ACADÉMIE DE MONTPELLIER

UNIVERSITÉ MONTPELLIER II

— SCIENCES ET TECHNIQUES DU LANGUEDOC —

Thèse

présentée à l'Université des Sciences et Techniques du Languedoc
pour obtenir le diplôme de DOCTORAT

SPÉCIALITÉ	: **Informatique**
Formation Doctorale	: **Informatique**
École Doctorale	: **Sciences pour l'Ingénieur**

Méthodes et algorithmes pour reconstruire les arbres de l'Évolution

par

Vincent BERRY

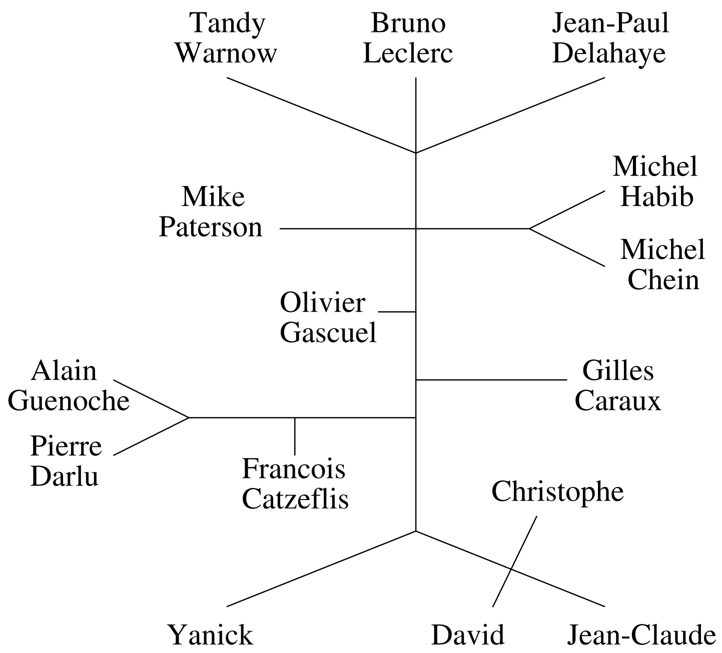
Soutenue le 18 décembre 1997 devant le Jury composé de :

M. Jean-Paul DELAHAYE, Professeur, Univ. Sciences et Technologies Lille, LIFL, ... Rapporteur
M. Bruno LECLERC, Maître de Conférence, E.H.E.S.S. Paris, C.A.M.S., ... Rapporteur
Mme. Tandy WARNOW, Professeur, Univ. of Pennsylvania, Philadelphie, USA, ... Rapporteur
M. Pierre DARLU, Directeur de Recherche, C.N.R.S.-I.N.S.E.R.M., Paris, ... Examinateur
M. Gilles CARAUX, Professeur, E.N.S.A. de Montpellier, ... Examinateur
M. Alain GUÉNOCHE, Chargé de Recherche, C.N.R.S., L.I.M., Marseille, ... Examinateur
M. Michel HABIB, Professeur, Univ. Montpellier II, L.I.R.M.M., ... Examinateur
M. Olivier GASCUEL, Chargé de Recherche, C.N.R.S., L.I.R.M.M., ... Directeur de Thèse

Remerciements

Au moment de remercier tous ceux qui m'ont aidé au cours de ces trois années de labeur, je me sens quelque peu dépossédé de cet ouvrage. Cette thèse est le produit d'une convergence de volontés, d'amitiés et de savoirs dont je n'ai été que l'acteur principal.

Impossible de nommer ici tout le monde, impossible de remercier à leur juste valeur toutes les contributions, impossible d'en proposer un classement raisonné ... Et pourtant, il faut bien se livrer à cet exercice difficile. Ne sachant plus penser sans construire des arbres, j'adopterai ici ce modèle :



Depuis cet arbre, l'ordre des remerciements s'obtient par un parcours simple depuis un sommet que je laisse au lecteur le soin d'identifier (les longueurs d'arêtes sont fixées selon le critère de l'esthétique maximum et n'ont aucune autre signification).

Je suis particulièrement reconnaissant à Bruno Leclerc, Jean-Paul Delahaye et Tandy Warnow qui ont accepté la lourde tâche de rapporteur. La rédaction finale de ma thèse a bénéficié de leur lecture très attentive et de la qualité de leurs remarques. Je les remercie pour l'intérêt qu'ils ont accordé à mon travail. Je

voudrais aussi remercier Tandy Warnow pour la confiance et le soutien actif qu'elle m'a accordés de longue date, ainsi que pour son enthousiasme communicatif.

Je remercie Michel Habib, qui a accepté de présider le jury de ma thèse, et Michel Chein pour les nombreuses discussions que nous avons pues avoir au détour des couloirs sur l'exploitation de mes résultats. Ils ont su me transmettre le virus de la recherche en informatique, et m'ont soutenu tout au long de ma thèse. Je n'oublie pas non plus tous les enseignants qui ont contribué à ma formation, ainsi que Philippe Ezequel qui a dirigé mes premiers pas en recherche.

Mike Paterson a eu la gentillesse de m'accueillir en Post-Doc à l'Université de Warwick alors que je finissais ma thèse. Je lui suis reconnaissant d'avoir accepté les nombreuses perturbations horaires et déplacements qu'impliquaient la fin de rédaction et la soutenance de cette thèse.

Au centre de l'arbre des remerciements, se trouve celui à qui ce travail de thèse doit le plus, celui qui a toujours su me pousser plus loin, me communiquant sa rigueur et son exigence, celui que je n'ai pas hésité à solliciter même certains dimanches (et j'en demande pardon à sa femme et ses quatre filles) : Olivier Gascuel. J'ai pu particulièrement apprécier ses vastes compétences scientifiques, sa vision globale de l'informatique et des mathématiques qui ont toujours su diriger mes efforts dans la bonne direction. Je ne parlerai pas ici des rapports amicaux que nous avons pu nouer au cours d'un certain nombre de périgrinations en dehors de toute activité scientifique, mais il va sans dire qu'ils m'ont été et me sont précieux. Je ne le remercierai jamais assez de l'attention toute particulière dont il a su entourer le déroulement de ma thèse.

Je dois encore à Olivier d'avoir rencontré Gilles Caraux, qui a su prendre la relève en son absence. Il a été pour moi une oreille attentive et amicale, ne comptant jamais le temps qu'il me consacrait. Je lui dois l'essentiel de ma formation en statistiques : les connaissances qu'il m'a gentiment et patiemment trasmises ont contribué, de façon non négligeable, à la pertinence de mes recherches. Tout comme Olivier, il a aussi su m'offrir de précieux moments de détente (autour d'un bon verre de vin), et n'a cessé de m'encourager. Je l'en remercie chaleureusement.

Olivier et Gilles m'ont toujours poussé à communiquer avec les autres chercheurs du domaine, m'encourageant régulièrement à exposer mes travaux lors de réunions des groupes de recherche auxquels ils participaient. A la faveur de ces réunions, j'ai rencontré certaines personnes passionnées et passionnantes, comme Alain Guénoche. Je le remercie pour son attitude chaleureuse ainsi que pour la gentillesse et la patience dont il a toujours fait preuve à mon égard. Les longues discussions que nous avons eues à maintes reprises m'ont beaucoup apporté.

À l'occasion de certaines de réunions, j'ai eu la chance de rencontrer de nombreux biologistes, dont Pierre Darlu. Il a grandement contribué à mon intérêt pour la Reconstruction Phylogénétique, par ouvrage interposé d'abord, puis par de nombreuses discussions sur l'histoire et les enjeux du domaine. La pertinence de mes recherches lui doit beaucoup. Je le remercie aussi pour avoir bien voulu porter la "casquette du biologiste" lors de ma soutenance de thèse.

Je remercie aussi François Catzeflis pour m'avoir laissé libre accès à sa bibliothèque (et à sa photocopieuse !), et pour avoir bien voulu écouter les divagations d'un informaticien lors de mes irruptions dans son bureau.

Enfin, il me faut remercier tous les assistants dont, dans l'ombre, l'aide a été aussi discrète qu'efficace : dans le désordre, je citerai David, assistant-linguistique, Jean-Claude, assistant-vulgarisateur, Christophe, assistant-communication (expression orale, costume), Yanick, assistante-expression, Guillaume, assistant-biologie, Babak, assistant-détente. Bien-sûr, je n'oublie pas les nombreux thésards de la zone technique dont le calme légendaire a contribué à l'ambiance feutrée de mon bureau, mes amis et ma famille qui ont souvent dû se plier à mes horaires particulières.

Cette thèse est dédiée à la personne qui compte le plus pour moi, qui a eu la patience de me supporter pendant ces longues années d'étude, et qui a contribué de multiples façons à cette thèse, Yanick Goustiaux-Berry.

"Un homme devint fou pour avoir réflechi trop profondément à l'action d'ouvrir et de fermer une porte [...] À côté était un autre fou qui cherchait à deviner si l'oeuf avait précédé la poule, ou si la poule avait précédé l'oeuf. [...]

Un fou est un homme qui voit un abîme et y tombe. Le savant l'entend tomber, prend sa toise, mesure la distance, fait un escalier, remonte, et se frotte les mains, après avoir dit à l'univers : "cet abîme a dit 1802 pieds de profondeur, la température du fond est de 2 degrés plus chaude que celle de notre atmosphère". [...]

Ici je serai toujours entre la toise du savant et le vertige du fou. [...]

Je parle pour les gens habitués à trouver de la sagesse dans la feuille qui tombe, des problèmes gigantesques dans la fumée qui s'élève, des théories dans les vibrations de la lumière, de la pensée dans les marbres, et le plus horrible des mouvements dans l'immobilité. Je me place au point précis où la science touche à la folie, et je ne puis mettre de garde-fous."

Théorie de la démarche,
Honoré de Balzac, (1833),
p. 20-21, Pandora Editions, 1978.

Résumé

Dans cette thèse, nous nous intéressons à la reconstruction d'arbres *évolutifs* (ou "*phylogénies*") retraçant l'histoire évolutive d'espèces vivantes. Afin d'obtenir une estimation aussi proche que possible d'une phylogénie inconnue, nous prenons le parti pris de produire une phylogénie ne contenant que des arêtes *fiables*, quitte à ne retrouver qu'une partie de la phylogénie recherchée. Les méthodes que nous proposons se divisent en deux catégories : les méthodes *à résolution décroissante* partent d'une phylogénie complète de laquelle elles enlèvent les arêtes non sûres ; les méthodes *à résolution croissante* infèrent d'emblée une phylogénie ne contenant que des arêtes jugées fiables.

Les méthodes à résolution décroissante que nous proposons reposent sur des techniques de rééchantillonnage, et plus particulièrement sur celle du bootstrap. Dans un premier temps, nous montrons comment détourner la procédure de bootstrap, telle qu'elle fut introduite en reconstruction phylogénétique, pour proposer une estimation fiable d'une phylogénie d'espèces. Puis, nous définissons une deuxième procédure bootstrap, dont l'originalité est d'estimer, en fonction de la qualité du jeu de données et de la méthode de reconstruction utilisée, le nombre d'arêtes que doit comporter la phylogénie proposée pour avoir une erreur minimum. Cette nouvelle procédure donne des résultats d'aussi bonne qualité que la précédente, en des temps plus courts.

Les méthodes à résolution croissante que nous proposons reposent sur un principe d'optimisation combinatoire lié aux quadruplets d'espèces : on infère d'abord des phylogénies sur quatre espèces, qu'on cherche ensuite à combiner en une phylogénie sur l'ensemble des espèces, de façon à optimiser un certain critère. Dans ce cadre, nous montrons que le problème du sous-ensemble maximum de quadruplets, équivalent à un arbre, peut être résolu de façon exacte en temps polynômial, $O(n^4)$. En associant l'algorithme obtenu avec un principe simple d'inférence de phylogénies sur 4 espèces, nous obtenons la méthode Q^*, ne proposant que des arêtes combinatoirement sûres. Nous caractérisons la boule de consistance de cette méthode, au sens de la norme L_∞, ainsi que sa vitesse de convergence. Nous obtenons aussi, la probabilité que chaque arête de la phylogénie inconnue soit retrouvée par la méthode Q^* (sous un certain modèle de l'évolution). Expérimentalement, nous observons que la méthode Q^* infère des phylogénies partiellement résolues mais comportant un pourcentage très faible d'arêtes incorrectes. Enfin, pour compléter l'arbre inféré par la méthode Q^*, parfois trop conservatrice, nous proposons un schéma agglomératif basé sur les quadruplets et de complexité $O(n^4)$.

Les simulations réalisées montrent que les méthodes (à résolution croissante et à résolution décroissante) que nous proposons obtiennent de meilleurs résultats que les méthodes usuelles pour estimer la structure d'une phylogénie inconnue.

Sommaire

Introduction générale

Reconstruire l'histoire évolutive des espèces est un formidable défi que les scientifiques de plusieurs domaines essayent de relever. Cette reconstruction se base sur le fait que certains caractères ont été hérités d'une espèce à l'autre, tandis que d'autres se sont transformés au cours du temps, laissant ainsi des *traces* de l'évolution. L'évolution des espèces est comprise comme un processus divergent, modélisé le plus souvent sous la forme d'un arbre dont les noeuds feuilles représentent les espèces contemporaines et les autres noeuds, les espèces ancestrales. On parle alors d'*arbre d'évolution* ou plus communément de *phylogénie*.

Pour définir toute l'importance du problème de la reconstruction phylogénétique, cédons la parole à Darlu et Tassy ([DT93], p. 228) :

> "la reconstruction phylogénétique est une étape obligée dans la compréhension des mécanismes évolutifs qui ont conduit à la diversité du monde vivant. Il est impossible de comprendre ou de localiser des changements de vitesse d'évolution sans une phylogénie préalable, impossible également de trancher sur la neutralité des gènes ou du rôle de la sélection si l'on n'a pas une image aussi claire que possible de la parenté entre les molécules de différents génomes. Impossible encore de préciser l'organisation du génome sans une histoire évolutive des éléments qui le constituent."

Parmi les applications de ce domaine, on compte la lutte contre les maladies infectueuses [Roo95], comme le virus du SIDA [BD92] ou la malaria [EHH96]. Par exemple, en classant les différentes souches d'un virus, on peut sélectionner les caractéristiques que doit posséder un remède destiné à une certaine souche de la maladie, en fonction des caractéristiques des remèdes s'appliquant aux souches voisines.

Dans cette thèse, nous abordons le problème de la reconstruction phylogénétique selon un point de vue lié à l'estimation statistique. L'objectif est de retrouver la phylogénie inconnue d'un groupe d'espèces étudiées. Les méthodes de reconstruction sont vues comme autant d'estimateurs possibles pouvant fournir une estimation arborée de cette phylogénie. Toute phylogénie inférée dans ce but peut être considérée comme un assemblage d'arêtes, ayant chacune une certaine probabilité d'être correcte (d'appartenir à la phylogénie recherchée). Comme les données dont on dispose sont bruitées, elles ne correspondent jamais exactement avec une phylogénie. Tout l'art des méthodes de reconstruction (et c'est ce qui les divise principalement) est alors de deviner quel est la phylogénie la plus vraisemblable en fonction des données dont on dispose. Ce faisant, un certain degré d'inexactitude est acceptable (p. ex. en ce qui concerne les taux d'évolution on fait parfois l'hypothèse de *l'horloge moléculaire*[1]). Toutefois, un problème important est que la plupart des méthodes infèrent une phylogénie trop spécifique des données, au sens où certaines de ses arêtes sont remises en cause par la phylogénie inférée depuis un autre jeu de données pour les mêmes espèces, ou depuis le même jeu de données auquel on a ajouté une nouvelle espèce.

Dans cette thèse, nous attachons une importance toute particulière à ce problème, en essayant de proposer une estimation fiable de la phylogénie des espèces, *i.e.*, une estimation ne contenant que des arêtes ayant une probabilité importante d'être correctes. Les méthodes que nous présentons recouvrent les deux démarches possibles pour atteindre cet objectif :

– l'approche *à résolution décroissante* consiste à partir d'une estimation complète de la phylogénie inconnue, de laquelle on cherche à enlever les éventuelles arêtes non-fiables.

– l'approche *à résolution croissante* consiste à essayer d'emblée de proposer une estimation de la phylogénie ne contenant que des arêtes fiables.

La rédaction de cette thèse est organisée selon trois parties. La partie I constitue une introduction au domaine de la reconstruction phylogénétique. Elle présente les notions de base sur l'évolution telle qu'on la modélise, et propose un panorama des méthodes de reconstruction usuelles.

Les apports de cette thèse se situent sur le plan méthodologique : les méthodes de reconstruction phylogénétique que nous proposons font l'objet des parties II et III.

La partie II présente des méthodes à résolution décroissante, basées sur des techniques de rééchantillonnage. La rédaction de cette partie prend donc davantage en compte l'aspect statistique que l'aspect algorithmique.

La partie III présente des méthodes à résolution croissante, basées sur un principe de quadruplets. Les problèmes et les méthodes abordés font ici partie du domaine de l'optimisation combinatoire.

1. *i.e.*, que les espèces étudiées ont toujours évoluées à la même vitesse.

Première partie

Premiers pas dans le domaine de la Reconstruction Phylogénétique

Introduction

Cette première partie de la thèse offre un panorama général sur le domaine de la reconstruction phylogénétique. Le lecteur désireux d'approfondir les notions qui y sont présentées pourra consulter les ouvrages très complémentaires de Barthélemy et Guénoche [BG88, BG91], de Darlu et Tassy [DT93], et de Swofford *et al* [SOWH96].

Le chapitre 1 de cette partie présente les fondements du domaine de la reconstruction phylogénétique en développant un certain nombre de notions liées à l'évolution et à sa modélisation. Ces notions sont utiles pour bien comprendre les problèmes rencontrés par les méthodes de reconstruction.

Le chapitre 2 détaille les différentes familles de méthodes de reconstruction, ainsi que les qualités recherchées (mais jamais toutes réunies) pour une méthode.

À la lumière de ce bilan, le chapitre 3 établit la problématique de cette thèse, en montrant comment l'approche proposée ici se distingue de l'approche traditionnelle du problème.

Notons enfin que, pour des raisons de concision, nous utiliserons l'abréviation "R.P." pour désigner la "reconstruction phylogénétique".

Chapitre 1

La reconstruction phylogénétique

Le terme « phylogénie » fut utilisé pour la première fois par E. Haeckel en 1866 pour définir l'enchaînement des espèces animales et végétales au cours du temps, concept que, jusque là, le terme « généalogie » recouvrait. Ce terme fut ensuite repris dans la dernière édition de *l'Origine des espèces* de C. Darwin, pour désigner "*les lignes généalogiques de tous les êtres organisés*". La *théorie de l'évolution* que proposait C. Darwin dans cet ouvrage avait pour principal postulat la descendance avec modification. Cette théorie dut sa reconnaissance au fait qu'elle donnait une interprétation générale à des phénomènes déjà admis comme l'homologie (héritage d'un ancêtre commun) et l'ordre de la nature (classification des espèces du vivant en différentes catégories). À cette époque fleurissaient déjà de nombreuses généalogies et classifications des espèces animales et végétales. La *théorie de l'évolution* ajouta une nouvelle dimension aux classifications : la dimension temporelle. Depuis cette époque, les **arbres** se sont imposés comme le support graphique naturel des phylogénies, permettant de représenter simultanément les groupements d'espèces et la dimension temporelle.

L'étude des phylogénies s'insère dans le domaine de la biologie comparative, aussi appelée systématique, dont le but est de **proposer une classification des organismes vivants**, afin de comprendre les causes de leur diversité. Le classement des espèces a d'abord été pratiqué sur la base d'observations morphologiques, comme la présence ou l'absence d'ailes, le nombre de pattes, etc. L'utilisation de données moléculaires remonte aux études immunologiques du début du siècle [Nut04], mais les phylogénies moléculaires ne furent vraiment acceptées que dans les années 60, suite à la publication de phylogénies, obtenues depuis des séquences d'ADN ou protéiques, possédant de "bonnes" congruences avec les classifications morphologiques [ZP62, FM67]. La comparaison des espèces, en fonction des différences constatées dans leur génome, connut alors un engouement qui n'a pas cessé d'augmenter depuis, et ce pour une double raison. Premièrement, les séquences étant des enchaînements de constituants chimiques simples (les nucléotides `A,C,G` et `T`), la comparaison de séquences moléculaires est dépourvue d'ambiguïté, alors que les données morphologiques sont parfois sujettes à diverses interprétations rendant les observations subjectives et donc critiquables. Deuxièmement, les millions de nucléotides, qui composent le génome des organismes, constituent un réservoir de données quasiment inépuisable en comparaison des quelques centaines de caractères morphologiques qu'on utilisait précédemment pour reconstruire des phylogénies.

Dans les années 70 et 80, les techniques de séquencage de l'ADN furent considérablement améliorées [SNC77, QMB83, MF87], augmentant de façon toujours exponentielle la quantité de données exploitables pour les analyses phylogénétiques. Ces données sont pour la grande majorité disponibles dans des banques de données accessibles à la communauté scientifique par l'intermédiaire du réseau *Internet* (*EMBL*, *GenBank*, *Swissprot*, etc). Ces immenses quantités de données ont créé le besoin de disposer d'outils de traitement systématiques, autant pour organiser cette connaissance (stockage) que pour l'exploiter (inférence de phylogénies). C'est ce qui conduisit les informaticiens à s'intéresser au problème de la reconstruction phylogénétique. Ils rejoignaient dans ce domaine les biologistes, mais aussi les mathématiciens, qu'intéressait déjà l'aspect analyse de données. En effet, depuis les données moléculaires ou morphologiques, on peut facilement évaluer des similitudes ou des dissimilitudes entre les espèces deux à deux. La reconstruction phylogénétique consiste alors en la représentation d'une matrice de dissimilarités (*i.e.*, de *distances évolutives*) par une structure arboreée, problème bien connu en mathématiques.

1.1 Les données traduisant l'évolution

Les données morphologiques et moléculaires dont on dispose initialement peuvent généralement être considérées comme des **caractères** statistiques discrets pouvant prendre plusieurs valeurs ou **états**. Dans le cas de données moléculaires, les caractères correspondent par exemple aux différents *sites nucléotidiques* de la séquence, dont les états possibles sont `A,C,G` et `T` (représentant respectivement l'adénine, la cytosine, la guanine et la thymine). Dans le cas de données morphologiques, les caractères sont déterminés par l'investigateur et peuvent être soit binaires (présence ou absence de dents) soit multivalués (nombre de dents). Chaque espèce est décrite par les valeurs qu'elle prend pour les différents caractères. D'un point de vue informatique, il s'agit d'une chaîne de caractères sur un alphabet limité (p. ex. à

4 lettres pour les 4 bases chimiques de l'ADN ou à 20 lettres pour coder les acides aminés des protéines).

L'**évolution** des espèces est généralement considérée comme un processus divergeant au cours du temps, *i.e.*, faisant se scinder les espèces en plusieurs lignées à divers points du temps, et traduisant un certain nombre de modifications dans les états pris par les espèces pour les différents caractères. Quand un changement d'état intervient pour un caractère, on parle de **substitution** ou encore de mutation. Ce sont ces changements d'état, indépendants d'une espèce à l'autre, qui contribuent à l'éloignement progressif des espèces les unes des autres. Ainsi, pour des données morphologiques, l'évolution se traduit comme l'héritage ou la transformation de certains états de caractères d'un ancêtre à ses descendants (p. ex. la disparition des ailes chez les reptiles au Crétacé Supérieur). Dans le cas de données moléculaires, la différenciation du génome des espèces résulte non seulement de substitutions (modification de l'état d'un site), mais aussi d'*insertions* (ajout de sites supplémentaires) et de *délétions* (suppression de sites de la séquence). Ces événements peuvent se produire en divers endroits de la séquence, modifiant sa longueur et changeant la place relative des caractères initiaux. Pour pouvoir retrouver d'une espèce à l'autre quels nucléotides correspondent aux mêmes sites (*i.e.*, caractères) ancestraux, on est obligé de passer par une étape d'*alignement* des séquences.

1.1.1 L'alignement de séquences moléculaires

Lorsqu'on envisage de reconstruire la phylogénie d'un groupe d'espèces sur la base de données moléculaires, on sélectionne, dans un premier temps, une ou plusieurs portions bien précises de leur génome, comme des gènes (cytochrome C, myoglobine, etc). Une deuxième étape, indispensable, consiste à aligner les séquences obtenues pour les différentes espèces, *i.e.*, à **mettre en correspondance les séquences, nucléotide par nucléotide**. Il est en effet fréquent que des bouts d'ADN divers se soient insérés (ou aient été supprimés) dans certaines des séquences, brouillant ainsi les correspondances initiales entre les sites des différentes séquences. Pour que la reconstruction de la phylogénie puisse être menée correctement, il est nécessaire d'identifier les parties des séquences qui correspondent à de tels événements. Prenons l'exemple donné dans Caraux *et al* [CGAL95], où on dispose des trois séquences

$$
\begin{aligned}
S_1 &= \mathrm{AGAATAGCCA} \\
S_2 &= \mathrm{AGGATAGGA} \\
S_3 &= \mathrm{AGTATGGA}
\end{aligned}
$$

un alignement possible est :

$$
\begin{array}{cccccccccccc}
 & & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\
S_1 & : & \mathrm{A} & \mathrm{G} & \mathrm{A} & \mathrm{A} & \mathrm{T} & \mathrm{A} & \mathrm{G} & \mathrm{C} & \mathrm{C} & \mathrm{A} \\
S_2 & : & \mathrm{A} & \mathrm{G} & \mathrm{G} & \mathrm{A} & \mathrm{T} & \mathrm{A} & \mathrm{G} & \mathrm{G} & . & \mathrm{A} \\
S_3 & : & \mathrm{A} & \mathrm{G} & \mathrm{T} & \mathrm{A} & \mathrm{T} & . & \mathrm{G} & \mathrm{G} & . & \mathrm{A}
\end{array}
$$

Cet alignement s'explique ainsi : en position 0, 1, 3, 4, 6 et 9, aucun événement mutationel n'est survenu ; en position 2, il y a eu 2 substitutions ; en position 5, il y a eu une délétion ; en position 7, une substitution ; en position 8, une insertion. Cet alignement peut aussi s'interpréter de façon différente : on peut aussi supposer qu'en position 8, il y a eu deux délétions. Toutefois, l'interprétation précédente de cette position est la plus parcimonieuse, et c'est celle qu'on retiendra. De la même manière, d'autres alignements que celui proposé ci-dessus sont possibles pour les séquences S_1, S_2 et S_3. On tendra toujours vers l'alignement le plus parcimonieux.

D'un point de vue informatique, ce problème s'apparente à la recherche de la "distance d'édition" entre les séquences. La définition de critères permettant d'évaluer la qualité des alignements, comme la recherche du meilleur alignement au sens de ces critères, constitue tout un domaine de recherche [MC91], généralement séparé de la R.P., à de rares exceptions près [Hei89, JL92]. Trouver le meilleur alignement pour un ensemble de n séquences est un problème NP-difficile en général.

La correspondance, établie entre les nucléotides des différentes séquences lors de l'alignement, définit les caractères (correspondant aux positions) qui servent de données à la R.P. Aussi, la justesse d'une phylogénie proposée sur la base de données moléculaires dépend-elle de la qualité de l'alignement. Le risque principal consiste à regrouper sous un même caractère des nucléotides provenant initialement de sites differents. Ceci explique que l'alignement soit une étape qui doit être particulièrement soignée, et pour laquelle l'expérience et la connaissance des biologistes sont primordiales. Ceux-ci utilisent généralement divers algorithmes pour proposer un premier alignement acceptable, qu'ils affinent ensuite "à la main", en se servant de logiciels interactifs. Ils veulent avant tout s'assurer que les états mis en correspondance sont *homologues*, *i.e.*, sont issus des mêmes sites pour toutes les espèces.

Dans les régions stables des séquences (*i.e.*, les régions dans lesquelles on peut proposer un alignement sans insertion-délétion), on peut sans grand risque supposer que les nucléotides, regroupés sous une même position, correspondent bien à un même site de la séquence ancestrale. Aussi ces régions sont-elles privilégiées dans l'alignement : dans le petit alignement donné plus haut, on estimera sans doute que la région 0-4 est bien conservée, tandis que la région 5-9 ne l'est pas et doit être écartée.

Finalement, seuls les caractères pour lesquels l'alignement attribue un état à toutes les espèces sont conservés pour la R.P. C'est ce qui fait le paradoxe de l'alignement : plus on étudie d'espèces et moins on dispose de caractères pour inférer la phylogénie, car les caractères définis sur toutes les séquences sont de moins en moins nombreux.

Dans le cadre de cette thèse, nous ne nous intéressons qu'à la phase de reconstruction de la phylogénie et non à l'obtention des données qui la sous-tendent. Aussi, lorsque nous évoquerons des séquences moléculaires décrivant les espèces étudiées, nous supposerons toujours qu'elles auront été alignées précédemment. Nous supposerons disposer du résultat de l'alignement sous la forme d'une matrice X de k caractères, dont les lignes correspondent bijectivement aux espèces et les colonnes bijectivement aux sites. Nous supposerons que pour tous les sites, nous connaissons l'état pris par chacune des espèces, *i.e.*, X est une matrice complète.

1.1.2 Le concept de similitude

Toute reconstruction phylogénétique est basée sur le concept de **similitude** : plus les espèces se ressemblent (par les états qu'elles prennent pour les différents caractères), plus leur degré de parenté est supposé important. On raisonne avant tout par **homologie**, *i.e.*, on suppose que des caractéristiques communes sont héritées d'un ancêtre commun. Par exemple pour un caractère donné, sur la base de l'observation d'états `A` ou `C` selon les espèces, on déduira que le caractère était initialement dans l'état `A` puis est devenu `C` (ou inversement) à une certaine date, pour une certaine lignée ; suivant ce raisonnement, toutes les espèces pour lesquelles on observe l'état `C` (ou inversement l'état `A`) sont supposées appartenir à la lignée ayant subie la substitution, et sont supposées former un groupe séparé des autres espèces dans la phylogénie (cf, Fig. 1.1-(a)).

Au fur et à mesure que les séquences évoluent dans le temps, la probabilité qu'une substitution se produise pour un caractère ayant déjà supporté une substitution, augmente. Quand cet événement se produit, le premier changement d'état peut être occulté (on ne dispose comme donnée que des séquences aux feuilles de

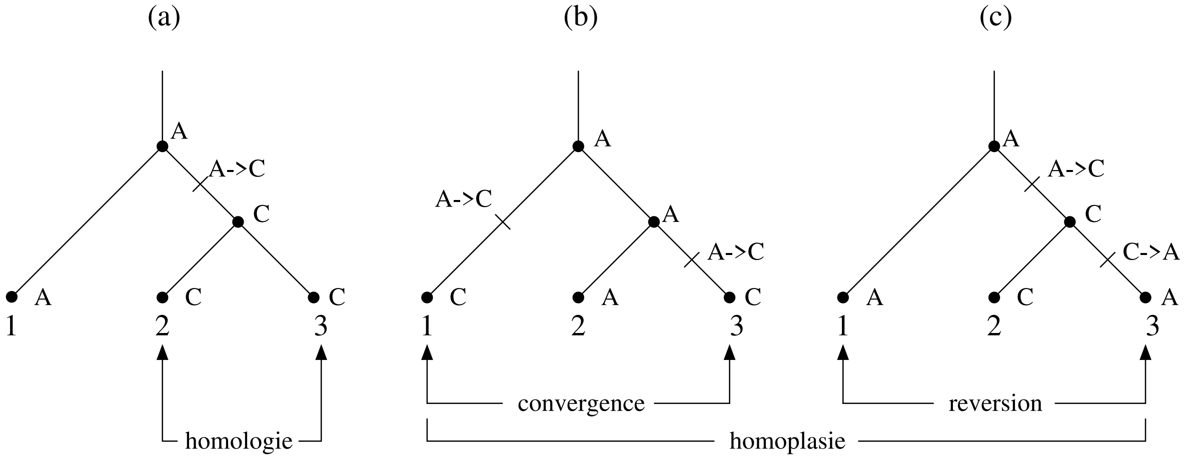


FIG. 1.1 – *Les différentes catégories de ressemblances suite à l'évolution d'un caractère : (a) ressemblance due à l'homologie, (b-c) ressemblance due à l'homoplasie suite à une convergence ou à une réversion (Fig. d'après [DT93]).*

la phylogénie et non d'un historique des changements d'états). On parle dans ce cas d'**homoplasie**, ou encore de *substitution cachée*. Les deux types d'homoplasies existantes sont les *réversions* (Fig. 1.1-(b)) et les *convergences* (Fig. 1.1-(c)).

Les homoplasies contredisent l'hypothèse selon laquelle les similitudes entre espèces indiquent une forte parenté commune, et risquent ainsi de mettre en péril les méthodes de R.P. Par exemple sur la Fig. 1.1-(b), toute méthode raisonnable ne disposant que des états observés aux feuilles de la phylogénie pour le caractère représenté aurait faussement regroupé les espèces 1 et 3. Les homoplasies sont très généralement en minorité par rapport aux homologies mais, même en petite quantité, elles peuvent induire les méthodes en erreur.

1.2 Petite histoire des méthodes de R.P.

Avant les années 60, la R.P. pouvait difficilement être considérée comme une *science* à part entière : les méthodes employées étaient le plus souvent obscures, tandis que le choix et l'interprétation des caractères retenus pour décrire les espèces étaient relativement subjectifs. Tout ceci rendait quasiment impossible la reproductibilité d'une classification par un savant autre que celui qui l'avait produite.

Dans le début des années 60, la connaissance de caractères moléculaires de plus en plus nombreux et peu soumis à la subjectivité, ainsi que l'avènement des premiers ordinateurs, correspondirent au désir des phylogénéticiens de préciser leurs méthodes et de se doter de bases conceptuelles explicites. Désormais, les chercheurs proposaient des méthodes de classification automatique, au sens où une classification devait résulter de l'application d'une certaine **méthode**, caractérisée par un algorithme précis.

De cette clarification méthodologique, émergèrent deux écoles de pensée, la **phénétique** et le **cladisme**, chacune possédant ses propres méthodes pour classer les espèces. Ces deux courants étaient opposés par la **signification** qu'ils donnaient aux classifications des espèces, aussi le débat fut-il essentiellement philosophique et, d'autant plus acharné, que les arguments théoriques étaient en nombre réduit. En pratique, chacune de ces écoles appréhendait de façon différente le concept de similitude, dont nous avons vu précédemment qu'il est à la base de toute classification.

1.2.1 La phénétique

Les principes de la première école furent exposés clairement par Sneath et Sokal dans leur ouvrage "*Principles of Numerical Taxonomy*" [SS63, SS73]. Les **phénéticiens** prônaient qu'une classification ne devait tendre à **regrouper les espèces que dans le but de résumer leurs similitudes et dissimilitudes**. Pour eux il n'était initialement pas question qu'une classification représente les liens de parentés entre espèces, ce n'était qu'accessoirement, en utilisant d'autres critères, que des inférences phylogénétiques pouvaient être tirées des regroupements d'espèces résultant de la classification.

Le concept de ressemblance ne pouvait être utilisé scientifiquement que du point de vue des **similitudes globales** entre espèces, obtenues sur la base d'un nombre aussi large que possible de caractères. Face à la critique selon laquelle les homologies et les homoplasies se retrouvaient ainsi mêlées, les adeptes de cette école répondaient que si on disposait de suffisamment de caractères, la "vraie" ressemblance (homologie) prenait le pas sur la "fausse" ressemblance (homoplasie). Simplement résumée, cette approche se fonde sur l'**analyse quantitative** du plus grand nombre de caractères disponibles suivant le principe «ce qui se ressemble s'assemble».

Construisant des classifications sur la base des similitudes globales, les phénéticiens avaient essentiellement recours à des **méthodes phénétiques**, *i.e.*, utilisant en entrée une matrice de similarités, ou de dissimilarités, entre les espèces deux à deux. Des algorithmes, dits de *cluster analysis*, regroupaient en un **taxon** (une classe) les espèces les plus similaires, puis itéraient le processus jusqu'à avoir obtenue une classification hiérarchique complète des espèces étudiées. Parmi les méthodes les plus célèbres, citons la méthode du lien moyen (U.P.G.M.A.) et ses variantes.

1.2.2 Le cladisme

Les **cladistes** furent opposés aux phénéticiens par l'idée qu'une classification des espèces devait avant tout **élucider les relations de parenté entre espèces** en retrouvant le scénario évolutif selon lequel elles ont divergé au cours du temps. Cette école de pensée fut fondée par Hennig [Hen50, Hen66] et son ouvrage "Phylogenetic systematics".

Pour les cladistes, la similitude globale seule ne constitue pas une base saine en raison des "fausses" similitudes que sont les homoplasies (convergences et réversions). Seule une **analyse qualitative des caractères**, les partitionnant en homologies et homoplasies, peut permettre de reconstruire correctement la phylogénie. Toutefois, le concept même d'homologie doit être raffiné en distinguant les états *primitifs* (plésiomorphes) et les états *dérivés* (apomorphes) des caractères homologues. Pour les cladistes, seul le partage par plusieurs espèces d'états dérivés (les synapomorphies) indique une parenté commune et peut permettre de retrouver les **clades**, *i.e.*, les groupes d'espèces appartenant à une même lignée. Le concept de similitude se trouve donc ici partitioné en trois catégories, dont une seule est vraiment informative.

Les méthodes de reconstruction employées par les cladistes reposent donc sur l'analyse séparée des caractères. Comme ils s'intéressent à la polarité des caractères (distinction entre état primitif et état dérivé), les classifications proposées sont enracinées. Seulement, l'état primitif des caractères n'est pas toujours connu (comme pour les caractères moléculaires), ce qui conduit à envisager tous les états ancestraux possibles pour chaque caractère. Selon l'état ancêtre choisi pour chaque caractère et selon la structure de la classification, le nombre de changements d'état, nécessaires pour expliquer les états observés aux espèces, varie. La solution retenue est celle demandant le moins de changements d'état. Il s'agit de la méthode du **maximum de parcimonie**, consistant à proposer pour la phylogénie des espèces, le scénario évolutif le plus économe. Pour les cladistes, un avantage non-négligeable de la parcimonie (et d'autres méthodes basées sur les caractères) est d'indiquer les états pris par les espèces ancestrales (les noeuds internes de la phylogénie) pour les différents caractères.

1.2.3 La place d'un modèle de l'évolution

La question philosophique du *sens* d'une classifications domina le débat entre phénéticiens et cladistes et finalement, dans les années 70, l'objectif cladiste s'imposa à celui des phénéticiens [EC80, dQG94]. Les systématiciens jugèrent finalement plus important de classer les espèces en fonction de leur histoire évolutive, ne serait-ce qu'en raison des hypothèses qu'on peut dès lors formuler en retour sur le processus de l'évolution (explication et modélisation des substitutions) et en raison du pouvoir prédictif de la classification obtenue. Par exemple, si une nouvelle espèce est découverte, pour laquelle certains caractères ne sont pas encore observés (voire inobservables), on sait que cette espèce a une forte probabilité de posséder le même état pour les caractères inobservés que les espèces du groupe d'espèces auquel on l'aura rattachée depuis ses caractères observables. Signalons aussi que, même si ce n'était pas leur but initial, les phénéticiens ont, eux aussi, essayé d'employer leurs méthodes dans le but de reconstruire la phylogénie des espèces, toutefois les résultats des méthodes de *cluster analysis* se sont généralement révélés d'un faible pouvoir prédictif [SBPH84].

Malgré la suprématie de l'objectif cladiste, le débat ne fut pas clos pour autant, il se transposa aussitôt sur la question d'un **modèle de l'évolution**.

Les cladistes de la première heure ne font l'hypothèse d'aucun modèle de l'évolution particulier. La seule hypothèse explicite sur laquelle reposent leurs analyses phylogénétiques est que “la vie a évolué”, et ce, “suivant un processus divergent” [Hen66, EC80]. Pour eux, l'évolution est décomposée en deux éléments distincts : une **histoire évolutive** (la phylogénie des espèces) et un **processus évolutif** (ensemble de lois qui régit l'évolution en général et les substitutions en particulier) [EC80]. Ainsi, les cladistes ne nient pas le fait qu'il existe un certain processus d'évolution, ni qu'on puisse le modéliser, ils refusent juste d'admettre qu'on puisse en tenir compte pour la construction d'hypothèses phylogénétiques. Pour eux, c'est la phylogénie qui peut permettre de découvrir et caractériser le processus évolutif.

Cette position tranchée s'explique par le fait que le cladisme naquit dans les années 60, époque où la connaissance des molécules était encore à un stade peu avancé, tandis qu'on pouvait difficilement envisager un modèle décrivant l'évolution des caractères morphologiques, imprévisibles et de natures très différentes les uns des autres.

En ce qui concerne les données moléculaires, on sait maintenant clairement qu'il existe de fortes contraintes agissant sur les gènes, selon leur fonction ou leur structure, et conduisant par exemple à des vitesses d'évolution différentes pour des parties différentes de la phylogénie des espèces, ou à un déséquilibre dans la composition des séquences (biais en `G+C`) [LLW85, WLS87, Nei87, MC91].

Depuis la fin des années 70, plusieurs travaux ont montré qu'en raison de telles contraintes, la méthode de parcimonie employée par les cladistes pouvait parfois être inconsistante, au sens où dans certains cas, disposant de plus en plus de caractères (donc d'information), elle peut converger de façon sûre vers une phylogénie incorrecte [Cav78, Fel78a, HP89, DeB92].

Ce constat suggéra à certains phylogénéticiens de poser un modèle de l'évolution, défini de façon explicite, préalablement à toute reconstruction phylogénétique. Un tel modèle permet, dans une certaine mesure, de tenir compte mathématiquement des contraintes évolutives exercées sur les séquences moléculaires, et de corriger en conséquence les distances évolutives observées entre les espèces deux à deux. De tels modèles statistiques de l'évolution se marient donc naturellement avec les méthodes de distances (ainsi qu'avec des méthodes probabilistes, comme nous le verrons plus tard).

Pour cette raison même, certains cladistes refusent toujours d'admettre qu'on puisse tenir compte d'un modèle de l'évolution pour inférer une phylogénie, prétextant que de tels modèles sont de toute façon trop simplificateurs [Far86, CGAS88]. Cependant, un tel avis dérive plus des suites de la longue lutte ayant opposé les cladistes aux phénéticiens (défenseurs des méthode de distances), que de la raison scientifique.

