages 235–247, 2000.

[13] A. Caprara. The reversal median problem. *INFORMS Journal on Computing*, 15: 93 - 113, 2003.

[14] W.C. Chang et O. Eulenstein. Reconciling gene trees with apparent polytomies. Dans D.Z. Chen et D.T. Lee, éditeurs, *Proceedings of the 12th Conference on Computing and Combinatorics (COCOON)*, volume 4112 de *Lecture Notes in Computer Science*, pages 235 - 244, 2006.

[15] C. Chauve et N. El-Mabrouk. New perspectives on gene family evolution : losses in reconciliation and a link with supertrees. Dans S. Batzoglou, éditeur, *RECOMB 2009*, volume 5441 de *Lecture Notes in Computer Science*, pages 46 - 58, 2009.

[16] C. Chauve, H. Gavranovic, A. Ouangraoua et E. Tannier. Yeast ancestral genome reconstructions : the possibilities of computational methods. *Plos Computational Biology*, 4(11):e1000234, 2008.

[17] C. Chauve et E. Tannier. A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes. *Plos Computational Biology*, 4(11):e1000234, 2008.

[18] K. Chen, D. Durand et M. Farach-Colton. Notung : Dating gene duplications using gene family trees. *Journal of Computational Biology*, 7:429 - 447, 2000.

[19] A. Doroftei et N. El-Mabrouk. Removing noise from gene trees. Dans *Proceedings of the 11th international conference on Algorithms in bioinformatics (WABI'11)*, Lecture Notes in Bioinformatics, pages 76 - 91, 2011.

[20] D. Durand, B. Haldórsson et B. Vernot. A hybrid micro-macroevolutionary approach to gene tree reconstruction. *Journal of computational biology*, 13:320 - 335, 2006.

[21] N. El-Mabrouk. Sorting signed permutations by reversals and insertions/deletions of contiguous segments. *Journal of Discrete Algorithms*, 1:105- 122, 2001.

[22] N. El-Mabrouk, J. Nadeau et D. Sankoff. Genome halving. Dans M. Farach, éditeur, *CPM 9th annual symposium*, volume 1448 de *Lecture Notes in Computer Science*, pages 235 - 250, 1998.

[23] N. El-Mabrouk et D. Sankoff. *Analysis of Gene Order Evolution beyond Single-Copy Genes*, volume Evolutionary Genomics : statistical and computational methods de *Methods in Molecular Biology*, chaptitre Part II. Springer (Humana). (to appear).

[24] N. El-Mabrouk et D. Sankoff. The reconstruction of doubled genomes. *SIAM Journal on Computing*, 32(1):754-792, 2003.

[25] The French-Italian Public Consortium for Grapevine Genome Characterization. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 449:463-468, 2007.

[26] D. Fulkerson et O. Gross. Incidence matrices and interval graphs. *Pac. J. Math.*, 15:835- 855, 1965.

[27] H. Gavranovic, C. Chauve, J. Salse et E. Tannier. Mapping ancestral genomes with massive gene loss : A matrix sandwich problem. *Bioinformatics*, 27(ISMB 2011): i257- i265, 2011.

[28] H. Gavranović et E. Tannier. Guided genome halving : provably optimal solutions provide good insights into the preduplication ancestral genome of Saccharomyces cerevisiae. volume 15 de *Pacific Symposium on Biocomputing*, pages 21-30, 2010.

[29] M.B. Gerstein, C. Bruce, J.S. Rozowsky et al. What is a gene, post-encode ? history and updated definition. *Genome Research*, 17:669-681, 2007.

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

[31] J.L. Gordon, K.P. Byrne et K.H. Wolfe. Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern saccharomyces cerevisiae genome. *PloS Genetics*, 5(5), 2009.

[32] P. Gorecki et J. Tiuryn. Dls-trees : a model of evolutionary scenarios. *Theoretical computer science*, 359:378 - 399, 2006.

[33] R. Guigó, I. Muchnik et T. Smith. Reconstruction of ancient molecular phylogeny. *Molecular Phylogenetics and Evolution*, 6:189 - 213, 1996.

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

[35] S. Hannenhalli et P.A. Pevzner. Transforming men into mice. Dans *Proceedings of the IEEE 36th Annual Symposium on Foundations of Computer Science*, pages 581–592, 1995.

[36] S.M. Hedtke, T.M. Townsend et D.M. Hillis. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. *Systematic Biology*, 55:522- 529, 2006.

[37] International Brachypodium Initiative. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*, 463:763- 768, 2010.

[38] H. Innan et F. Kondrashov. The evolution of gene duplications : classifying and distinguising between models. *Nature Reviews Genetics*, 11:97- 108, 2010.

[39] W. James Kent, Robert Baertsch, Angie Hinrichs, Webb Miller et David Haussler. Evolution's cauldron : duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci U S A*, 100(20):11484–11489, Sep 2003.