Autant cette position était compréhensible à une époque où les caractères moléculaires étaient moins employés, autant elle est maintenant difficilement défendable. Conscients de l'importance de la prise en compte du processus évolutif pour l'inférence phylogénétique, certains cladistes ont essayé d'incorporer à la méthode de parcimonie des hypothèses sur la nature de l'évolution par le biais des schémas de pondération des caractères. Toutefois cette approche n'a pas donné de résultats satisfaisants pour l'instant [WF90, FY91]. D'autres cladistes, encore, ont essayé de caractériser le modèle de l'évolution sur lequel repose la parcimonie [Far77] ou de justifier cette méthode au sens du maximum de vraisemblance [Sob83, Sob85]. Là aussi, les résultats ne sont pas satisfaisants, puisque le modèle proposé par Farris [Far77] colle de trop près à la méthode de parcimonie pour caractériser son comportement statistique, tandis que les conclusions de Sobber sont erronées, comme il l'a lui-même reconnu plus tard [FS86].

La démarche ci-dessus peut sembler singulière : une méthode est posée (le maximum de parcimonie) et on étudie ensuite les hypothèses qu'elle sous-tend sur l'évolution ; toutefois, cette méthode donnant des résultats satisfaisants en pratique, elle est toujours utilisée. La démarche opposée semble plus justifiable : on pose explicitement un modèle de l'évolution, et si ce modèle est juste, les méthodes de reconstruction raisonnant dans le cadre de ce modèle sont asymptotiquement consistantes, *i.e.*, ont une probabilité de retrouver la phylogénie correcte qui tend vers 1 quand le nombre de caractères tend vers l'infini. Les modèles de l'évolution reposent avant tout sur un mécanisme décrivant la façon dont les substitutions se produisent dans les séquences ; le mécanisme adopté est généralement Markovien (*i.e.*, sans mémoire) et suppose souvent que les substitutions sont "*indépendantes et identiquement distribuées*" (i.i.d.). Autrement dit, les changements d'états se produisant à différents endroits de la séquence et n'importe où dans la phylogénie sont indépendants et tous les sites ont la même probabilité de changer d'état n'importe où dans la phylogénie. Les modèles de l'évolution sont parfois critiqués en raison des hypotèses simplificatrices qu'ils font généralement, car il existe des jeux de données pour lesquels certaines de ces hypotèses ne sont pas raisonnables. Toutefois, ces modèles servent maintenant de base à la plupart des méthodes de distances, ainsi qu'à certaines méthodes probabilistes (reposant en même temps sur la distribution des états observés pour les espèces sur les différents caractères [Hen91]).

1.2.4 La hiérarchie actuelle des méthodes

Le point de vue adopté vis-à-vis du modèle de l'évolution divise donc les méthodes actuelles de R.P. en trois catégories :

- les méthodes cladistes, essentiellement des **variantes de la méthode du maximum de parcimonie**, pour lesquelles aucun modèle explicite de l'évolution n'est posé. Comme le notent Swofford *et al* [SOWH96], ceci ne signifie pas que ces méthodes soient libres de toute hypothèse, par exemple elles nécessitent que les homoplasies soient peu fréquentes en comparaison des homologies.

- les **méthodes de distances**, corrigeant dans un premier temps selon un modèle de l'évolution explicite les distances globales observées entre espèces, puis inférant ensuite une phylogénie suivant un principe algorithmique basé sur les distances. On peut encore parler ici de méthodes "phénétiques" [1], puisque pour les *phylogénéticiens* ce terme recouvre toute méthode basée sur une mesure de (di)similitude globale entre espèces. Toutefois, pour les *généticiens*, ce terme a un sens légèrement différent : il désigne toute méthode basée sur la (di)similitude *observée* entre les espèces et ne recouvre donc pas les méthodes utilisant des distances entre espèces *corrigées* par un modèle de l'évolution. Pour éviter toute confusion, nous utiliserons l'appellation "méthodes de distances".

1. du grec "phéno-" : briller, éclairer.

– les **méthodes probabilistes**, ayant elles aussi recours à un modèle de l'évolution explicite, mais inférant une phylogénie sur la base de l'analyse individuelle des caractères, et non des distances globales entre espèces. Le modèle mathématique de l'évolution permet d'obtenir les probabilités d'occurrence des différentes substitutions (passage d'un certain état à un autre). Ces probabilités permettent d'évaluer la vraisemblance de chaque phylogénie possible pour les espèces en fonction des caractères observés, et de choisir la phylogénie la plus vraisemblable. Cette approche du maximum de vraisemblance fut introduite par Felsenstein [Fel81], puis reprise par d'autres auteurs [Sai88, Sai90], des variantes étant aussi proposées [Hen91, HPS94].

Entre les deux premiers types de méthodes, le débat fait toujours rage, comme en témoigne une série d'articles parus récemment dans la revue *Nature* [Ste93, Sid94, Ste94b, HHS94]. En revanche, le dernier type de méthodes semble actuellement obtenir un consensus relatif au sein du monde des phylogénéticiens, car, comme le remarque Galtier [Gal97], il satisfait d'un côté le courant "modélisateur" en tenant compte d'un modèle de l'évolution, et satisfait d'un autre côté les cladistes en étant basé sur les caractères.

1.3 Modèles mathématiques de l'évolution

Les règles complexes selon lesquelles évoluent les caractères moléculaires sont loin d'être connues, mais on peut toutefois modéliser la dynamique de l'évolution suivant un certain nombre de principes simples. Un modèle de l'évolution décrit l'apparition aléatoire de substitutions entre deux séquences, en caractérisant les probabilités des différents changements d'états.

On pose généralement que les taux de substitutions (nombre de substitutions par unité de temps) des différents changements d'état sont des paramètres spécifiques du modèle. L'ensemble des taux de substitution peut être décrit par une matrice carrée, notée R, où chaque entrée R_{ij} décrit le taux auquel un état i se transforme en un état j. Les modèles de l'évolution font généralement un certain nombre d'hypothèses communes [CGAL95] :

– *Les sites évoluent selon un processus de Markov homogène :* l'évolution est vue comme un processus sans mémoire, qui peut être caractérisé en décrivant pour un site quelconque seulement le passage de l'état courant à l'état suivant dans le temps. L'hypothèse d'homogénéité indique que les taux de substitutions sont constants dans le temps, et donc identiques sur toutes les arêtes de la phylogénie. Ceci implique que la probabilité d'un certain changement d'état (p. ex. A→C) entre deux séquences ne dépende que de la durée de l'intervalle de temps qui les sépare, et non de la position de cet intervalle dans le temps. Il s'agit d'une hypothèse forte, signifiant que les différentes lignées de la phylogénie évoluent à la même vitesse.

– *Les substitutions sont indépendantes et identiquement distribuées sur les sites (i.i.d.)*: les substitutions affectant un site ne dépendent pas de sa position dans la séquence, ni des substitutions affectant les autres sites. Ceci permet de poser que la probabilité qu'une séquence se transforme en une autre séquence au cours d'un certain intervalle de temps, est le produit de la probabilité du changement d'état correspondant pour chacun des sites. De plus, tous les sites suivent un processus d'évolution identique, *i.e.*, ont la même probabilité de changer d'état au cours d'un intervalle de temps t.
– *Le processus évolutif est réversible*: le sens de l'évolution est indifférent, la probabilité pour un site de passer d'un état i à un état j dans un intervalle de temps t est identique à sa probabilité de passer d'un état j à un état i dans le même laps de temps.
– *La probabilité d'une substitution durant un intervalle de temps infinitésimal, dt, est proportionnelle à la durée.* La constante de proportionnalité est le taux de substitution déjà mentionné précédemment. On peut ainsi écrire $dP = Rdt$. On pose aussi que la probabilité que deux substitutions aient lieu durant l'intervalle dt, est négligeable vis-à-vis de la probabilité qu'il y en ait au plus une.

Les différents modèles issus de ces hypothèses se distinguent généralement par les paramètres auxquels ils ont recours pour décrire la matrice R des taux de substitutions.

1.3.1 Les premiers modèles de l'évolution

Pour les séquences nucléotidiques, le premier modèle reconnu fut celui de Jukes et Cantor [JC69]. Il s'agit du modèle le plus simple puisqu'il n'utilise qu'un seul paramètre, noté α. La matrice des taux de substitution associée à ce modèle est :

$$R = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}$$

où les entrées horizontales et verticales de la matrice sont données dans l'ordre A, G, C, T. Cette matrice exprime simplement que tous les changements d'état sont équiprobables. Par convention, les entrées R_{ii} sont fixées de façon à ce que $\sum_j R_{ij} = 0$.

Un modèle plus réaliste fut ensuite proposé par Kimura [Kim80] : ce modèle permet de tenir compte de deux types différents de substitutions, les *transitions* (changement entre A et G ou entre C et T), dont le taux est noté α, et les *transversions* (les autres changements), dont le taux est noté β et que l'on sait moins fréquentes ($\alpha > \beta$). La matrice des taux de substitution est donc la suivante :

$$R = \begin{pmatrix} -\alpha - 2\beta & \alpha & \beta & \beta \\ \alpha & -\alpha - 2\beta & \beta & \beta \\ \beta & \beta & -\alpha - 2\beta & \alpha \\ \beta & \beta & \alpha & -\alpha - 2\beta \end{pmatrix}.$$

Notons que, pour $\alpha = \beta$, on retrouve le modèle précédent.

Les taux de la matrice R sont définis pour un intervalle de temps infinitésimal, dt. Pour utiliser ces taux sur un intervalle de temps t fini, afin d'obtenir la matrice P_t des probabilités de substitution en un site donné au cours de cet intervalle de temps, il faut résoudre une équation différentielle simple. À l'issue de cette résolution, on obtient $P_t = e^{Rt}$ [KH90]. Par exemple, la probabilité $p_{ij}(t)$ que le nucléotide i se transforme en nucléotide j, $j \neq i$, après une période de temps de longueur t pour le modèle de Jukes et Cantor est :

$$p_{ij}(t) = \frac{1}{4}(1 - e^{-4\alpha t})$$

tandis que pour le modèle de Kimura à deux paramètres, cette probabilité est :

$$p_{ij}(t) = \tfrac{1}{4}(1 - 2e^{-2(\alpha+\beta)t} + e^{-4\beta t}) \quad \text{si } ij \text{ est une différence de type transition,}$$

$$p_{ij}(t) = \tfrac{1}{4}(1 - e^{-4\beta t}) \quad \text{si } ij \text{ est une différence de type transversion.}$$

Lorsque deux séquences ont divergé depuis un temps t d'une espèce ancestrale commune, on peut considérer qu'elles sont séparées par un temps $2t$ (le processus de l'évolution est supposé réversible). On en déduit la probabilité $p_{xy}(t)$ d'observer pour un site donné deux états différents dans les deux séquences x et y, par exemple pour le modèle de Jukes et Cantor :

$$p_{xy}(t) \quad = \quad \frac{3}{4}(1 - e^{-8\alpha t}) \tag{1.1}$$

et pour le modèle de Kimura à deux paramètres :

$$p_{xy}(t) \;=\; \frac{1}{4}(3 - 2e^{-4(\alpha+\beta)t} - e^{-8\beta t}) \; . \qquad (1.2)$$

Depuis les modèles de Jukes et Cantor [JC69] et de Kimura [Kim80], de nombreux autres modèles plus sophistiqués ont été proposés. Certain permettent de s'accommoder de taux de substitutions différents selon les sites (loi gamma [JN90]), tandis que d'autres permettent de tenir compte des insertions et délétions de nucléotides dans les séquences [TN84].

1.3.2 Le modèle de l'évolution de Neyman-Cavender-Farris

Un autre modèle de l'évolution bien connu, sur des caractères binaires et non plus des nucléotides, est celui proposé indépendamment par Neyman [Ney71], Cavender [Cav78] et Farris [Far73]. La simplicité de ce modèle en fait l'outil idéal pour démontrer des propriétés statistiques sur les méthodes de R.P., propriétés, qui sont ensuite étendues à d'autres modèles de l'évolution (cf. section 2.3).

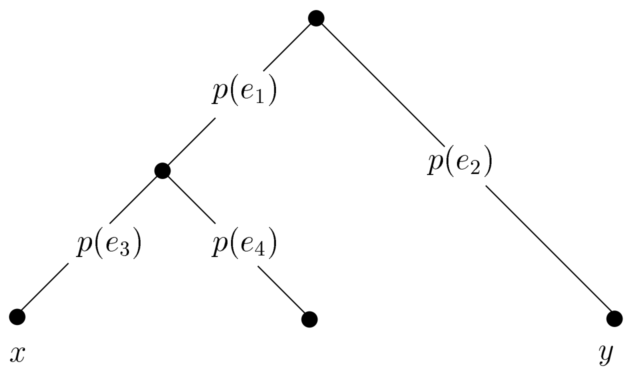


FIG. 1.2 – *Phylogénie dont les séquences évoluent suivant le modèle NCF. À chaque arête e_i est associée une probabilité $p(e_i)$ qu'une substitution affecte tout caractère entre les deux extrémités de l'arête.*

Comme les modèles décrits précédemment, le modèle *NCF* (pour *Neyman-Cavender-Farris*) fait l'hypothèse que les différents caractères des séquences évoluent de façons identique et indépendante (i.i.d.) le long d'une phylogénie enracinée, suivant un processus de Markov. La différence est qu'il considère l'évolution de caractères binaires, *i.e.*, pouvant prendre 0 ou 1 comme état.

Une séquence d'états est associée à chaque noeud de la phylogénie. Les différences entre ces états d'une espèce sur l'autre traduisent l'évolution subie par les caractères. Les caractères évoluent depuis la racine de la phylogénie (où les deux états 0 et 1 sont équiprobables) jusqu'aux feuilles de la phylogénie. À chacune des arêtes e de la phylogénie est associée une matrice de transition $P(e)$ décrivant la probabilité d'observer deux états différents (*i.e.*, la probabilité d'un nombre impair de substitutions) aux deux extrémités de l'arête, et ce pour n'importe quel caractère : $P(e)_{01} = P(e)_{10} = p(e)$ t.q. $0 < p(e) < 0,5$. La probabilité $p(e)$ est fonction du taux d'évolution le long de l'arête e et de la longueur t de cette arête dans le temps, toutefois, contrairement aux modèles précédents, on ne s'intéresse pas ici à la décomposition de $p(e)$ en ces deux quantités.

Un résultat bien connu pour ce modèle est qu'on peut facilement obtenir la probabilité p_{xy} de constater deux états différents pour un caractère aux deux extrémités d'un chemin $[xy]$ dans la phylogénie, depuis les probabilités de différence observée $p_1, p_2, \dots, p_k$ associées aux arêtes $e_1, \dots, e_k$ de ce chemin [FK96] :

$$p_{xy} = \frac{1}{2}\left(1 - \prod_{i=1}^{k}(1 - 2p_i)\right) \tag{1.3}$$

1.3.3 Correction des distances observées

Depuis les résultats précédents, on peut caractériser l'évolution de plusieurs séquences le long des différentes arêtes d'une phylogénie, et déduire une estimation $\hat{D}_{xy}$ de la **distance évolutive** D_{xy} entre deux espèces quelconques x et y aux feuilles de la phylogénie, c'est-à-dire une estimation du nombre de substitutions moyen par site séparant les deux séquences. Si pour tout couple d'espèces x et y, l'estimation $\hat{D}_{xy}$ de D_{xy} obtenue sur la base d'un certain modèle de l'évolution est suffisamment précise, alors la matrice de distances $\hat{D}$ permet de reconstruire sans ambiguïté la phylogénie correcte (comme nous le verrons à la section 1.5.6). Mais il nous faut d'abord décrire comment obtenir $\hat{D}$ sur la base d'un modèle de l'évolution et des séquences de caractères décrivant les espèces.

Supposons que nous disposons d'un jeu de données X constitué de n séquences de k caractères. Pour tout couple d'espèces x et y, si $H(x, y)$ dénote la distance de Hamming entre les séquences de x et y[2], la quantité $f_{xy} = H(x, y)/k$ dénote la **distance évolutive observée** entre les deux espèces x et y. Autrement dit, f_{xy} est la proportion de sites n'ayant pas le même état dans les deux séquences, aussi en espérance, $p_{xy}(t) = E(f_{xy})$, où t est le temps d'évolution séparant les

2. *i.e.*, le nombre de caractères pour lesquels x et y ont des états différents.

deux séquences. La distance observée f_{xy} sous-estime la vraie distance évolutive D_{xy} entre les espèces x et y, puisque elle ne prend pas en compte les substitutions multiples (convergences, parallélismes, etc) s'étant éventuellement produites entre les deux espèces sur un même site. Toutefois, en ayant recours à un modèle de l'évolution, on peut **corriger** la distance observée entre deux espèces afin d'obtenir une estimation de leur distance évolutive.

Par exemple, pour le modèle de Jukes et Cantor, si x et y sont deux espèces ayant divergé d'une même espèce ancestrale depuis un temps t, de par la définition de R, chaque site a subi un taux de substitution de 3α entre la séquence ancestrale et chacune des séquences associées aux espèces. Ainsi l'espérance du nombre de substitutions par site entre x et y est $D_{xy} = 2 \times 3\alpha t$. En utilisant la formule (1.1), on obtient facilement pour le modèle de Jukes et Cantor :

$$D_{xy} = -\frac{3}{4}\ln(1 - \frac{4}{3}p_{xy}(t)) \ ,$$

et en estimant $p_{xy}(t)$ par f_{xy}, la distance observée entre x et y, on obtient l'estimation $\hat{D}_{xy}$ de D_{xy} :

$$\hat{D}_{xy} \quad = \quad -\frac{3}{4}\ln(1 - \frac{4}{3}f_{xy}) \ . \tag{1.4}$$

De même, en réinterprétant le modèle NCF comme un processus poissonien (et en considérant que les taux d'évolution sont identiques sur toutes les arêtes de la phylogénie), on obtient la formule suivante de correction des données :

$$\hat{D}_{xy}(t) \quad = \quad -\frac{1}{2}\ln(1 - 2f_{xy}) \ . \tag{1.5}$$

Pour le modèle de Kimura à deux paramètres, on utilise, non pas f_{xy}, mais s_{xy}, la proportion de transitions observées entre les séquences de x et y, ainsi que v_{xy}, la proportion de transversions observées. Par le même raisonnement que ci-dessus, on obtient depuis (1.2) :

$$\hat{D}_{xy} \quad = \quad -\frac{1}{2}\ln\left((1 - 2s_{xy} - v_{xy})\sqrt{1 - 2v_{xy}}\right) \ . \tag{1.6}$$

Ainsi l'expression analytique utilisée pour corriger les distances observées dépend du modèle de l'évolution adopté. Rappelons que si le modèle employé est juste, lorsque le nombre de caractères disponibles pour les séquences augmente, la distance estimée entre deux espèces tend vers leur véritable distance évolutive.

Dans ce cas, toute méthode de R.P. raisonnable, basée sur ces distances corrigées, retrouve la phylogénie correcte avec une probabilité qui tend vers 1. Ceci est dû au fait que la distance évolutive induite par une phylogénie entre les espèces la caractérise de façon unique (cette distance ne peut être réalisée sur aucune autre phylogénie, cf. section 1.5.6), et que la plupart des méthodes retrouvent la phylogénie correspondant à une telle distance. Pour conclure cette section, signalons aussi la méthode de correction des distances dite du *logdet*, que Steel [Ste94a] a montrée consistante pour la plupart des modèles de l'évolution, sous certaines hypothèses raisonnables. Pour un exposé plus détaillé sur les modèles d'évolution, le lecteur pourra se reporter aux ouvrages : [SOWH96, And97, Gal97].

1.4 La R.P. vue comme une procédure d'estimation statistique

Dans cette thèse, nous adopterons la plupart du temps un point de vue propre à l'estimation statistique. La R.P. peut en effet être vue comme une procédure d'estimation, ainsi Swofford *et al* ([SO90],p. 412) écrivent :

> "(When inferring a phylogeny...) we are making a "best estimate" of an evolutionary history based on incomplete information. In the context of molecular systematics, we generally do not have direct information about the past - we have access only to contemporary species and molecules. Because we can postulate evolutionary scenarios by which any chosen phylogeny could have produced the oberved data, we must have some basis for selecting one or more preferred trees from among the set of possible phylogenies"

Cette base, ce moyen par lequel nous pouvons sélectionner une ou plusieurs phylogénies, est une méthode de reconstruction. Les différentes méthodes sont autant d'*estimateurs* disponibles pour reconstruire la *phylogénie inconnue* des espèces. Si l'on note t cette phylogénie, la phylogénie inféré par toute méthode de reconstruction est vu comme une *estimation*, $\hat{t}$, de t.

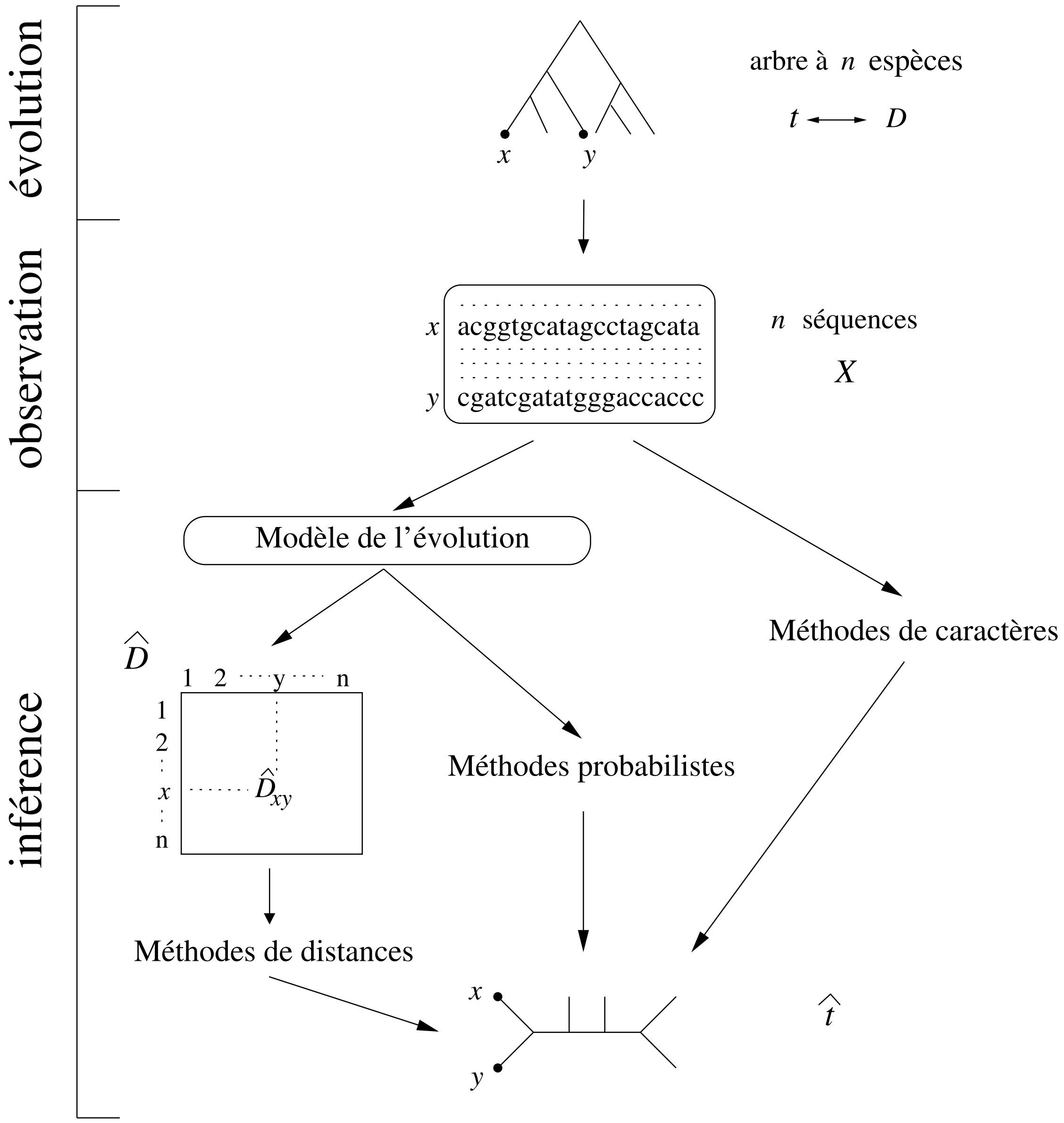


FIG. 1.3 – *La reconstruction phylogénétique vue comme une procédure d'estimation statistique.*

La Fig. 1.3 présente la reconstruction phylogénétique selon un point de vue lié à l'estimation statistique. L'histoire évolutive d'un ensemble E de n espèces y est représentée par une phylogénie enracinée t dont les feuilles correspondent aux espèces étudiées (les espèces contemporaines) et les noeuds internes aux espèces ancestrales hypothétiques. À chaque arête de cette phylogénie, on peut attribuer une valeur, ou une *longueur*, représentant une distance dans le temps (p. ex. en dizaines de milliers d'années) ou une distance évolutive (nombre de substitutions par unité de temps). Ainsi, entre tout couple d'espèces x et y, la phylogénie t induit une *distance* D_{xy} égale à la somme des longueurs des arêtes sur le chemin entre x et y. Nous noterons $D = (D_{xy})$ la matrice de distances $n \times n$ obtenue entre les espèces suivant la phylogénie t.

Le point de départ de la R.P. est la description des espèces par un certain nombre k d'états qu'elles prennent pour des caractères homologues (p. ex. les sites issus de leur gêne respectif de la myoglobine, préalablement alignés afin que les états relevés séparément pour les différentes espèces correspondent vraiment aux mêmes sites initiaux). Les séquences observées pour les espèces sont représentées sous la forme d'une matrice X dont les lignes sont associées bijectivement aux espèces étudiées et les colonnes aux caractères (p. ex. des sites moléculaires). Chaque entrée X_{xc} de cette matrice décrit l'état pris par l'espèce x pour le caractère c.

La matrice X peut être exploitée directement pour produire une phylogénie estimée $\hat{t}$ par l'intermédiaire d'une méthode basée sur les caractères, comme nous le verrons dans le chapitre suivant. Une autre approche consiste à poser explicitement un modèle de l'évolution, utilisé en conjonction avec la matrice X soit par des méthodes probabilistes [Fel81, Hen91] (raisonnant sur la distribution des états observés), soit par des méthodes de distances obtenant depuis X une matrice $\hat{D} = (\hat{D}_{xy})$ de distances entre espèces. La distance $\hat{D}_{xy}$ entre les espèces x et y est obtenue sur la base des similitudes et des dissimilitudes observées entre leurs descriptions dans la matrice X et en fonction du modèle de l'évolution choisi. Ce modèle permet de corriger les distances observées (depuis X) afin de tenir compte de substitutions muettes (les homoplasies), que X ne traduit pas. Depuis la matrice $\hat{D}$, diverses méthodes de reconstruction basées sur les distances peuvent être utilisées pour obtenir une estimation $\hat{t}$ de la phylogénie t. La plupart de ces méthodes essayent d'approximer $\hat{D}$ au plus près (par la matrice de distances induite par la phylogénie $\hat{t}$), sur la base du fait que $\hat{D}$ est un estimateur consistant de D, si le modèle de l'évolution choisi est correct.

Notre préoccupation principale dans cette thèse est de retrouver les différents groupes selon lesquels se répartissent les espèces étudiées, sans précision des dates auxquelles ces différents groupes se sont séparés, ni des quantités d'évolution les séparant. Autrement dit, nous nous intéressons avant tout à la *structure* de t, *i.e.*, au schéma de branchement suivant lequel les espèces se sont séparées. En ce sens, nous adoptons une *approche discrète de la R.P.*, cherchant simplement à savoir quelles sont les arêtes de la phylogénie inconnue, sans souci de leurs longueurs respectives. Nous délaissons l'évaluation des longueurs d'arêtes, non seulement parce qu'une fois la structure de la phylogénie fixée, des méthodes efficaces existent pour fixer les longueurs de ses arêtes en fonction des distances entre espèces, mais aussi parce que la mesure exacte de ces distances dépend du contexte (p. ex. du type ou de la provenance des caractères initiaux) ou encore du modèle de l'évolution utilisé pour corriger les distances observées.

La qualité d'une estimation $\hat{t}$ est mesurée selon sa capacité à retrouver les arêtes de la phylogénie inconnue, t, et à ne pas proposer d'arêtes incorrectes. Selon le point de vue de l'estimation statistique, on distingue pour toute estimation $\hat{t}$ de t une erreur de 1ère espèce (les arêtes incorrectes inférées par $\hat{t}$, aussi appelées "*false positive*") et une erreur de 2ème espèce (les arêtes de t non inférées par $\hat{t}$, aussi appelées "*false negative*"). Comme en R.P. nous ne connaissons généralement pas de façon certaine la phylogénie correcte, il est très difficile de déterminer objectivement quelles arêtes inférées sont incorrectes et doivent être rejetées, et quelles arêtes peuvent être considérées comme des hypothèses fortement probables, retraçant l'évolution des espèces. C'est pourquoi on a souvent recours en R.P. à des études comparant les différentes méthodes sur la bases de simulations (où la phylogénie t est connue), la *meilleure* méthode est alors celle qui montre le taux d'erreur le plus faible.

1.5 Définitions préliminaires

Nous définissons ici plusieurs concepts que nous rencontrerons fréquemment tout au long de cette thèse. Certaines des notions évoquées ci-dessous font appel à la théorie des graphes pour laquelle nous renvoyons le lecteur aux ouvrages : [Ber70, AU87, CLR94].

1.5.1 Phylogénies

Définition 1 *Une* **phylogénie** P *pour un ensemble d'espèces* $E = \{1, 2, \ldots, n\}$ *est un arbre* $P = (S, A)$ *sans sommet de degré 2 et dont les feuilles sont bijectivement associées aux espèces de* E.

Les feuilles de la phylogénie, représentant généralement des espèces contemporaines, sont aussi appelées **noeuds externes**. Les autres noeuds de l'arbre, aussi appelés **noeuds internes**, correspondent aux espèces ancestrales hypothétiques des espèces étudiées. On distingue aussi deux types d'arêtes : les **arêtes externes**

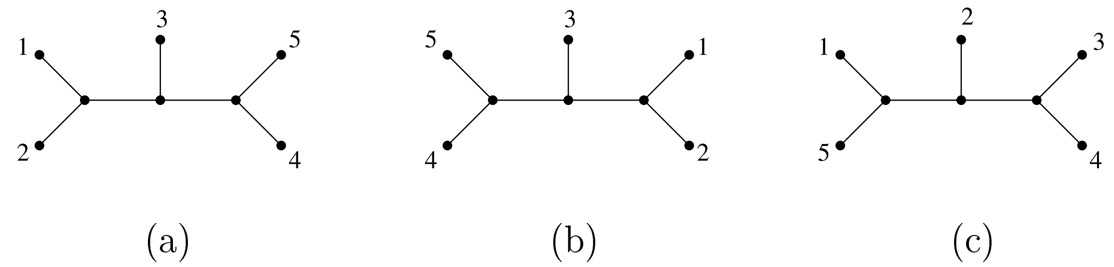


FIG. 1.4 – *Exemple de phylogénies. Les arbres (a) et (b) représentent la même phylogénie, différente de la phylogénie (c).*

connectant les feuilles au reste de la phylogénie, et les **arêtes internes** connectant deux noeuds internes. Notons que deux phylogénies sur E ayant le même arbre pour support, mais telles que les espèces de E sont réparties de façon différente entre les feuilles, ne sont pas forcément identiques. Par exemple, sur la Fig. 1.4, les phylogénies (a) et (b) sont identiques, mais différentes de la phylogénie (c).

1.5.2 Le problème de la racine

La définition d'une phylogénie que nous donnons ci-dessus ne prend pas en compte l'existence d'une racine, puisque la plupart des méthodes de reconstruction, ne connaissant pas la polarité des caractères (quel état est dérivé? quel état est ancestral?), proposent des phylogénies non-enracinées. Il est toutefois facile de passer du formalisme de phylogénie non-enracinée à celui de phylogénie enracinée. Ainsi une **phylogénie enracinée** admet-elle la même définition qu'une phylogénie, à la nuance près qu'elle contient un sommet distingué, r, de degré 2, appelé **racine** de la phylogénie. Obtenir une phylogénie enracinée depuis une phylogénie usuelle P est une opération simple (cf. Fig. 1.5) : il suffit d'insérer un noeud r sur n'importe quelle arête. Cette arête est ainsi scindée en deux et le noeud r est le seul de la phylogénie à posséder un degré 2. Depuis une phylogénie P on peut ainsi obtenir autant de phylogénies enracinées que P possède d'arêtes. Cependant, une phylogénie enracinée correspond à une seule phylogénie non-enracinée. Le problème du positionnement de la racine dans une phylogénie est parfois résolu en joignant au groupe d'espèces étudiées une espèce de nature différente, nommée *outgroup* (p. ex. un rongeur si on étudie les hominidés). Après avoir inféré une phylogénie non-enracinée par une méthode usuelle, on déduit que la racine est située sur l'arête reliant l'*outgroup* au reste de la phylogénie (p. ex. sur la phylogénie (B) de la fig 1.5, si l'outgroup est l'espèce 5, on déduit que la phylogénie (C) représente la phylogénie enracinée des espèces). Toutefois, l'ajout d'un outgroup aux espèces étudiées peut aussi rendre la reconstruction bien plus difficile (p. ex. augmentation de la distance évolutive maximum, ...).

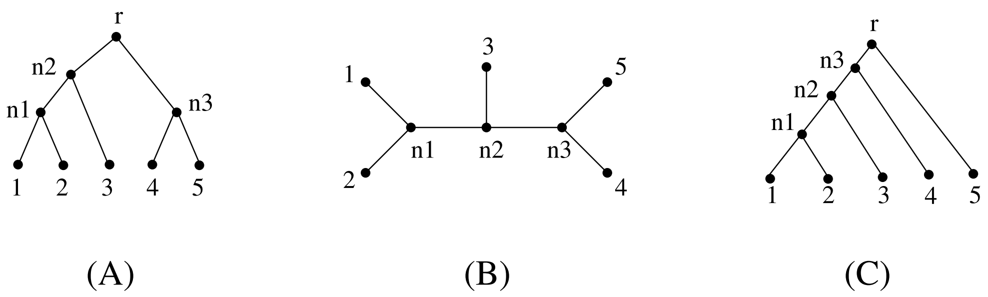


FIG. 1.5 – *Enracinement d'une phylogénie. La phylogénie enracinée (A), resp. (C), est obtenue en insérant une racine sur l'arête* (n_2, n_3)*, resp.* $(n_3, 5)$ *de la phylogénie (B).*

1.5.3 Degré de résolution

Les phylogénies $P = (S, A)$ que nous cherchons à reconstruire sont supposées **binaires** (*i.e.*, $\forall x \in S$, $d(x) = 1$ ou $d(x) = 3$)[3], conformément à l'hypothèse généralement admise que les phénomènes de multispéciations (un ancêtre donnant naissance à plus de deux descendants à la fois) sont improbables. Toutefois, nous ne nous interdisons pas de proposer comme estimation de ces phylogénies, des phylogénies ayant des noeuds internes de degré plus grand que 3. De tels noeuds expriment l'incertitude quant aux relations liant certains groupes d'espèces, et ne signifient en aucun cas qu'un phénomène de multispéciation soit envisagé. Tout noeud de degré plus grand que 3 est qualifié d'**irrésolu**, étant donné qu'il ne tranche pas entre les différentes résolutions possibles pour les groupes d'espèces qui lui sont reliés. Une résolution de ce noeud est obtenue en le remplaçant par 2 noeuds de moindre degré reliés entre eux et se partageant son voisinage, puis en répétant cette opération jusqu'à ne disposer que de noeuds de degré 3 (p. ex. sur la Fig. 1.6, la phylogénie (c) est obtenue en résolvant le noeud u de la phylogénie (b)). Si une phylogénie P est obtenue en résolvant un ou plusieurs noeuds irrésolus d'une phylogénie P', on dit que P est un **raffinement** de P', et P' une **contraction** de P. Tout noeud de degré 3 est dit **résolu**, car il ne peut pas être subdivisé

3. $d(x)$ dénote le degré du noeud x.

en deux noeuds de degré 3. Une phylogénie dont tous les noeuds internes sont résolus est dite **totalement résolue** (ou binaire) (p. ex. Fig. 1.6(c)), autrement elle est dite **partiellement résolue** (p. ex. Fig. 1.6(a,b)). La phylogénie ne contenant qu'un seul noeud interne est nommée phylogénie **étoile** en raison de sa forme (Fig. 1.6(a)) : un seul noeud interne auquel sont directement reliées toutes les feuilles. Il s'agit de la phylogénie la moins résolue qu'on puisse proposer sur les espèces.

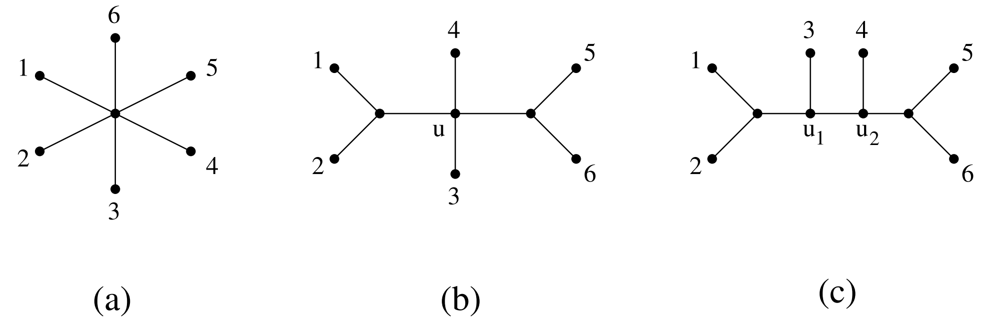


FIG. 1.6 – *Trois exemples de* phylogénies *sur les espèces* $1, \ldots, 6$ *: (a) la phylogénie étoile, (b) une autre phylogénie* partiellement *résolue, (c) une phylogénie* totalement *résolue correspondant à une résolution possible du noeud u de la phylogénie précédente.*

1.5.4 Bipartitions

Soient $P = (S, A)$ une phylogénie et $e \in A$ une de ses arêtes. Supprimer l'arête e sépare P en deux composantes connexes. Soient σ et $\overline{\sigma}$ les sous-ensembles d'espèces appartenant respectivement à ces deux composantes. $\sigma|\overline{\sigma}$ constitue une **bipartition** (parfois appelée *split*) sur les espèces de E. On dit que la bipartition $\sigma|\overline{\sigma}$ est **induite** par l'arête e, ou par la phylogénie P. Par exemple, l'arête (u_1, u_2) de la phylogénie (c) de la Fig. 1.6 induit la bipartition $\{1, 2, 3\}|\{4, 5, 6\}$. Tout au long de cette thèse, c'est cette information structurelle que nous retiendrons pour une arête. Si $|\sigma| < 2$ ou $|\overline{\sigma}| < 2$, alors la bipartition $\sigma|\overline{\sigma}$ est dite **triviale**, sinon elle est **non-triviale**. Les ensembles σ et $\overline{\sigma}$ sont appelés les **composantes** de la bipartition. Dans le cas de phylogénies enracinées, les composantes non-triviales correspondant à des sous-arbres n'incluant pas la racine sont aussi appelées **clades** (p. ex. les clades de la phylogénie (A) Fig. 1.5 sont $\{1, 2\}$, $\{1, 2, 3\}$ et $\{4, 5\}$). Les bipartitions triviales induites par une phylogénie correspondent à ses arêtes externes et les bipartitions non-triviales à ses arêtes internes, d'où le fait que toutes les phylogénies sur le même ensemble d'espèces induisent les mêmes bipartitions triviales. Pour cette raison, on s'intéresse uniquement aux bipartitions non-triviales induites par une phylogénie, et on compare les méthodes de reconstruction sur leur capacité à retrouver les arêtes internes d'une phylogénie.

Un résultat fondamental, montré par Buneman [Bun71], est que toute phylogénie est désignée de façon unique par l'ensemble des bipartitions (non-triviales) qu'elle induit sur les espèces. Ainsi l'ensemble des bipartitions d'une phylogénie constitue en quelque sorte un codage de sa structure. C'est la caractérisation que nous retiendrons avant tout pour une phylogénie puisque, dans cette thèse, nous nous intéressons principalement à l'aspect structurel des phylogénies. Dans la suite, nous noterons $B(P)$ l'ensemble des bipartitions induites par une phylogénie P.

Si toute phylogénie correspond à un ensemble de bipartitions, tout ensemble de bipartitions n'est pas forcément représentable par une phylogénie : un ensemble de bipartitions est dit **compatible** si et seulement si il existe une phylogénie t telle que, chaque bipartition est induite par une arête de t. Le résultat suivant est célèbre :

Théorème 1 *[LeQ69, Bun71, EJM76]*

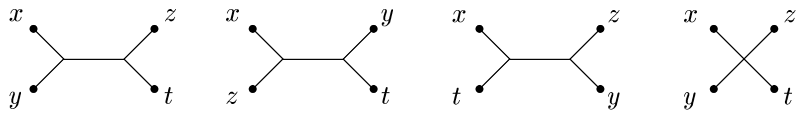
- *Deux bipartitions $\sigma_1|\overline{\sigma}_1$ et $\sigma_2|\overline{\sigma}_2$ sont* **compatibles** *ssi au-moins l'une des intersections $\sigma_1 \cap \sigma_2$, $\sigma_1 \cap \overline{\sigma}_2$, $\overline{\sigma}_1 \cap \sigma_2$, $\overline{\sigma}_1 \cap \overline{\sigma}_2$ est vide.*

- *Un ensemble de bipartitions est compatible ssi ses bipartitions sont compatibles deux à deux.*

Ainsi, déterminer si un ensemble B de bipartitions sur n espèces est compatible, demande au plus $O(|B|^2 n)$ (voir Meacham [Mea81] et Gusfield [Gus91a] pour des algorithmes linéaires, *i.e.*, en $O(|B|.n)$).

1.5.5 Quadruplets et r4-arbres

Soit E un ensemble d'espèces étudiées, un **quadruplet** sur E désigne tout groupe de 4 espèces distinctes de E. Un **r4-arbre** est une phylogénie totalement résolue sur un quadruplet d'espèces de E. Tout quadruplet d'espèces x, y, z, t peut être **résolu** de trois façons différentes (cf. Fig. 1.7, phylogénies (a), (b) et (c)). Ces trois r4-arbres sont notés $xy|zt$, $xz|yt$ et $xt|yz$, indiquant de quelle façon l'arête centrale sépare les 4 espèces du quadruplet en deux paires. L'information que nous retenons pour un r4-arbre est précisément la partition en deux paires qu'il induit sur quatre espèces, aussi du point de vue des notations $xy|zt \equiv yx|zt \equiv zt|yx$. La



(1) r4-arbre $xy|zt$ (2) r4-arbre $xz|yt$ (3) r4-arbre $xt|yz$ (4) irrésolue

FIG. 1.7 – *Phylogénies possibles pour un quadruplet d'espèces.*

phylogénie étoile sur quatre espèces (Fig. 1.7, phylogénie (d)) n'est pas un r4-arbre puisqu'elle possède un noeud irrésolu.

Soit E un ensemble d'espèces, les fonctions $\mathcal{R}$ et $\mathcal{Q}$ nous permettent de passer de la notion de r4-arbre à celle de quadruplet :

- Soit $q = x, y, z, t$ un quadruplet sur E, $\mathcal{R}(q)$ désigne l'ensemble des trois r4-arbres possibles pour q, *i.e.*, $\{xy|zt, xz|yt, xt|yz\}$.

- Soit $r = xy|zt$ un r4-arbre sur E, $\mathcal{Q}(r)$ désigne le quadruplet d'espèces concerné par r, *i.e.*, x, y, z, t.

En R.P. (comme dans d'autres domaines), les ensembles Q de r4-arbres considérés contiennent au plus un r4-arbre par quadruplet, *i.e.*, $\forall\, r, r' \in Q,\ \mathcal{Q}(r) \neq \mathcal{Q}(r')$. Un ensemble de r4-arbres Q est dit **complet** pour E s'il contient un r4-arbre pour chaque quadruplet d'espèces de E, sinon il est dit **incomplet**.

Tout ensemble Q de r4-arbres peut être utilisé pour construire une phylogénie P sur l'ensemble des espèces. Les r4-arbres de Q sont alors vus comme autant de contraintes structurelles que P doit respecter. Face à un ensemble Q incomplet, deux solutions sont possibles pour les quadruplets non-représentés dans Q : soit on considère que l'absence de décision pour ces quadruplets ne porte pas à conséquence et qu'ils peuvent être résolus de n'importe quelle façon dans la phylogénie proposée ; soit on considère qu'il est préférable de ne pas les résoudre, *i.e.*, que l'absence de décision représente la contrainte correspondant à la phylogénie étoile (Fig. 1.7-(d)). C'est l'interprétation la plus raisonnable quand l'erreur de la phylogénie inféré pour estimer la phylogénie réelle des espèces est évaluée sur la base des r4-arbres qu'il résout différemment de Q. Dans un tel cas, la non-résolution des quadruplets concernés semble en effet la meilleure solution, puisqu'en proposant une quelconque résolution au hasard, on aurait 2 chances sur 3 (le nombre de r4-arbres) de se tromper.

Toute phylogénie P sur les espèces E peut être caractérisée par un ensemble de r4-arbres : soit $x, y, z, t \in E$, on dit que P **induit** le r4-arbre $xy|zt$ ssi les chemins

$[xy]$ et $[zt]$ ont une intersection vide dans P. En effet, dans ce cas, le sous-arbre de P restreint aux espèces x, y, z, t, correspond exactement au r4-arbre $xy|zt$, puisque tous deux possèdent une arête partitionnant de la même façon les quatres espèces en deux paires, x, y et z, t (p. ex. la phylogénie (b) de la Fig. 1.5 induit les r4-arbres $12|34, 12|35, 12|45, 13|45, 23|45$). On dit aussi dans ce cas que P **satisfait** le r4-arbre $xy|zt$. À l'inverse, il **contredit** les r4-arbres $xz|yt$ et $xt|yz$, car aucune arête de P ne sépare x, z de y, t et x, t de y, z. L'ensemble des r4-arbres induits par une phylogénie P est noté Q_P. Remarquons que si P est une phylogénie binaire (totalement résolue), l'ensemble Q_P est complet.

Nous avons vu précédemment qu'une phylogénie peut aussi être caractérisée par l'ensemble de bipartitions non-triviales qu'elle induit sur les espèces de E, selon ses arêtes internes. À toute bipartition non-triviales $b = \sigma|\overline{\sigma}$ d'espèces de E nous pouvons associer l'ensemble de r4-arbres $Q_b = \{xy|zt \text{ t.q. } x, y \in \sigma \text{ et } z, t \in \overline{\sigma}\}$ qu'elle **induit**. On voit donc clairement qu'il y a une relation naturelle entre les concepts de r4-arbres, de bipartition et de phylogénie : toute phylogénie peut être considéré comme un ensemble de bipartitions et toute bipartition comme un ensemble de r4-arbres. La réciproque n'est pas toujours vérifiée, tout ensemble de r4-arbres ne correspond pas forcément à un ensemble de bipartitions et tout ensemble de bipartitions ne correspond pas forcément à une phylogénie. Nous avons vu précédemment qu'un ensemble de bipartitions est *compatible* en une phylogénie si et seulement si il existe une phylogénie dont les arêtes sont en bijection avec ces bipartitions. En terme de r4-arbres, il y a deux façons d'exprimer la compatibilité en une phylogénie :
Un ensemble de r4-arbres Q est dit :

- **arbre-consistant** (ou *tree-consistent* [Ste92]) ssi $\exists\, P$ t.q. $Q \subseteq Q_P$
- **arboré** (ou *tree-like* [BD86]) ssi $\exists\, P$ t.q. $Q = Q_P$.

De ces définitions, il apparaît que la notion d'arbre-consistance est moins stricte que la notion d'arboricité, la première ne nécessitant que l'existence d'une phylogénie *impliquant* (induisant) les r4-arbres de Q et la seconde nécessitant l'existence d'une phylogénie *équivalente* à l'ensemble de r4-arbres. Tout ensemble de r4-arbres arboré est donc arbre-consistant. La Fig. 1.8 montre des exemples d'ensembles de r4-arbres vérifiant ces propriétés.

1.5.6 Valuation d'une phylogénie

Une phylogénie $P = (S, A)$, où S désigne les sommets de P et A ses arêtes, peut éventuellement être valuée en lui associant une fonction $v : A \to \mathbb{R}^+$ qui attribue

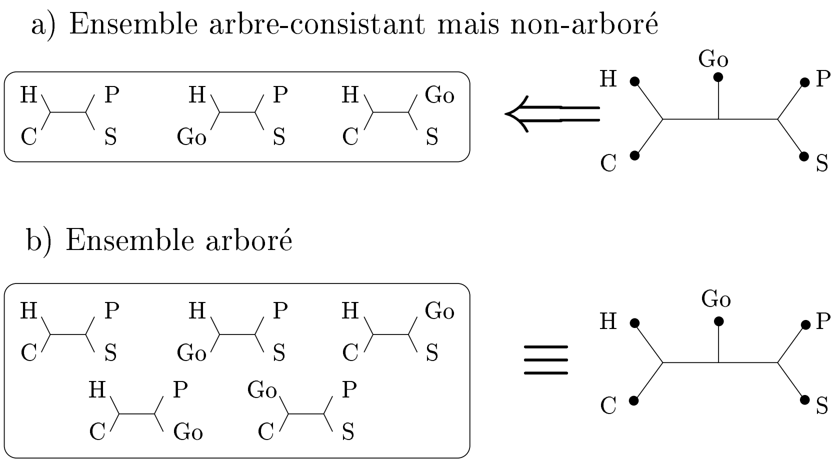


FIG. 1.8 – *Propriétés d'ensembles de r4-arbres .*

une **longueur** à chacune de ses arêtes. Cette longueur représente généralement une quantité d'évolution (proportionnelle au nombre de changements d'état d'une séquence entre ses deux extrémités).

Toute phylogénie valuée définit une distance $D = (D_{xy})$ sur l'ensemble E des espèces : la distance D_{xy} entre deux espèces x et y de E est obtenue en sommant les valeurs des arêtes de l'unique chemin qui relie ses deux feuilles dans la phylogénie. On dit dans ce cas que les longueurs d'arêtes sont **additives**, au sens où elles peuvent s'additionner. En conséquence, la mesure de distance D obtenue depuis une phylogénie valuée est appelée **distance additive d'arbre**, ou plus simplement, **distance d'arbre** (on parle aussi parfois de *distance arborée*).

Ainsi, toute phylogénie valuée correspond à une distance d'arbre sur E. Inversement, toute distance d'arbre est représentée par une phylogénie unique [Smo69] (pouvant être exhibé en temps polynômial [HY64]). En revanche, une distance $\hat{D}$ *quelconque* sur E, p. ex. obtenue depuis la matrice de caractères X et sur la base d'un certain modèle de l'évolution, ne correspond pas forcément à une phylogénie. De plus, si on parle de "distance", c'est au sens large du terme, car en R.P., $\hat{D}$ ne vérifie généralement pas l'inégalité triangulaire ($\hat{D}_{xy} \leq \hat{D}_{xz} + \hat{D}_{zy}$). En revanche, elle est toujours positive ($\hat{D}_{xy} \geq 0$), symétrique ($\hat{D}_{xy} = \hat{D}_{yx}$) et le plus souvent réflexive ($\hat{D}_{xy} = 0 \Leftrightarrow x = y$). Pour désigner $\hat{D}$, on dit fréquemment que c'est une *dissimilarité*. En raison de ces propriétés, $\hat{D}$ est usuellement représentée par une demi-matrice inférieure (*i.e.*, $(\hat{D}_{ij})$ où $i > j$).

Un résultat fondamental, obtenu indépendamment par de nombreux auteurs (citons entre autres [Zar65, Bun71]) et connu sous le nom de **condition des quatre points**, permet de caractériser les distances d'arbres :

Propriété 1 *Une mesure de distances $\hat{D} = (\hat{D}_{xy})$, vérifiant les conditions de positivité, de réflexivité et de symétrie, est une distance d'arbre ssi pour tout quadruplet (x, y, z, t), les deux plus grandes des trois sommes $\hat{D}_{xy}+\hat{D}_{zt}$, $\hat{D}_{xz}+\hat{D}_{yt}$, $(\hat{D}_{xt}+\hat{D}_{yz}$, sont égales, ce qui s'exprime aussi par :*

$$\hat{D}_{xy} + \hat{D}_{zt} \leq max\,(\hat{D}_{xz} + \hat{D}_{yt}, \hat{D}_{xt} + \hat{D}_{yz}) \ .$$

De plus, si $\hat{D}_{xy} + \hat{D}_{zt}$ est la plus petite des trois double-sommes, alors il existe au moins une arête séparant x, y de z, t dans la phylogénie P correspondant à $\hat{D}$.

Autrement dit, si $\hat{D}_{xy} + \hat{D}_{zt}$ est la plus petite double-somme, le r4-arbre $xy|zt$ est induit par la phylogénie P. Dans ce cas, on sait aussi que

$$(\hat{D}_{xy}+\hat{D}_{zt}) - (\hat{D}_{xz}+\hat{D}_{yt}) = (\hat{D}_{xy}+\hat{D}_{zt}) - (\hat{D}_{xt}+\hat{D}_{yz}) = 2l \ ,$$

où l est la longueur du chemin valué séparant les chemins $[x, y]$ et $[z, t]$ dans la phylogénie.

1.5.7 Un aperçu de la combinatoire des phylogénies

Une phylogénie non-enracinée sur $|E| = n$ espèces contient toujours n noeuds externes et entre 0 et $n-2$ noeuds internes ; en termes d'arêtes, une phylogénie sur n espèces contient toujours n arêtes externes, et entre 0 et $n-3$ arêtes internes. La phylogénie étoile est l'unique phylogénie ne possédant aucune arête interne. A l'autre extrême, les phylogénies totalement résolues ont $n-3$ arêtes internes et sont en nombre exponentiel en n. Ceci s'explique par l'argument simple suivant : toute phylogénie totalement résolue peut être obtenue en partant de la phylogénie à trois espèces sur laquelle sont successivement greffées les autres espèces ; greffer une espèce sur une arête quelconque augmente de deux le nombre d'arêtes de la phylogénie. Ainsi, on peut choisir entre 3 arêtes possibles pour la première *greffe* (donnant à chaque fois une phylogénie différente), puis entre 5 arêtes pour la 2ème, et ainsi de suite jusqu'à pouvoir greffer la dernière espèce (la $n-3$ème) sur une des $2n-5$ arêtes de la phylogénies. Le nombre de phylogénies binaires possibles est donc :

$$\prod_{k=3}^{n}(2k-5) = (2n-5)\,!! \quad , \tag{1.7}$$

où la notation "p!!" signifie $1 \times 3 \times 5 \times \ldots \times p$ (ce résultat se retrouve dans de nombreux travaux [Har71, Fel78b], etc). L'ajout d'une racine à une phylogénie non-enracinée augmente de 1 le nombre de noeuds et d'arêtes internes. Puisque la racine peut être placée sur n'importe laquelle des $2n-3$ arêtes d'une phylogénie binaire non-enracinée, le nombre de phylogénies binaires enracinées est $2n-3$ fois celui de l'éqution (1.7).

Depuis cette équation, on constate, par exemple, qu'il existe 3 phylogénies binaires non-enracinées pour 4 espèces, 15 pour 5 espèces, plus de 2 millions pour 10 espèces, $2,2 \times 10^{20}$ pour 20 espèces, etc. Dès lors, on devine à quel point l'importante combinatoire des phylogénies constitue un problème majeur auquel sont confrontées les méthodes de R.P.

Chapitre 2

Critères et méthodes de R.P.

Comme nous en avons déjà pris conscience, inférer une phylogénie est une réelle procédure d'estimation ; nous essayons de fournir la meilleure estimation d'une histoire évolutive pour laquelle nous ne disposons que de données incomplètes. Les méthodes de R.P. sont autant d'estimateurs possibles pour choisir une phylogénie parmi un nombre exponentiel de possibilités (cf. (1.7)).

La plupart des méthodes de R.P. peuvent être vues comme des procédures d'optimisation, chacune étant caractérisée par deux points :

– par le **critère** d'optimalité (aussi appelé *critère de reconstruction* ou fonction objective) qu'elle considère pour évaluer les différentes phylogénies possibles. Ceci permet ainsi potentiellement de classer ces phylogénies par ordre de préférence pour estimer la phylogénie inconnue.

– par l'**algorithme** qu'elle emploie pour rechercher une phylogénie optimale au sens du critère choisi. Bien souvent, cet algorithme ne considère pas toutes les phylogénies possibles, auquel cas il est dit **heuristique** ; autrement il est dit **exact**.

Les critères de reconstruction peuvent être classés suivant différents principes. Le plus célèbre d'entre eux, le principe de **parcimonie** date des années 60. Selon ce principe, l'estimation la plus plausible de l'histoire des espèces est la phylogénie qui correspond au scénario le plus économe en terme de changements évolutifs. Dans le cadre des méthodes de caractères [Hen50, ECS64, CS65, ED66, Hen66, KF69, Far70], ceci correspond à la phylogénie qui induit un nombre minimum de substitutions (on parle alors de "*maximum de parcimonie*"), tandis que pour les méthodes de distances [ECS63, CSE67, FM67, Far72, TNT82, RN92, RN93], ceci correspond à la phylogénie dont la somme des longueurs estimées des arêtes est la plus faible (on parle alors de "*quantité d'évolution minimale*").

L'idée de proposer la phylogénie la plus courte comme estimation de la phylogénie inconnue repose sur principalement deux justifications.

La première est d'ordre philosophique : comme dans tout domaine scientifique, lorsque l'on cherche les causes d'un phénomène naturel, on privilégie toujours l'explication minimisant le nombre d'hypothèses *ad hoc*. En R.P., les changements d'états des caractères le long des arêtes de la phylogénie inférée constituent autant d'hypothèses sur l'évolution. Selon ce raisonnement, la phylogénie la plus courte, constitue la meilleure estimation possible, puisqu'elle minimise le nombre d'hypothèses nécessaires pour expliquer les données observées.

Un autre argument en faveur de la phylogénie la plus courte a été donné par Cavalli-Sforza et Edwards [CSE67], dans un cadre géométrique : l'espace multidimensionnel des caractères est considéré, où chaque caractère est associé à une dimension, et où chaque espèce occupe une position déterminée par l'état de ses caractères. Si on ajoute à cet espace la dimension temporelle (Fig. 2.1), normale en tout point à cet espace, le cours de l'évolution peut alors être vu comme un arbre, dont les branches divergent au cours du temps et dont l'intersection avec le plan "actuel" désigne les espèces contemporaines. Comme on ne dispose de données que pour ces espèces, on ne peut espérer reconstruire que la projection de la phylogénie sur le plan "actuel" (cf. Fig. 2.1) (ce qui explique que toute information sur la position de la racine soit perdue). Dès lors que les probabilités de changement d'état sont faibles, la phylogénie la "plus courte" pour relier les espèces étudiées est l'estimation la plus plausible de la projection de leur phylogénie réelle.

L'idée des critères de parcimonie et de quantité d'évolution minimum, consistant à rechercher la phylogénie la plus courte, fait naturellement penser au problème de l'arbre de Steiner. Sous sa forme géométrique, ce problème remonte au XVIIème siècle [dF36, Tor46] et peut être défini comme la recherche de l'arbre le plus court joignant un ensemble de points dans un espace métrique donné [CR61]. Depuis l'étude originale de ce problème dans l'espace euclidien, de nombreuses variantes ont été proposées, s'accommodant de divers autres espaces métriques. Le problème dans son ensemble a généré une vaste bibliographie (cf. [GP68]), certains travaux établissant le lien entre des variantes du problème de l'arbre de Steiner et des critères de reconstruction phylogénétiques [CSE67, Tho73, ST77, Fou84, Gus91b]. Notons que la plupart des variantes du problème de l'arbre de Steiner ont été montrées NP-difficiles [GJ79], nous laissant peu d'espoir d'optimiser de façon exacte, et en temps polynômial, les critères de R.P. basés sur ce problème. Plus généralement, tous les critères de R.P., basés ou non sur le principe de parcimonie, ont été montrés NP-difficiles (généralement par William Day [Day86, DS86, Day87]).

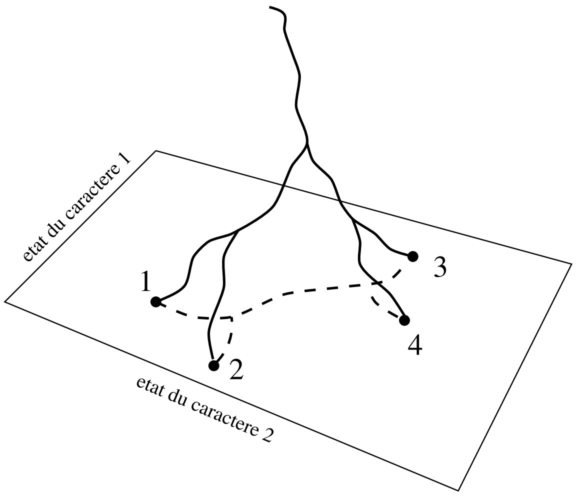


FIG. 2.1 – *(empruntée à Cavalli-Sforza et Edwards [CSE67]). Une phylogénie et sa projection sur le plan "actuel". Par souci de simplification, seules deux dimensions associées à des caractères sont représentées (la dimension temporelle est représentée verticalement). .*

Nous présentons maintenant un certain nombre de méthodes de R.P. Comme il n'est pas possible de présenter les quelques dizaines de méthodes qui existent, nous nous limiterons aux méthodes les plus caractéristiques. Ces méthodes sont classées suivant leur principe de base : caractères ou distances. Le lecteur désireux d'en savoir plus pourra avantageusement consulter les ouvrages [BG88, BG91, DT93, SOWH96].

2.1 Les méthodes de caractères

Les méthodes cladistiques les plus connues sont la méthode du Maximum de Parcimonie et la méthode de Compatibilité.

Tous les critères basés sur les caractères supposent que les données sont disponibles sous la forme d'une matrice de données X assignant un état X_{xc} à chaque espèce x pour chaque caractère c. La caractéristique principale des méthodes basées sur les caractères est l'attribution de séquences aux noeuds internes de la phylogénie. On essaye à la fois de reconstruire la structure de la phylogénie et de retrouver l'état des séquences ancestrales.

2.1.1 Le Maximum de Parcimonie (MP)

Nous précisons ici le critère de reconstruction, dit du *Maximum de Parcimonie* (MP), que nous avons déjà évoqué rapidement à plusieurs reprises. Nous évoquons ensuite deux méthodes pour optimiser ce critère.

Définition du critère

La **valeur de parcimonie** d'une phylogénie désigne le nombre minimum de changements d'états nécessaires pour expliquer les séquences associées à ses différents noeuds. Connaître la valeur de parcimonie d'une phylogénie dont on connait les séquences associées aux feuilles et aux noeuds internes est immédiat : il suffit de compter le nombre total de différences entre toute paire de séquences séparées par une arête dans la phylogénie. Par exemple sur la Fig. 2.2, on observe une dif-

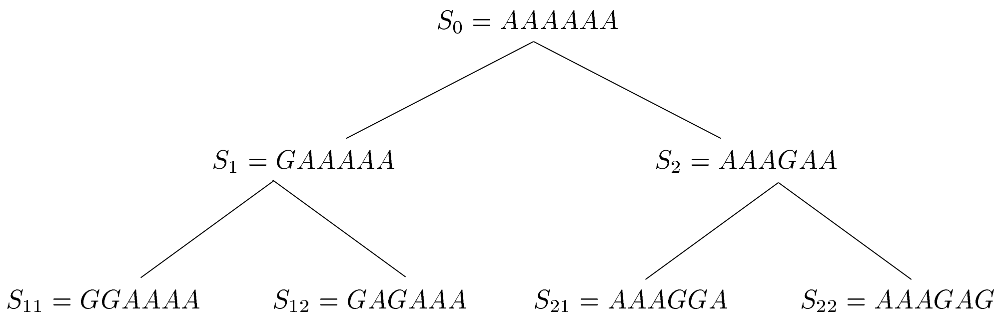


FIG. 2.2 – *Phylogénie d'un ensemble de séquences (d'après [CGAL95]).*

férence entre les séquences S_0 et S_1, une différence entre S_1 et S_{11}, etc., si bien que la valeur de parcimonie de cette phylogénie est 6. Dit autrement, 6 changements d'états seulement suffisent à expliquer les différences entre les séquences (p. ex. le 1er caractère s'est transformé de `A` en `G` entre S_0 et S_1) ; bien sûr d'autres explications sont possibles : entre S_0 et S_1, le 1er caractère peut très bien avoir évolué de `A` en `C` puis en `G`, demandant ainsi 2 changements. Cependant, le principe de parcimonie consiste à toujours choisir l'explication faisant le moins d'hypothèses, on considère donc qu'un seul changement a eu lieu pour le 1er caractère et qu'aucun changement n'a affecté les autres caractères entre ces deux séquences, si bien que l'arête séparant S_0 et S_1 se voit attribuer un coût de 1.

Dans le cas général, nous ne disposons initialement que des séquences associées aux feuilles de la phylogénie (les espèces contemporaines). Il existe théoriquement un grand nombre de possibilités pour attribuer des séquences aux noeuds internes, chacune impliquant une quantité d'évolution (une longueur) différente pour la phylogénie. De par sa définition (cf. ci-dessus), la valeur de parcimonie d'une telle phylogénie est obtenue en retenant la combinaison de séquences assignées aux noeuds internes qui induit le nombre minimum de substitutions. Calculer la valeur de parcimonie d'une phylogénie dans le cas général semble un problème difficile au regard du grand nombre de combinaisons possibles pour assigner des séquences aux noeuds internes; cependant le problème est simplifié par le fait que les différents caractères peuvent être considérés indépendamment les uns des autres, et que, pour chaque caractère, seul un nombre restreint de combinaisons est à examiner. Un algorithme linéaire, *i.e.*, en $O(kn)$, fut proposé par Fitch [Fit71] et prouvé par Hartigan [Har72].

Disposant d'un ensemble de séquences pour n espèces, il existe une multitude de phylogénies possibles (cf. (1.7)), ayant chacune leur valeur de parcimonie. Le critère de R.P. dit de **Maximum de Parcimonie** (MP) consiste à désigner la phylogénie ayant la valeur de parcimonie minimale comme meilleure estimation de la phylogénie qu'on cherche à reconstruire. Suivant ce critère, l'ordre défini entre les phylogénies par leur valeur de parcimonie reflète leur qualité pour estimer la phylogénie inconnue.

Notons que la phylogénie qui minimise le nombre total de changements d'état ne minimise pas forcément le nombre de changements d'état pour chacun des caractères: en conséquence, l'histoire de certains caractères est expliquée par des homoplasies. Cependant, le critère MP tend à minimiser le nombre d'homoplasies dans la phylogénie qu'il propose, car les caractères qui en sont affectés sont plus coûteux en nombre de substitutions.

Notons aussi que des phylogénies non-enracinées sont généralement considérées pour résoudre ce problème car l'insertion d'une racine, sur quelque arête que ce soit, ne modifie pas la valeur de parcimonie de la phylogénie.

Méthodes d'optimisation de MP

Le problème MP a été montré NP-difficile [Day86]. L'importante combinatoire associée au problème rend difficilement envisageable une recherche exhaustive, par exemple pour 10 espèces, il existe déjà 2 millions de phylogénies possibles.

Une méthode de *branch-and-bound*, proposée par Hendy et Penny [HP82], permet d'obtenir la (ou les) phylogénie(s) ayant une valeur de parcimonie optimum sans explorer toutes les phylogénies possibles. Cette méthode permet de traiter des jeux de données contenant plus de 30 espèces.

Les nombreux algorithmes heuristiques proposés [Swo90, Fel93] restent les plus utilisés pour optimiser le critère MP. Ils sont généralement basés sur le principe glouton proposé par Wagner et publié par Farris [Far70] : on relie initialement trois des espèces considérées (choisies arbitrairement), puis on fait croître cette phylogénie en ajoutant successivement les espèces restantes ; à chaque étape, on ajoute une nouvelle espèce sur l'une des arêtes de la phylogénie courante, en choisissant l'arête qu'on estime conduire au maximum de parcimonie. Cet algorithme rapide, en $O(kn^2)$, est souvent complété par une phase d'amélioration itérative, où on essaye de faire changer de place dans la phylogénie chacune des espèces à son tour. Les heuristiques basées sur ce principe ont des performances très acceptables pour optimiser le critère MP, comme nous le verrons à la fin de la partie introductive.

2.1.2 La méthode de Compatibilité

La méthode de compatibilité repose essentiellement sur les travaux de Lequesne [LeQ69, Leq72] et Estabrook [Est72, EJM76], elle est définie pour des caractères *binaires*. Chaque caractère binaire induit une bipartition sur l'ensemble E des espèces étudiées, suivant l'état, 0 ou 1, pris par les espèces pour ce caractère. Le principe de la méthode repose sur le constat suivant : les caractères binaires qui nous renseignent le mieux sur l'évolution sont ceux qui ne se sont transformés qu'une seule fois au cours du temps ; ils correspondent chacun à une bipartition des espèces induite par une arête de la phylogénie des espèces, t. Au contraire, les caractères qui se sont transformés plusieurs fois (homoplasiques) risquent d'amener des regroupements erronés d'espèces, en désignant des bipartitions ne correspondant à aucune arête de la phylogénie correcte.

Les bipartitions correspondant aux arêtes de la phylogénie t sont *compatibles* puisqu'elles peuvent être combinées ensembles en une même phylogénie. En conséquence, si on cherche le plus grand nombre de caractères compatibles les uns avec les autres, il y a de fortes chances pour que les caractères non-homoplasiques soient retrouvés. Chacun des caractères retenus a ainsi une probabilité raisonnable d'indiquer une arête de la phylogénie inconnue, suivant la bipartition qu'il induit sur les espèces (en fonction de leur état, 0 ou 1).

La **méthode de compatibilité** correspond donc à la version d'optimisation du problème de compatibilité des bipartitions (cf. chap. précédent, section 1.5.4) : étant donné un ensemble de caractères binaires, on cherche le sous-ensemble maximum de caractères qui sont compatibles. Cet ensemble définit une phylogénie qui est proposée comme estimation de la phylogénie inconnue. Ce problème est NP-difficile [DS86].

Tous les caractères rejetés par cette procédure nécessitent plusieurs substitutions pour expliquer leurs états aux espèces sur la phylogénie proposée, ils sont donc jugés homoplasiques. Selon Darlu et Tassy ([DT93], p.148), la méthode de compatibilité peut être considérée comme une méthode utilisant le principe de parcimonie, en ce sens qu'elle retient comme phylogénie celle qui *minimise le nombre* de caractères jugés homoplasiques (les plus coûteux à expliquer).

Plusieurs variantes de ce problème ont été étudiées récemment (notamment pour des caractères pouvant prendre plus de 2 états) : citons le problème de la *phylogénie parfaite* [BFW92, War94], et celui du *nombre phylogénétique* [GGP+96].

2.1.3 La méthode du Maximum de Vraisemblance (MV)

L'idée de la méthode du Maximum de Vraisemblance fut présente en R.P. dès les années 60-70 [ECS64, Ney71], mais ne fut véritablement envisageable pour des données moléculaires que suite aux travaux de Felsenstein [Fel81]. Le principe de cette méthode est relativement simple. Étant donné un modèle de l'évolution M et un jeu de données X, la vraisemblance d'une phylogénie $\hat{t}$ est la probabilité du jeu de données étant donnés la phylogénie et le modèle, $P(X;\hat{t},M)$, considérée comme une fonction de la phylogénie [Fel88]. La somme des probabilités de tous les jeux de données doit être 1, mais quand le jeu de données est gardé constant, tandis que la phylogénie varie, les différentes valeurs $P(X;\hat{t},M)$ ne doivent pas donner 1 quand on les additionne, et sont appelées **vraisemblances** plutôt que probabilités. La méthode du Maximum de Vraisemblance (MV) choisit simplement la phylogénie $\hat{t}$ qui maximise la vraisemblance, maximisant ainsi la probabilité que le jeu de données soit observé [Fel88].

Dans le cadre des modèles de l'évolution utilisés en R.P. (cf. section 1.3), on peut obtenir la probabilité $p_{ij}(t)$ qu'un site quelconque d'une séquence passe de l'état i à l'état j au cours d'une période de temps de longueur t. Cette probabilité sert de base à l'établissement de la vraisemblance d'une phylogénie.

Considérons d'abord le cas simple d'une phylogénie contenant uniquement deux séquences S_1 et S_2 de k caractères, séparées par une arête de longueur t. La probabilité que S_1 se transforme en S_2, en une période de temps t, *i.e.*, la **vraisemblance** attachée à l'évolution de S_1 en S_2 au cours d'une période de temps t est, [CGAL95, SOWH96] :

$$L(S_1,S_2;t) \;=\; \prod_{c=1}^{k} \pi_{S_{1c}} p_{S_{1c}S_{2c}}(t) \;, \tag{2.1}$$

où S_{1c} (resp. S_{2c}) dénote l'état du site c de la séquence S_1 (resp. S_2), et $\pi_{S_{1c}}$ est la probabilité d'observer l'état S_{1c} dans la séquence S_1. Le seul paramètre à estimer ici est la longueur t de l'arête séparant S_1 de S_2. La méthode MV

consiste à choisir la valeur $\hat{t}$ maximisant l'équation 2.1. Remarquons qu'en raison de l'hypothèse de réversibilité des modèles de l'évolution, $L(S_1, S_2; t) = L(S_2, S_1; t)$. Ainsi la vraisemblance est indépendante du sens de l'évolution, et donc de la position de la racine.

Si on considère maintenant le cas d'une phylogénie comportant plus de deux séquences, après avoir fixé la racine arbitrairement sur une séquence S_0, il nous faut évaluer la vraisemblance de chaque état pour chacun des caractères de chacune des séquences de la phylogénie en fonction de sa structure. En raison de l'hypothèse d'indépendance entre les différents sites, on peut raisonner indépendamment sur chaque site c. Si la séquence racine S_0 a divergé pour donner les séquences S_1 et S_2, on montre la formule de récurrence suivante :

$$L(S_{0c}{=}i) \quad = \quad \left[\sum_{j\in A,C,G,T} p_{ij}(t_{S_0S_1})L(S_{1c}{=}j)\right]\left[\sum_{j\in A,C,G,T} p_{ij}(t_{S_0S_2})L(S_{2c}{=}j)\right]$$

où i, j sont des états de caractères, $t_{S_0S_1}$ (resp. $t_{S_0S_2}$) est le temps de divergence entre S_0 et S_1 (resp. S_2) en terme de distance évolutive. Ceci exprime que la vraisemblance de la phylogénie dont S_0 est la racine se calcule en fonction de la vraisemblance des deux sous-arbres de racine S_1 et S_2. Si S_1 (ou de façon équivalente S_2) est une feuille de la phylogénie, alors $L(S_{1c}{=}j){=}1$ si l'état j est observé pour le site c de S_1, sinon $L(S_{1c}{=}j){=}0$.

On peut ainsi calculer la vraisemblance de toute phylogénie $\hat{t}$ dont on connait la structure et les longueurs d'arêtes. Malheureusement, en R.P., on ne connait ni la structure de la phylogénie réelle des espèces, ni la valuation de ses arêtes, ni l'état des séquences aux noeuds internes (X ne donne que les séquences aux feuilles de la phylogénie !). Pour obtenir la phylogénie de vraisemblance maximum on est obligé d'évaluer la vraisemblance de chaque structure possible, en essayant pour chacune toutes les combinaisons d'états possibles pour les séquences aux noeuds internes, sans oublier que, pour chaque combinaison, les longueurs d'arêtes doivent, elles aussi, être estimées !

On comprend dès lors les limites de cette approche. Même si la méthode MV repose en pratique sur une heuristique permettant d'inférer une phylogénie sans examiner le nombre exponentiel de phylogénies [Fel81, Sai88], et si elle est reconnue comme la méthode donnant les résultats les plus fiables, elle reste la méthode de R.P. la plus lente et n'est guère utilisée pour plus de 10 espèces.

2.2 Les méthodes de distances

Les méthodes évoquées ci-dessous sont basées sur des distances entre espèces. En ce sens, elles intéressent un grand nombre de domaines où on cherche à reconstruire un arbre pour des objets entre lesquels on connait une distance : recherche opérationnelle, traitement de l'information, biologie moléculaire, biologie des populations, psychométrie, linguistique, archéologie, etc. En R.P. ces méthodes sont généralement appliquées sur des matrices de distances obtenues depuis des matrices de caractères, même si elles sont aussi adaptées à d'autres types de données (immunologiques, hybridation ADN-ADN).

Dans le cas où l'on dispose initialement de séquences de caractères, nous avons décrit précédemment comment pour deux espèces x et y, on pouvait obtenir une estimation $\hat{D}_{xy}$ de leur distance évolutive, sur la base d'un modèle de l'évolution. Les méthodes de distances sont basées sur la donnée de la matrice $(\hat{D}_{xy})$. Le fondement de ces méthodes est que toute phylogénie valuée définit aussi une mesure de distance entre les espèces (cf. section 1.5.6, chap. précédent). Rappelons qu'une telle mesure est dite *distance d'arbre* et vérifie la condition des quatre points. Les dissimilarités $\hat{D}$ dont on dispose en R.P. (comme dans d'autres domaines) ne vérifient proverbialement pas cette condition, aussi le but des méthodes de distances s'énonce ainsi [BG91] : étant donné un ensemble E d'espèces et une dissimilarité $\hat{D}$ sur E, on cherche une phylogénie valuée, dont l'ensemble des feuilles est contenu dans E, et telle que les longueurs des chemins entre deux éléments de E constituent une «bonne approximation» de $\hat{D}$. La différence entre les diverses méthodes de distances consiste principalement à définir ce qu'est une bonne approximation.

Nous présentons dans cette section les méthodes de distances les plus connues en R.P. Les deux dernières ont pour particularité de ne pouvoir être appliquées qu'à quatre espèces, et leur intérêt réside essentiellement dans l'inférence de r4-arbres. Ces r4-arbres peuvent ensuite être combinés en une phylogénie sur l'ensemble des espèces par une méthode de *quadruplets* (cf. part. *III*).

2.2.1 La méthode de représentation multidimensionnelle (MDS)

Edwards et Cavalli-Sforza [ECS63, CSE67] furent parmi les premiers à appliquer une approche de distances en R.P. Ils disposaient de données consistant en des fréquences de gènes observées pour différentes espèces, qu'ils proposaient de convertir en une matrice de dissimilarités $\hat{D}$. Leur idée était de représenter, sur la base de $\hat{D}$, l'ensemble des n espèces étudiées par un ensemble de n points dans un espace euclidien à $n-1$ dimensions. La longueur d'une phylogénie reliant les espèces est simplement obtenue en sommant la longueur de ses différentes arêtes, chacune étant définie comme la distance euclidienne séparant ses deux extrémités. Cavalli-Sforza et Edwards proposaient de rechercher l'arbre de Steiner de longueur minimum pour relier les points associés aux espèces. Autrement dit, on cherche ici la phylogénie dont la quantité d'évolution est minimale, puisque la distance euclidienne entre deux points représente leur distance évolutive.

La représentation euclidienne d'une dissimilarité connait toujours un essor important, sous le nom de MDS (Multi-Dimensional Scaling). Certaines dissimilarités $\hat{D}$ ne possèdent pas de représentation euclidienne de n points respectant les distances qu'elles imposent entre les espèces. Une condition nécessaire et suffisante pour que $\hat{D}$ admette une représentation est qu'elle vérifie l'inégalité triangulaire. Si ce n'est pas le cas, on peut avoir recours à la méthode de la *constante additive*: à toutes les entrées de la matrice $\hat{D}$, on ajoute une certaine constante, choisie assez grande pour que toutes les inégalités triangulaires soient vérifiées. Cette pratique est particulièrement adaptée en R.P., puisque si $\hat{D}$ est une distance d'arbre, la phylogénie désignée ne change pas (seules les longueurs d'arêtes sont modifiées), et si $\hat{D}$ n'est pas une distance d'arbre, on peut montrer que la phylogénie inférée par la plupart des méthodes de reconstruction ne change pas.

La méthode MDS, introduite en R.P. dans les années 60, a été délaissée depuis au profit d'autres méthodes proposant elle-aussi d'inférer la phylogénie la plus courte, mais selon une approche différente (cf. section 2.2.4).

2.2.2 Les méthodes d'ajustement

Ces méthodes cherchent à produire la distance d'arbre $\hat{T}_{xy}$ la plus proche d'une mesure de distance $\hat{D}$ donnée, au sens d'un certain critère, généralement de la forme:

$$\sum_{x<y} W_{xy} \left(\hat{D}_{xy} - \hat{T}_{xy}\right)^{\alpha}$$

où W_{ij} est une fonction de poids. Pour $\alpha=1$, on utilise l'expression $|\hat{D}_{xy} - \hat{T}_{xy}|$. Pour $\alpha=2$ et $W_{xy}=1$, on obtient le critère des *moindres carrés ordinaires* (MC), sans doute le plus répandu. L'optimisation de ce critère a été montrée NP-difficile [Kri86].

L'approche exhaustive [FM67] ne permet pas de traiter plus d'une dizaine d'espèces, aussi les praticiens se sont-ils tournés vers des méthodes donnant des solutions approchées. Parmi celles-ci, citons l'approche issue de la programmation mathématique [dS83] et l'approche de réduction [Rou88, GL96].

Signalons qu'en pratique les moindres carrés sont très souvent employés pour évaluer les longueurs d'arêtes d'une phylogénie, même si cette phylogénie a été produite par une autre méthode de reconstruction.

2.2.3 Les méthodes de scores

Les méthodes de score fonctionnent suivant un **principe agglomératif** : partant de la phylogénie étoile (Fig. 2.3-A), à chaque étape deux espèces connectées au noeud central sont regroupées en un même sous-arbre (Fig. 2.3-B). Ce sous-arbre est désormais considéré comme une nouvelle espèce (*i.e.*, une espèce virtuelle) connectée au noeud central à la place des deux espèces agglomérées. Puis, on itère le processus, généralement jusqu'à ce que le noeud central ne soit plus connecté qu'à trois espèces, *i.e.*, que la phylogénie soit totalement résolue (Fig. 2.3-C).

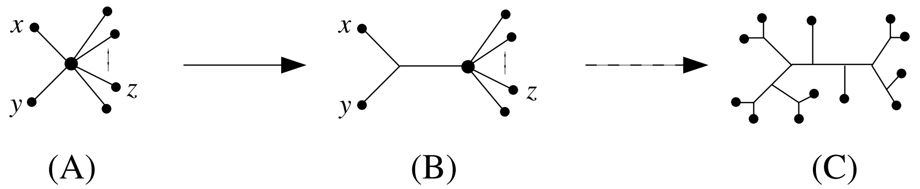


FIG. 2.3 – *Processus agglomératif employé pour construire une phylogénie.*

À chaque étape on regroupe les deux espèces qui sont les plus proches l'une de l'autre au sens d'un certain *score*. Ce score est calculé pour toute paire d'espèces connectées au noeud central, suivant un principe qui varie d'un algorithme à l'autre. *Addtree* [ST77] est la méthode de scores la plus connue, et possède de nombreuses variantes (cf. [BG91] pour un survol). Cet algorithme prend en entrée une matrice de distances $\hat{D}$ qu'il utilise à chaque étape pour évaluer le score des paires d'espèces, selon le principe suivant : $x, y \in E$ sont voisins relativement à $z, t \in E$ ssi :

$$\hat{D}_{xy} + \hat{D}_{zt} < min(\hat{D}_{xz} + \hat{D}_{yt}\ ,\ \hat{D}_{xt} + \hat{D}_{yz})\ .$$

Le score de la paire xy est simplement le nombre de paires zt relativement auxquelles xy sont voisins. Une fois déterminée la paire xy de score maximum, *Addtree* met à jour la matrice $\hat{D}$ pour tenir compte de la nouvelle *espèce* $z \in E$ (regroupant x et y). Pour ceci, il évalue la distance séparant z de tout autre espèce u connectée au noeud central. Cette distance est estimée par la formule suivante :

$$\hat{D}_{zu} = \frac{1}{2}\left(\hat{D}_{xu} + \hat{D}_{yu} - 2\hat{D}_{xy}\right)\ .$$

Notons que cette distance peut aussi être estimée aux moindres carrés, selon la formule donnée par Gascuel [Gas97b].