[40] E.V. Koonin. Orthologs, paralogs and evolutionary genomics. *The Annual Review of Genetics*, 39:309- 338, 2005.

[41] S. Lin et B.W. Kernighan. An effective heuristic algorithm for the traveling salesman problem. *Operations Research*, 21:498- 516, 1973.

[42] B. Ma, M. Li et L. Zhang. From gene trees to species trees. *SIAM Journal on Computing*, 30:729 - 752, 2000.

[43] J. Ma, A. Ratan, B.J. Raney, B.B. Suh, L. Zhang, W. Miller et D. Haussler. Dupcar : Reconstructing contiguous ancestral regions with duplications. *Journal of Computational Biology*, 15(8):1- 21, 2008.

[44] 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, 2007.

[45] M. Marron, K.M. Swenson et B.M.E. Moret. Genomic distances under deletions and insertions. *Theor. Computer Science*, 325:347- 360, 2004.

[46] B. Moret, L. Wang, T. Warnow et S. Wyman. New approaches for reconstructing phylogenies from gene order data. *Bioinformatics*, 17:S165–S173, 2001.

[47] M. Muffato, A. Louis, C.E. Poisnel et H. Roest Crollius. Genomicus : a database and a browser to study gene synteny in modern and ancestral genomes. *Bioinformatics*, 26(8):1119- 1121, 2011.

[48] F. Murat, J.H. Xu, E. Tannier, M. Abrouk, N. Guilhot, C. Pont, J. Messing et J. Salse. Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. *Genome Research*, 2010.

[49] S. Ohno. *Evolution by gene duplication*. Springer, Berlin, 1970.

[50] S. Ohno, U. Wolf et N.B. Atkin. Evolution from fish to mammals by gene duplication. *Hereditas*, 59:169 - 187, 1968.

[51] A. Ouangraoua, E. Tannier et C. Chauve. Reconstructing the architecture of the ancestral amniote genome. *Bioinformatics*, 27(19):2664- 2671, 2011.

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

[53] 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, 1994.

[54] I. Pe'er et R. Shamir. The median problems for breakpoints are NP-complete. volume 5 de *Electronic colloquium on computational complexity*, 1998.

[55] M. Sanderson et M. McMahon. Inferring angiosperm phylogeny from est data with widespread gene duplication. *BMC Evolutionary Biology*, 7:S3, 2007.

[56] D. Sankoff. Reconstructing the history of yeast genomes. *PLoS Genet.*, 5: e1000483, 2009.

[57] D. Sankoff et M. Blanchette. The median problem for breakpoints in comparative genomics. Dans T. Jiang et D.T. Lee, éditeurs, *Computing and Combinatorics, Proceeedings of COCOON '97*, numbéro 1276 dans Lecture Notes in Computer Science, pages 251–263, Berlin, 1997. Springer.

[58] D. Sankoff et M. Blanchette. Multiple genome rearrangement and breakpoint phylogeny. *Journal of Computational Biology*, 5:555–570, 1998.

[59] D. Sankoff, G. Sundaram et J. Kececioglu. Steiner points in the space of genome rearrangements. *International J. Foundations Comput. Sci.*, 7:1- 9, 1996.

[60] I. Schubert et M.A. Lysak. Interpretation of karyotype evolution should consider chromosome structural constraints. *Trends in Genetics*, 27(6):207-216, 2011.

[61] A.U. Sinha et J. Meller. Cinteny : flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. *BMC Bioinformatics*, 8:82, 2009.

[62] E. Tannier. Yeast ancestral genome reconstructions : the possibilities of computational methods. 5817:1 - 12, 2009.

[63] G. Tesler. Efficient algorithms for multichromosomal genome rearrangements. *Journal of Computer and System Sciences*, 65(3):587-609, 2002.

[64] A.H. Paterson *et al.* The sorghum bicolor genome and the diversification of grasses. *Nature*, 457:551- 556, 2009.

[65] S. Ouyang *et al.* The tigr rice genome annotation resource : improvements and new features. *Nucleic Acids Research*, 35:D883- D885, 2007.

[66] J.D. Watson et F.H.C. Crick. A structure for deoxyribose nucleic acid. *Nature*, 171: 737-738, 1953.

[67] K.H. Wolfe et D.C. Shields. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387:708 - 713, 1997.

[68] S. Yancopoulos, O. Attie et R. Friedberg. Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics*, 21:3340-3346, 2005.

[69] C. Zheng et D. Sankoff. On the Pathgroups approach to rapid small phylogeny. *BMC Bioinformatics*, 12:S4, 2011.

[70] C. Zheng, Q. Zhu, Z. Adam et D. Sankoff. Guided genome halving : hardness, heuristics and the history of the hemiascomycetes. ISMB, pages 96 - 104, 2008.

[71] C. Zheng, Q. Zhu et D. Sankoff. Descendants of whole genome duplication within gene order phylogeny. *Journal of Computational Biology*, 15(8):947-964, 2008.