2.2.4 La méthode d'évolution minimale (ME)

Cette méthode adopte le principe consistant à chercher la phylogénie la plus courte, *i.e.*, la phylogénie à laquelle est associée la quantité d'évolution minimale. On utilise donc l'idée que les longueurs d'arêtes d'une phylogénie peuvent être additionnées pour indiquer une quantité d'évolution. Toutefois, on ne reprend pas l'idée consistant à représenter les espèces dans un espace métrique. Les longueurs d'arêtes ne sont ainsi plus mesurées comme la distance entre deux points dans un espace donné, mais sont ici évaluées aux moindres carrés. Plus précisément, le critère de "**minimum evolution**" (ME) est défini ainsi : soit $\hat{D}$ la matrice de données dont on dispose, on recherche la phylogénie dont les longueurs d'arêtes sont évaluées aux moindre carrés vis-à-vis de $\hat{D}$, et qui soit la plus courte (au sens de la somme des longueurs de ses arêtes). Autrement dit, pour toute phylogénie τ, on choisit ses longueurs d'arêtes $w(e)$ de façon à minimiser $L^2(D^\tau, \hat{D})$ puis on choisit la phylogénie

$$\hat{t} = min_\tau(\sum_{e\in\tau} w(e)) \ .$$

Ce problème n'a pas encore été montré NP-difficile, mais tout laisse à penser qu'il l'est.

Le bien-fondé du critère ME a été prouvé par Rzhetsky et Nei [RN93]. Ils montrent, sous l'hypothèse que les distances utilisées sont non-biaisées (*i.e.*, pour tout couple d'espèces x et y, $E(\hat{D}_{xy}) = D_{xy}$), que si les longueurs d'arêtes sont évaluées aux moindres carrés, alors la phylogénie la plus courte est en espérance la phylogénie correcte. Cette bonne propriété ainsi que l'existence d'heuristiques peu coûteuses a contribué à répandre largement l'utilisation de ce critère de reconstruction.

Une heuristique pour ME

Examiner toutes les phylogénies possibles n'est pas raisonnable en raison de leur nombre exponentiel (cf. équation (1.7)). Les phylogénéticiens ont plutôt recours à un algorithme heuristique présenté par Saitou et Nei [SN87]. Cet algorithme est appelé *Neighbor-Joining* (NJ), puisqu'il adopte, comme *Addtree* (cf. section 2.2.3), un principe agglomératif le conduisant à *joindre*, à chaque étape, des noeuds *voisins* dans la phylogénie. L'algorithme NJ fonctionne donc selon le même principe qu'*Addtree* à la différence qu'à chaque étape, il choisit de regrouper les deux espèces x et y, de sorte que la longueur de la phylogénie résultant de leur agglomération en un sous-arbre commun, estimée aux moindres carrés, soit minimale. En ce sens, l'algorithme optimise à chaque étape le critère ME. Toutefois, cette optimisation est approximative puisque l'algorithme est glouton, et rien n'indique qu'ajouter la plus courte arête à chaque étape conduise à la phylogénie la plus courte.

Cet algorithme fut initialement décrit avec une complexité en $O(n^5)$ [SN87]. L'année d'après, Studier et Keppler [SK88] ont montré qu'une version en $O(n^3)$ pouvait être obtenue en simplifiant les calculs effectués à chaque étape pour choisir la paire de sommets à agglomérer. Cette amélioration de la complexité a grandement contribué au succès de cette méthode, qui est une des plus utilisées aujourd'hui.

Expérimentalement, de grandes similitudes ont été souvent constatées entre les résultats des méthodes NJ et *Addtree* [Nei91]. Ces similitudes ne sont pas surprenantes puisque les deux méthodes partagent le même principe agglomératif de reconstruction, tout en estimant les longueurs d'arêtes aux moindres carrés ; toutefois elles ne choisissent pas les paires d'espèces voisines de la même façon. En fait, Gascuel a montré qu'*Addtree* optimise un critère de voisinage correspondant à une version discrétisée du critère employé par NJ [Gas94].

2.2.5 La méthode FPM

Cette méthode est liée à la condition des quatre points (cf. chap. 1), caractérisant les distances qui correspondent exactement avec une phylogénie valuée. Comme nous l'avons déjà indiqué, les distances analysées en R.P. ne vérifient presque jamais la condition des quatre points en raison d'un bruit d'échantillonnage affectant les distances observées entre les espèces. Théoriquement, le recours à une méthode de correction des distances (cf. chap. 1) permet de retrouver les distances correctes entre les espèces, mais uniquement si le modèle de l'évolution posé est juste et si on dispose d'un nombre infini de caractères. Cependant, il peut se trouver certains cas où les distances obtenues entre les espèces (avec ou sans correction) sont relativement proches des distances correctes, si bien que, même si la condition des quatre points n'est pas exactement vérifiée, on peut espérer retrouver la phylogénie des espèces. C'est le point de vue adopté par la méthode FPM (*Four-Point condition Modified*) [Bun71, ST77, Fit81, BD86, DvHK86, ESSW97b], inférant une phylogénie pour quatre espèces sur la base d'une variante de la condition des quatre points.

Plus précisément, la condition des quatre points assure qu'une distance $\hat{D}$ sur quatre espèces x, y, z, t correspond exactement à une phylogénie si et seulement si les deux plus grandes des trois sommes $\hat{D}_{xy} + \hat{D}_{zt}$, $\hat{D}_{xz} + \hat{D}_{yt}$, $\hat{D}_{xt} + \hat{D}_{yz}$ sont égales ; de plus, si $\hat{D}_{xy}+\hat{D}_{zt}$ est la plus petite somme, alors la phylogénie désignée est la structure binaire sur x, y, z, t, dont l'arête interne sépare les espèces x et y des espèces z et t, *i.e.*, le r4-arbre noté $xy|zt$ (Fig. 1.7, chap. 1). La méthode FPM utilise la variante suivante de la condition des quatre points pour choisir la phylogénie des quatre espèces x, y, z, t :

$$\left(\hat{D}_{xy}+\hat{D}_{zt} < \left\{ \begin{array}{l} \hat{D}_{xz}+\hat{D}_{yt} \\ \hat{D}_{xt}+\hat{D}_{yz} \end{array} \right. \right) \quad \Leftrightarrow \quad xy|zt \ .$$

Cette règle a été montrée bien fondée au sens de plusieurs critères de reconstruction [SN87, GL96].

Si l'on étudie plus de quatre espèces, on peut utiliser la méthode FPM sur tous les quadruplets d'espèces, ce qui aboutit à la production d'un ensemble de r4-arbres, qu'on peut ensuite combiner en une phylogénie sur l'ensemble des espèces, au moyen d'un algorithme de reconstruction basé sur les quadruplets. Lorsque l'on étudie plus de quatre espèces et que l'on dispose d'une distance d'arbre $\hat{D}$, si $\hat{D}_{xy}+\hat{D}_{zt}$ est la plus petite des trois sommes sur x, y, z, t, la condition des quatre points indique qu'il existe au moins une arête dans la phylogénie correspondant à $\hat{D}$ qui sépare x, y de z, t. Cette contrainte structurelle correspond au r4-arbre $xy|zt$. Quand $\hat{D}$ n'est pas une distance d'arbre, utiliser la méthode FPM revient à faire l'hypothèse que même si les distances sont bruitées, elles ne le sont pas suffisamment pour changer l'ordre entre les doubles sommes, au point de faire que la plus petite des trois double-sommes ne reste pas inférieure aux deux autres. Si cette hypothèse est vérifiée, les r4-arbres inférés par FPM sont exactement ceux de la phylogénie correcte, qui est alors retrouvée par n'importe quel algorithme de quadruplets raisonnable. Si le bruit affectant les données est relativement important, FPM est conduite à inférer un r4-arbre erroné pour certains quadruplets, et il ne reste plus qu'à espérer que la proportion de r4-arbres incorrects ne soit pas trop importante, auquel cas une méthode de quadruplets doit encore pouvoir retrouver la phylogénie correcte, ou sa majeure partie.

Nous verrons au chapitre 2, de la partie III, que la méthode FPM a de bonnes propriétés statistiques.

2.2.6 La méthode ordinale

Cette méthode repose sur l'ordre existant entre les différentes distances $\hat{D}_{xy}$ entre espèces. Elle a été étudiée par Kannan et Warnow [KW95], puis par Guénoche [Gué97a, Gué97b] et Kearney [KHM97, Kea97] (nous utiliserons ici le vocabulaire proposé par A. Guénoche).

La méthode ordinale suppose que le bruit évolutif n'a pas trop perturbé l'ordre des distances entre espèces, *i.e.*, que l'importance *relative* des distances entre elles est conservée. Cette hypothèse est plus faible que celle faite par la méthode FPM qui suppose, elle, que l'importance *absolue* des distances n'est pas trop perturbée. Guénoche introduit le vocabulaire suivant :

- On note $\mathcal{P}(\hat{D})$ la préordonnance associée à une dissimilarité $\hat{D}$, obtenue en ordonnant les valeurs de $\hat{D}$. Si $\hat{D}_{xy} < \hat{D}_{zt}$, on aura $xy < zt$ dans la préordonnance, ce qui peut être noté par le *couple* $(xy < zt)$.

- Une préordonnance $\mathcal{P}(\hat{D})$ est dite *arborée* ssi il existe une distance d'arbre D t.q. $\mathcal{P}(D) = \mathcal{P}(\hat{D})$.

- On dit que les couples $(xy < zt)$ et $(xz < yt)$ sont *emboîtés* si les *paires* de l'un ou de l'autre couple sont les valeurs extrêmes des quatre paires, *i.e.*, on a dans $\mathcal{P}(\hat{D})$ l'un des deux ordres $xy < xz < yt < zt$ ou $xz < xy < zt < yt$ (cf. Fig. 2.4-(a)).

- On dit que le couple $(xz < yt)$ *domine* le couple $(xy < zt)$ si dans $\mathcal{P}(\hat{D})$ on a $xy < xz$ et $zt < yt$ (cf. Fig. 2.4-(b)).

- On appelle *condition ordinale des quatre points* la condition telle que tout quadruplet d'espèces x, y, z, t ait deux couples emboîtés et que le troisième couple ne domine pas les deux autres.

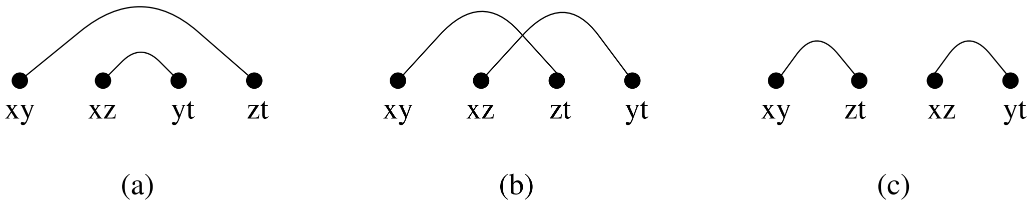


FIG. 2.4 – *Différentes configurations pour les couples $(xy < zt)$ et $(xz < yt)$. (a) Les deux couples sont emboîtés. (b) et (c) le couple $(xz < yt)$ domine le couple $(xy < zt)$.*

Guénoche [Gué97b] montre qu'une préordonnance vérifie la condition ordinale des quatre points si et seulement si elle est arborée. Même si une matrice de données $\hat{D}$ ne définit pas toujours une préordonnance qui vérifie la condition ordinale des quatre points pour tous les quadruplets, on peut toujours chercher à reconstruire une phylogénie sur la base des quadruplets vérifiant cette condition par la règle suivante [Gué97b, KHM97, Kea97] :

le quadruplet x, y, z, t est t.q. le couple $(xy < zt)$
(équiv. $(zt < xy)$) est dominé par les deux autres $\Leftrightarrow$ $xy|zt$.
couples, et ceux-ci sont emboîtés

2.3 Propriétés souhaitables pour une méthode de reconstruction

Du corpus de publications du domaine [PHS92, SOWH96], on peut facilement dégager un certain nombre de propriétés qu'une méthode de reconstruction phylogénétique doit *idéalement* posséder.

2.3.1 Complexité polynomiale

Un premier facteur discriminant les méthodes en pratique est la *rapidité d'exécution*. Pour les biologistes, il existe en effet une frontière très nettement marquée entre les méthodes "*efficaces*", *i.e.*, de complexité polynomiale en n (le nombre d'espèces), et les méthodes "non-raisonnables", *i.e.*, de complexité exponentielle en n. Comme nous l'avons vu précédemment, la plupart des critères de R.P. sont NP-difficiles à optimiser, le plus souvent en raison du nombre exponentiel de phylogénies possibles pour n espèces (1.7). Les méthodes de reconstruction étiquetées "efficaces" sont donc le plus souvent des algorithmes heuristiques, *i.e.*, optimisant de façon approchée un certain critère, tandis que les méthodes considérées comme irraisonnables sont essentiellement des algorithmes exacts, donnant la phylogénie optimum au sens du critère retenu.

2.3.2 Convergence

Les biologistes sont prêts à accepter que, pour des petits jeux de données (comportant un faible nombre de caractères), l'échantillonnage des caractères ait une influence non-négligeable sur la phylogénie proposée par une méthode de reconstruction, au sens où le choix de la phylogénie dépend de la composition du jeu de données. Toutefois, pour des jeux de données de taille importante, une propriété qu'on peut raisonnablement demander est la stabilité de la phylogénie proposée, *i.e.*, qu'on retrouve toujours les mêmes arêtes, indépendamment des caractères décrivant les espèces. La propriété correspondante s'appelle la *convergence* : une méthode est dite convergente si elle infère une certaine phylogénie avec une probabilité de plus en plus élevée lorsque le nombre de caractères disponibles augmente.

2.3.3 Consistance, *i.e.*, convergence vers la phylogénie corecte

Depuis la notion de convergence, on peut formuler une autre propriété encore plus souhaitable pour une méthode, la *consistance* : une méthode est dite consistante quand elle converge vers la phylogénie corecte. Depuis le chapitre précédent, on connait déjà une condition de consistance pour les méthodes de distances : ces méthodes font l'hypothèse explicite d'un certain modèle de l'évolution afin de corriger les distances évolutives observées initialement entre les espèces ; ainsi, toute méthode de distances raisonnable est consistante si le modèle qu'elle utilise est juste. En effet, si le modèle est juste, lorsque le nombre de caractères k augmente, la matrice de distances corrigées, $\hat{D}$, se rapproche de la véritable matrice de distances, D, qui est une distance d'arbre, or toute méthode raisonnable retrouve la phylogénie associée à une telle distance.

Conditions sur les longueurs relatives des arêtes

L'intérêt des phylogénéticiens pour la consistance des méthodes remonte à la fin des années 70, quand la très populaire méthode MP fut montrée inconsistante, sous le modèle NCF, pour une phylogénie de quatre espèces seulement [Fel78a, Cav78]. Une telle phylogénie est caractérisée par des probabilités de substitution très élevées sur deux arêtes *opposées* (d'un côté et de l'autre de l'arête centrale) vis-à-vis des probabilités des autres arêtes (p. ex. sur la Fig. 2.5, quand $p_1 = p_3 >> p_2 = p_4 = p_5$ sur la phylogénie (a)). Ce phénomène, qui affecte aussi les phylogénies de plus de quatre espèces possédant deux très longues arêtes séparées par une courte arête, est connu sous le nom de **problème d'attraction des longues branches** [Fel78a, HP89, PHS91, Ste94b, TN94a] : la probabilité d'une seule substitution sur la courte arête est plus faible que la probabilité de plusieurs substitutions sur les arêtes longues ; en conséquence les séquences au bout des longues arêtes ont tendance à être similaires, et sont ainsi regroupées sous un même noeud par les méthodes de reconstruction. Par exemple sur la Fig. 2.5-(a), si les probabilités p_1 et p_3 sont importantes vis-à-vis de p_5, la phylogénie incorrecte (b), regroupant les espèces 1 et 3, est inférée. Sur la base de ce phénomène d'attraction des longues branches, MP a été montrée inconsistante sur des phylogénies de 5 espèces et plus possédant pourtant des taux d'évolution identiques sur toutes les arêtes (dans ce cas, les différences entre les longueurs d'arêtes sont dûes à des intervalles de temps inégaux). Ceci a été montré sous le modèle NCF [PHH87, HP89] et sous celui de Jukes et Cantor [TN94a]. Signalons que plusieurs jeux de données réels, pour lesquels la phylogénie supposée est soumise au problème d'attraction des longues branches, ont été exhibés [Lak86, PHS91].

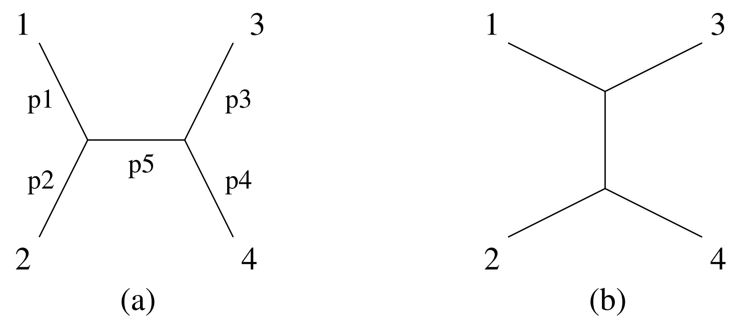


FIG. 2.5 – *(a) Phylogénie sur 4 espèces, où à chaque arête est affectée une probabilité p_i que tout site subisse une substitution le long de cette arête. (b) Phylogénie incorrecte inférée par MP quand $p_1 = p_3 >> p_2 = p_4 = p_5$.*

Poursuivant les travaux précédents sous le modèle NCF, Steel a caractérisé les cas où diverses méthodes de R.P. sont, ou ne sont pas, consistantes pour des phylogénies à quatre espèces, en fonction des longueurs de leurs arêtes [Ste88b, PHS91]. En se basant sur l'équation (1.3) et en définissant $\omega_i = (1 - 2p_i)$ selon la terminologie de la Fig. 2.5, il montre les résultats suivants :

– Méthodes de *cluster analysis* : toute méthode regroupant les deux espèces x et y ayant la plus petite distance observée est consistante ssi :

$$\omega_5 < \frac{max(\omega_1\omega_2, \omega_3\omega_4)}{max(\omega_1\omega_3, \omega_1\omega_4, \omega_2\omega_3, \omega_2\omega_4)} \ .$$

On peut montrer qu'il s'agit d'un cas particulier de la condition suivante.

– Méthodes MP, de *Compatibilité* et méthodes de distances reposant sur le principe FPM ; ces méthodes sont consistantes ssi :

$$\omega_5 < min\left\{\frac{\omega_1\omega_2 + \omega_3\omega_4}{\omega_1\omega_3 + \omega_2\omega_4}, \frac{\omega_1\omega_2 + \omega_3\omega_4}{\omega_1\omega_4 + \omega_2\omega_3}\right\} \ .$$

Pour des taux d'évolution identiques sur les différentes arêtes, cette condition est toujours vérifiée [HP89, PHS91].

Penny *et al* [PHS91] affirment aussi que la méthode MV est toujours consistante sur quatre espèces [PHS91]. Cette affirmation pose problème car, d'une part, elle n'est pas prouvée, et d'autre part, on sait que la méthode MV est non-étudiable en général [Gas97a]. Felsenstein [Fel78a] a lui aussi annoncé qu'on pouvait montrer la consistance de MV dans le cas du modèle NCF ou de modèles voisins, toutefois il n'a jamais publié ce résultat.

Conditions sur la distance L_∞ entre $\hat{D}$ et D

Dès lors que la méthode MP a été montrée inconsistante sous les diverses conditions données ci-dessus, cette méthode a cessé d'être si populaire, et ce, au profit des méthodes de distances, que l'on sait consistantes, sous l'hypothèse que le modèle de l'évolution qu'elles utilisent soit correct. Lorsque le modèle utilisé par une méthode est effectivement correct, on a la garantie que cette méthode retrouve la phylogénie correcte avec une probablité de 1 lorsqu'un nombre infini de caractères est disponible. Toutefois, un nombre infini de caractères n'est pas forcément requis pour que les méthodes retrouvent la phylogénie correcte avec une probabilité de 1 : ceci est possible dès que $\hat{D}$ et D sont suffisamment *proches*. Plusieurs travaux récents se sont intéressés au degré de proximité nécessaire entre $\hat{D}$ et D pour que différentes méthodes de distances retrouvent la phylogénie correcte avec une probabilité de 1. Ces travaux sont basés sur la norme L_∞, définie entre deux matrices de dissimilarités quelconques, D et D', comme la plus grande différence entre le contenu de deux mêmes entrées dans les matrices :

$$L_\infty(D, D') = max_{xy}\left|D_{xy} - D'_{xy}\right| .$$

Erdös *et al* [ESSW97b] ont montré que les méthodes de distances reposant sur le principe FPM retrouvent la phylogénie corecte t dès que $L_\infty(\hat{D}, D) < \frac{l(e)}{2}$, où $l(e)$ est la longueur de la plus courte arête interne de t. Les méthodes *Addtree* et NJ possèdent aussi la même garantie [Att97b]. En revanche, la méthode du $SinglePivot$ (SP) [ABF+96] ne possède pour l'instant de garantie de retrouver la phylogénie correcte que si $L_\infty(\hat{D}, D) < \frac{l(e)}{8}$; et l'on sait qu'une telle garantie ne peut être étendue aux cas $L_\infty(\hat{D}, D) \leq \frac{l(e)}{6}$, en raison d'un contre-exemple donné par Erdös *et al* [ESSW97b]. Autrement dit, il existe des jeux de données pour lesquelles NJ, *Addtree* et les méthodes basées sur FPM, retrouvent la phylogénie corecte, alors que SP non.

On sait aussi que pour toute distance d'arbre $D = t$, il existe une matrice quelconque $\hat{D}$ et une distance d'arbre $D' = t'$ tels que $L_\infty(\hat{D}, D) = \frac{l(e)}{2}$ et $L_\infty(\hat{D}, D') = \frac{l(e)}{2}$ [ESSW97b, Att97c]. Ceci caractérise donc le rayon le plus large autour de t, et au sens de L_∞, dans lequel on peut garantir qu'une méthode raisonnable retrouve la phylogénie corecte avec une probabilité de 1. En effet, pour toute matrice de données $\hat{D}$, dès que $L_\infty(\hat{D}, D) > \frac{l(e)}{2}$, t n'est plus la phylogénie (unique) la plus proche de $\hat{D}$ au sens de L_∞. Ce dernier résultat implique que les méthodes de distances basées sur FPM ainsi que les méthodes *Addtree* et NJ ont des garanties de retrouver la phylogénie corecte de façon sûre, aussi larges que possible au sens de L_∞ !

Le côté paradoxal des résultats basés sur L_∞, est que toute méthode *exacte* pour le problème L_∞ (*i.e.*, inférant la phylogénie la plus proche de $\hat{D}$ au sens de L_∞) n'a pour l'instant été montrée garantie retrouver la phylogénie corecte avec une probabilité de 1 que lorsque $L_\infty(\hat{D}, D) < \frac{l(e)}{4}$ [ESSW97b] ! Les avis divergent sur le fait qu'on puisse montrer pour une telle méthode exacte le même résultat que pour les méthodes de distances simples citées ci-dessus. Quoiqu'il en soit, on sait déjà que le résultat ne peut être meilleur.

Au cours de l'étude de la méthode Q^* (chap. 2, part. III), nous reviendrons plus en détails sur les résultats au sens de la distance L_∞.

2.3.4 Puissance statistique

Très récemment aussi, un certain nombre de travaux ont abordé le problème de la *puissance statistique* des méthodes, *i.e.*, de leur **vitesse (ou taux) de convergence**, définie comme le nombre k de caractères nécessaires à une méthode pour retrouver la phylogénie correcte avec une forte probabilité.

Résultats empiriques

Les travaux de Hillis et Huelsenbeck [HH93, HHS94] sont à l'origine de la discussion actuelle sur la puissance des méthodes. Ces auteurs font remarquer qu'une méthode *consistante* n'est d'aucune utilité si elle requiert un nombre exponentiel de caractères pour retrouver la phylogénie corecte avec une probabilité acceptable. Cette remarque est étayée par des simulations, où la vitesse de convergence de diverses méthodes est étudiée pour des phylogénies à quatre espèces, et pour divers modèles d'évolution. Le résultat principal (toutefois attendu) est que la méthode MP, quand elle est consistante, converge souvent plus rapidement vers la phylogénie correcte que les méthodes de distances. C'est ce qui fait que la parcimonie est encore très utilisée en pratique : la longueur des séquences disponibles (après alignement) est, le plus souvent, la ressource critique, conduisant à privilégier une méthode qui requiert un minimum de caractères pour avoir de bonnes chances de retrouver la phylogénie correcte. Ces auteurs montrent aussi que les phylogénies soumises au problème d'attraction des longues branches sont difficiles à reconstuire même pour les méthodes de distances : celles-ci demandent un nombre irréaliste de caractères pour avoir ne serait-ce qu'une chance sur deux de retrouver la phylogénie correcte [HH93, HHS94].

Résultats théoriques

De cet intérêt accru pour la puissance des méthodes, ont découlé un certain nombre de résultats théoriques, donnant des *bornes inférieures* sur les taux de convergence de diverses méthodes de distances. Pour obtenir de tels résultats, on se fixe d'abord une métrique entre les phylogénies, pour laquelle on montre ensuite que la phylogénie inférée par la méthode étudiée se rapproche de la phylogénie correcte, quand le nombre de caractères disponibles augmente.

Farach et Kannan [FK96] ont été les premiers à caractériser la vitesse de convergence d'une méthode ; ils introduisent pour cela une distance *variationnelle* entre les phylogénies, dans le cadre du modèle NCF : soit t une phylogénie à n feuilles évoluant sous le modèle NCF, pour $v \in \{0,1\}^n$ on peut définir la probabilité $\mathcal{P}_t(v)$ que t engendre v aux feuilles (le i^{eme} bit de v représente l'état à la feuille i). Farach et Kannan définissent la distance variationnelle $V(t,t')$ entre deux phylogénies t et t' comme

$$V(t,t') = ||t-t'||_1 = \sum_{x\in\{0,1\}^n} |\mathcal{P}_t(x) - \mathcal{P}_{t'}(x)| \ .$$

Sur la base de cette distance, ils montrent que la méthode SP [ABF$^+$96] a un taux de convergence différant seulement polynomialement du meilleur taux de convergence possible. Le taux de convergence obtenu pour SP est polynômial en n pour des conditions d'évolution raisonnables.

Le modèle de l'évolution NCF a aussi été employé en conjonction avec la métrique L_∞ par Erdös *et al* [ESSW97b, ESSW97a], pour montrer (par l'intermédiaire de l'inégalité de Azuma-Hoeffding) la vitesse de convergence de leurs méthodes, SQM et DCM :

$$k \ \in \ O\left(\frac{\log n}{f^2(1-2g)^{min(d,4p-5)}}\right) \ , \tag{2.2}$$

où pour toute arête e de la phylogénie correcte t, $p(e) \in [f,g]$, d représente le diamètre[1] de t, et p sa profondeur[2]. Ces auteurs donnent aussi (avec l'aide de

1. nombre maximum d'arêtes entre deux feuilles.

2. nombre maximum d'arêtes entre un noeud interne et la feuille la plus proche ; notion aussi connue sous le nom de *rang* de la phylogénie.

Farach et Kannan) la vitesse de convergence de la méthode SP au sens de L_∞ :

$$k \in O\left(\frac{\log n}{f^2(1-2g)^d}\right) . \tag{2.3}$$

En ce qui concerne les méthodes usuelles de R.P., Atteson [Att97c] a montré une borne pour le taux de convergence des méthodes NJ et *Addtree*, du même ordre que celle de SP, *i.e.*, dépendant elle aussi du diamètre de la phylogénie (mais toutefois meilleure d'un facteur ≈ 16).

La principale différence entre des vitesses de convergence (2.2) et (2.3) montrées pour les différentes méthodes tient au fait que dans un cas la borne repose sur le diamètre de la phylogénie et que dans l'autre cas elle repose sur sa profondeur. Si on quitte la théorie pour revenir à la pratique, cette distinction entre la profondeur et le diamètre de la phylogénie n'a pas vraiment lieu d'être. Ces deux notions ne sont employées que pour borner la plus grande distance évolutive D_{max} entre deux espèces considérées conjointement par la méthode de reconstruction. Or, en R.P., la plus grande distance évolutive considérée entre deux espèces étudiées (même à l'opposé l'une de l'autre dans la phylogénie) est bornée par une petite constante (p. ex. $D_{max} < 1$ [Nei91])[3]. Des distances trop élevées traduisent une forte saturation et suggerent un rapport signal/bruit faible dans les données. De plus, si le nombre réel de substitutions par site est trop élevé, les estimations des distances évolutives ne sont pas très précise (grande variance), ce qui conduit à des erreurs dans la reconstruction [Gal98]. Considérer que D_{max} est bornée par une petite constante, implique que toutes les méthodes mentionnées ci-dessus ont une vitesse de convergence du même ordre.

La longueur f de la plus petite arête de la phylogénie est un facteur pénalisant les méthodes en pratique : plus f est faible, et plus la phylogénie a des chances importantes de sortir de la boule de consistance des diverses méthodes de reconstruction (la phylogénie est plus sensible au problème d'attraction des longues branches). Vis-à-vis du facteur f, les vitesses de convergence évoquées ci-dessus, sont toutes du même ordre, *i.e.*, $k > O(\frac{\log n}{f^2})$ (toutefois les constantes de complexité peuvent être grandement différentes d'une méthode à l'autre). Pour la plupart des jeux de données étudiés en pratique, on peut raisonnablement poser que la plus petite longueur d'arête est inversement proportionnelle au nombre d'espèces de la phylogénie, ce que l'on peut noter $f = O(1/n)$ par abus de notation. Dans un tel cas, la vitesse de convergence est polynomiale. On sait par ailleurs

3. lorsque les distances sont caculées en tenant compte de la variabilité des vitesses de substitution entre sites, la valeur de D_{max} est quelque peu augmentée.

que toute méthode déterministe ou probabiliste, reposant sur n'importe quel modèle de l'évolution, doit disposer d'au-moins $k > \Omega(\log n)$ caractères [ESSW97b] [4]. Ainsi les bornes supérieures obtenues pour les méthodes *Addtree*, NJ, SQM, SP semblent à première vue relativement "serrées". En fait, il n'en est rien : quand on instancie les expressions analytiques exactes de ces bornes selon les paramètres de phylogénies réalistes, le nombre k de caractères indiqué pour que la phylogénie correcte soit retrouvée avec une probabilité élevée est le plus souvent irréaliste. Ceci résulte du fait que les bornes citées ci-dessus sont obtenues par une analyse basée sur le pire des cas (plus précisément, comme une combinaison de pire des cas, p. ex. l'inégalité de Azuma-Hoeffding).

2.3.5 Robustesse

La dernière propriété que nous évoquerons ici est la *robustesse* d'une méthode de reconstruction, *i.e.*, le fait qu'une méthode soit consistante, même quand plusieurs hypothèses qu'elle fait sur l'évolution sont invalidées. Par exemple, la plupart des méthodes font l'hypothèse que les sites moléculaires sont distribués identiquement et indépendamment (hypothèse "*i.i.d.*") ; on sait pourtant de façon sûre que ce n'est pas toujours le cas (p. ex. biais en `G+C` [Gal97], évolution conjointe de sites voisins [WH88], etc). La robustesse des méthodes a été encore très peu étudiée, seules quelques études expérimentales ont été conduites récemment [Nei91, HH92, KF94, RW97, RSWY97], certaines tendant à montrer que les méthodes MP et MV sont plus robustes que les méthodes de distances.

2.3.6 Facteurs d'approximation

La plupart des problèmes de R.P. sont NP-difficiles, comme nous l'avons vu tout au début de ce chapitre, ce qui explique que bon nombre de méthodes de R.P. heuristiques aient été développées. Les plus intéressantes sont généralement regroupées dans des *packages* [Swo90, Fel93]. Une propriété importante, qui permet de distinguer les heuristiques les plus intéressantes, est la preuve d'une borne d'approximation garantie. On cherche ici à connaître par quelle fraction, soit $\rho = \frac{l_h}{l_{opt}} \geq 1$, on peut diviser dans le pire des cas la longueur l_h de la phylogénie inférée par l'heuristique pour obtenir la longueur l_{opt} de la phylogénie la plus courte (*i.e.*, la phylogénie optimale). Bien sûr, plus ρ est faible, plus l'heuristique est de qualité.

4. Par exemple, si l'on dispose de caractères binaires, pour pouvoir coder de façon distincte les $(2n-5)!!$ phylogénies différentes qui existent pour n espèces, depuis n séquences de k caractères, il faut forcément $(2n-5)!! < 2^{kn}$, *i.e.*, $k > \Omega(\log n)$ caractères.

Un certain nombre de problèmes de R.P. se définissent comme des variantes du problème de l'arbre de Steiner (comme MDS et MP) et bénéficient ainsi des facteurs d'approximation connus pour ce problème dans différents espaces métriques. Par exemple, la variante de MP où les espèces sont décrites par k caractères binaires (ayant 0 et 1 pour état) correspond à la variante *rectilinéaire* du problème de Steiner, où l'espace est considéré comme une grille qu'il faut suivre pour passer d'un point à l'autre [Gus91b]. Plus précisément, l'espace métrique considéré, noté $(2^k, d_M)$, est l'hypercube à k dimension (chacune ne possédant que 0 et 1 comme valeurs possibles) muni de la distance de Manhatan d_M, *i.e.*, la distance entre deux séquences x et y est $d_M(x,y)=\sum_{c=1}^{k} |X_{xc} - X_{yc}|$. Si l'on dispose de séquences moléculaires (à 4 ou 20 états), le problème MP devient équivalent au problème de Steiner dans l'espace métrique (A^k, d_H), où A désigne l'alphabet sur lequel sont définis les caractères (p. ex. $\{\texttt{A},\texttt{C},\texttt{G},\texttt{T}\}$) et d_H désigne la distance de Hamming: $d_H(x,y)=card(\{c\in 1..k \text{ t.q. } X_{xc}\neq X_{yc}\})$ [SR75, Fou84].

Une heuristique célèbre pour le problème de l'arbre de Steiner est de résoudre le problème polynômial de l'arbre recouvrant minimum (on cherche ici aussi l'arbre le plus court, mais cette fois, dont tous les noeuds, même internes, sont attachés à une espèce). On peut facilement montrer que l'arbre inféré en $O(n^2)$ par cette heuristique est t.q. $\rho \leq 2$ (la preuve initiale est dûe à E.F. Moore, selon Gilbert et Pollack [GP68]). Cette borne peut être affinée à $\rho \leq 2\frac{n-1}{n} < 2$ [Gus91b], et s'applique dans le cas des graphes dont les arêtes sont à valeur dans $\mathbb{R}^+$, ainsi que dans n'importe quel espace métrique en général [GP68]. Elle s'applique donc aux variantes de MP mentionnées ci-dessus (ainsi qu'à MDS) [SR75, Fou84, Gus91b].

Pour le problème de Steiner *rectilinéaire* (la grille de dimension p associée à la distance de Manhattan) il existe une conjecture [HRW92] selon laquelle, dans l'espace rectilinéaire à p dimensions, $\rho = \frac{p+1}{p}$, *i.e.*, la borne se rapproche de 1 quand le nombre de dimensions augmente. Ceci laisse présager de très bonnes performances pour les heuristiques de R.P. s'attaquant au problème MP pour des caractères binaires.

L'heuristique SP (de complexité $O(n^2)$) exhibée récemment par Agarwala *et al* [ABF$^+$96] est la première méthode ayant une garantie de performance pour un problème de R.P. du type "ajustement" (cf. section 2.2.2). Il s'agit ici du problème consistant à approcher au plus près, au sens de la norme L_∞, une matrice de distances par une phylogénie. Si on note $\hat{D}$ la matrice de distance, t la phylogénie la plus proche de cette matrice et $\hat{t}$ la phylogénie proposée par l'heuristique, Agarwala *et al* montrent que $L_\infty(\hat{D},\hat{t}) \leq 3L_\infty(\hat{D},t)$. De plus, ils prouvent qu'approcher t à $\frac{9}{8}$ (au lieu du facteur 3) est un problème NP-difficile. Signalons aussi que, récemment, Chepoi et Fichet [CF97] simplifient grandement les preuves des résultats de

Agarwala *et al.*

En résumé, ces différents travaux montrent que des heuristiques efficaces existent pour la plupart des critères de R.P., même si ceux-ci sont NP-difficiles. Soulignons que les résultats théoriques cités ci-dessus sont établis *dans le pire des cas*, et qu'en pratique, les résultats obtenus *en moyenne* par les heuristiques sont bien meilleurs.

Chapitre 3

Motivations et problématique de la thèse

Comme nous l'avons vu au cours du chapitre précédent, la plupart des méthodes de reconstruction sont des méthodes d'optimisation. Les critères qu'elles optimisent sont définis sur la base de principes simples, liés à des hypothèses sur l'évolution. Non seulement, il est vraisemblable que ces principes ne capturent pas entièrement les mécanismes complexes de l'évolution, mais on sait de plus que les hypothèses qu'ils font ne sont pas toujours vérifiées [Yan93, LSHP94, SSH94, Gal97]. Ainsi, ce n'est pas parce qu'on dispose d'heuristiques approximant relativement bien tel ou tel critère de reconstruction que l'on a plus de chances, pour autant, de retrouver la phylogénie inconnue des espèces. Il faut avant tout que le critère de reconstruction optimisé soit consistant pour le jeu de données étudié et, de plus, qu'il ait une vitesse de convergence assez rapide pour que la taille des séquences, dont on dispose, soit suffisamment importante pour retrouver la phylogénie correcte. Dans la grande majorité des cas, cette situation idéale n'est pas atteinte, et la phylogénie correcte ne peut pas être *entièrement* retrouvée. En revanche, on peut raisonnablement espérer retrouver la *partie* de cette phylogénie sur laquelle les données possèdent des informations (généralement, les arêtes internes longues).

Le problème est que les méthodes usuelles de reconstruction infèrent des phylogénies totalement résolues (binaires), pour des raisons liées au critère qu'elles cherchent à optimiser (en quelques mots, l'ajout de toute nouvelle arête à une phylogénie partiellement résolue ne peut qu'améliorer sa valeur pour le critère optimisé). En conséquence, la phylogénie proposée par une de ces méthodes, notée $\hat{t}$, contient souvent un certain nombre d'arêtes incorrectes et il est clair que la phylogénie partiellement résolue comportant uniquement les arêtes correctes de $\hat{t}$, (*i.e.*, la phylogénie $\hat{t}$ où les arêtes incorrectes ont été écartées) représente un meilleur estimateur que $\hat{t}$ de la phylogénie inconnue. Cette réflexion constitue le thème conducteur du travail présenté ici : **il est préférable d'inférer une phylogénie partiellement résolue dont les arêtes sont sûres**, *i.e.*, de proposer une estimation stable, plutôt que de proposer une phylogénie complètement résolue mais comportant des arêtes non-sûres, et par là extrêmement variables.

La meilleure preuve de la **grande variabilité des arêtes de la phylogénie $\hat{t}$ complètement résolu** est que l'ajout d'une espèce aux données, ou même l'utilisation d'une autre méthode de reconstruction, fait varier une proportion significative des arêtes de $\hat{t}$ [Fel88, PHS91, LPLG93, KF94, BG96a]. Cette variabilité importante peut, à juste titre, faire douter le biologiste de la phylogénie reconstruite. Autrement dit, une méthode de reconstruction qui serait détectée comme étant trop variable, ne serait d'aucune utilité au biologiste, puisqu'il ne pourrait *a priori* faire confiance à aucune des arêtes inférées, se demandant pour chacune si elle est uniquement spécifique du jeu de données ou, si elle sera retrouvée depuis d'autres jeux de données.

Comme nous le verrons plus tard, proposer des phylogénies partiellement résolues permet effectivement de réduire la variabilité des estimations, *i.e.*, d'augmenter la stabilité de l'estimation. Il ne faut toutefois pas tomber dans l'autre extrême et proposer des phylogénies trop peu résolues : une méthode proposant toujours la phylogénie étoile (*i.e.*, la phylogénie la moins résolue, n'induisant aucun groupement d'espèces) n'est d'aucun intérêt pour le biologiste, puisqu'elle n'a aucun pouvoir de résolution. **Il y a une juste limite à trouver entre degré de résolution et stabilité**. Comme souvent en théorie de l'estimation, ces deux notions sont liées à la complexité du modèle proposé (ici, du degré de résolution de la phylogénie inférée). Jusqu'à présent, les recherches entreprises en R.P. ont porté essentiellement sur le cas "extrême" des phylogénies complètement résolues. Ceci résulte simplement du fait que la phylogénie inconnue est supposée complètement résolue, et que l'on a naturellement tendance à suivre son modèle. Au cours de cette thèse, nous envisageons l'inférence de phylogénies non-totalement résolues visant à obtenir des estimations plus robustes de la phylogénie des espèces.

Pour inférer des phylogénies (partiellement résolues) ne possédant que des arêtes fiables, on peut adopter deux stratégies. L'une, qualifiée de stratégie **à résolution croissante** (ou encore de *forward*), consiste à partir d'un degré nul de résolution et à ajouter des arêtes jusqu'à ce que plus aucune arête, dont la qualité soit satisfaisante au sens d'un certain critère, ne puisse être ajoutée. La stratégie inverse, dite **à résolution décroissante** (ou *backward*), consiste à partir d'une phylogénie complètement résolue, dont on élimine les arêtes les moins sûres jusqu'à un certain point. La partie suivante de cette thèse présente deux méthodes apparentées à cette stratégie. Enfin la troisième partie abordera le cas de la stratégie à résolution croissante.

Deuxième partie

Une approche statistique

Introduction

Dans cette partie, nous adoptons une *approche statistique* de la R.P. : c'est pourquoi nous aurons recours à des outils de rééchantillonnage et d'analyse statistiques. Comme outil de rééchantillonnage, nous considèrerons essentiellement la technique du bootstrap. Comme outils d'analyse, nous aurons recours aux notions d'erreur de 1ère espèce et de 2ème espèce, de biais et de variance.

Dans le chapitre 1 de cette partie, nous présentons la technique du bootstrap telle qu'elle fut introduite en R.P. par Felsenstein [Fel85]. Cette technique permet d'attribuer une valeur de bootstrap, *bp*, à toute arête de la phylogénie inférée. L'apport de ce chapitre se répartit principalement sur deux points :

- Nous montrons (sur la base de simulations) qu'une certaine utilisation du principe de bootstrap permet d'obtenir des estimations plus précises de la phylogénie inconnue que les estimations obtenues usuellement par l'emploi de diverses méthodes de reconstruction. Le problème principal consiste à choisir la valeur de bootstrap à partir de laquelle les arêtes de la phylogénie inférée sont jugées fiables.

- Pour analyser les résultats des simulations, nous montrons une décomposition de l'espérance d'erreur d'un estimateur quelconque en termes de biais et de variance. L'importance de cette décomposition réside dans son indépendance vis-à-vis du domaine considéré (ici la R.P.) : elle peut être appliquée telle quelle à tout domaine où l'on cherche à estimer un processus arboré.

Dans le chapitre 2, nous proposons une deuxième façon d'utiliser la technique du bootstrap en R.P. La procédure bootstrap qui en résulte a pour originalité de laisser les données fixer le degré de résolution *adéquat* en fonction de la quantité d'information qu'elles contiennent. Dans cette approche, le nombre de paramètres à fixer est moins important que dans l'approche précédente, ce qui laisse présager, pour une même précision des résultats, un temps de calcul plus court.

Chapitre 1

La technique du bootstrap en reconstruction phylogénétique

Nous effectuons dans un premier temps un rapide survol des études concernant la technique du bootstrap en R.P (section 1.1). Le problème principal qui s'en dégage est le choix d'un seuil au-delà duquel les arêtes de la phylogénie inférée doivent être raisonnablement écartées. Dans la section 1.2 nous donnons ensuite les mesures d'erreur employées pour rendre compte de l'erreur commise par une phylogénie inférée pour estimer la phylogénie correcte. Puis nous montrons (section 1.3) comment on peut approximer le seuil optimal de sélection des arêtes dans un arbre bootstrap. Enfin, nous présentons les résultats empiriques de l'arbre bootstrap obtenu par approximation du seuil optimal (section 1.4).

1.1 État de l'art sur le bootstrap

Pour estimer la fiabilité des arêtes d'une phylogénie inférée par une méthode de reconstruction, on peut avoir recours à une **technique de rééchantillonnage** des caractères. La technique du **bootstrap**, dûe à Efron [Efr79], est la technique de rééchantillonnage des données la plus utilisée en R.P., depuis son introduction dans le domaine par Felsenstein [Fel85]. Étant donnée la matrice X de k caractères (colonnes) sur n espèces (les lignes) dont on dispose initialement, la technique bootstrap procède par rééchantillonnage des colonnes de X. Plus précisément une matrices X^* de même taille que X est engendrée, ses lignes référençant les mêmes espèces que X (et dans le même ordre) et ses colonnes étant obtenues par *tirage aléatoire avec remise des caractères de* X. Comme le tirage s'effectue avec remise, certains caractères de X sont présents plusieurs fois dans la matrice X^* alors que d'autres en sont absents. Un grand nombre b de jeux bootstrappés X^* sont ainsi générés (généralement de 100 à 2000). Ces jeux sont aussi appelés **pseudojeux de données** par contraste avec le jeu de données initial X.

Le jeu de données X et les pseudojeux X^* sont séparément analysés par une même méthode de reconstruction, ce qui aboutit à la production de la phylogénie $\hat{t}$ et d'autant d'**arbres bootstrap** $\hat{t}^*$ que de pseudojeux. La phylogénie $\hat{t}$ est aussi appelée **arbre direct**, puisque, contrairement aux phylogénies $\hat{t}^*$, elle est obtenue depuis le jeu de données initial.

Pour chaque arête de $\hat{t}$, on obtient une **proportion bootstrap** (notée bp) en relevant le pourcentage de fois où elle apparaît dans les différentes phylogénies $\hat{t}^*$. Nous appellerons **arbre bootstrap moyen** la phylogénie $\bar{t}^*$ de même structure que $\hat{t}$ et dont les arêtes internes sont valuées par les bp (rappelons que seules les arêtes internes nous intéressent).

Sur la base des caractères dont on dispose, la technique du bootstrap simule la variabilité du processus d'échantillonnage par lequel les caractères observés ont été obtenus depuis le génome des espèces. Une propriété fondamentale du bootstrap est que la variabilité de la distribution statistique des caractères obtenue par rééchantillonnage se rapproche asymptotiquement de la variabilité obtenue par échantillonnage direct.

Les bp ont pour but d'estimer la fiabilité des arêtes [Fel85], toutefois la signification statistique précise de cette mesure n'est pas encore établie clairement. Suite aux travaux de Hillis et Bull [HB93], le débat s'est intensifié, et on peut distinguer trois points de vue :

1. Pour Efron [Efr79] et Felsenstein [Fel85], le bootstrap est une **mesure de répétabilité**. L'idée principale est que la distribution de la phylogénie $\hat{t}^*$ (pour les multiples jeux bootstrappés) observée autour de la phylogénie initialement inférée $\hat{t}$ permet d'inférer la distribution inobservable de la phylogénie $\hat{t}$ autour de la phylogénie correcte t. En d'autres termes, quand la bp d'une arête de $\hat{t}$ est élevée, nous pouvons être presque sûrs que cette arête serait inférée à nouveau si l'on disposait d'un autre jeu de données initial et qu'on l'analysait avec la même méthode de reconstruction. Quand la proportion bootstrap est faible, la répétabilité est jugée faible et l'on peut penser que l'arête correspondante est erronée. Considérer la proportion bootstrap comme une mesure de répétabilité conduit ainsi à écarter les arêtes qui ne seraient pas inférées suffisamment fréquemment depuis d'autres jeux de données. Notons que cette signification justifie déjà amplement l'utilisation de la technique du bootstrap. Celle-ci est en effet capable de détecter les cas où

les données contiennent peu de "signal phylogénétique" et où la méthode de reconstruction choisit les arêtes d'une façon presque aléatoire. La procédure bootstrap choisira alors une irrésolution pour ces arêtes, ce qui est nettement mieux qu'une fausse résolution.

2. La seconde interprétation d'une proportion bootstrap [San89, ZL92a, ZL92b, HB93] est plus ambitieuse que la première. Elle consiste à considérer la *bp* comme une **mesure d'exactitude**, *i.e.*, comme un estimateur de la probabilité que l'arête de $\hat{t}$ correspondante appartienne à la phylogénie correcte t. Il est évident que cette interprétation requiert que la méthode de reconstruction soit consistante. Ceci est important car une méthode inconsistante (mais convergente) associée à des données ayant un fort signal phylogénétique (cf. [Fel78a]), proposerait de façon répétée les mêmes mauvaises arêtes ; ces arêtes auraient alors en même temps une forte valeur de bootstrap et une probabilité nulle d'être correctes.

   Les simulations de Hillis et Bull [HB93] pour la méthode MP montrent que pour des conditions réalistes d'évolution, où cette méthode de reconstruction est consistante, les plus hautes *bp* sous-estiment la probabilité que les arêtes soient correctes. Ce comportement avait été prédit par l'étude théorique de Zharkikh et Li [ZL92a, ZL92b] pour le cas de quatre espèces. En réponse à l'article de Hillis et Bull, Felsenstein et Kishino [FK93] ont expliqué ce phénomène sur la base d'un modèle simplifié, tandis que Efron *et al* [EHH96] ont montré que ce phénomène ne pouvait pas être vu comme une propriété de la technique du bootstrap en général.

3. Finalement, la proportion bootstrap peut aussi être assimilée au seuil de confiance d'un test statistique d'hypothèses usuel [FK93, EHH96, ZL95]. Remarquons que cette interprétation d'une *bp* requiert, elle aussi, que la méthode de reconstruction soit consistante. La nuance importante qui est faite par rapport à l'interprétation précédente d'une *bp* est qu'on parle ici de probabilité conditionnelle d'une décision, *i.e.*, que l'arête soit correcte sous une certaine hypothèse nulle (par exemple que l'arête est incorrecte), au lieu de simplement considérer la probabilité que l'arête soit correcte.

   La difficulté avec cette troisième approche est que la procédure bootstrap initiale, telle qu'elle fut employée par Felsenstein en 1985, n'est pas directement liée au test d'hypothèses au sens statistique habituel [EHH96]. Toutefois, Efron *et al* [EHH96] montrent que la *bp* est en fait une approximation au premier ordre de la variable de décision d'un tel test. De plus, ils proposent une procédure bootstrap plus élaborée (et plus coûteuse en temps calcul) pour obtenir une approximation du second ordre de ce seuil. De même, sur la base d'une analyse simplifiée du problème et du nombre de résolutions alternatives pour une arête, Zharkikh et Li [ZL95] proposent eux aussi de recourir à un double bootstrap, selon une méthode qu'ils appellent le bootstrap "complet et partiel". Ils montrent que les approximations obtenues sont satisfaisantes, d'un point de vue statistique, dans des cas très simples (5 espèces, horloge moléculaire).

Malgré la variété des interprétations détaillées ci-dessus, un point commun apparaît : **écarter les arêtes de $\hat{t}$ ayant une faible *bp* est une pratique bien-fondée**. On s'attend à ce que ces arêtes aient une faible chance d'être correctes et que la plupart des arêtes incorrectes soient détectées de cette façon. On peut ainsi espérer qu'un gain (*i.e.*, une réduction d'erreur) soit obtenu pour estimer la phylogénie inconnue, si au lieu de considérer la phylogénie $\hat{t}$ en entier, on ne considère que ses arêtes ayant une *bp* suffisamment élevée. Dans la suite de la présentation, S t.q. $0 \leq S \leq 1$ dénotera le seuil de sélection utilisé pour distinguer les arêtes conservées des arêtes écartées. $\hat{t}^*_S$ dénotera **l'arbre bootstrap réduit** au seuil S, *i.e.*, la phylogénie $\bar{t}^*$ (ou de façon équivalente $\hat{t}$) dont les arêtes ayant une *bp* inférieure à S auront été écartées.

Le seuil S doit être fixé précautionneusement : plus il est haut et plus on a de chances d'écarter des arêtes erronées, mais plus on a aussi de chances d'écarter des arêtes correctes dont la *bp* ne serait pas assez élevée. Ce dilemme se retrouve dans l'expression de l'erreur d'estimation décomposée en **erreur de 1ère espèce** (ou *faux positif* et **erreur de 2ème espèce** (ou faux négatif).

En effet, on commet une erreur de 1ère espèce à chaque fois qu'on accepte une arête incorrecte et on commet une erreur de 2ème espèce à chaque fois qu'on rejette une arête correcte. Ainsi, plus le seuil S est élevé plus le nombre d'erreurs de 1ère espèce diminue, mais plus le nombre d'erreurs de 2ème espèce augmente. En conséquence, le seuil optimal de sélection des arêtes dépendra du poids relatif accordé à chacun de ces types d'erreur dans la mesure d'erreur totale. Dans la section suivante nous formalisons cette notion d'erreur d'estimation et introduisons une autre décomposition possible de cette erreur, en terme de biais et de variance. Nous montrons ensuite comment on peut fixer le seuil S de sélection des arêtes pour minimiser l'erreur de la phylogénie $\hat{t}^*_S$, en fonction de l'expression retenue pour l'erreur d'estimation.

1.2 Mesures d'erreur

Rappelons que nous ne nous intéressons dans cette étude qu'à la structure des phylogénies, aussi aurons-nous recours uniquement à des mesures d'erreur structurelle, *i.e.*, basées sur les différences structurelles entre phylogénies.

1.2.1 Décomposition de l'erreur en termes de 1ère et de 2ème espèce

La décomposition de l'erreur d'estimation en termes de 1ère et de 2ème espèce est liée à la distance de Robinson et Foulds [RF81, Ste88a], comme nous allons le voir. Rappelons que toute arête d'une phylogénie partitionne les espèces en deux ensembles disjoints, et que toute phylogénie peut ainsi être considérée comme un ensemble de bipartitions. On peut donc comparer deux phylogénies sur la base des bipartitions qu'elles induisent. Plus particulièrement, nous nous intéresserons uniquement aux bipartitions non-triviales, puisque les phylogénies que nous comparerons seront définies sur un même ensemble d'espèces, induisant ainsi les mêmes bipartitions non-triviales. La distance de Robinson et Foulds entre deux phylogénies t_1 et t_2 est définie comme le nombre minimum de contractions d'arêtes[1] et d'extensions de noeuds[2] qu'il est nécessaire d'effectuer pour passer de la structure de t_1 à la structure de t_2. Cette distance est symétrique et on montre aussi qu'elle est égale au nombre de bipartitions induites par t_1 qui ne sont pas induites par t_2, plus le nombre de bipartitions induites par t_2 qui ne sont pas induites par t_1 [RF81]. Dans notre cas, nous nous intéressons à la distance d'une phylogénie inférée $\hat{t}$ (ou $\hat{t}^*_S$) à la phylogénie correcte t. L'erreur de 1ère espèce est donc

$$e_1(\hat{t}) = \left|\{b \in \hat{t} \ / \ b \notin t\}\right| ,$$

où b dénote une bipartition. L'erreur de 2ème espèce est définie symétriquement par

$$e_2(\hat{t}) = \left|\{b \in t \ / \ b \notin \hat{t}\}\right| .$$

Par exemple, si l'on dispose de 10 espèces, une phylogénie complètement résolue (binaire) induit 7 bipartitions non-triviales. Comme la phylogénie correcte est habituellement supposée complètement résolue, e_1 et e_2 prennent leur valeur dans $\{1, \ldots, 7\}$. Notons que si $\hat{t}$ est complètement résolue, elle inclut le même nombre

1. Une arête est contractée en confondant ses deux extrémités dans la phylogénie.
2. Un noeud est étendu en le remplaçant par deux noeuds reliés entre eux se partageant son ancien voisinage.

de bipartitions non-triviales que t, ce qui entraine $e_1(\hat{t}) = e_2(\hat{t})$: chaque mauvaise arête qu'elle propose implique qu'elle ne pourra inférer l'une des arêtes de t. Notons aussi qu'en termes de bipartitions $B(\hat{t}_S^*) \subseteq B(\hat{t})$, donc nous aurons toujours par la suite $e_1(\hat{t}) \geq e_1(\hat{t}_S^*)$ et $e_2(\hat{t}) \leq e_2(\hat{t}_S^*)$.

Les deux termes d'erreur e_1 et e_2 peuvent être combinés de plusieurs façons pour exprimer l'erreur d'estimation totale. La combinaison la plus simple consiste à leur donner le même poids, ce qui correspond à la distance de Robinson et Foulds :

$$e_{RF}(\hat{t}) = e_1(\hat{t}) + e_2(\hat{t}) \ .$$

Les valeurs prises par e_{RF} quand on dispose de 10 espèces sont donc comprises dans $\{1, \dots, 14\}$.

Une généralisation possible consiste à attribuer des poids différents aux termes e_1 et e_2. En R.P., on accorde généralement plus d'importance au terme e_1 puisque oublier d'inférer une arête correcte est préférable à l'inférence d'une mauvaise arête. Ce point de vue conduit à exprimer l'erreur totale comme

$$\lambda\, e_1(\hat{t}) \ + \ e_2(\hat{t}) \quad \text{où} \quad \lambda \geq 1 \,,$$

traduisant le fait qu'on donne λ fois plus d'importance à l'erreur de 1ère espèce qu'à l'erreur de 2ème espèce. Cependant, pour normaliser l'étendue des valeurs prises par l'erreur d'estimation généralisée, notée $\hat{e}_{RF}^{\lambda}$, nous utiliserons l'expression

$$e_{RF}^{\lambda}(\hat{t}) = \frac{2}{\lambda+1} \Big(\lambda\, e_1(\hat{t}) + e_2(\hat{t}) \Big) \ . \tag{1.1}$$

Ainsi l'erreur d'estimation varie dans le même intervalle que $e_1 + e_2$ (p. ex. , est comprise entre 0 et 14 pour 10 espèces) et la distance de Robinson et Foulds [RF81] est obtenue pour $\lambda = 1$. Dans la suite nous considèrerons des valeurs de λ prises dans $[1, 10]$. Des valeurs plus élevées, avec $\lambda = \infty$ pour extrême, rendraient négligeable l'erreur de 2ème espèce et le meilleur estimateur de la phylogénie inconnue t serait alors invariablement la phylogénie étoile, ne proposant aucun regroupement d'espèces.

1.2.2 Décomposition de l'erreur en termes de biais et de variance

Pour déterminer la qualité d'un estimateur d'une certaine statistique, la variance et le biais sont des indicateurs importants. Le biais mesure la distance entre la valeur moyenne des estimations et la valeur correcte à estimer. La variance mesure la dispersion des différentes estimations autour de cette valeur moyenne. Les expressions analytiques de ces indicateurs sont bien connues dans le cas d'estimateurs usuels (moyenne, variance empirique, etc), et sont généralement obtenues dans le cadre d'une métrique, p. ex. L_2 (somme des carrés des écarts). Toutefois, dans le cas des arbres, les termes de variance et de biais sont plus difficiles à déterminer et il est vraisemblable qu'ils puissent être définis de plusieurs manières différentes.

Kuhner et Felsenstein [KF94] ont proposé récemment une première façon de mesurer le biais des méthodes de reconstruction. Soit t la phylogénie qu'on cherche à retrouver, leur idée est de considérer qu'une méthode est biaisée pour t, si le nuage des phylogénies $\hat{t}$, inférées par la méthode depuis les différents jeux de données, n'est pas détecté comme centré autour de la phylogénie t. Cette situation est décelée en comparant $\bar{d}_{RF}(\hat{t}, t)$, la distance moyenne de Robinson et Foulds [RF81] entre les diverses phylogénies $\hat{t}$ et la phylogénie correcte t (autrement dit $\bar{e}_{RF}(\hat{t})$), et $\bar{d}_{RF}(\hat{t}, \hat{t}')$, la distance moyenne au sens de [RF81] entre deux phylogénies $\hat{t}$ et $\hat{t}'$ proposées par la méthode. Si $\bar{e}_{RF}(\hat{t})$ est observée plus grande que $\bar{d}_{RF}(\hat{t}, \hat{t}')$ alors le nuage de points est jugé comme non centré sur la phylogénie correcte et la méthode est dite biaisée pour cette phylogénie.

Cette mesure de biais n'est pas adaptée au cas de méthodes inférant des phylogénies partiellement résolues, comme par exemple la procédure bootstrap détaillée plus haut, proposant $\hat{t}^*_S$ comme estimation de t. La raison en est que ces phylogénies comportent en moyenne moins d'arêtes que la phylogénie correcte, *i.e.*, $B(\hat{t}^*_S) \leq B(t)$. En conséquence, $\bar{e}_2(\hat{t}^*_S)$, et donc $\bar{e}_{RF}(\hat{t}^*_S)$ est supérieure à $|B(t) - B(\hat{t}^*_S)|$. Cette augmentation artificielle de $\bar{e}_{RF}(\hat{t}^*_S)$ conduit bien souvent à la situation où $\bar{e}_{RF}(\hat{t}^*_S) > \bar{d}_{RF}(\hat{t}, \hat{t}')$, *i.e.*, où la méthode est biaisée, même si les arbres bootstrap $\hat{t}^*_S$ sont centrés autour de t. Dans un tel cas, la mesure de Kuhner et Felsenstein perd de son intérêt, puisqu'elle mesure avant tout le biais structurel, information dont on est déjà conscient par le fait de proposer un modèle plus simple (une phylogénie partiellement résolue) que la variable estimée (la phylogénie des espèces, totalement résolue). Dans la suite nous présentons une nouvelle mesure de biais, pouvant être appliquée à des méthodes proposant des phylogénies partiellement résolues aussi bien qu'à des méthodes proposant des phylogénies complètement résolues.

En théorie statistique de l'estimation, la notion de biais est naturellement liée à celle de variance. Par exemple, si l'on considère l'erreur aux moindres carrés comme fonction de coût, un résultat bien connu est que l'espérance de l'erreur de tout estimateur numérique est la somme d'un terme de biais et d'un terme de variance. Une règle empirique est que l'importance de ces deux termes doit être suffisamment équilibrée si l'on veut minimiser l'espérance de l'erreur totale. Une importante littérature a été consacrée aux estimateurs sans biais (et de variance minimum). Toutefois, il est bien connu ([KS73], p.21-22) que de meilleurs estimateurs (*i.e.*, ayant une espérance d'erreur plus faible) peuvent être obtenus en acceptant un certain biais. Par exemple, considérons une variable aléatoire normale x d'espérance nulle et de variance inconnue (*i.e.*, $x \approx \mathcal{N}(0, \sigma)$). Le meilleur estimateur sans biais de la variance σ^2 est $\sum x_i^2/s$, où s est la taille de l'échantillon. Cependant, il peut être facilement montré que le meilleur estimateur, même s'il est légèrement biaisé, est $\sum x_i^2/(s+2)$. Cet estimateur est plus performant que le premier en raison d'une plus faible variance ([KS73], p34).

Le même phénomène peut être attendu en terme d'estimation d'arbre, et plus précisément de phylogénies. Étant donnée une fonction de coût, ici e_{RF} ou plus généralement e_{RF}^{λ}, on peut effectivement montrer une décomposition similaire en termes de biais et de variance.

Soit t la phylogénie sur n espèces qu'on cherche à estimer. Soit $n_t = |B(t)|$ le nombre de bipartitions non-triviales qu'elle induit sur les espèces. Supposons que nous disposons de plusieurs jeux de données de même taille, issus, par un certain processus d'échantillonnage, de la distribution des caractères induite par la phylogénie t ; et soit une certaine méthode de reconstruction à laquelle on s'intéresse. Pour chaque jeu de données analysé, cette méthode produit une nouvelle estimation $\hat{t}$ de t. Notre but est de caractériser la distribution des phylogénies $\hat{t}$. Soit $\mathcal{B}$ l'ensemble de toutes les bipartitions non-triviales sur n espèces. Le fait de considérer chaque phylogénie $\hat{t}$ du point de vue de son ensemble de bipartitions $B(\hat{t})$ nous amène à introduire l'ensemble des événements associés aux bipartitions de $\mathcal{B}$ (précisons que dans la suite de cette section nous abrègerons $b \in B(\hat{t})$, resp. $b \in B(t)$, en $b \in \hat{t}$, resp. $b \in t$, en raison de la fréquence élevée des occurrences de ces expressions). L'**événement** $\{b \in \hat{t}\}$ est vrai quand la bipartition b est représentée par une arête de $\hat{t}$, et faux sinon. On peut maintenant définir le **poids** $\mathcal{P}(b)$ d'une bipartition b comme la probabilité de l'événement $\{b \in \hat{t}\}$, *i.e.*,

$$\mathcal{P}(b) = Pr(\{b \in \hat{t}\}).$$

Notons que la distribution des poids $\mathcal{P}$ n'est pas une distribution de probabilité sur l'ensemble des événements (comme nous le montrerons un peu plus tard). Soit $I_{\{b \in \hat{t}\}}$ la fonction indicatrice attachée à l'événement $\{b \in \hat{t}\}$, *i.e.*, $I_{\{b \in \hat{t}\}} = 1$ quand l'événement est vrai et $I_{\{b \in \hat{t}\}} = 0$ quand il est faux. En utilisant ces notations, le nombre de bipartitions d'une phylogénie donnée $\hat{t}$ est simplement la somme sur $\mathcal{B}$ des valeurs de toutes ses fonctions indicatrices. De ceci découle que l'espérance de la taille de $\hat{t}$, $E(\left|\hat{t}\right|)$, est

$$E\left(\left|\,\hat{t}\,\right|\right) = E\left(\sum_{b\in\mathcal{B}} I_{\{b\in\hat{t}\}}\right) = \sum_{b\in\mathcal{B}} E(I_{\{b\in\hat{t}\}}) = \sum_{b\in\mathcal{B}} \mathcal{P}(b)\,, \tag{1.2}$$

i.e., est simplement la somme de tous les poids $\mathcal{P}(b)$. Ceci illustre que $\mathcal{P}$ n'est pas une distribution de probabilité ($\sum \mathcal{P}(b) \neq 1$). En utilisant ces poids, on peut exprimer les espérances des erreurs de 1ère et de 2ème espèces, notées $E\left(e_1(\hat{t})\right)$ et $E\left(e_2(\hat{t})\right)$:

- l'erreur de 1ère espèce est la somme des fonctions indicatrices des événements $\{b \in \hat{t} \wedge b \notin t\}$, d'où

$$E\left(e_1(\hat{t})\right) = E\left(\sum_{b\in\mathcal{B}-t} I_{\{b\in\hat{t}\}}\right) = \sum_{b\in\mathcal{B}-t} \mathcal{P}(b)\,. \tag{1.3}$$

- L'erreur de 2ème espèce est la somme des fonctions indicatrices des événements $\{b \notin \hat{t} \wedge b \in t\}$, d'où

$$E\left(e_2(\hat{t})\right) = E\left(\sum_{b\in t} \left(1 - I_{\{b\in\hat{t}\}}\right)\right) = n_t - \sum_{b\in t} \mathcal{P}(b)\,. \tag{1.4}$$

En utilisant la distribution $\mathcal{P}$ des poids, on peut aussi définir des termes de biais et de variance. Pour ceci nous avons d'abord besoin d'une notion de centralité de la distribution, $\mathcal{M}$, que nous définissons comme étant la somme des n_t poids les plus importants. Autrement dit, $\mathcal{M}$ est égale à la somme des poids des n_t bipartitions les plus probables, et représente en ce sens la partie centrale de la distribution.

Nous pouvons maintenant définir un terme de **variance** (représentant plus généralement un indicateur de dispersion, de variabilité[3]) pour la méthode de reconstruction. Ce terme est défini comme la tendance de la méthode de reconstruction à inférer quelque chose de différent du point central de la distribution :

$$V = \sum_{b\in\mathcal{B}} \mathcal{P}(b) - \mathcal{M}\,.$$

Une méthode qui propose toujours les mêmes bipartitions, quel que soit le jeu de données, a un terme de variance V nul. A l'autre extrême, une méthode inférant

3. nous employons le terme de “variance” pour bien marquer l'opposition biais-variance.

toutes les bipartitions avec la même probabilité induit une distribution uniforme sur les poids. Cette méthode produit donc des résultats ayant une variance maximum, puisque les n_t poids sommés dans $\mathcal{M}$ ne représentent qu'une partie négligeable de la distribution, composée d'un nombre exponentiel en n de bipartitions de même poids.

Nous pouvons définir un terme de **biais**, A, exprimant l'**acentralité** de la phylogénie correcte t (considérée comme un ensemble de bipartitions) dans la distribution des bipartitions inférées, de la façon suivante :

$$A = \mathcal{M} - \sum_{b \in t} \mathcal{P}(b)\,.$$

Si la phylogénie $\hat{t}$ proposée par la méthode est régulièrement proche de t, les bipartitions de t ont des poids élevés et contribuent ainsi de façon significative au terme $\mathcal{M}$, si bien que le terme A est faible. Si, au contraire, la phylogénie inférée contient rarement des bipartitions de t, celles-ci sont marginale dans la distribution des poids, et A est alors proche de $\mathcal{M}$. Nous voyons donc clairement que le terme A correspond à l'idée générale du biais, *i.e.*, comme exprimé précédemment, à la distance entre le point central des estimations et la valeur estimée.

Nous pouvons aussi définir un terme de **biais structurel**, noté BS. Ce terme exprime le manque d'adéquation de la structure inférée à représenter la structure de t, en mesurant la différence en nombre de bipartitions. La valeur de ce terme est obtenue en utilisant l'espérance de la taille de $\hat{t}$ (1.2) :

$$BS = n_t - E\left(|\hat{t}|\right) = n_t - \sum_{b \in \mathcal{B}} \mathcal{P}(b)\,.$$

BS est nul si toutes les phylogénies inférées ont le même nombre n_t de bipartitions que t. Le biais structurel augmente quand des phylogénies moins résolues sont proposées, comme c'est le cas pour les arbres bootstrap réduits.

En utilisant les trois termes de biais et de variance définis ci-dessus, nous pouvons dériver de façon simple une décomposition de l'espérance de l'erreur d'estimation e_{RF}, en réécrivant les expressions analytiques des erreurs de 1ère espèce (1.3) et 2ème espèce (1.4) :

Lemme 1

$$E(e_1(\hat{t})) \quad = \quad A + V \tag{1.5}$$

Preuve :

$$\begin{aligned}
E(e_1(\hat{t})) &= \sum_{b \in \mathcal{B}-t} \mathcal{P}(b) \\
&= \sum_{b \in \mathcal{B}-t} \mathcal{P}(b) - \left(\sum_{b \in \mathcal{B}} \mathcal{P}(b) - \mathcal{M} \right) + \left(\sum_{b \in \mathcal{B}} \mathcal{P}(b) - \mathcal{M} \right) \\
&= \left(\mathcal{M} - \sum_{b \in \mathcal{B}} \mathcal{P}(b) + \sum_{b \in \mathcal{B}-t} \mathcal{P}(b) \right) + \left(\sum_{b \in \mathcal{B}} \mathcal{P}(b) - \mathcal{M} \right) \\
&= \qquad\qquad A \qquad\qquad + \qquad V\,.
\end{aligned}$$

□

Lemme 2

$$E(e_2(\hat{t})) \quad = \quad BS + A + V \tag{1.6}$$

Preuve :

$$\begin{aligned}
E(e_2(\hat{t})) &= n - \sum_{b \in t} \mathcal{P}(b) \\
&= n - \sum_{b \in t} \mathcal{P}(b) - \left(\sum_{b \in \mathcal{B}} \mathcal{P}(b) - \mathcal{M} \right) + \left(\sum_{b \in \mathcal{B}} \mathcal{P}(b) - \mathcal{M} \right) \\
&= \left(n - \sum_{b \in \mathcal{B}} \mathcal{P}(b) \right) + \left(\mathcal{M} - \sum_{b \in t} \mathcal{P}(b) \right) + \left(\sum_{b \in \mathcal{B}} \mathcal{P}(b) - \mathcal{M} \right) \\
&= \qquad BS \qquad + \qquad A \qquad + \qquad V\,.
\end{aligned}$$

□

Des lemmes précédents et de l'équation (1.1), nous obtenons l'expression de l'espérance de l'erreur généralisée :

Théorème 2

$$E(e^{\lambda}_{RF}(\hat{t})) = 2\left(V + A + \frac{BS}{\lambda+1}\right) . \tag{1.7}$$

Si l'on donne le même poids aux erreurs de 1ère et de 2ème espèces (*i.e.*, $\lambda = 1$), comme dans la distance de Robinson et Foulds [RF81], l'espérance de l'erreur est $2V + 2A + BS$. Le terme de biais structurel se voit accordé moins d'importance que les deux autres. Ce phénomène s'accentue si l'on donne plus d'importance à l'erreur de 1ère espèce (*i.e.*, en augmentant λ).

1.3 Choix du seuil de sélection des arêtes

Étant donné l'expression de l'erreur généralisée (1.1), et une valeur de λ donnée, la question naturelle qui se pose est celle du choix du seuil optimal, noté $S(\lambda)$, qui doit être utilisé pour réduire l'arbre bootstrap moyen $\bar{t}^*$, afin de minimiser l'espérance d'erreur.

Pour montrer comment on peut déterminer ce seuil, considérons l'analogie qu'on peut établir avec le cas d'un jeu simple. Un joueur parie de façon répétée sur l'issue d'un événement, ayant vrai ou faux comme résultat possible. La probabilité p que l'événement soit vrai, et donc la probabilité $1-p$ qu'il soit faux, sont supposées connues et fixes. Le joueur commet une erreur de 1ère espèce, e_1, à chaque fois qu'il parie sur "vrai" et que l'événement est faux. Inversement il commet une erreur de 2ème espèce quand il parie sur "faux" et que l'événement est "vrai". Supposons que nous disposons d'une fonction d'erreur, e^{λ}, de la forme $e^{\lambda} = \lambda e_1 + e_2$, *i.e.*, du type de l'expression (1.1), où le terme de normalisation $\frac{2}{\lambda+1}$, ici inutile, a été enlevé.

Dans le cadre de ce jeu, la question que l'on se pose naturellement est de savoir quelle ligne de conduite (basée sur le choix d'un seuil) doit adopter le joueur pour minimiser ses pertes, étant donné p et λ. En d'autres termes, comment déterminer le seuil optimal $S(\lambda)$ pour que l'espérance de l'erreur, $E(e^{\lambda})$, soit minimisée. Pour un seuil fixe S, il y a deux possibilités :

- `cas (1)` : si $p \geq S$ alors le joueur parie toujours sur "vrai", si bien que les seules erreurs possibles sont de 1ère espèce, et sont commises avec une probabilité $1-p$. Ainsi l'espérance d'erreur est $E(e^{\lambda}) = \lambda(1-p)$;
- `cas (2)` : si $p < S$ alors le joueur parie toujours sur "faux", les seules erreurs possibles sont de 2ème espèce et sont commises avec une probabilité p. Ainsi $E(e^{\lambda}) = p$.

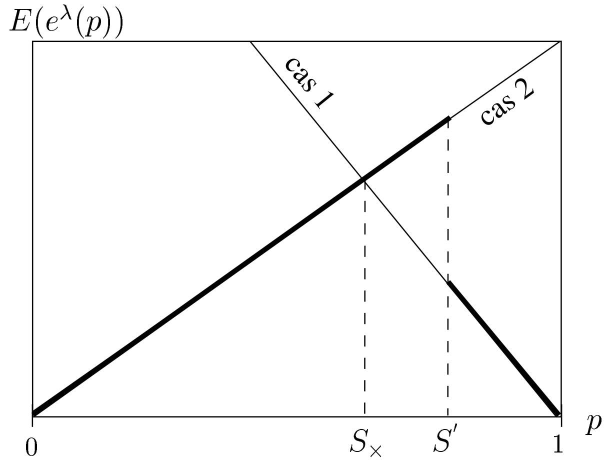


FIG. 1.1 – *Espérance de l'erreur, $E(e^{\lambda}(p))$, dans un jeu où un joueur parie de façon répétitive sur le résultat d'un événement ("vrai", "faux") de probabilité p. Le joueur utilise le seuil S pour choisir sa réponse : si $p \geq S$ (cas 1) il parie toujours sur "vrai", et sinon (cas 2) sur "faux". L'espérance de l'erreur dans le cas 1 est donnée par la droite $E(e^{\lambda}(p)) = \lambda(1-p)$, et dans le cas 2 par la droite $E(e^{\lambda}(p)) = p$. Le seuil optimum, i.e., minimisant l'espérance de l'erreur est l'abscisse S_x du point d'intersection des deux droites.*

La Fig. 1.1 représente l'espérance d'erreur en fonction de la probabilité p. Les deux lignes droites figurent l'espérance d'erreur dans le cas `(1)` et dans le cas `(2)`, respectivement. Soit S_x le seuil correspondant à l'intersection de ces deux droites, et S' t.q. $0 \leq S' \leq 1$ un seuil arbitraire. La partie en caractères gras de chaque droite représente l'espérance de l'erreur si S' est utilisé par le joueur dans la règle de décision définie ci-dessus (*i.e.*, si on fixe $S = S'$). On observe facilement que si le seuil S_x était utilisé à la place du seuil S', l'espérance d'erreur serait moindre sur l'intervalle $[S_x, S']$, et identique sur le reste de l'intervalle $[0, 1]$. Ainsi il apparaît de façon évidente que le seuil optimal $S(\lambda)$ que doit choisir le joueur pour minimiser son espérance d'erreur est le seuil situé à l'intersection des deux droites. On obtient analytiquement ce seuil en résolvant l'équation $S(\lambda) = \lambda(1 - S(\lambda))$, ce qui mène à $S(\lambda) = \frac{\lambda}{\lambda+1}$.

En R.P., le problème qui nous occupe est tout-à-fait similaire : on cherche à décider quelles arêtes de la phylogénie inférée $\hat{t}$ on doit garder (cas `(1)`) et quelles arêtes on doit rejeter (cas `(2)`). Considérons le cas d'une certaine arête, et soit p la probabilité que cette arête soit vraie (*i.e.*, appartienne à la phylogénie correcte). Le problème ici est que nous ne connaissons pas p, à l'inverse du joueur dans le jeu simple précédemment exposé. Cependant, nous pouvons l'estimer, dans une certaine mesure, par la proportion bootstrap, bp, obtenue pour cette même arête. Si nous posons $p = bp$, l'analyse du jeu précédent nous indique qu'il faut rejeter les arêtes (*i.e.*, parier sur "faux") de $\hat{t}$ t.q. $bp < \frac{\lambda}{\lambda+1}$. On dispose ainsi d'une première approximation du seuil optimum de sélection des arêtes, applicable à bp :

$$S_1(\lambda) = \frac{\lambda}{\lambda + 1} \ .$$

En fait, plusieurs auteurs ont montré que, pour des données sur lesquelles la méthode de reconstruction utilisée est consistante, il y a effectivement une corrélation entre p et bp, mais ces deux mesures ne sont toutefois pas identiques [ZL92a, ZL92b, FK93, HB93]. Ainsi, on peut soupçonner que $S_1(\lambda)$ ne soit qu'une approximation grossière du seuil optimum. Supposons maintenant que nous disposions d'une fonction (croissante) f, telle que $f(bp)$ puisse être considérée comme une estimation de p. Dans ce cas là, les arêtes de $\hat{t}$ peuvent être conservées dès que

$$f(bp) \geq \frac{\lambda}{\lambda + 1}$$

ou de façon équivalente

$$bp \geq f^{-1}\left(\frac{\lambda}{\lambda + 1}\right) ,$$

et écartées sinon. Les auteurs cités ci-dessus ont réalisé des simulations, dans des cas simples, montrant que la forme générale de la fonction f est sinusoïdale (pour le critère de reconstruction MP). En reproduisant ces expériences, on obtient la fonction f représentée sur la Fig. 1.2 pour deux méthodes différentes : une optimisant le critère MP, basé sur les caractères, et l'autre optimisant le critère ME, basé sur les distances. On retrouve effectivement la forme sinusoïdale pour MP,

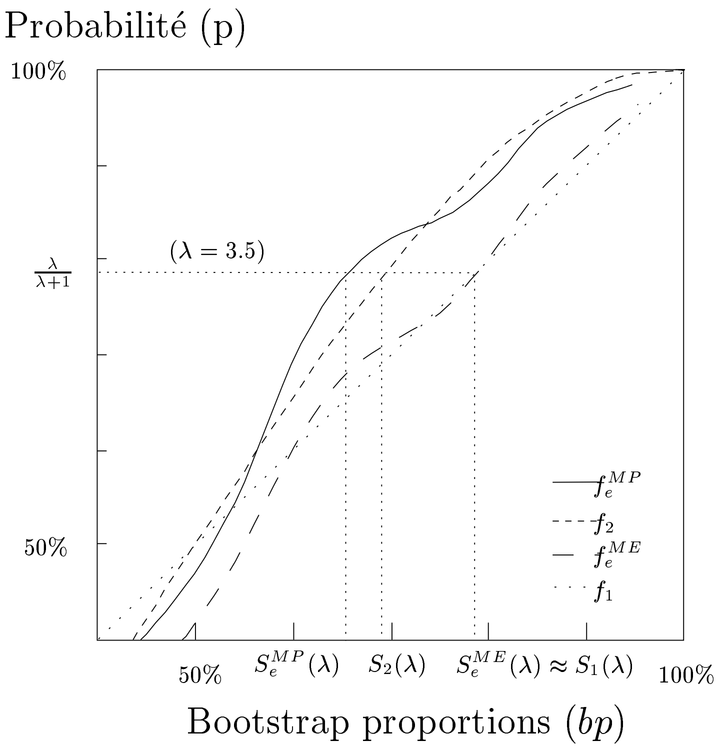


FIG. 1.2 – *Probabilité (p) qu'une arête de $\hat{t}$ soit correcte en fonction de sa proportion bootstrap (bp). f_e^{ME} et f_e^{MP} sont les fonctions empiriques obtenues pour les méthodes de reconstuction ME et MP respectivement, tandis que $f_1(x) = x$ et $f_2(x) = \frac{-cos(\pi x)+1}{2}$ sont les estimations proposées pour ces fonctions. Les approximations $S_1(\lambda)$, $S_2(\lambda)$ et $S_e(\lambda)$ du seuil optimum de sélection des arêtes sont obtenues en appliquant l'inverse de ces fonctions à $\frac{\lambda}{\lambda+1}$.*

et de façon légèrement moins marquée pour ME. Nous pouvons donc penser raisonnablement que la fonction cosinus peut être employée pour fonder une bonne estimation de $S(\lambda)$:

$$S_2(\lambda) = \arccos\left(1 - \frac{2\lambda}{\lambda+1}\right)/\pi.$$

Notons que pour l'erreur de Robinson et Foulds ($\lambda = 1$), les deux approximations $S_1(\lambda)$ et $S_2(\lambda)$ indiquent un seuil de 50%. Ce seuil correspond au célèbre *arbre majoritaire*, constitué des arêtes apparaissant dans plus de 50% des cas, et souvent utilisé pour obtenir le consensus d'un ensemble de phylogénies. On s'attend ainsi à ce que la phylogénie $\bar{t}^*_{50\%}$ soit plus ou moins le meilleur estimateur de la phylogénie inconnue (au sens de l'erreur de Robinson et Foulds).

Une alternative à ces approximations analytiques peut être obtenue sur la base de simulations. Pour toute méthode de reconstruction étudiée, sous un certain nombre c de conditions d'évolution diverses, on peut retenir le seuil optimum constaté pour chaque condition i, soit $S^i(\lambda)$. Depuis ces seuils optimaux observés pour les différentes conditions, on peut définir une approximation empirique de $S(\lambda)$ comme :

$$S_e(\lambda) \;=\; \frac{1}{c}\sum_{i=1..c} S^i(\lambda)\,. \tag{1.8}$$

Pour estimer l'efficacité des différentes approximations de $S(\lambda)$ proposées ci-dessus ($S_1(\lambda)$, $S_2(\lambda)$ et $S_e(\lambda)$) en général, on peut choisir un certain nombre de valeurs de λ (soit v valeurs) et mesurer la perte de qualité maximum (sur l'ensemble des valeurs de λ et des conditions d'évolution) suivant les formules suivantes :

$$L_A(S_a) = max_{\substack{\lambda\in\{1,\dots,v\}\\ i\in\{1,2,\dots,c\}}} \{\bar{e}^{\lambda}_{RF}(\bar{t}^*_{S_a(\lambda)}) - \bar{e}^{\lambda}_{RF}(\bar{t}^*_{S^i(\lambda)})\} \tag{1.9}$$

et

$$L_R(S_a) = max_{\substack{\lambda\in\{1,\dots,v\}\\ i\in\{1,2,\dots,c\}}} \left\{\frac{\bar{e}^{\lambda}_{RF}(\bar{t}^*_{S_a(\lambda)}) - \bar{e}^{\lambda}_{RF}(\bar{t}^*_{S^i(\lambda)})}{\bar{e}^{\lambda}_{RF}(\bar{t}^*_{S^i(\lambda)})}\right\} \tag{1.10}$$

où $a \in \{1, 2, e\}$. En d'autres termes, on mesure ainsi l'augmentation *absolue* maximale (1.9), resp. l'augmentation *relative* maximale (1.10), de l'erreur de l'arbre bootstrap réduit, dûe à l'utilisation du seuil approximé plutôt que du seuil optimum, sur l'ensemble des valeurs λ et des conditions d'évolution examinées. Plus l'augmentation observée est faible, plus l'approximation est efficace.

1.4 Étude empirique

Pour étudier les performances de l'arbre bootstrap réduit, nous avons eu recours à des simulations reproduisant les conditions d'évolution de séquences moléculaires. Il faut noter que cette démarche est fréquente en R.P., où elle permet de palier le fait que la phylogénie des espèces étudiées est rarement connue avec certitude. Des études comparatives basées sur des simulations paraissent régulièrement pour identifier parmi les nouvelles méthodes de reconstruction proposées celles qui sont pertinentes [TNT82, SI89, Nei91, KF94, TN94b, BG96a].

Nous avons vu précédemment comment estimer le seuil optimal S_{opt} de sélection des arêtes. Nous examinons ici l'efficacité pratique des approximations proposées ainsi que le gain obtenu en estimant la phylogénie correcte t par l'arbre bootstrap réduit plutôt que par la phylogénie $\hat{t}$ initialement proposée.

Étant donné un seuil S de sélection des arêtes, nous évaluerons avant tout :

- la diminution de l'erreur de 1ère espèce, *i.e.*, le nombre d'arêtes incorrectes de $\hat{t}$ écartées.
- l'augmentation de l'erreur de 2ème espèce, *i.e.*, le nombre d'arêtes correctes présentes dans $\hat{t}$ mais qui sont écartées.
- la réduction de l'erreur totale qui résulte du rejet des arêtes de faible *bp* ($bp < S$).
- la qualité des approximations $\hat{S}_1$ et $\hat{S}_2$, *i.e.*, l'erreur totale induite par les arbres $\hat{t}^*_{\hat{S}_1}$, $\hat{t}^*_{\hat{S}_2}$ vis-à-vis de l'erreur induite par l'arbre $\hat{t}^*_{\hat{S}_{opt}}$ réduit selon le meilleur seuil empiriquement constaté.

1.4.1 Schéma de simulation

Nous avons effectué des simulations suivant un schéma classique en R.P. : pour des conditions d'évolution définies, on génère aléatoirement des phylogénies depuis lesquelles on obtient des jeux de données moléculaires suivant les conditions d'évolution fixées. Plus précisément, nous avons essayé de reproduire au plus près les simulations de Kuhner et Felsenstein [KF94].

La figure 1.3 décrit la façon dont les phylogénies $\hat{t}$ et $\hat{t}^*_S$ sont obtenus depuis la phylogénie t. Pour chaque condition d'évolution, 50 phylogénies t sont aléatoirement générées et servent de support à la production de 100 jeux de données dits *originels*. Chacun de ces jeux X est, d'une part, analysé par une certaine méthode de reconstruction inférant directement une phylogénie $\hat{t}$, et, d'autre part, est bootstrappé pour générer 100 pseudojeux de données X^*. Ces pseudojeux sont séparemment analysés par la même méthode de reconstruction que pour X, ce qui aboutit à la production de 100 arbres bootstrap $\hat{t}^*$.

L'arbre bootstrap moyen $\bar{t}^*$ est obtenu en reproduisant la structure de $\hat{t}$ et en affectant à chaque arête interne une *bp* représentant le nombre de fois où elle est présente dans les différents arbres bootstrap $\hat{t}^*$. Avec 100 réplications, l'écart-type des *bp* obtenues atteint au maximum 5% (quand $bp = 50\%$), ce qui est suffisant pour notre propos tout en restant réaliste d'un point de vue calculatoire. L'arbre

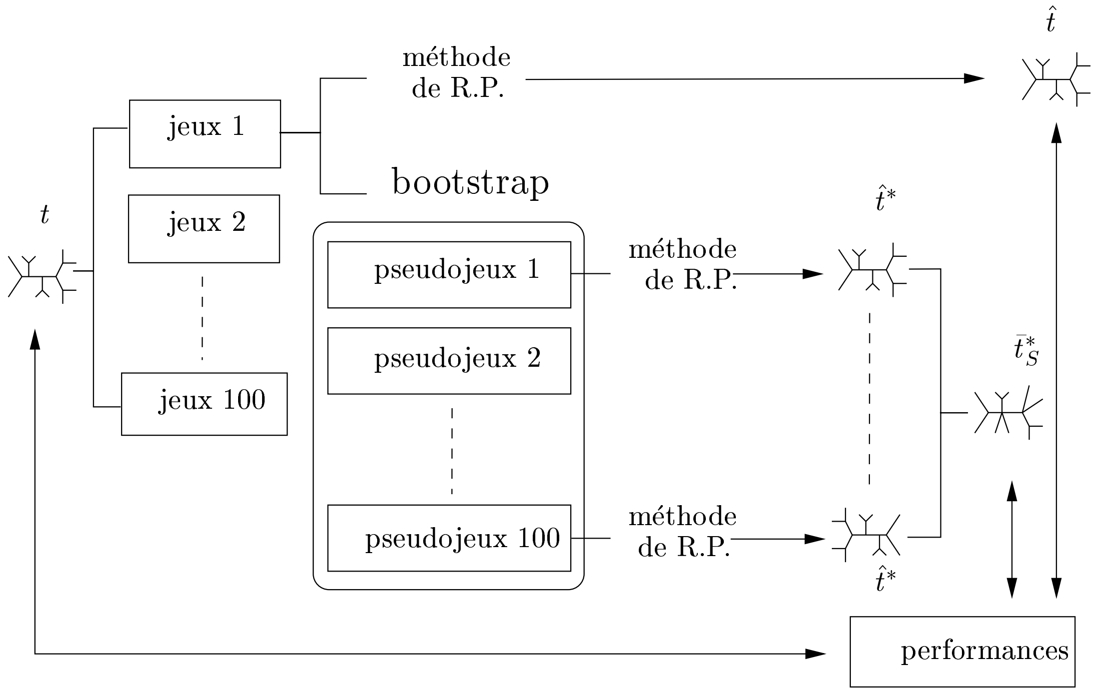


FIG. 1.3 – *Schéma de simulations. La phylogénie initialement inférée (i.e., sans rééchantillonnage), notée $\hat{t}$, et l'arbre bootstrap réduit au seuil S (pour un seuil S fixé), noté $\bar{t}^*_S$, sont comparés sur la base de leur efficacité à retrouver la phylogénie correcte t.*

bootstrap réduit $\hat{t}^*_S$ est ensuite obtenu en écartant les arêtes internes de $\bar{t}^*$ ayant une *bp* inférieure au seuil S fixé. Pour certains seuils S bien choisis, nous avons comparé les performances respectives des phylogénies $\hat{t}$ et $\hat{t}^*_S$ pour estimer la structure de la phylogénie correcte t. Ces performances ont été évaluées en termes d'erreur de 1ère espèce et de 2ème espèce, ainsi qu'en termes de biais et de variance.

Génération de phylogénies modèles

Nous générons des phylogénies sur 10 espèces qui sont totalement résolues, par analogie aux phylogénies réelles, supposées binaires. Pour générer chaque phylogénie, nous avons recours au processus itératif de branchement présenté par Kuhner et Felsenstein [KF94]. On considère initialement une racine (un sommet) à partir de laquelle deux "lignées" (arêtes pendantes) divergent. Chacune de ces deux lignées est autorisée à se séparer en deux nouvelles lignées et ainsi de suite, jusqu'à disposer d'une phylogénie possédant dix lignées *pendantes*, *i.e.*, 10 lignées ne s'étant pas encore séparées. Au cours de ce processus, quand on ne dispose pas encore de 10 lignées pendantes, la prochaine lignée pendante qui va se séparer en deux est désignée au hasard (avec une probabilité uniforme). Ce mécanisme, dit de Yule-Harding [Har71, Bro94, Ald95], est reproduit jusqu'à obtenir une phylogénie à 10 lignées pendantes, *i.e.*, à 10 feuilles.

Tout au long de ce processus on associe aussi une longueur aux arêtes de la phylogénie, *i.e.*, une durée à chaque lignée. On fixe tout au long du processus une probabilité constante dt qu'une lignée se sépare dans le prochain court intervalle de temps de longueur dt. À chaque étape du processus, on allonge les k lignées pendantes d'une même durée d avant de procéder à la prochaine séparation d'une lignée en deux. La durée d est obtenue depuis une distribution exponentielle d'espérance $\frac{1}{k}$.

L'ensemble du processus se résume donc ainsi : on dispose initialement de 2 lignées descendant d'un ancêtre commun depuis une durée d tirée d'une loi exponentielle d'espérance $\frac{1}{2}$; au bout de cette durée d, une des 2 lignées se sépare en 2 nouvelles lignées, et on dispose alors de 3 lignées, qui sont étendues chacune d'une durée d' tirée d'une loi exponentielle d'espérance $\frac{1}{3}$; et ainsi de suite jusqu'à ce que la séparation d'une lignée pendante porte à 10 le nombre de lignées pendantes de la phylogénie, auquel cas ces lignées sont toutes prolongées d'une durée tirée d'une loi exponentielle d'espérance 1/10, après quoi le processus est interrompu.

L'espérance de la hauteur de la phylogénie (somme des longueurs d'arêtes sur un chemin de la racine à une feuille) est donc 1,93, *i.e.*, $\frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ldots + \frac{1}{10}$.

Génération des séquences depuis une phylogénie modèle

Sur chacune des phylogénies ainsi obtenue, on peut simuler l'évolution d'une séquence moléculaire de la racine aux feuilles de la phylogénie, suivant un certain modèle de l'évolution. On retient ici le modèle de Kimura à deux paramètres [Kim80] (cf. chap. 1, part. *I*), où un taux de transitions (α) deux fois plus important que le taux de transversions (β) est choisi. Les sites de la séquence ancestrale sont répartis en `A`, `C`, `G` et `T` de façon à ce que chacun de ces nucléotides soit également représenté. La probabilité d'une substitution le long d'une arête est le produit de la longueur de l'arête (exprimée en unité de temps suite au processus de génération de la phylogénie) et d'un taux d'évolution donné. Les deux taux que nous avons considérés sont un taux *faible* ($0,01$ mutation par unité de temps) et un taux *élevé* ($0,1$ mutation par unité de temps).

Le temps est étalonné en fonction du modèle de l'évolution de façon à ce que le temps moyen écoulé dans une lignée, avant que celle-ci ne se sépare en deux, corresponde à 1 mutation par site. Ainsi, chaque site d'une séquence évoluant sur une arête à taux faible a une probabilité $0,01$ de subir une substitution le long de cette arête (ce qui correspond au plus petit taux d'évolution généralement considéré [KF94]). L'espérance du nombre de substitutions par site de la racine à une feuille est ainsi compris entre 0,0193 et 0,193, selon les taux de substitution affectés aux arêtes de la phylogénie. Les jeux de données X sont composés en prenant les 10 séquences de nucléotides obtenues aux feuilles de la phylogénie. Lors des simulations nous avons eu recours aux programmes RANTREE et DNATREE pour générer les phylogénies, resp. les jeux de données. Ces programmes nous ont été communiqués par M.K. Kuhner et J. Felsenstein ; ce sont ceux qu'ils ont utilisé dans leur étude [KF94].

Nous avons généré des jeux de données X pour quatre conditions d'évolution différentes :

- taux faible de substitution sur toutes les arêtes
- taux élevé de substitution sur toutes les arêtes
- taux faible de substitution pour la moitié des arêtes et taux élevé pour l'autre moitié
- taux faible de substitution pour la moitié des sites et taux élevé pour l'autre moitié

Pour chacune de ces conditions nous avons engendré des séquences de 100, 300, 1000 et 3000 nucléotides. Pour assurer l'indépendance des résultats vis-à-vis de la structure de la phylogénie, nous avons généré des phylogénies différentes pour chaque combinaison d'une condition d'évolution et d'une longueur de séquences.

Nous avons étudié à la fois une méthode de reconstruction basée sur les caractères et une méthode basée sur les distances. La première consistait en une heuristique pour le critère MP implémentée dans le programme DNAPARS du package PHYLIP [Fel93] (dans des simulations préliminaires, nous avions en effet observé que cette heuristique et un algorithme optimal pour le critère avaient la même efficacité pour estimer t comme similairement constaté par Kuhner et Felsenstein [KF94]). Pour méthode de distances nous avons utilisé le programme NEIGHBOR de PHYLIP [Fel93] apparenté à la très populaire heuristique de Neighbor-Joining (NJ) [SN87]. Celle-ci optimise le critère de reconstruction ME.

Comme Kuhner et Felsenstein [KF94], nous avons donné en entrée à l'heuristique NEIGHBOR les distances corrigées par le modèle de l'évolution de Kimura [Kim80] avec le taux correct de transition/transversion (*i.e.*, 2). Un certain nombre d'autres programmes ont été nécessaires pour ces simulations, ceux-ci étant pris dans PHYLIP version 3.5 [Fel93] ou réécrits quand nécessaire. Tous ces programmes sont bien-sûr disponibles sur demande. Au final, plus de 16 millions de phylogénies sur 10 espèces ont été inférées, représentant plusieurs mois de calcul sur une station SUN SPARC 5.

1.4.2 Résultats des simulations

Les résultats des simulations sont décrits précisément dans l'annexe C. Nous résumons ici les traits marquants de ces résultats.

Seuil optimum de sélection des arêtes

Rappelons que le seuil S de sélection des arêtes est utilisé pour obtenir l'arbre bootstrap réduit $\bar{t}^*_S$, en écartant un certain nombre d'arêtes jugées non-fiables (*i.e.*, celles t.q. $bp < S$) de l'arbre bootstrap moyen $\bar{t}^*$. Une question importante consiste à déterminer le seuil S optimal, *i.e.*, celui qui conduit à minimiser l'espérance d'erreur de la phylogénie $\bar{t}^*_S$.

Les estimations du seuil optimum proposés à la section 1.3 apparaissent de très bonne qualité dans les résultats des simulations, au sens où utiliser ces estimations plutôt que le seuil otimal a conduit, dans le pire des cas, à une très faible augmentation de l'erreur : pour 10 valeurs de λ et sur les 16 conditions d'évolution, l'augmententation d'erreur maximum constatée est seulement de $0,46$ sur une echelle de $[0,14]$.

La grande qualité observée pour les estimations du seuil optimal s'explique par l'aspect de la courbe représentant l'erreur de l'arbre bootstrap réduit en fonction du seuil de sélection des arêtes employé. La Fig. 1.4 représente cette fonction pour le cas où $S_2(\lambda)$ a été observé le moins performant en terme d'erreur aboslue, pour la méthode MP. La zone de la courbe autour du seuil de sélection optimum est relativement plate, impliquant que même une estimation un peu lâche de ce seuil optimum, conduit seulement à une faible augmentation de l'erreur de l'arbre bootstrap réduit.

Sur l'ensemble des simulations, l'estimation empirique $S_e(\lambda)$ est la plus efficace des trois estimations du seuil optimum de sélection des arêtes. Toutefois les deux approximations analytiques simples, $S_1(\lambda)$ et $S_2(\lambda)$, sont presques aussi efficaces, chacune se révélant plus appropriée pour une des méthodes de R.P. testées : $S_1(\lambda)$

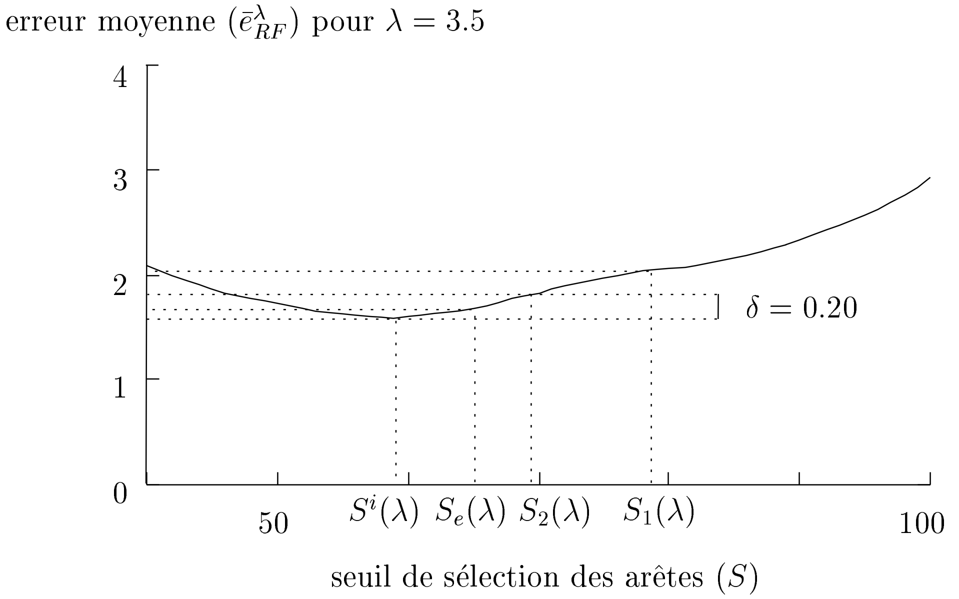


FIG. 1.4 – *Erreur moyenne ($\bar{e}^{\lambda}_{RF}$) de l'arbre bootstrap réduit $\bar{t}^*_S$, representée comme une fonction du seuil S de sélection des arêtes ($\lambda = 3.5$, faible taux de substitution, 300 nucléotides, MP). $S^i(\lambda)$ est le seuil optimum à employer dans ce cas avec MP (le seuil optimum pour ME n'est pas figuré). $S_2(\lambda)$ est l'approximation du seuil recommandée pour la méthode MP. $S_e(\lambda)$ est le seuil empirique. $S_1(\lambda)$ est la plus simple approximation possible. δ est la perte de qualité absolue observée pour $S_2(\lambda)$.*

est plus adaptée quand les phylogénies sont inférées par la méthode ME, tandis que $S_2(\lambda)$ est plus adaptée quand les phylogénies sont inférées par la méthode MP. En conséquence, dans la suite des simulations, nous avons estimé le seuil optimal par $S_1(\lambda)$ quand la méthode ME était employée et par $S_2(\lambda)$ quand MP était employée.

Réduction d'erreur dûe à la suppression des arêtes de faible *bp*

En estimant la phylogénie t, non plus par la phylogénie $\hat{t}$ inférée directement depuis le jeu de données initial, mais par l'arbre bootstrap réduit $\bar{t}^*_{S(\lambda)}$ (*i.e.*, la phylogénie $\hat{t}$ de laquelle ont été enlevées les arêtes t.q. $bp < S(\lambda)$), nous espérions diminuer l'erreur d'estimation.

Les résultats des simulations montrent le bien fondé de nos espérances, puisque sur l'ensemble des conditions d'évolution et des longueurs de séquences étudiées, l'erreur moyenne de $\bar{t}^*_{S(\lambda)}$ pour estimer t, $\bar{e}^{\lambda}_{RF}(\bar{t}^*_{S(\lambda)})$, a *toujours* été constatée inférieure à celle de $\hat{t}$.

La différence entre ces deux erreurs est d'autant plus importante qu'on donne un poids important à l'erreur de 1ère espèce (*i.e.*, que λ augmente). En R.P., on considère $\lambda \geq 1$, or pour $\lambda = 1$ (*i.e.*, pour l'erreur standard de Robinson et Foulds), l'arbre bootstrap induit déjà une erreur plus faible en moyenne que celle de l'arbre direct $\hat{t}$. Ceci résulte d'une importante réduction de l'erreur de 1ère espèce, tandis que l'erreur de 2ème espèce augmente faiblement. Autrement dit, les arêtes de la phylogénie initialement inférée, écartées grâce au processus de bootstrap, sont essentiellement des arêtes incorrectes, attestant ainsi du succès de cette approche.

Dans les simulations effectuées, l'étendue de la réduction d'erreur dûe à l'utilisation de la technique du bootstrap dépend grandement de la longueur des séquences : pour toutes les conditions d'évolution testées, la réduction d'erreur est d'autant moins importante qu'on dispose de beaucoup de caractères pour décrire les espèces. La cause de ce phénomène est double : l'erreur de $\hat{t}$ diminue quand la longueur des séquences décrivant les espèces augmente, laissant ainsi moins de marge à une amélioration possible des résultats ; ensuite, pour des séquences plus longues, le processus de rééchantillonnage crée moins de variabilité vis-à-vis du jeu de données initial, si bien que les phylogénies inférés depuis les jeux bootstrappés sont relativement semblables à celles inférées depuis le jeu de données initial.

La diminution d'erreur de $\hat{t}$, observée lorsque le nombre de caractères augmente, traduit le fait que les méthodes de R.P. employées (ME ou MP) sont *globalement* consistantes pour les conditions d'évolution étudiées, *i.e.*, que pour un nombre suffisant de caractères, elles retrouvent toutes les arêtes, ou une grande partie des arêtes, des phylogénies modèles.

Même si les méthodes de reconstruction sont globalement consistantes sur toutes les conditions d'évolution étudiées, elles convergent plus ou moins rapidement vers la phylogénie correcte selon les conditions. Ainsi, pour des séquences évoluant à des taux faibles constants, ou à des taux inégaux selon les arêtes, les méthodes de reconstruction requièrent plus de caractères pour obtenir le même niveau d'efficacité (au sens de e^{λ}_{RF}) que celui constaté pour des séquences évoluant selon un taux constant mais élevé.

Décomposition de l'erreur en termes de biais et de variance

Sur l'ensemble des simulations, un faible biais fondamental (mesuré par le terme A) a toujours été constaté pour les méthodes d'inférence de phylogénies, que ce soit pour la méthode directe (proposant la phylogénie $\hat{t}$ depuis le jeu de données initial), ou la procédure bootstrap (proposant $\bar{t}^*_{S(\lambda)}$ sur la base de plusieurs jeu de données rééchantillonnés).

Les mesures de biais structurel (terme BS) et de variance (terme V) ont des valeurs moyennes totalement opposées pour les deux approches : l'approche directe ne possède aucun biais structurel ($\hat{t}$ comporte toujours le même nombre d'arêtes que la phylogénie modèle), mais est victime d'une forte variance ; inversement, l'approche bootstrap possède un biais structurel non négligeable, mais en retour une faible variance. La réduction importante de variance induite par l'approche bootstrap compense largement son déficit en terme de biais structurel, ce qui explique les meilleurs résultats moyens obtenus par la phylogénie $\bar{t}^*_{S(\lambda)}$, comparé à $\hat{t}$ pour estimer la phylogénie des espèces.

1.4.3 Application sur un jeu de données réel

Nous montrons ici le résultat de l'application de la technique du bootstrap avec un seuil de sélection de 50% sur un jeux de données de 11 mammifères fourni par D. Penny, et utilisé dans une série d'articles (p. ex. [PHS91]). Ce jeu de données (détaillé à l'annexe G) est constitué de 11 séquences de 466 nucléotides. L'arbre majoritaire résultant de l'analyse bootstrap est présenté sur la Fig. 1.5.

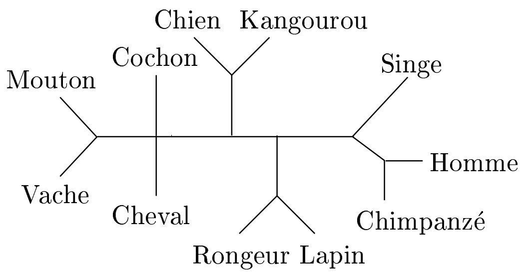


FIG. 1.5 – *Arbre bootstrap majoritaire (parcimonie),* $\bar{t}^*_{50\%}$.

On retrouve ainsi correctement le groupe des primates ((homme, chimpanzé),singe) et celui des ongulés (mouton, vache, cochon, cheval), toutefois le sous-groupe (vache, mouton, cochon) n'est pas ici retrouvé. Le groupement (chien, kangourou), est jugé incorrect par une majorité de biologistes, toutefois aucune conclusion ferme n'a été retenue pour l'instant [PHS91]. Selon Penny *et al* [PHS91], le groupe (lapin, rongeur, primates) est correct, mais le groupe (lapin, rongeur) est erroné : il résulte du problème d'attraction des longues branches [Fel78a].

1.5 Conclusion

Nous avons montré comment on pouvait utiliser la technique de rééchantillonnage du bootstrap [Efr79, Fel85] pour obtenir une meilleure estimation d'une phylogénie inconnue t que celle donnée par la phylogénie $\hat{t}$, inférée directement depuis le jeu de données initial, selon une méthode de reconstruction usuelle. La technique du bootstrap propose une phylogénie partiellement résolue, $\bar{t}^*_S$, obtenu en éliminant les arêtes de $\hat{t}$ ayant une proportion bootstrap inférieure au seuil S. Le point central consiste à déterminer le seuil optimal de sélection des arêtes, au sens où un maximum d'arêtes non fiables sont écartées, tout en conservant un maximum d'arêtes correctes. Au cours de ce chapitre,

(1) nous avons proposé deux approximations analytiques du seuil optimal, fonctions de la mesure d'erreur retenue (exprimée sous la forme d'une distance entre la phylogénie estimée t et l'estimation proposée). Nous avons montré sur la base de simulations poussées que ces deux approximations sont presque optimales, l'une pour la méthode MP, l'autre pour la méthode ME, puisqu'elles conduisent à des erreurs seulement très légèrement plus importantes que l'erreur induite par l'utilisation du seuil optimal, empiriquement constaté ;

(2) nous avons montré qu'une réduction importante de l'erreur pouvait être obtenue quand on estime t par l'arbre bootstrap $\bar{t}^*_S$ (pour un seuil S adéquat) plutôt que par la phylogénie $\hat{t}$: jusqu'à 39% pour la méthode MP et 33% pour la méthode ME. Ce gain important est expliqué par une grande réduction de l'erreur de 1ère espèce, compensant largement la faible augmentation en erreur de 2ème espèce. Le gain dû à l'utilisation du bootstrap est d'autant plus important qu'on dispose de séquences courtes et qu'on se situe dans des conditions d'évolution difficiles ;

(3) nous avons aussi donné une décomposition de l'espérance de l'erreur d'estimation en un terme de biais fondamental (souvent négligeable), un terme de variance et un terme de biais structurel, ces deux derniers étant antagonistes. Sur cette base, nous avons montré que le procédé bootstrap admet, en tant qu'estimateur de t, un biais structurel assez important (mesurant le degré moyen d'irrésolution des phylogénies proposées) mais largement compensé par la réduction en variance. De plus, le procédé bootstrap n'induit aucun biais fondamental supplémentaire, relativement à la méthode de reconstruction sur laquelle il est basé.

Ces divers résultats confirment l'idée que pour estimer une phylogénie inconnue, un gain significatif est obtenu en proposant une phylogénie partiellement résolue plutôt qu'une phylogénie complètement résolue, pouvant contenir des parties extrêmement variables (*i.e.*, non fiables).

Pour obtenir des phylogénies partiellement résolues, nous avons ici utilisé la technique du bootstrap telle qu'elle fut introduite par Felsenstein [Fel85]. On peut cependant imaginer d'autres procédures statistiques ou combinatoires pour écarter les arêtes incertaines de la phylogénie $\hat{t}$ originellement inférée. Dans le chapitre suivant, nous proposons une autre méthodologie pour appliquer la technique du bootstrap en R.P..

Chapitre 2

Une nouvelle méthode pour appliquer la technique du bootstrap en reconstruction phylogénétique

La problématique de cette thèse consiste à montrer que de meilleures estimations d'une phylogénie inconnue peuvent être obtenues en inférant des pyhlogénies partiellement résolues, plutôt que des phylogénies totalement résolues. Dans le chapitre précédent, nous avons proposé une première méthode possible pour inférer des phylogénies partiellement résolues, basée sur l'utilisation de la technique du bootstrap telle qu'elle fut introduite en R.P. par Felsenstein [Fel85]. Les résultats théoriques et expérimentaux ont montré le bien fondé de notre démarche. Nous proposons maintenant une deuxième méthodologie pour obtenir des phylogénies partiellement résolues, sur la base d'une nouvelle utilisation du principe du bootstrap en R.P.

L'utilisation du bootstrap introduite en R.P. par Felsenstein repose principalement sur deux points :

a) l'affectation d'une proportion bootstrap (*bp*) à chaque arête de la phylogénie $\hat{t}$ initialement inférée, pour mesurer la fiabilité de cette arête ;

b) le choix d'un seuil S tel que les arêtes ayant une *bp* inférieure à S sont écartées.

Les deux points précédents portent à discussion. En effet, a) la signification statistique des *bp* n'est pas encore clairement établie (cf. le débat [FK93, HB93, EHH96]), et plusieurs alternatives à cette mesure ont même été proposées [EHH96, ZL95] ; b) le fait de devoir fixer le seuil de sélection des arêtes n'est pas satisfaisant, car c'est un paramètre de plus à estimer. Nous avons vu qu'on pouvait approcher ce seuil de façon relativement fiable, toutefois les approximations proposées ne tiennent pas compte des conditions d'évolution et de la méthode de reconstruction employée, or les simulations ont montré que ces facteurs pouvaient parfois avoir une influence sur la valeur du seuil optimal. De tels facteurs peuvent difficilement être pris en compte de façon analytique, limitant ainsi la qualité des approximations pour ce paramètre.

La méthode que nous proposons dans ce chapitre pour appliquer la technique du bootstrap en R.P. évite les deux problèmes évoqués ci-dessus. L'utilisation de la technique du bootstrap introduite par Felsenstein [Fel85] consiste à décider *quelles* arêtes doivent être éliminées. Nous considérons ici le problème sous un autre angle, essayant plutôt de savoir *combien* d'arêtes doivent être éliminées. Autrement dit, nous essayons de connaître le degré de résolution que doit adopter la phylogénie inférée dans le but de fournir la meilleure estimation possible. En termes de statistiques inférentielle, ceci correspond à déterminer la meilleure **complexité** pour le modèle proposé.

2.1 Approche à complexité variable

La nouvelle procédure de bootstrap que nous proposons repose sur le principe initial du bootstrap, *i.e.*, "*utiliser les différences observées entre* $\hat{t}^*$ *et* $\hat{t}$ *(i.e., entre l'arbre bootstrap et l'arbre direct*[1]*) pour inférer les différences inobservables entre* $\hat{t}$ *et* t *(i.e., entre l'arbre direct et l'arbre vrai)*" [EHH96]. Nous appliquons ici ce principe à la détermination de la complexité la plus *adéquate* pour estimer la phylogénie t, étant donné le jeu de données (soit X) dont on dispose pour les espèces de cette phylogénie, et la méthode de reconstruction (soit M) qu'on utilise. Par complexité *adéquate* nous entendons la complexité à laquelle est attachée la probabilité d'erreur théorique la plus faible. Par exemple, si le jeu de données est extrêmement bruité ou si la méthode de reconstruction infère une phylogénie complètement au hasard, la complexité nulle (correspondant à la phylogénie étoile) est la plus adéquate. À l'autre extrême, si le jeu de données est de très bonne qualité et si la méthode de reconstruction est consistante, une complexité maximale (correspondant à une phylogénie complètement résolue) est la plus adéquate.

Nous appellerons cette utilisation du bootstrap, **l'approche à complexité variable** : aucune complexité n'est fixée initialement, c'est la technique du bootstrap

1. *selon la terminologie définie dans le chapitre précédent.*

qui la détermine. En contraste, la méthode de Felsenstein sera appelée **l'approche à seuil fixe**, en regard du seuil de sélection des arêtes, fixé manuellement en début de procédure.

L'approche à complexité variable suit le schéma suivant :

1. Une séquence $\hat{s}$ de phylogénies de complexités décroissantes est inférée depuis la matrice de données initiale X:

$$X \longrightarrow \hat{s} = [\hat{t}_{max}, \dots, \hat{t}_c, \hat{t}_{c-1}, \dots, \hat{t}_0],$$

   où $\hat{t}_c$ dénote la phylogénie ayant c arêtes internes (*i.e.*, de complexité c) et $\hat{t}_{max} = \hat{t}$, la phylogénie complètement résolue inférée par la méthode de reconstruction M depuis X. Notons que M peut aussi bien être une méthode basée sur les caractères (exploitant X directement), qu'une méthode basée sur les distances (obtenant depuis X une matrice $\hat{D}$ de distances entre espèces, depuis laquelle $\hat{t}$ est produit). Quoi qu'il en soit, M doit produire une phylogénie $\hat{t}$ qui soit une représentation aussi fidèle que possible du jeu de données X, pouvant être vu, en quelque sorte, comme la (meilleure) représentation arborée de X. La méthode M repose donc sur l'optimisation d'un critère de reconstruction δ mesurant la *qualité d'ajustement* de $\hat{t}$ à X.

   La phylogénie $\hat{t}$ est identifiée à $\hat{t}_{max}$ et la séquence $\hat{s}$ est complétée de façon à ce que les phylogénies $\hat{t}_{max-1}, \dots, \hat{t}_0$ optimisent δ tout en respectant leur niveau de complexité ($max - 1, \dots$). En R.P., la plupart du temps, δ est d'autant mieux optimisé qu'on dispose d'un niveau de complexité (*i.e.*, d'un degré de liberté) élevé. En conséquence les phylogénies $\hat{t}_{max}, \dots, \hat{t}_0$ de $\hat{s}$ sont d'efficacité décroissante pour optimiser δ.

2. La matrice initiale de données X est ensuite rééchantillonnée par un processus de bootstrap, pour produire des pseudojeux de données X^* qui servent de base à la production d'autant de séquences de phylogénies, $\hat{s}^*$, au moyen de la méthode M :

$$X^* \longrightarrow \hat{s}^* = [\hat{t}^*_{max}, \dots, \hat{t}^*_0],$$

3. Pour chaque séquence $\hat{s}^*$ obtenue depuis un jeu de données X^*, on évalue l'efficacité de chacune de ses phylogénies pour approximer le jeu de données initial X, au sens du critère δ. La meilleure complexité $\hat{\alpha}^*$ pour inférer une phylogénie depuis X^* est déduite comme étant celle associée à la phylogénie de $\hat{s}^*$ qui approxime le mieux X, *i.e.*, la phylogénie $\hat{t}^*_{\hat{\alpha}^*}$ tel que $\hat{\alpha}^* = argmin_{\alpha_i}(\delta(X, \hat{t}^*_{\alpha_i}))$.

4. Les deux étapes précédentes (de la constitution d'un pseudojeu X^* à la désignation d'une complexité $\hat{\alpha}^*$) sont répétées un grand nombre r de fois, correspondant au nombre de réplications bootstrap. Puis, la complexité $\hat{\alpha}$, *i.e.*, la complexité estimée la plus adéquate pour minimiser l'erreur de $\hat{t}$, est obtenue en prenant simplement la moyenne des complexités $\hat{\alpha}^*$, jugées les plus adéquates sur la base des différences observées (au sens de δ) entre X et les phylogénies des séquences $\hat{s}^*$. L'estimateur de la phylogénie t, proposé suite à toute cette procédure, est la phylogénie $\hat{t}_{\hat{\alpha}}$, *i.e.*, la phylogénie de complexité $\hat{\alpha}$ de la séquence $\hat{s}$ inférée depuis la matrice de données initiale X.

   Si l'on considère non pas X, mais $\hat{D}$ son équivalent en terme de distances, on retrouve dans cette méthode l'idée initiale du bootstrap, *i.e.*, estimer $\hat{t}-t$ depuis $\hat{t}^*-\hat{t}$: ici on estime la meilleure complexité $\hat{\alpha}$ pour que $\hat{t}$ estime $D=t$, par la meilleure complexité $\hat{\alpha}^*$ pour que $\hat{t}^*$ estime $\hat{D}$.

   Pour déduire $\hat{\alpha}$ on peut utiliser des fonctions plus complexes que la moyenne, toutefois celle-ci donne déjà des résultats très satisfaisants (cf. section 2.3). Notons aussi qu'utiliser une moyenne géométrique ou arithmétique donne expérimentalement des résulats du même ordre.

Comme l'approche à seuil fixe, l'approche présentée ci-dessus peut être employée en accord avec n'importe quel principe de reconstruction usuel, qu'il soit basé sur les caractères, sur les distances entre espèces ou autre. Le principe retenu déterminera le choix de δ de M, et la production des séquences de phylogénies. Nous donnons ci-dessous un exemple d'instanciation de ce schéma de bootstrap, en considérant un critère et une méthode de reconstruction apparentés aux quadruplets.

2.2 Exemple d'utilisation de l'approche à complexité variable

2.2.1 Le critère δ

Le critère δ doit permettre d'évaluer la qualité d'approximation d'une phylogénie $\hat{t}$ (ou $\hat{t}_{\hat{\alpha}}$) pour représenter les données. Ces données sont initialement présentes sous la forme d'une matrice de caractères. Cependant, afin de tenir compte des substitutions cachées, nous convertirons systématiquement toute matrice de caractères en matrice de dissimilarités entre espèces, que nous corrigerons par un modèle de l'évolution. La matrice obtenue depuis X et estimant les distances entre espèces est notée $\hat{D}$; de même on note $\hat{D}^*$ la matrice de dissimilarités obtenue depuis un jeu de données bootstrappé X^*.

Comme dans le reste de cette thèse, nous nous intéressons avant tout à l'aspect structurel (groupes d'espèces induits) de la phylogénie correcte t, plus qu'à son aspect numérique (distances entre espèces). Nous délaissons donc les critères usuels de distances (moindres carrés, quantité d'évolution minimum) et cherchons plutôt à définir le critère δ comme une mesure de la qualité d'adéquation *structurelle* entre la phylogénie inférée et la matrice de données $\hat{D}$. Pour ce faire, nous exploitons le fait que des contraintes structurelles sur des quadruplets d'espèces peuvent être définies aussi bien depuis une phylogénie (cf. section 1.5, chap. 1, part. I) que depuis une matrice de dissimilarités (p. ex. par la méthode FPM, section 2.2, chap. 2, part. I). Nous obtenons ainsi pour toute phylogénie τ inféré, un ensemble de r4-arbres Q_τ, et pour toute matrice de distances $\hat{D}$, un ensemble de r4-arbres $Q_{\hat{D}}$.

Robinson et Foulds [RF81] définissent la distance structurelle entre deux arbres comme la différence symétrique entre les ensembles de bipartitions qui les représentent. De façon similaire, nous pouvons définir δ comme la différence symétrique entre $\hat{D}$ et τ, du point de vue des ensembles de r4-arbres Q_τ et $Q_{\hat{D}}$ qui leur sont associés :

$$\delta_1(\hat{D},\tau) = \Big|\{q \in Q_\tau \text{ t.q. } q \notin Q_{\hat{D}}\}\Big|,$$
$$\delta_2(\hat{D},\tau) = \Big|\{q \in Q_{\hat{D}} \text{ t.q. } q \notin Q_\tau\}\Big|,$$

$$\delta(\hat{D},\tau) = \delta_1(\hat{D},\tau) + \delta_2(\hat{D},\tau). \tag{2.1}$$

Ainsi δ mesure le degré de désaccord structurel entre $\hat{D}$ et τ, *i.e.*, plus $\delta(\hat{D},\tau)$ est faible, plus la phylogénie τ est de bonne qualité pour approximer $\hat{D}$. Notons que si on applique ce critère à la matrice D de distances correctes, induite par la phylogénie t, alors $\delta(D,\tau)=0$ ssi $B(\tau) = B(t)$.

2.2.2 La méthode de reconstruction M optimisant δ

Pour optimiser δ on peut avoir recours, comme souvent en R.P., à une heuristique agglomérative[2]. Le principe est très simple : partant de la phylogénie étoile, de complexité nulle, à chaque étape on ajoute une arête dans la phylogénie (par

2. cf. description de Addtree section 2.2.3, chap. 2, part. I, pour une explication de ce procédé.

regroupements de sous-arbres connectés au noeud central), augmentant ainsi sa complexité de une unité. À chaque étape on choisit d'ajouter l'arête qui améliore le plus la qualité de la phylogénie pour approximer les données $\hat{D}$ (ou $\hat{D}^*$) au sens du critère δ. Le processus agglomératif s'arrête quand la phylogénie construite, $\hat{t}$, est totalement résolue, *i.e.*, de complexité maximale.

Au chapitre 3, part. III, nous montrerons comment un tel algorithme peut être conçu en temps $O(n^4)$.

2.2.3 Obtention d'une séquence de phylogénies

Pour obtenir une séquence de phylogénies $\hat{s}$ (ou de façon équivalente $\hat{s}^*$) depuis une phylogénie $\hat{t}$ (resp. $\hat{t}^*$) totalement résolue, on se restreint aux sous-arbres de cette phylogénie. La production de la séquence est réalisée de façon symétrique à la production de $\hat{t}$: partant de $\hat{t}_{max} = \hat{t}$, à chaque étape on enlève de la phylogénie courante, $\hat{t}_c$ de complexité c, l'arête interne de moindre qualité, pour obtenir la phylogénie suivante de la séquence, $\hat{t}_{c-1}$. Plus précisément, l'arête enlevée est choisie de façon à ce que la phylogénie $\hat{t}_{c-1}$ soit de qualité maximum (pour son degré de complexité) pour approximer le jeu de données $\hat{D}$ (resp. $\hat{D}^*$ pour la phylogénie $\hat{t}^*_{c-1}$), *i.e.*, de façon à ce que $\hat{t}_{c-1} = argmin_{e \in \hat{t}_c}(\delta(\hat{D}, B(\hat{t}_c) - \{e\}))$.

Les phylogénies de chaque séquence constituent donc ici une suite de phylogénies incluses les unes dans les autres, puisque chacune contient toutes les arêtes de la phylogénie précédente sauf une. De plus, comme chaque arête interne permet généralement d'optimiser davantage δ, les phylogénies $\hat{t}_{max}, \ldots, \hat{t}_0$ sont de qualité décroissante pour approximer le jeu de données au sens de δ.

Remarquons que les arêtes de $\hat{t}$ ne sont pas forcément éliminées dans l'ordre inverse avec lequel elles ont été inférées, puisque $\hat{t}$ est obtenu de façon heuristique, par une procédure gloutonne.

On peut voir une certaine similarité entre la façon décrite ci-dessus pour obtenir une séquence, et l'algorithme CART de Breiman *et al.* [BFOC84] bien connu en apprentissage. Cet algorithme obtient, dans une première phase (de "construction"), un arbre $\hat{t}_{max}$, et dans une deuxieme phase ("d'élagage de sous-arbres") une séquence d'arbres de complexité décroissante, chaque arbre $\hat{t}_{c-1}$ étant obtenue depuis l'arbre $\hat{t}_c$. Toutefois, l'élagage pratiqué par CART peut impliquer d'éliminer des feuilles de l'arbre, ce que nous interdisons ici. Cet algorithme possède la propriété intéressante suivante : chaque arbre de la séquence est le sous-arbre[3] de

3. dans CART, le terme *sous-arbre* prend le sens de *sous-ensemble connexe d'arêtes de l'arbre initial*, différent du sens que nous lui donnons dans cette thèse, à savoir *composante connexe résultant de la suppression d'une arête de l'arbre initial.*

l'arbre initial qui est optimal, étant donnée sa complexité, pour approximer $\hat{D}$ au sens du critère de δ (différent de celui utilisé ici). Dans notre cas, nous ne pouvons qu'espérer que cette propriété soit vérifiée (sans en être assurés). Toutefois, notre approche a l'avantage de produire une séquence de phylogénies complète (contenant un arbre par niveau de complexité), alors que CART produit parfois des séquences relativement creuses (p. ex. sans intermédiaire entre un arbre de 20 arêtes et un arbre de 4 arêtes), dont l'intérêt est d'autant limité.

2.2.4 Validation croisée

Le schéma d'application du bootstrap à la R.P. présenté dans ce chapitre peut facilement être adapté à d'autres techniques de rééchantillonnage que le bootstrap. Pour exemple, nous détaillons ici le cas de la validation croisée.

Chaque rééchantillonnage du jeu de données X nécessite son partage en deux demi-matrices de même taille, X_1 et X_2. Les colonnes de X (les caractères définis sur les espèces) sont aléatoirement attribuées soit à X_1 soit à X_2. Soit $\hat{D}_1$ et $\hat{D}_2$ les deux matrices de distances obtenues depuis X_1 et X_2 respectivement. La matrice $\hat{D}_1$ est utilisée à l'étape 2 du schéma (cf. section 2.1) pour produire la séquence $\hat{s}^*$. La matrice $\hat{D}_2$ est utilisée à l'étape 3 : la meilleur complexité $\hat{\alpha}^*$ est déduite comme étant celle de la phylogénie de la séquence $\hat{s}^*$ qui approxime le mieux $\hat{D}_2$ au sens de δ, *i.e.*, $\hat{\alpha}^* = argmin_c\ \delta(\hat{D}_2, \hat{t}_c^*)$. Le processus menant de X à $\hat{\alpha}^*$ est répété un nombre r' de fois et $\hat{\alpha}$ est obtenue par la moyenne des complexités $\hat{\alpha}^*$ issues des différents jeux (X_1, X_2). La phylogénie proposée par la procédure de validation croisée est, comme précédemment, la phylogénie de la séquence $\hat{s}$ (obtenue depuis le jeu de données initial X) ayant une complexité $\hat{\alpha}$.

On peut donc adapter de façon simple la méthodologie d'inférence de phylogénies proposée dans ce chapitre à la validation croisée. Toutefois, on peut douter que cette technique de rééchantillonnage soit, dans ce cadre, aussi efficace que celle du bootstrap. En effet, comme les phylogénies des séquences sont construites en ne considérant que la moitié des caractères initiaux (*i.e.*, X_1), et sont évaluées par l'autre moitié (*i.e.*, X_2), elles sont jugées d'autant plus sévèrement que la composition de X_1 et X_2 diffère. Ainsi pour des jeu de données X comportant peu de caractères, la méthode risque d'être trop pessimiste dans son jugement des phylogénies, et proposer une complexité $\hat{\alpha}$ plus faible que nécessaire. Pour des jeux X de taille importante, ce phénomène devrait progressivement s'atténuer car X_1 et X_2 contiennent de plus en plus de caractères en commun, et dans des proportions de plus en plus similaires.

2.3 Validation empirique

Nous avons réalisé quelques simulations pour comparer l'efficacité relative de l'approche à seuil fixe, de l'approche à complexité variable et de l'approche directe (sans rééchantillonnage). Ces simulations suivent principalement le schéma définit dans le chapitre précédent : différentes phylogénies modèles sont générées aléatoirement, puis des jeux de données X sont obtenus en simulant l'évolution de séquences moléculaires sur ces phylogénies ; enfin les performances des différentes approches sont comparées sur la base de leur efficacité à retrouver la phylogénie modèle depuis un jeu de données.

Les résultats sont détaillés en annexe D. Globalement, on peut en retenir que les deux approches utilisant le bootstrap (approche à seuil fixe et approche à complexité variable) proposent en moyenne des estimateurs meilleurs que la méthode directe, pour les conditions d'évolution testées, et que dans le cadre de la deuxième approche, la technique de validation croisée est significativement moins efficace que la technique du bootstrap.

Les deux approches employant la technique du bootstrap ont des résultats similaires. La seule différence est que, pour l'approche à seuil fixe, on peut fixer un nombre de réplications bootstrap moins important (un seul paramètre est estimé au lieu des $O(n)$ valeurs *bp* propres à l'autre appproche), ce qui aboutit ainsi à des temps calculs plus faibles.

2.4 Conclusion

Nous avons présenté dans ce chapitre une nouvelle méthode pour employer la technique du bootstrap en R.P afin d'estimer la phylogénie d'un groupe d'espèces. Cette méthode évite le problème de la méthode proposée par Felsenstein [Fel85] pour laquelle on doit déterminer *a priori* un des paramètres de l'estimation (le seuil S), indépendamment de la qualité du jeu de données et de la méthode employée.

La méthode bootstrap que nous avons introduite ici est appliquée à l'estimation de la complexité que doit adoter la phylogénie inférée, toutefois d'autres paramètres peuvent aussi être estimés suivant le même schéma. On pourrait par exemple estimer la valeur - au sens du critère de reconstruction - pour laquelle les arêtes doivent être rejetées ou conservées dans la phylogénie inférée. Si on s'intéresse aux longueurs des arêtes, on peut de la même façon déterminer quelle est la longueur minimale jugée nécessaire pour que les arêtes soient significatives ...

Troisième partie

Une approche combinatoire

Introduction

Les résultats de la partie précédente nous ont donné confirmation du fait qu'on pouvait obtenir de meilleurs estimateurs d'une phylogénie en proposant des phylogénies partiellement résolues, plutôt qu'en inférant des phylogénies totalement résolues. Pour inférer de telles phylogénies partiellement résolues, nous avons précédemment adopté une approche *à résolution décroissante*, *i.e.*, partant d'une phylogénie totalement résolue, nous cherchions à savoir quelles arêtes, ou combien d'arêtes, de cette phylogénie devaient être écartées pour obtenir une estimation plus juste de la phylogénie inconnue. Nous allons maintenant explorer l'approche inverse, dite *à résolution croissante*. Rappelons que son principe consiste à inférer d'emblée une phylogénie contenant uniquement des arêtes fiables, quitte à ce qu'elle soit partiellement résolue. Le processus de reconstruction est en effet stoppé dès que la qualité des arêtes que l'on pourrait encore rajouter peut être mise en doute.

Les algorithmes et critères de reconstruction que nous présentons dans cette partie de la thèse sont tous basés sur un même principe : les quadruplets. Comme nous le verrons au cours des chapitres suivants, ce principe permet facilement d'adopter une approche à résolution croissante.

Les **méthodes de reconstruction basées sur les quadruplets** comportent deux étapes :

1) on considère d'abord tous les sous-ensembles de quatre espèces pour lesquels on choisit indépendamment une phylogénie non-enracinée (pour moins de quatre espèces il n'y a qu'une phylogénie possible) ;

2) ces phylogénies sur quatre espèces constituent en quelque sorte un ensemble de contraintes structurelles, appelées *r4-arbres*[4], qu'on essaye ensuite de combiner au mieux en une phylogénie sur l'ensemble des espèces.

La deuxième étape est liée à l'optimisation combinatoire, puisque l'on cherche la

4. cf. section 1.5.5, chap. 1, part. *I*

phylogénie sur n espèces dont les sous-arbres restreints à 4 espèces correspondent autant que possible aux r4-arbres de départ.

Pour mieux comprendre le principe des méthodes de quadruplets, on peut introduire une analogie avec un jeu bien connu : les r4-arbres peuvent être vues comme les *pièces* d'un *puzzle* qu'on cherche à reconstuire (*i.e.*, la phylogénie des espèces), sachant que certaines pièces ont été perdues (il est possible que pour certains quadruplets, on ne puisse choisir un r4-arbre plutôt qu'un autre), et que certaines pièces ont été déformées (pour certains quadruplets on peut inférer le mauvais r4-arbre), et ne s'emboîtent donc plus forcément avec les autres.

Plusieurs raisons existent pour recourir à une méthode de quadruplets. Premièrement, certains principes de reconstruction ont une efficacité reconnue mais sont inapplicables en pratique pour plus d'une quinzaine d'espèces, car de complexité exponentielle en n (le nombre d'espèces) : ainsi de la méthode du Maximum de Vraisemblance [Fel81], de la méthode du *Closest Tree* [Hen91] ou de la méthode de *Statistical Geometry* [EWOD88, SOWH96]. L'approche basée sur les quadruplets permet de contourner cet obstacle puisque appliquer un de ces principes bien-fondés sur seulement quatre espèces reste raisonnable en temps calcul, et les r4-arbres ainsi inférés sont ensuite combinés en une phylogénie par une heuristique polynomiale (cf. [ST77, SvH96, ESSW97b] ou chap. 3).

Il existe aussi des principes de reconstruction qui ne peuvent pas s'appliquer sur plus de quatre espèces pour des raisons mathématiques, comme les invariants de Lake [Lak87] et de Cavender [Cav81]. Ces principes sont toutefois intéressants car leur consistance a été montrée dans des cas où bon nombre des méthodes de R.P. sont inconsistantes [HH93]. Pour appliquer ces principes à plus de quatre espèces il n'y a pas d'autre solution que de recourir aux méthodes de quadruplets.

Une autre justification des méthodes de quadruplets se trouve dans l'étude de congruence de phylogénies [Kim93, Lap97], où l'on cherche à savoir si des phylogénies inférées depuis différents critères de reconstruction (basés sur les caractères, les distances, ou les probabilités) "racontent la même histoire évolutive" pour les espèces. Comme nous l'avons dit en introduction, une méthode de R.P. est généralement définie par un critère de reconstruction et par un algorithme optimisant ce critère. Cependant l'algorithme employé est généralement différent d'un critère à l'autre, rendant difficile toute comparaison objective des critères en eux-même. Pour obtenir une telle comparaison, on peut adopter un principe de quadruplets : des ensembles de r4-arbres sont inférés séparemment par chacun des critères à comparer, et sont ensuite chacun convertis en phylogénies en appliquant à chaque fois le même algorithme de reconstruction basé sur les quadruplets. On évite ainsi l'aléa introduit autrement par l'utilisation d'algorithmes de reconstruction généralement différents pour chacun des critères examinés.

D'autres raisons encore, toutefois moins importantes, conduisent à utiliser les méthodes de quadruplets ; nous renvoyons le lecteur à l'article de Steel [Ste92] pour une énumération de ces raisons.

Une heuristique reposant sur le principe des quadruplets a déjà été mentionnée rapidement au chap 2, part. II. Ce type de méthodes fait maintenant l'objet d'un traitement plus systématique. Dans le chap. 1 de la présente partie, nous donnons un certain nombre de notions et de résultats sur la reconstruction d'une phylogénie depuis un ensemble de r4-arbres. Nous introduisons aussi le problème SMA (Sous-ensemble Maximum Arboré), et montrons qu'il est pertinent pour la R.P. La spécificité d'une solution à ce problème est de définir une phylogénie dont chaque arête est combinatoirement sûre, au sens où elle satisfait un nombre très important de r4-arbres de données. Puis, le chap. 2 détaille un algorithme pour résoudre le problème SMA. Nous présentons d'abord une version simple de cet algorithme, de complexité $O(n^5)$, puis nous montrons comment cette complexité peut être ramenée à $O(n^4)$ en assurant que chaque r4-arbre n'est examiné qu'une seule fois. Le chap. 3 conclura cette partie de la thèse en présentant un schéma heuristique générique pour reconstruire une phylogénie depuis des quadruplets selon un principe agglomératif. Ce schéma heuristique peut être employé avec une variété de critères basés sur les quadruplets, tout en ne nécessitant qu'une complexité $O(n^3 + |Q|)$. Cette heuristique possède aussi la propriété intéressante d'accepter que des contraintes structurelles soient fixées *a priori* pour la reconstruction (pour une faible augmentation de sa complexité, alors en $O(n^4)$). Le biologiste peut ainsi imposer des regroupements d'espèces déjà connus avec certitude et profiter de l'information amenée par les autres espèces étudiées pour déterminer les relations de parenté à l'intérieur de ces groupes d'espèces. Par exemple, on peut utiliser le schéma heuristique, suite à l'algorithme résolvant le problème SMA, présenté au chap. 2.

Chapitre 1

Reconstruction de phylogénies depuis des r4-arbres

Au chapitre 1 de la partie I, nous avons défini, pour un ensemble Q de r4-arbres, les notions d'arbre-consistance et d'arboricité. En correspondance avec ces notions, on peut définir les deux problèmes suivants :

Problème ACR **:** Arbre-Consistance d'un ensemble de R4-arbres
Entrée : un ensemble Q de r4-arbres
Question : est-ce que Q est arbre-consistant ?

Problème AR **:** Arboricité d'un ensemble de R4-arbres
Entrée : un ensemble Q de r4-arbres
Question : est-ce que Q est arboré ?

Ces deux problèmes sont très différents d'un point de vue opérationnel puisque AR se résout en temps polynômial [CS81, BD86], tandis que ACR est NP-complet [Ste92]. L'algorithme permettant de résoudre le problème AR pour un ensemble de r4-arbres Q repose sur les propriétés structurelles suivantes [CS81, BD86] : [1]

1. Dans leur présentation de ces propriétés, Colonius et Schulze [CS81], ainsi que Bandelt et Dress [BD86], considéraient non pas un ensemble de r4-arbres, mais une relation quaternaire R définie sur $E^2 \times E^2$, d'où le fait que leur formulation imposait aussi que la relation quaternaire soit symétrique (*i.e.* $xyRzt \Leftrightarrow ztRxy$) et 2-antisymétrique (*i.e.*, au plus une des trois expressions $xyRzt, xzRyt, xtRyz$ est vérifiée). De par la définition des r4-arbres, ces deux propriétés sont toujours vérifiées.

Propriété 2 *Les trois propositions suivantes sont équivalentes :*

- Q *est arboré.*
- $\forall\, x, y, z, t, u,\ \{xu|yz, xu|zt\} \cap Q \neq \emptyset$ *et* $xy|zt \in Q\ \Rightarrow yu|zt \in Q$ *[CS81].*
- $\forall\, x, y, z, t,\ xy|zt \in Q\ \Rightarrow \forall\, u \neq \{x, y, z, t\},\ \{xy|zu, xu|zt\} \cap Q \neq \emptyset$ *[BD86].*

Ainsi, il suffit de vérifier que l'ensemble Q respecte une de ces propriétés pour résoudre le problème AR. Par exemple, par la troisième proposition, on sait en $O(n^4)$ si Q est arboré. On utilise pour celà une propriété supplémentaire liée à cette proposition [BD86], *i.e.*, le fait qu'on peut fixer une des espèces, soit x : si pour toutes les espèces y, z, t, u, l'ensemble Q vérifie la proposition, alors il vérifie aussi la proposition pour tous les x autres que celui fixé. Il suffit ainsi de faire parcourir à y, z, t, u l'ensemble des espèces E, ce qui demande $O(n^4)$, et de vérifier à chaque fois en $O(1)$ que les r4-arbres requis sont présents dans Q.

Le problème ACR a été montré NP-complet [Ste92], mais un certain nombre de cas particuliers de ce problème sont polynômiaux (dans P) :

- quand Q est complet : car on est alors ramené au problème AR.
- quand les r4-arbres de Q référencent tous une même espèce x : on est alors ramené au cas de triplets enracinés, pour lequel on peut appliquer l'algorithme polynômial de Aho *et al* [ASSU81] (cf. [Ste92]).

On peut aussi montrer [Ste97] que le dernier de ces cas particuliers ne s'étend pas à 2 espèces : s'il existe deux espèces x et y telles que tout r4-arbre de Q référence soit x, soit y, le problème ACR reste NP-complet.

Disposant d'un ensemble de r4-arbres Q sur les espèces E, si Q correspond à une phylogénie (*i.e.*, si Q est arboré ou arbre-consistant), on peut naturellement penser proposer une telle phylogénie pour estimer la phylogénie t des espèces E. Malheureusement, en raison du bruit évolutif affectant généralement les données (certaines informations nous arrivent déformées), il est peu vraisemblable qu'un ensemble de r4-arbres issu de données réelles possède une de ces propriétés. Aussi, pour estimer une phylogénie depuis un ensemble de r4-arbres Q, vient naturellement l'idée de chercher le plus grand sous-ensemble de Q qui vérifie une de ces propriétés. On aboutit ainsi aux problèmes d'optimisation suivants :

Problème $SMAC$ **:** Sous-ensemble Maximum Arbre-Consistant
Entrée : un ensemble Q de r4-arbres
Sortie : un sous-ensemble de Q de cardinal maximum qui soit arbre-consistant.

Problème SMA **:** Sous-ensemble Maximum Arboré
Entrée : un ensemble Q de r4-arbres
Sortie : le sous-ensemble Q^* de Q de cardinal maximum qui soit arboré.

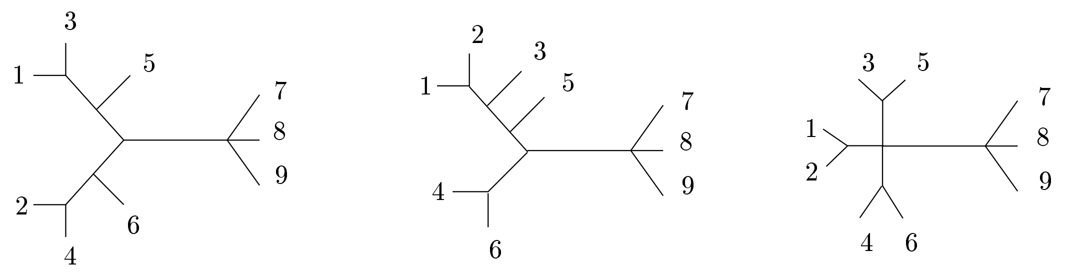
Le problème $SMAC$ est NP-difficile en raison de la NP-complétude du problème ACR. Il ne pourra donc être raisonnablement résolu que de façon approchée. Diverses heuristiques peuvent être employées pour ce problème, comme *Addtree* [ST77], *Quartet Puzzle* [SvH96], *SQM* [ESSW97b], ou encore un algorithme de recuit simulé [DvHK86].

Pour le problème SMA, aucun algorithme approché ou exact n'était connu jusqu'aux travaux menés dans le cadre de cette thèse [BG97], même si plusieurs auteurs conjecturaient que le problème devait être polynômial [Ban96, Ste97]. Pour résoudre ce problème (chap. 2) nous utilisons une caractérisation de Q^* donnée par Bandelt et Dress [BD86] (rappelons que Q_b désigne les r4-arbres induits par la bipartition b) :

Définition 2 ([BD86]) *Soit* B^* *l'ensemble de bipartitions* $\{b=\sigma|\overline{\sigma} \text{ t.q. } Q_b \subseteq Q\}$*, alors* $Q^* = \bigcup_{b\in B^*} Q_b$.

Bandelt et Dress prouvent que la définition ci-dessus implique que l'ensemble Q^* est l'unique sous-ensemble arboré de cardinalité *maximum* de Q. La preuve repose sur le fait que B^* est un ensemble de bipartitions compatibles en une phylogénie (notée T^*), et qu'il existe une bijection entre les sous-ensembles de bipartitions de B^* et les sous-ensembles de Q arborés. Les sous-ensembles arborés de Q sont ainsi organisés en sup demi-treillis dont l'unique élément maximum est l'ensemble Q^* correspondant à l'ensemble B^* en entier.

Notons aussi que, trivialement, quand Q est arboré, $Q^* = Q$. À l'autre extrême, nous pouvons aussi avoir $Q^* = \emptyset$, auquel cas B^* contient uniquement des bipartitions triviales et T^* est la phylogénie étoile sur E. Curieusement, l'ensemble Q^* n'est jamais un sous-ensemble arbre-consistant *maximum* de Q (sauf quand $Q^* = Q$), puisque ce n'est même pas un sous-ensemble arbre-consistant *maximal* de Q : $\forall r \in Q - Q^*$, $\{r\} \cup Q^*$ est arbre-consistant. De plus, on peut aisément trouver un contre-exemple montrant que Q^* (le sous-ensemble arboré maximum de Q) n'est pas toujours contenu dans le sous-ensemble arbre-consistant maximum de Q :



t_1 (53 r4-arbres satisfaits) t_2 (52 r4-arbres satisfaits) t_3 (30 r4-arbres satisfaits)

FIG. 1.1 – *Exemple montrant que l'ensemble Q^* n'est pas forcément inclus dans le sous-ensemble* arbre-consistant *maximum de Q.*

Exemple 1 *Soit Q l'ensemble composé des r4-arbres suivants :*

- $\{12|xy, \forall\{x,y\} \subseteq \{3,\dots,9\}, x \neq y\}$
- $\{13|xy, \forall\{x,y\} \subseteq \{4,5,6,7,8,9\}, x \neq y\}$
- $\{15|xy, \forall\{x,y\} \subseteq \{4,6,7,8,9\}, x \neq y\}$
- $\{24|xy, \forall\{x,y\} \subseteq \{3,5,6,7,8,9\}, x \neq y\}$
- $\{26|xy, \forall\{x,y\} \subseteq \{3,5,7,8,9\}, x \neq y\}$

L'ensemble B^* contient la bipartition $b = \{1,2\}|\{3,\dots,9\}$. Les phylogénies t_2 et t_3 de la Fig. 1.1 contiennent chacune une arête qui induit cette bipartition. Parmi les phylogénies contenant une arête correspondant à cette bipartition, la phylogénie t_2 de la figure est celle qui satisfait le plus grand sous-ensemble de r4-arbres de Q (à égalité avec les trois phylogénies qui résolvent son noeud irrésolu, connectant 7,8,9 au reste de la phylogénie). La phylogénie t_1 ne contient pas d'arête induisant la bipartition b, une analyse exhaustive indique que c'est pourtant la phylogénie qui satisfait le plus grand sous-ensemble de r4-arbres de Q (à égalité avec les trois phylogénies qui résolvent le noeud irrésolu de t_1). Les r4-arbres $\{12|xy$ t.q. $x,y \in \{3,4,5,6\}, x \neq y\}$ appartiennent à Q^* mais ne sont pas consistants avec t_1, ainsi le sous-ensemble arboré *maximum de Q, i.e., Q^*, n'est pas ici inclus dans le sous-ensemble* arbre-consistant *maximum de Q.*

Signalons aussi qu'au contraire du sous-ensemble maximum arboré, un sous-ensemble arbre-consistant maximum n'est pas forcément unique :

Exemple 2 *L'ensemble* $Q = \{12|34,\ 12|35,\ 14|25\}$ *n'est pas arbre-consistant, et on peut aisément vérifier que tout sous-ensemble de deux r4-arbres de* Q *est arbre-consistant. En conséquence, un sous-ensemble arbre-consistant maximum d'un ensemble de r4-arbres n'est pas forcément unique.*

Obtenir le sous-ensemble maximum arboré d'un ensemble de r4-arbres constitue un critère de R.P. intéressant pour plusieurs raisons. D'abord ce sous-ensemble, noté Q^*, correspond exactement à une phylogénie, T^* ; et passer d'une des représentations T^*, Q^* ou B^* à l'autre est peu coûteux en temps calcul. Par exemple, plusieurs algorithmes [Mea81, Gus91a] permettent d'obtenir T^* en $O(n.b)$ depuis les b bipartitions de B^* définies sur les n espèces de E ; obtenir Q^* depuis T^* est en $O(|Q^*|)$, etc.

Comme l'ensemble Q^* est unique, et que tout ensemble de r4-arbres arboré correspond à une et une seule phylogénie [CS81], T^* est la seule phylogénie correspondant à la plus grande partie des données qui soit représentable sous cette forme. Autrement dit, la phylogénie T^* ne contredit aucune partie des données (*i.e.*, aucun r4-arbre de Q) et ne nécessite aucune hypothèse autre que celles faites par les données (*i.e.*, aucun r4-arbre n'appartenant pas à Q).

Enfin, **la phylogénie T^* ne possède que des arêtes combinatoirement sûres** au sens où chacune est soutenue par $O(n^2)$ à $O(n^4)$ r4-arbres des données. En effet, par définition, tous les r4-arbres induits par une arête e doivent être présents dans les données, or selon la répartition des espèces autour de e, aux extrêmes $\sigma_e = 2$, $\overline{\sigma}_e = n-2$, et $\sigma_e = \frac{n}{2}$, $\overline{\sigma}_e = \frac{n}{2}$, il y a entre $\binom{n-2}{2}$ et $\binom{n/2}{2}\binom{n/2}{2}$ façons de choisir deux espèces de chaque côté de l'arête, donc entre $O(n^2)$ et $O(n^4)$ r4-arbres induits. Ainsi, il est vraisemblable qu'aucune arête incorrecte ne sera inférée dans T^* en raison d'erreurs aléatoires résultant de l'obtention des séquences de caractères, de leur alignement ou encore de l'inférence des r4-arbres. Cependant T^* n'est pas à l'abri d'erreurs systématiques affectant les données qu'il analyse, comme à la suite de l'inconsistance de la méthode d'inférence des r4-arbres.

En proposant ainsi une estimation stable de la phylogénie, le problème SMA s'inscrit dans la démarche propre à l'approche à résolution croissante : on propose d'emblée des arêtes ayant une forte probabilité d'être correctes, quitte à n'inférer qu'une phylogénie partiellement résolue (on peut en effet douter que T^* soit binaire, ce qui nécessiterait que le jeu de données Q soit complet et arboré !).

Dans le chapitre suivant, nous montrons comment on peut calculer $Q^*(\equiv T^*)$ et ainsi résoudre le problème SMA.

Chapitre 2

Résolution du problème du Sous-ensemble Maximum Arboré

Au chap. 1 nous avons défini le problème SMA consistant à trouver le Sous-ensemble Maximum Arboré, Q^*, d'un ensemble de r4-arbres Q. Nous avons montré que ce sous-ensemble possède des propriétés intéressantes pour la R.P. Nous montrons dans ce chapitre comment Q^* peut être calculé de façon exacte en temps polynômial.

Dans un premier temps nous montrons qu'une approche incrémentale est correcte pour calculer Q^*, puis nous donnons un premier algorithme simple, de complexité $O(n^5)$ adoptant une telle approche. Nous montrons ensuite comment on peut encore améliorer ce résultat, tout en conservant une approche incrémentale, ce qui conduit à un algorithme de complexité $O(n^4)$.

2.1 Calcul de Q^* par un principe incrémental

L'idée d'utiliser un algorithme incrémental pour calculer Q^* vient des travaux de Bandelt et Dress [BD92] qui mentionnent un algorithme de ce type pour obtenir les $O(n^2)$ *d-splits*[1] d'une matrice de distances d en temps $O(n^6)$. Pour le problème SMA, adopter une démarche incrémentale nous permettra d'obtenir ici un premier algorithme simple calculant les $O(n)$ bipartitions de B^* en temps $O(n^5)$. L'approche incrémentale consiste à ne considérer initialement que quatre espèces, pour lesquelles on calcule l'ensemble $B^* \equiv Q^*$, puis à prendre en compte les espèces restantes une par une. À chaque fois qu'on incorpore une nouvelle espèce, l'ensemble B^* est mis à jour facilement grâce à une relation simple le liant à celui de l'étape précédente. Quand on considère les quatre espèces initiales, un seul

1. les d-splits de d sont les splits $A|B$ t.q. $A \cup B = E$ et $\forall a, a' \in A,\ b, b' \in B,\ max\{(d_{ab} + d_{a'b'}), (d_{a'b} + d_{ab'})\} - (d_{aa'} + d_{bb'}) > 0$.

r4-arbre de Q, celui référençant ces espèces, a besoin d'être examiné pour calculer l'ensemble B^*. Puis à chaque fois qu'on prend en compte une nouvelle espèce, on élargit l'ensemble des r4-arbres de Q examinés aux r4-arbres référençant cette nouvelle espèce, et ainsi de suite, jusqu'à avoir examiné tout l'ensemble Q pour insérer la dernière espèce.

Définition 3 *Nous supposons établi un ordre arbitraire* $(1, 2, \dots, n)$ *sur les espèces de* E *et notons* $[i]$ *les* i *premières espèces.*

- $Q_{[i]}$ *dénote le sous-ensemble de r4-arbres de* Q référençant *uniquement des espèces dans* $[i]$*, i.e.,* $\forall\, xy|zt \in Q_{[i]}$*, on a* $xy|zt \in Q$ *et* $\{x, y, z, t\} \subseteq [i]$.
- Q_i *dénote le sous-ensemble de r4-arbres de* $Q_{[i]}$ *référençant l'espèce* i*, i.e.,* $\forall\, xy|zt \in Q_i,\ i \in \{x, y, z, t\}$.
- B_i^* *est l'ensemble de bipartitions* $b = \sigma|\overline{\sigma}$ *sur* $[i]$ *t.q.* $Q_b \subseteq Q_{[i]}$*. Ainsi* B_i^* *contient au moins les* $i+1$ *bipartitions triviales* $\{1\}|\{2, \dots, i\}, \dots,$ $\{i\}|\{1, \dots, i-1\}$*,* $\emptyset|\{1, \dots, i\}$[2]*, qui n'induisent aucun r4-arbre.*
- $Q_i^* = \bigcup_{b \in B_i^*} Q_b$.

Dans les définitions précédentes, B_i^* et Q_i^* sont les équivalents de B^* et Q^* quand on restreint E aux espèces $[i]$. Ainsi ces définitions impliquent que Q_i^* est l'unique sous-ensemble maximum arboré de $Q_{[i]}$ (T_i^* dénote la phylogénie correspondante) et nous avons $Q_n^* = Q^*$.

Pour le principe incrémental, les espèces sont considérées dans l'ordre arbitraire $(1, 2, \dots, n)$ et à chaque étape i, une nouvelle espèce est considérée (soit i), et l'ensemble B_i^* est obtenu depuis B_{i-1}^* en fonction de $Q_{[i]}$. Chaque bipartition $\sigma|\overline{\sigma} \in B_{i-1}^*$, définie sur $[i-1]$, peut *a priori* être étendue en deux bipartitions définies sur $[i]$: $b = \sigma \cup \{i\}|\overline{\sigma}$ et $b' = \sigma|\overline{\sigma} \cup \{i\}$. B_i^* contiendra une, deux ou aucune de ces bipartitions selon que $Q_b \subseteq Q$ et que $Q_{b'} \subseteq Q$. Nous montrons maintenant la relation existant entre les ensembles B_{i-1}^* et B_i^*.

2.1.1 Justification de l'approche incrémentale

Définition 4 *Soit* b *une bipartition sur* $[i]$*,* $b_{\setminus i}$ *dénote sa* restriction *aux espèces* $[i-1]$*. Dans un sens plus large,* $B_{\setminus i}^*$ *dénote les bipartitions de* B_i^* *restreintes aux espèces* $[i-1]$*.*

2. *cette* bipartition *artificielle est nécessaire pour la suite.*

Par exemple, si $b=\{1,2,3\}|\{4,5,6\}$, alors $b_{\backslash 6}=\{1,2,3\}|\{4,5\}$.

Lemme 3 $B^*_{\backslash i} \subseteq B^*_{i-1}$.

Preuve : Soit $b=\sigma|\overline{\sigma}$ une bipartition de B^*_i. Si b est une bipartition triviale, alors $b_{\backslash i}$ est triviale et appartient ainsi à B^*_{i-1}. Autrement, supposons que $i \in \sigma$ (de façon équivalente $i \in \overline{\sigma}$). Le fait que $b \in B^*_i$ implique que $\forall x, y \in \sigma$ et $z, t \in \overline{\sigma}$, on a $\{x, y, z, t\} \subseteq [i]$ et $xy|zt \in Q_{[i]}$. Ainsi, nous avons deux possibilités :

1. $b_{\backslash i}=\sigma-\{i\}|\overline{\sigma}$ est triviale ($|\sigma-\{i\}|=1$) ce qui implique $b_{\backslash i} \in B^*_{i-1}$.

2. $b_{\backslash i}$ est non-triviale et $\forall x, y \in \sigma-\{i\}$ et $z, t \in \overline{\sigma}$ nous avons $\{x, y, z, t\} \subseteq [i-1]$ et $xy|zt \in Q_{[i]}$, c'est-à-dire plus précisément $xy|zt \in Q_{[i-1]}$. D'où par définition, $b_{\backslash i} \in B^*_{i-1}$.

□

Le lemme précédent ne peut pas être complété pour affirmer que $B^*_{\backslash i} = B^*_{i-1}$, comme le montre l'exemple suivant : soit $Q=\{12|34, 15|23, 15|24, 15|34\}$, d'où B^*_4 contient $\{12\}|\{34\}$ et les bipartitions triviales sur [4], B^*_5 contient $\{15\}|\{234\}$ et les bipartitions triviales sur [5], et $B^*_{\backslash 5}$ contient uniquement les bipartitions triviales sur [4]. Nous avons ainsi $b = \{1,2\}|\{3,4\} \in B^*_4$ mais $b \notin B^*_{\backslash 5}$ [3] montrant que dans certains cas $B^*_{\backslash i} \neq B^*_{i-1}$.

Corollaire 1 *Toute bipartition de B^*_i peut être directement obtenue depuis B^*_{i-1}, i.e., $\forall b \in B^*_i$, $\exists\, b'=\sigma|\overline{\sigma} \in B^*_{i-1}$ t.q. $b=\sigma \cup \{i\}|\overline{\sigma}$ ou $b=\sigma|\overline{\sigma} \cup \{i\}$.*

Preuve : $\forall b \in B^*_i$, par le lemme précédent nous savons $b_{\backslash i} = \sigma|\overline{\sigma} \in B^*_{i-1}$ et par définition de $b_{\backslash i}$, b est $\sigma|\overline{\sigma} \cup \{i\}$ ou $\sigma \cup \{i\}|\overline{\sigma}$. Dans les deux cas, il existe une bipartition de B^*_{i-1} qui peut être étendue en b. □

Ce corollaire montre le bien fondé de l'approche incrémentale pour obtenir $B^*_n = B^*$, et ce, quel que soit l'ordre dans lequel les espèces sont considérées (B^* et Q^* sont uniques). Ce corollaire a aussi une implication opérationnelle sur le sous-ensemble de r4-arbres de Q qu'il est nécessaire d'examiner à chaque étape :

Propriété 3 *A l'étape i, seuls les r4-arbres $xi|yz \in Q$ t.q. $\{x, y, z\} \subseteq [i-1]$ ont besoin d'être examinés.*

3. Pour que $b \in B^*_{\backslash 5}$ il faudrait $\{125\}|\{34\} \in B^*_5$ (impossible car $\{25|34, 12|35\} \notin Q$) ou $\{12\}|\{345\} \in B^*_5$ (impossible car $\{12|35, 12|45\} \notin Q$).

En effet, soit b une bipartition obtenue par extension d'une bipartition $b' \in B^*_{i-1}$. Pour savoir si $b \in B^*_i$, nous n'avons pas à vérifier dans Q la présence de tous les r4-arbres de Q_b, puisque nous savons déjà que $Q_{b'} \subseteq Q$. Comme $Q_{b'} \subseteq Q_b$, nous avons juste à examiner les nouveaux r4-arbres $(Q_b - Q_{b'})$, induits par l'ajout de i dans $b' = \sigma|\overline{\sigma}$. Ainsi, pour savoir si $b = \sigma \cup \{i\}|\overline{\sigma}$ (ou de façon équivalente si $b = \sigma|\overline{\sigma} \cup \{i\}$) appartient à B^*_i, nous devons juste vérifier que le sous-ensemble Q_i de Q contient les r4-arbres $xi|yz$ pour lesquels $x \in \sigma$ et $y, z \in \overline{\sigma}$.

2.1.2 Un algorithme incrémental simple

L'algorithme considère initialement les quatre premières espèces et obtient trivialement l'ensemble B^*_4 : cet ensemble contient les quatre bipartitions triviales sur $[4]$, plus éventuellement la bipartition non-triviale sur ces quatre espèces qui correspond au r4-arbre éventuellement présent dans Q (p. ex. la bipartition $\{1, 2\}|\{3, 4\}$ si $12|34 \in Q$). L'algorithme élargit ensuite progressivement l'ensemble $[i]$ des espèces considérées et étend conjointement l'ensemble B^*_i (cf. Algorithme 1). À l'étape où l'espèce i est considérée, pour chaque bipartition $b = \sigma|\overline{\sigma} \in B^*_{i-1}$, on cherche à savoir si $b' = \sigma \cup \{i\}|\overline{\sigma}$ et $b'' = \sigma|\overline{\sigma} \cup \{i\}$ se qualifient comme bipartitions de B^*_i. De par la propriété 3, il suffit seulement pour cela de vérifier que $(Q_{b'} - Q_b) \subseteq Q_i$ et $(Q_{b''} - Q_b) \subseteq Q_i$, respectivement. Le même traitement est répété étape après étape, jusqu'à ce que l'espèce n ait été prise en compte, *i.e.*, qu'on dispose de $B^*_n = B^*$.

L'initialisation de l'algorithme (lignes 1-2) demande $O(1)$ opérations. On supposera en effet que les r4-arbres de Q sont répartis dans une matrice quadridimensionnelle (et triangulaire inférieure), chaque entrée correspondant à un quadruplet x, y, z, t et pouvant prendre quatre valeurs possibles pour indiquer quel r4-arbre de $\mathcal{R}(q)$ appartient à Q ou, qu'aucune résolution n'est présente dans Q pour ce quadruplet. Notons que si Q n'est initialement disponible que sous la forme d'une liste de r4-arbres, un prétraitement en $O(n^4)$ est nécessaire pour initialiser la matrice à quatre dimensions, ce qui n'augmente toutefois pas la complexité globale de l'algorithme.

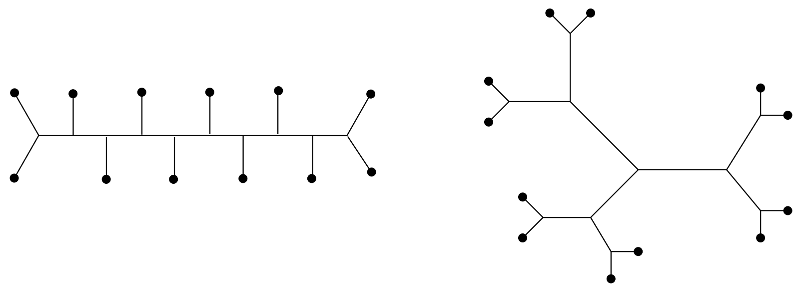
L'algorithme 1 calcule B^* en $O(n)$ étapes (ligne 3). Chacune examine les deux extensions possibles des $O(n)$ bipartitions de B^*_i. Le traitement de chaque bipartition nécessite d'interroger Q dans le pire des cas pour les $O(n^3)$ r4-arbres référençant i (lignes 4-5).

On obtient ainsi un algorithme simple en $O(n^5)$. Une analyse plus fine de cet algorithme montre que même dans le cas où T^*_i est complètement résolue à chaque étape, la complexité oscille en fait entre $O(n^5)$, si T^*_i est une phylogénie *chenille* (*caterpillar*) (Fig. 2.1-(1)), et $O(n^4 \log n)$, si T^*_i est une phylogénie binaire équilibrée (Fig. 2.1-(2)).

Algorithme 1: Calcul de $B^* \equiv Q^*$ par application directe du principe incrémental

Données: Un ensemble de r4-arbres Q
Résultat: Le plus grand sous-ensemble arboré Q^* de Q

```
/* Initialisations */
B*_4 ← { {1}|{2,3,4}, {2}|{1,3,4}, {3}|{1,2,4}, {4}|{1,2,3} }
si ∃ r = xy|zt, {x,y,z,t} = {1,2,3,4} t.q. r ∈ Q alors
    B*_4 ← B*_4 ∪ { {x,y}|{z,t} }
/* Ajout progressif des autres espèces */
pour i ∈ 5 ... n
    B*_i ← { {i}|{1,...,i−1} }
    pour chaque σ|σ̄ ∈ B*_{i−1}
        Ext_g, Ext_d ← vrai
        pour chaque x ∈ σ, y ∈ σ̄ et tant que Ext_g ou Ext_d est vrai
            pour chaque z ∈ σ̄, z ≠ y et tant que Ext_g = vrai
                si xi|yz ∉ Q alors Ext_g ← faux
            pour chaque z ∈ σ, z ≠ x et tant que Ext_d = faux
                si xz|yi ∉ Q alors Ext_d ← faux
        si Ext_g = vrai alors B*_i ← B*_i ∪ { σ ∪ {i}|σ̄ }
        si Ext_d = vrai alors B*_i ← B*_i ∪ {σ|σ̄ ∪ {i}}
```



(1) (2)

FIG. 2.1 – *Différentes formes de phylogénies. (1) phylogénie chenille ; (2) phylogénie équilibrée.*

2.2 IQ^* un algorithme en $O(n^4)$ pour résoudre SMA

Nous présentons ici l'algorithme IQ^*, améliorant la complexité de l'algorithme précédent pour le problème SMA. Ce deuxième algorithme repose lui aussi sur un principe incrémental, mais on s'intéresse maintenant plus à $T_i^* \equiv Q_i^*$ qu'à $B_i^* \equiv Q_i^*$. En quelques mots, à chaque étape la question n'est plus de savoir "quelles bipartitions de B_{i-1}^* s'étendent en bipartitions de B_i^* par ajout de l'espèce i", mais "**où doit être greffée l'espèce i dans la phylogénie T_{i-1}^* pour obtenir T_i^***". Pour répondre à cette question on examine l'insertion de i dans tous les sous-arbres de T_{i-1}^* et, des réponses positives et négatives selon les sous-arbres, on déduit l'emplacement exact où greffer i. De même que certaines bipartitions de B_{i-1}^* ne pouvaient précédemment être étendues en bipartitions de B_i^*, certaines arêtes de T_{i-1}^* ne supportent l'insertion de l'espèce i dans aucun des deux sous-arbres qui leur sont associés. En conséquence, ces arêtes sont éliminées pour obtenir T_i^* (*i.e.*, leurs extrémités sont confondues). Par un parcours spécifique des sous-arbres de T_i^* et par un partage d'information entre ces sous-arbres, on garantit que chaque r4-arbre n'est examiné qu'une seule fois. Ainsi, à chaque étape, T_i^* est obtenu depuis T_{i-1}^* en $O(n^3)$, alors que l'algorithme simple décrit précédemment obtenait B_i^* depuis B_{i-1}^* en $O(n^4)$.

Ci-dessous, nous expliquons plus formellement et plus longuement la mécanique de ce nouvel algorithme ; puis nous décrivons de quelle façon les différents sous-arbres peuvent partager de l'information, avec pour seul prérequis qu'ils soient parcourus dans un ordre spécifique. Nous décrivons ensuite comment cet ordre de parcours des sous-arbres peut être réalisé. Enfin nous détaillons l'algorithme résultant de l'application de ces divers principes et analysons sa complexité.

2.2.1 Principe détaillé de l'algorithme

Définition 5 *Plaçons-nous à l'étape i (i.e., où l'espèce i est considérée) et soit* $b{=}\sigma|\overline{\sigma}$ *une bipartition de* B_{i-1}^*

- σ *et* $\overline{\sigma}$ *sont appelées les* **parties** *ou les* **composantes** *de la bipartition b.*
- *Un r4-arbre* $xi|yz$ *est* **requis** *par la composante* σ *de* b *s'il est induit par l'ajout de* i *dans cette partie de* b*, i.e., ssi* $x \in \sigma$ *et* $y, z \in \overline{\sigma}$*. En d'autres termes, la présence de ce r4-arbre dans* Q *est* **requise** *pour que* $\sigma \cup \{i\}|\overline{\sigma}$ *soit une bipartition valide de* B_i^**. Notons que cette définition n'englobe que les r4-arbres qui sont "nouvellement"* requis *(i.e., ceux de* Q_i*), puisque on*

sait par la propriété 3 que les autres r4-arbres de $Q_{[i]}$ n'ont pas besoin d'être examinés. L'ensemble des r4-arbres requis par la composante σ est noté R_σ.

- *Un r4-arbre est dit* **traité** *quand sa présence dans Q a été examinée.*
- *Dans la suite de ce chapitre le terme* **sous-arbre** *d'une phylogénie quelconque dénotera uniquement une composante connexe de cette phylogénie pouvant être obtenue en enlevant une des arêtes de la phylogénie. Ainsi, à toute arête d'une phylogénie sont associés les deux sous-arbres correspondant aux deux parties de la phylogénie déconnectées si l'arête est supprimée. Chaque sous-arbre induit par une arête correspond à une composante de la bipartition associée à cette arête (cf. Fig. 2.2).*

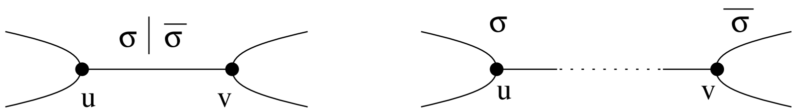


FIG. 2.2 – *L'arête (u, v) définit deux sous-arbres (enracinés resp. en u et en v). Chacun correspond à une composante (σ resp. $\overline{\sigma}$) de la bipartition $\sigma|\overline{\sigma}$ associée à cette arête.*

D'après la définition précédente, chaque r4-arbre peut être requis par les composantes de plusieurs bipartitions. Par exemple, dans la Fig. 2.3, le r4-arbre $xi|yz$ est requis par la composante σ_1 de la bipartition $\sigma_1|\overline{\sigma}_1$ mais aussi par la composante σ_2 de $\sigma_2|\overline{\sigma}_2$ puisque ajouter l'espèce i à ces deux composantes, *i.e.*, ajouter i à la gauche de e_1 dans la phylogénie, conduirait la phylogénie à induire le r4-arbre $xi|yz$.

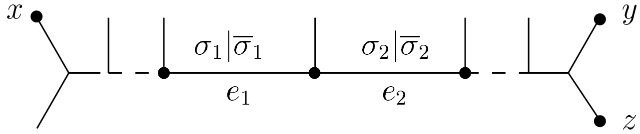


FIG. 2.3 – *Relations d'inclusion entre les composantes de bipartitions.*

Comment pouvons-nous connaître toutes les composantes qui requièrent un r4-arbre donné? Simplement, en exploitant des relations d'inclusion existant entre les composantes. Par exemple, sur la Fig. 2.3, $x \in \sigma_1 \subset \sigma_2$, d'où par définition, tous les r4-arbres $xi|yz$ qui référencent des paires d'espèces $y, z \in \overline{\sigma}_2$ sont requis par la composante σ_2, mais comme $\overline{\sigma}_2 \subseteq \overline{\sigma}_1$, ils sont aussi requis par σ_1. Ceci se comprend aisément, puisque les espèces x, y, z sont réparties de la même façon vis-à-vis de e_1 que vis-à-vis de e_2 (dans les deux cas, x, y sont séparés de z par l'arête).

Pour systématiser dans l'algorithme l'utilisation des relations d'inclusion à chaque étape i, nous avons besoin de connaître aisément toutes les relations de ce type entre composantes des diverses bipartitions de B^*_{i-1}. Puisque ces bipartitions sont compatibles en une phylogénie (T^*_{i-1}), leurs composantes sont en bijection avec les sous-arbres de cette phylogénie (dans la suite de l'exposé nous confondrons parfois une composante avec le sous-arbre auquel elle correspond). On peut donc tirer profit de la structure de T^*_{i-1}, pour obtenir très simplement les relations d'inclusion entre composantes depuis celles existant entre les sous-arbres de cette phylogénie. Par exemple, sur la Fig. 2.3, chaque arête sur la droite de e_1 a sa composante (son sous-arbre) de droite contenu dans la composante (le sous-arbre) de droite de e_1, tandis que sa composante de gauche contient la composante de gauche de e_1.

La Fig. 2.4 illustre la situation autour d'un noeud interne r quelconque, connecté aux arêtes $e_\sigma, e_1 \dots e_l$. À ces arêtes correspondent les composantes $\sigma, \overline{\sigma}, \sigma_1, \overline{\sigma}_1, \dots, \sigma_l, \overline{\sigma}_l$.

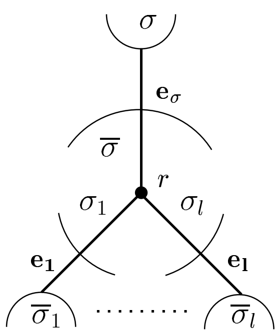


FIG. 2.4 – *Un noeud interne et les arêtes qui l'entourent.*

Grâce aux relations d'inclusion entre sous-arbres, nous savons que tout r4-arbre $xi|yz$ requis par une des composantes $\sigma_j, j \in 1 \dots l$ et t.q. $x \in \sigma$, est aussi requis par σ (puisque $y, z \in \overline{\sigma}_j \subset \overline{\sigma}$. Savoir si l'espèce i peut être ajoutée à la composante σ de la bipartition $\sigma|\overline{\sigma} \in B^*_{i-1}$, revient à savoir si l'espèce i peut être greffée dans le sous-arbre σ de la phylogénie T^*_{i-1}. Nous voyons donc aisément l'intérêt d'un traitement ordonné des composantes des bipartitions de B^*_{i-1} pour obtenir B^*_i : supposons que les composantes $\sigma_1, \dots, \sigma_l$ soient traitées avant σ, l'information consistant à savoir si l'espèce i peut être ajoutée aux composantes $\sigma_1, \ldots, \sigma_l$, peut être utilisée pour savoir si i peut être ajoutée à σ, sans tester Q à nouveau sur la présence des r4-arbres que toutes ces composantes requièrent communément. Plus généralement, toute composante contenant σ (pas seulement $\sigma_1, \ldots, \sigma_l$) requiert certains r4-arbres que σ requiert aussi. Ceci indique que toutes les composantes contenant σ doivent être considérées avant σ si l'on veut éviter au maximum d'interroger Q quand σ est considérée. Ceci nous amène tout naturellement à envisager un parcours récursif de la phylogénie T^*_{i-1}, puisque les relations d'inclusion entre composantes suivent les relations de descendance entre sous-arbres. Nous détaillerons plus loin comment un ordre satisfaisant de traitement des composantes peut être obtenu de cette façon. Toutefois, pour que cet ordre nous soit de quelque utilité, il nous faut d'abord montrer qu'on peut aisément caractériser les r4-arbres requis par une composante σ et encore non *traités*, étant donné que toutes les composantes contenant σ ont déjà été considérées. C'est ce que nous étudions maintenant.

Définition 6 *Une composante σ est dite* **maximalement contenue** *dans une autre composante σ_j ssi $\sigma_j \supset \sigma$ et $\nexists\, \theta$ t.q. $\sigma_j \supset \theta \supset \sigma$. En termes de sous-arbres, σ est* **sous-arbre maximal** *de σ_j. Si σ est maximalement contenue dans σ_j, on note $\sigma \sqsubset \sigma_j$ ou de façon équiv. $\sigma_j \sqsupset \sigma$.*

Notons que pour deux bipartitions $\sigma|\overline{\sigma}$ et $\sigma_j|\overline{\sigma}_j$, si $\sigma \sqsubset \sigma_j$, alors $\overline{\sigma} \sqsubset \overline{\sigma}_j$. Dans un but de simplification de l'exposé et sans perte de généralité, nous nous limiterons désormais à considérer, pour toute bipartition $\sigma|\overline{\sigma}$ de B^*_{i-1}, uniquement les composantes σ_j contenant maximalement σ et leur composante *opposée* $\overline{\sigma}_j$ contenues maximalement dans $\overline{\sigma}$. En d'autres termes, quand nous étudierons les sous-arbres associés à une arête e, nous ne prendrons en considération que les sous-arbres associés aux arêtes voisines de e.

Définition 7 *On appelle r4-arbres* **croisés** *de R_σ les r4-arbres de l'ensemble $R^c_\sigma = \bigcup_{j,k} R^{j,k}_\sigma$ ($j \neq k \in 1..l$), où $R^{j,k}_\sigma = \{xi|yz$ t.q. $x \in \sigma,\ y \in \overline{\sigma}_j,\ z \in \overline{\sigma}_k\}$.*

Les r4-arbres $xi|yz$ de R^c_σ sont dits *croisés*, car ils référencent respectivement deux espèces y, z appartenant à deux sous-arbres différents $\overline{\sigma}_j$ et $\overline{\sigma}_k$ de $\overline{\sigma}$, autrement dit, ils *croisent* les sous-arbres contenus dans $\overline{\sigma}$ (cf. Fig. 2.4).

Sur la base des définitions 5 et 7, ainsi que des relations d'inclusion entre composantes, on peut directement établir la propriété suivante :

Propriété 4

- $R_\sigma^c \subseteq R_\sigma$.
- *Si les composantes $\sigma_1, \ldots, \sigma_l \sqsupset \sigma$ ont été considérées avant la composante σ, alors tous les r4-arbres de R_σ ont été traités, excepté ceux de R_σ^c.*

Preuve : Triviale, par décomposition de l'ensemble R_σ. □

Cette propriété lie l'ensemble de r4-arbres requis par un sous-arbre et les ensembles de r4-arbres requis par les sous-arbres qui le contiennent. Nous avons en effet

$$R_\sigma = \left(\bigcup_{\sigma_i \sqsupset \sigma} R'_{\sigma_i} \right) \cup R_\sigma^c \,, \quad \text{où}$$

$$\forall i,\ R'_{\sigma_i} \subseteq R_{\sigma_i} \ .$$

où les R'_{σ_i} sont :

$$R'_{\sigma_i} = \{xi|yz \in R_{\sigma_i} \text{ t.q. } x \in \sigma\} \ .$$

De par la propriété précédente, à l'étape i, quand on cherche à savoir si on peut ajouter l'espèce i à une composante σ, on a juste besoin d'interroger Q sur les r4-arbres croisés R_σ^c requis par σ. On peut maintenant caractériser le processus récursif traitant les composantes des bipartitions de B_{i-1}^* comme suit : *a)* Toujours considérer une composante *après* les composantes qui la contiennent. *b)* Quand une composante σ est considérée, décider si l'espèce i peut lui être ajoutée (*i.e.*, si $\sigma \cup \{i\}|\overline{\sigma}$ est une bipartition de B_i^*) en fonction de $R_\sigma^c \subseteq Q$ *et* de l'information dérivée pour les autres r4-arbres de R_σ, précédemment examinés au cours de l'étape i. Il nous reste donc à spécifier respectivement : *a)* Comment on obtient un ordre de traitement des composantes assurant qu'une composante n'est considérée qu'après les composantes la contenant. *b)* Comment l'information est transmise entre les composantes, pour que la composante couramment examinée puisse connaître le résultat des interrogations précédentes de Q au sujet de r4-arbres qu'elle requiert. Ces deux points sont examinés dans les deux sous-sections suivantes.

2.2.2 Transmission de l'information

Comme nous l'avons vu précédemment, si $\sigma \sqsubset \sigma_j$, alors certains r4-arbres requis par σ_j sont aussi requis par σ. Cependant, tous les r4-arbres de R_{σ_j} ne sont pas pertinents pour σ. Plus précisément, les r4-arbres $xi|yz \in R_{\sigma_j}$ tels que $x \in \sigma_j - \sigma$ ne sont pas requis par σ. En conséquence, si σ_j ne peut pas être étendu, on ne peut pas en déduire systématiquement la même réponse pour σ, puisque la réponse négative obtenue pour σ_j peut très bien ne provenir que des r4-arbres non requis par σ. Nous noterons $L_j = \{x \in \sigma \text{ t.q. } \exists\, xi|yz \in R_{\sigma_j} - Q\}$ l'ensemble de toutes les espèces $x \in \sigma$ référencées dans un des r4-arbres requis par σ_j mais qui n'appartient pas à Q. On a $L_j \neq \emptyset$ *ssi* σ_j ne peut pas être *étendue* pour contenir l'espèce i, *i.e.*, $\sigma_j \cup \{i\}|\overline{\sigma}_j \notin B_i^*$. Quand σ_j ne peut pas être étendue, on peut en déduire que σ ne peut pas être étendue seulement si $L_j \cap \sigma \neq \emptyset$. Ceci implique que lorsque la composante σ est considérée, nous devons connaître les ensembles L_j des composantes σ_j t.q. $\sigma \sqsubset \sigma_j$. Lorsque σ est considérée, ces ensembles sont théoriquement disponibles, puisque, selon le principe du processus récursif, toutes les composantes contenant σ ont été précédemment traitées. Nous pouvons donc adopter la règle de décision suivante :

Propriété 5 *Si les composantes $\sigma_1, \ldots, \sigma_l \sqsupset \sigma$ ont été considérées avant σ, la composante σ peut être étendue pour contenir l'espèce i ssi $\bigcup_j L_j \cap \sigma = \emptyset$ et $R_\sigma^c \subseteq Q$.*

Notons que la situation apparemment paradoxale, où toutes les composantes contenant σ peuvent être étendues et σ non, est possible, puisque la réponse négative pour σ peut être dûe uniquement à l'absence de r4-arbres de R_σ^c dans Q. Sur la Fig. 2.4, cette situation particulière correspond au cas où l'espèce i est connectée au noeud interne r (par l'intermédiaire d'une arête externe).

Une autre situation particulière se produit quand une composante (ou plusieurs) σ_j ne peut pas être étendue pour inclure i, et où pourtant σ t.q. $\sigma \sqsubset \sigma_j$ peut, elle, être étendue. Si on considère cette situation du point de vue des arbres, ceci semble encore paradoxal, puisque cela signifie que l'espèce i peut être greffée dans le sous-arbre σ mais pas dans le sous-arbre σ_j contenant pourtant σ. Cependant, comme nous l'avons remarqué à la section 2.1, certaines bipartitions de B_{i-1}^* peuvent ne pas être étendues en bipartitions de B_i^*, ce qui se traduit par l'élimination de l'arête correspondante dans la phylogénie (en joignant ses deux extrémités). C'est ici le cas de la bipartition $\sigma_j|\overline{\sigma}_j$, qui, quand elle disparaît, résout la contradiction apparente.

L'algorithme 2 détaille le traitement d'une composante σ, lorsque les composantes $\sigma_1, \ldots, \sigma_l$ contenant σ ont déjà été traitées. L'ensemble L_σ, obtenu comme un sous-produit de l'algorithme, sera utilisé pour le traitement des composantes suivantes. C'est pourquoi, même si on sait dès l'union des L_j à la première ligne de l'algorithme que i ne peut être ajouté à σ, il faut quand même examiner tous les r4-arbres de R_σ^c afin que L_σ soit correct.

Algorithme 2: Savoir si une composante σ d'une bipartition $\sigma|\overline{\sigma}$ de B^*_{i-1} peut être étendue pour inclure l'espèce i.

$L_\sigma \leftarrow \bigcup_j L_j \cap \sigma$
Pour chaque *paire* $\overline{\sigma}_j, \overline{\sigma}_k$ *de sous-arbres maximaux de* $\overline{\sigma}$
  Pour chaque $x \in \sigma$, $y \in \overline{\sigma}_j$, $z \in \overline{\sigma}_k$ *(i.e., pour chaque* $xi|yz \in R^{j,k}_\sigma$*)*
    si $xi|yz \notin Q$ **alors** $L_\sigma \leftarrow L_\sigma \cup \{x\}$

si $L_\sigma = \emptyset$ **alors** répondre `OUI` **sinon** répondre `NON`

Il est facile de voir que $L_\sigma = \emptyset$ dès que $\overline{\sigma}$ est une feuille (correspondant à une espèce w). Ceci signifie que toute bipartition triviale, $\sigma|\{w\} \in B^*_{i-1}$, peut toujours être étendue en la bipartition $\sigma \cup \{i\}|\{w\}$ dans B^*_i (et aussi potentiellement en la bipartition $\sigma|\{w, i\}$). Ainsi T^*_n contiendra au moins les arêtes correspondant à la phylogénie étoile.

2.2.3 Ordre d'exploration des composantes

Nous considérons maintenant le problème consistant à déterminer un ordre de traitement sur les composantes des bipartitions de B^*_{i-1}. Cet ordre doit respecter la contrainte qu'une composante ne soit traitée qu'après les composantes qui la contiennent. Pour satisfaire cette contrainte, les deux composantes d'une même bipartition ne peuvent généralement pas être considérées consécutivement, impliquant qu'un seul parcours de l'arbre T^*_{i-1} n'est pas suffisant pour traiter toutes les composantes. Nous montrons cependant ci-dessous que deux parcours seulement sont suffisants.

Pour simplifier la présentation, nous considérons (sans perte de généralité) un problème plus simple, où la même contrainte d'ordre est imposée. Supposons que nous ayons à établir, pour chaque sous-arbre d'une phylogénie T donné, l'ensemble des espèces qu'il contient, avec la contrainte qu'un sous-arbre ne soit parcouru qu'après les sous-arbres qui le contiennent. On peut tirer parti de cette contrainte pour connaître les espèces contenues dans un sous-arbre : il suffit de prendre l'intersection des ensembles d'espèces contenues dans les sous-arbres qui le contiennent.

Pour résoudre ce problème, nous enracinons d'abord la phylogénie T en choisissant arbitrairement une de ses feuilles pour racine (soit l'espèce 1). Ceci a pour effet d'orienter la phylogénie au sens où chaque arête définit maintenant un **sous-arbre inférieur** et son complémentaire, un **sous-arbre supérieur**.

Puis nous effectuons un premier parcours récursif de la phylogénie, implémenté comme un parcours postfixe, pour calculer l'information associée aux sous-arbres supérieurs, utilisant pour chacun l'information déjà obtenue pour ses sous-arbres **fils**. Par exemple, dans la Fig. 2.5, nous savons que le sous-arbre supérieur, défini par l'arête numéro 4, contient les espèces $\{1,2,4,5\} \cap \{1,2,3,5\}$. Les seuls cas particuliers échappant à cette règle sont les sous-arbres supérieurs dont le sous-arbre opposé (*i.e.*, inférieur) est une feuille : ils contiennent simplement toutes les espèces sauf celle de la feuille en question. Par exemple sur la Fig. 2.5, le sous-arbre supérieur défini par l'arête 1, connectant l'espèce 2 à la phylogénie contient les espèces $\{1,3,4,5\}$. La Fig. 2.5 montre un exemple de phylogénie et le résultat du premier parcours récursif de cette phylogénie. Les sous-arbres supérieurs sont numérotés selon l'ordre dans lequel ils sont traités. Chacun est étiqueté dans la figure par l'information calculée lors de son traitement, *i.e.*, l'ensemble des espèces qu'il contient.

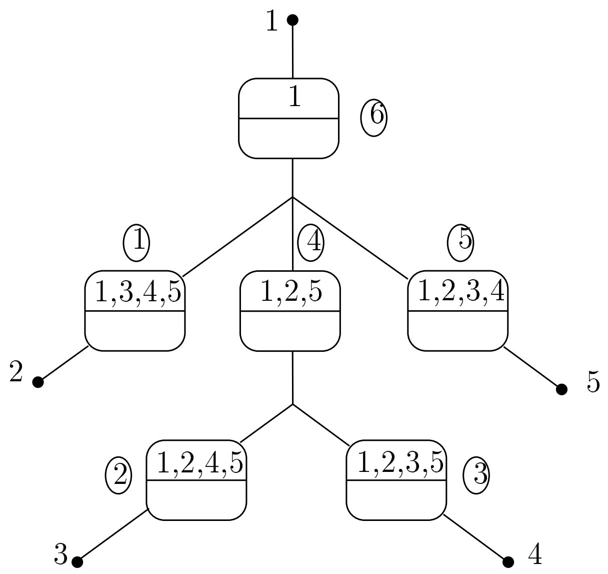


FIG. 2.5 – *Parcours postfixe en profondeur, traitant les sous-arbres supérieurs (les nombres entourés indiquent l'ordre dans lequel ils sont considérés). L'information concernant les sous-arbres supérieurs transite des feuilles vers la racine. Les sous-arbres sont étiquetés par l'information qui est calculée pour eux.*

Le second parcours récursif de la phylogénie permet de traiter ses sous-arbres inférieurs. Il est implémenté comme un parcours préfixe (Fig. 2.6), pour que l'information associée à un sous-arbre soit, ici aussi, obtenue depuis l'information des sous-arbres qui le contiennent. À la différence du premier parcours, chaque sous-arbre considéré n'est pas uniquement inclus dans des sous-arbres du même type (*i.e.*, inférieurs). En effet, chaque sous-arbre inférieur associé à une arête e est contenu dans un seul autre sous-arbre inférieur (celui défini par l'arête **père** de l'arête e) et dans plusieurs sous-arbre supérieurs (ceux définis par les arêtes **soeurs** de l'arête e). Par exemple sur la Fig. 2.6, le sous-arbre inférieur défini par l'arête numérotée ③[4] contient les espèces $\{2, 3, 4, 5\} \cap \{1, 3, 4, 5\} \cap \{1, 2, 3, 4\}$. Ici, le seul cas particulier est le sous-arbre inférieur dont le sous-arbre opposé est la racine. Ce sous-arbre est traité comme les cas particuliers du premier parcours, *i.e.*, il contient toutes les espèces sauf celle associée à la racine.

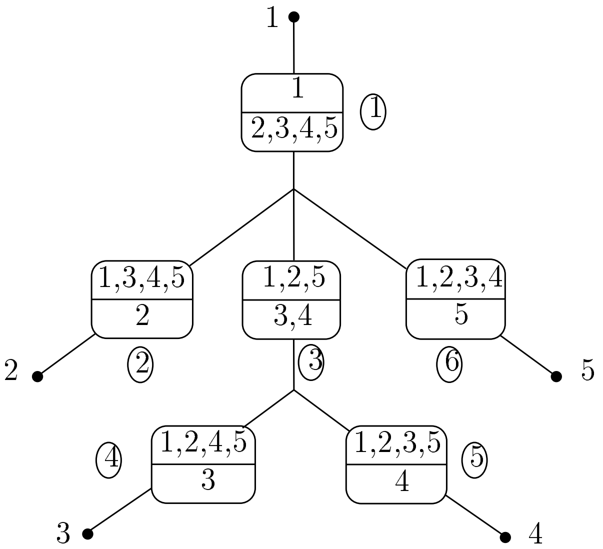


FIG. 2.6 – *Parcours préfixe en profondeur, traitant les sous-arbres inférieurs (les nombres entourés indiquent l'ordre dans lequel ils sont considérés). L'information concernant les sous-arbres inférieurs transite de la racine vers les feuilles.*

2.2.4 L'algorithme IQ^*

Nous donnons maintenant l'algorithme IQ^* de complexité $O(n^4)$, obtenu en combinant les différents résultats obtenus précédemment dans cette section. La différence principale avec l'algorithme simple en $O(n^5)$ de la section précédente réside dans la façon de traiter les composantes des bipartitions de B^*_{i-1}. À chaque étape i, l'algorithme IQ^* explore la phylogénie T^*_{i-1} selon les deux parcours récursifs décrits ci-dessus. Ceci contribue à établir un ordre de traitement sur les

4. numérotée ④précédemment car les deux parcours ne définissent bien sûr pas le même ordre sur les arêtes.

composantes, adapté au partage d'information entre les diverses composantes, évitant ainsi de réinterroger Q plusieurs fois pour le même r4-arbre. L'algorithme IQ^* est résumé ci-dessous.

Algorithme 3: Partie principale de IQ^*.

---

input : Un ensemble Q de r4-arbres définis sur les espèces $1, 2, \dots, n$
output: La phylogénie $T^* \equiv Q^*$
/ Initialisations */*
$B_4^* \leftarrow \{\ \{1\}|\{2,3,4\},\ \{2\}|\{1,3,4\},\ \{3\}|\{1,2,4\},\ \{4\}|\{1,2,3\}\ \}$
si $\exists\, r = xy|zt,\ \{x,y,z,t\} = \{1,2,3,4\}$ *et t.q.* $r \in Q$ **alors**
  $B_4^* \leftarrow B_4^* \cup \{\ \{x,y\}|\{z,t\}\ \}$, $T_4^* \leftarrow$ structure de r
sinon $T_4^* \leftarrow$ phylogénie étoile sur les espèces $\{1,2,3,4\}$
/ Insère progressivement les espèces restantes */*
Pour chaque *étape* $i \leftarrow 5$ *à* n
  pour chaque *sous-arbre* $\sigma \in T_{i-1}^*$ **faire**
    calcule l'ensemble X_σ d'espèces contenues dans σ
  Soit e l'arête incidente à la racine de T_{i-1}^*
  `Sous-Arbre-Sup(`e`)` */* traite les composantes associées aux sous-arbres* supérieurs *de* T_{i-1}^* **/*
  `Sous-Arbre-Inf(`e`)` */* traite les composantes associées aux sous-arbres* inférieurs *de* T_{i-1}^* **/*
  / Déduit* B_i^* *et* T_i^* **/*
  $B_i^* \leftarrow \{\ \{i\}|\{1, \dots, i-1\}\ \}$
  Pour chaque *bipartition* $\sigma|\overline{\sigma} \in B_{i-1}^*$
    si Ext_σ **alors** $B_i^* \leftarrow B_i^* \cup \{\sigma \cup \{i\}|\overline{\sigma}\}$
    si $Ext_{\overline{\sigma}}$ **alors** $B_i^* \leftarrow B_i^* \cup \{\sigma|\overline{\sigma} \cup \{i\}\}$
  Construire T_i^* depuis B_i^*

Dans l'algorithme 4 nous détaillons comment IQ^* effectue à l'étape i le parcours postfixe de la phylogénie T_{i-1}^*. Les notations de la Fig. 2.4 sont utilisées et la racine de la phylogénie est supposée située dans le sous-arbre σ. Le parcours préfixe (algorithme 5), est symétrique du parcours postfixe (algorithme 4) : la composante σ est cette fois traitée avant d'effectuer l'appel récursif ; de plus, les arêtes e_j, e_k ne sont plus des arêtes filles de l'arête associée au sous-arbre traité, mais sont choisies parmi ses arêtes soeurs et son arête père (par exemple sur la Fig. 2.4, si l'appel `Sous-Arbre-Inf(`e_1`)` est effectué, $\{j,k\}$ sont choisis parmi $\{\sigma, 2 \dots l\}$).

Algorithme 4: $Sous\!-\!Arbre\!-\!Sup(e_\sigma)$: effectue un parcours postfixe en profondeur du sous-arbre $\overline{\sigma}$, incident à l'arête e_σ. Ceci permet de savoir si la composante σ associée au sous-arbre supérieur peut être étendue pour inclure l'espèce i, *i.e.*, si $\sigma \cup \{i\} | \overline{\sigma} \in B_i^*$.

si *le sous-arbre inférieur associé à* e_σ *est une feuille* **alors** $L_\sigma \leftarrow \emptyset$
sinon
  /* *traite les composantes qui contiennent* σ */
  pour chaque e_j *t.q.* $j \in 1..l$, *arête fille de* e_σ *dans* T^*_{i-1} **faire**
    `Sous-Arbre-Sup`(e_j)
  /* *traite* σ */
  /* *1) réutilise les résultats obtenus pour les r4-arbres* $r \in R_\sigma - R_\sigma^c$ */
  $L_\sigma \leftarrow \bigcup_{j \in 1..l} L_j \cap X_\sigma$
  /* *2) examine les r4-arbres* $r \in R_\sigma^c$ */
  pour chaque *paire* $j, k \in 1..l$ *d'arêtes filles de* e_σ **faire**
    /* *examine les r4-arbres de* $R_\sigma^{j,k}$ */
    pour chaque *r4-arbre* $xi|yz$ *t.q.* $x \in X_\sigma,\ y \in X_{\sigma_j},\ z \in X_{\sigma_k}$ **faire**
      si $r \notin Q$ **alors**
        $L_\sigma \leftarrow L_\sigma \cup \{x\}$ /* *mémorise l'espèce appartenant à* σ *et à un r4-arbre responsable de la non-extension de* σ */
  /* *décide si la composante* σ *peut être étendue* */
  si $L_\sigma = \emptyset$ **alors**
    $Ext_\sigma =$ `OUI` /* *tous les r4-arbres requis par* σ *pour inclure* i *sont présents dans* Q */
  sinon $Ext_\sigma =$ `NON` /* *certains r4-arbres requis par* σ *ne sont pas dans* Q */

Algorithme 5: $Sous-Arbre-Inf(e_\sigma)$: effectue un parcours préfixe en profondeur du sous-arbre $\overline{\sigma}$, incident à l'arête e_σ. Ceci permet de savoir si la composante $\overline{\sigma}$ associée au sous-arbre inférieur peut être étendue pour inclure l'espèce i, *i.e.*, si $\sigma|\overline{\sigma}\cup\{i\} \in B_i^*$.

```
si le sous-arbre inférieur associé à e_σ est une feuille alors L_σ̄ ← ∅
sinon
    /* traite σ̄ */
    /* 1) réutilise les résultats obtenus pour les r4-arbres r ∈ R_σ̄ − R^c_σ̄ */
    L_σ̄ ← ⋃_{j∈1..l} L_j ∩ X_σ̄, où j parcourt les composantes supérieures des arêtes
    soeurs de e_σ et la composante inférieure de l'arête père de e_σ
    /* 2) examine les r4-arbres r ∈ R^c_σ̄ */
    pour chaque paire j,k ∈ 1..l d'arêtes soeurs ou père de e_σ faire
        /* examine les r4-arbres de R^{j,k}_σ̄ */
        pour chaque r4-arbre xi|yz t.q. x ∈ X_σ̄, y ∈ X_σj, z ∈ X_σk faire
            si r ∉ Q alors
                L_σ̄ ← L_σ̄ ∪ {x}   /* mémorise l'espèce appartenant à σ̄ et à un
                                     r4-arbre responsable de la non-extension de σ̄ */
    /* décide si la composante σ̄ peut être étendue */
    si L_σ̄ = ∅ alors
        Ext_σ̄ = OUI        /* tous les r4-arbres requis par σ̄ pour inclure i sont
                               présents dans Q */
    sinon Ext_σ̄ = NON  /* certains r4-arbres requis par σ̄ ne sont pas dans Q */
    /* traite les composantes contenues par σ̄ */
    pour chaque e_j t.q. j ∈ 1..l, arête fille de e_σ dans T*_{i−1} faire
        Sous-Arbre-Inf(e_j)
```

2.2.5 Complexité de l'algorithme IQ^*

Nous montrons ici que la complexité de l'algorithme IQ^* est $O(n^4)$. Ce résultat repose principalement sur le fait que chacun des $3\binom{n}{4}$ r4-arbres possibles n'est examiné qu'une seule fois, comme établi ci-dessous :

Lemme 4 *L'algorithme IQ^* examine chaque r4-arbre possible une et une seule fois.*

Preuve : chaque r4-arbre possible $r = xi|yz$ n'est considéré qu'à l'étape où l'espèce de plus haut rang qu'il référence est ajoutée à la phylogénie. Soit i cette étape (*i.e.*, $\{x, y, z\} \subseteq [i-1]$). Lors de cette étape (comme à chaque étape) seuls des r4-arbres de type croisé sont examinés, et ce, pendant le parcours postfixe et le parcours préfixe de T^*_{i-1} (ligne 12 de l'algorithme 4). Ces deux parcours assurent que chaque composante σ est considérée une et une seule fois. D'où, par l'algorithme 3, chaque ensemble $R^{j,k}_\sigma$ de r4-arbres croisés n'est considéré qu'une et une seule fois. Puisque, par définition, $r = xi|yz$ appartient à un et un seul de ces ensembles (l'ensemble $R^{j,k}_\sigma$ t.q. $x \in \sigma, y \in \overline{\sigma}_j, z \in \overline{\sigma}_k, j \neq k$), r est considéré une et une seule fois durant l'ensemble de l'algorithme IQ^*. □

Théorème 3 *B^*, T^* et donc Q^* peut être calculé en temps $O(n^4)$.*

Preuve : L'algorithme IQ^* utilise relativement peu de structures de données. Le tableau Ext, indexé par les diverses composantes, indique pour chacune si elle peut être étendue de l'étape i courante à la suivante. Pour chaque composante σ on retient aussi les espèces qu'elle contient dans un ensemble X_σ, implémenté comme une liste chaînée. Les ensembles L_σ sont codés sous la forme de tableaux binaires de taille n, indiquant pour chaque espèce si elle appartient à σ et en même temps à un r4-arbre requis par σ mais absent de Q. Les bipartitions de B^*_i sont aussi stockées sous la forme de tableaux binaires.

L'algorithme IQ^* est en fait composé des algorithmes 3, 4 et 5. Nous montrons maintenant que les algorithmes 3 et 4 requièrent chacun au plus $O(n^4)$ opérations (le comportement de l'algorithme 5 est similaire à celui de l'algorithme 4, aussi ces deux algorithmes ont-ils le même coût).

Algorithme 3 :

- Les *initialisations* demandent seulement un temps constant.

- Lors de chacune des $O(n)$ étapes, les ensembles X_σ sont initialisés (ligne 6) par un simple parcours récursif de la phylogénie T^*_{i-1}. Ceci demande $O(n^2)$ opérations à chaque fois, en proportion avec l'information à retenir, et ainsi $O(n^3)$ sur l'ensemble de l'algorithme.

- Un parcours postfixe et un parcours préfixe de la phylogénie T^*_{i-1} sont lancés à chaque étape de la boucle principale. Le coût de ces parcours est détaillé ci-dessous (Algorithme 4).

- Initialiser l'ensemble B_i^* pour indiquer qu'il contient la bipartition $\{i\}|[i-1]$ requiert $O(n)$ opérations, puisqu'il faut pour cela décrire à quelle partie de la bipartition appartient chaque espèce. Depuis le tableau Ext, on sait en $O(1)$ si chacune des $O(n)$ composantes des bipartitions de B_{i-1}^* peuvent être étendues (**ligne** `7`). Dans le pire des cas, toutes les bipartitions de B_{i-1}^* peuvent être étendues, mais une au plus peut supporter l'ajout de l'espèce i dans ses deux composantes : la bipartition correspondant à l'arête sur laquelle l'espèce i est greffée dans la phylogénie (si i n'est pas greffée sur un noeud r, cf. Fig. 2.4). Ainsi, les bipartitions de B_i^* sont obtenues simplement en ajoutant l'espèce i en $O(1)$ aux $O(n)$ bipartitions de B_{i-1}^* et en faisant une copie d'au plus une d'entre elles en $O(n)$, avant de lui ajouter l'espèce i. Ainsi, obtenir B_i^* demande $O(n)$ opérations à chaque étape, *i.e.*, $O(n^2)$ sur l'ensemble de l'algorithme.

- La phylogénie T_i^* est obtenue en temps $O(n^2)$ depuis B_i^* en ayant recours à un algorithme linéaire [Mea81, Gus91a].

Algorithme 4 :

La phylogénie T_{i-1}^* contient $O(n)$ arêtes, chacune étant traitée deux fois à l'étape i : une fois durant le parcours postfixe et une fois durant le parcours préfixe. Ainsi, d'un point de vue global, $O(n^2)$ arêtes e_σ sont considérées sur l'ensemble de l'algorithme, *i.e.*, l'appel à l'algorithme 4 sur un sous-arbre est effectué $O(n^2)$ fois. Nous considérons maintenant les diverses opérations effectuées à chaque exécution de cet algorithme.

- **ligne** `8` : Vider L_σ quand une feuille de la phylogénie est rencontrée (ligne `8`) coûte seulement $O(n)$, ainsi cette opération est en $O(n^3)$ sur l'ensemble de l'algorithme.

- **ligne** `9` : $O(n^2)$ appels récursifs sont effectués.

- **ligne** `10` : L'affectation d'un L_σ est réalisée en $O(n^2)$, puisque une composante σ est maximalement contenue dans au maximum $O(n)$ composantes σ_j (en raison du degré maximal d'un noeud) et l'union de L_j à L_σ requiert $O(n)$ opérations. Puisque l'algorithme 3 est effectué $O(n^2)$ fois, cette ligne coûte $O(n^4)$ sur l'ensemble de l'algorithme.

- **lignes** `11-13` : Cet ensemble d'opérations se réduit à un examen unique de chacun des $3\binom{n}{4}$ r4-arbres possibles (comme établi par le lemme 4). Les diverses requêtes adressées à Q pour les r4-arbres sont chacune émise en temps

$O(1)$, puisqu'elles sont obtenues lors du parcours des ensembles X_σ (codés sous forme de listes chaînées), et qu'aucun élément de ces ensembles n'est parcouru inutilement, *i.e.*, il correspond à l'émission d'une requête pour au moins un r4-arbre. De plus, savoir si un certain r4-arbre est contenu dans Q (ligne 13) est immédiat puisque nous faisons l'hypothèse raisonnable que Q est une matrice quadri-dimensionnelle ; si ce n'est pas le cas, un simple pré-traitement en $O(n^4)$ permet de coder Q par une telle matrice (cf. algorithme 1).

En conséquence, les lignes `11-13` demandent $O(n^4)$ opérations sur l'ensemble de l'algorithme.

– **ligne** `14` : Décider si une composante peut être étendue requiert l'examen complet de L_σ, *i.e.*, $O(n)$ opérations. Ext_σ est modifié en $O(1)$. Ainsi, sur l'ensemble des $O(n^2)$ exécutions de l'algorithme 4, cette ligne requiert $O(n^3)$ opérations.

L'analyse ci-dessus montre que IQ^* calcule T^* et B^* en temps $O(n^4)$, les opérations `10` et `11-13` étant les plus coûteuses. De plus, la constante affectée au facteur $O(n^4)$ de la complexité est faible, puisque une éventuelle initialisation de la partie triangulaire inférieure de la matrice Q demande $(\frac{n}{2})^4 + \binom{n}{4} < n^4$ opérations ; de même l'interrogation de Q (ligne `13`) demande $3\binom{n}{4} < n^4$, et l'affectation des ensembles L_σ (lignes `8` et `10`) demande $\sum_{i=5}^{n}\sum_{\sigma=1}^{i}(i-1)i < n^4$ avec opérations[5].

$\square$

L'algorithme IQ^* possède donc une complexité linéaire si l'entrée Q est codée sous la forme d'une matrice, et quasi-linéaire si Q est codée sous la forme d'une liste de r4-arbres ($|Q| \in O(n^4)$ dans le pire des cas). L'étude menée sur IQ^* laisse à penser qu'on ne peut pas obtenir un algorithme incrémental calculant Q^* avec une complexité inférieure (*i.e.*, en $O(|Q|)$ ou même en $O(|Q| + n^3)$). De même, il est peu vraisemblable qu'une autre approche (autre qu'incrémentale) permette d'obtenir une meilleure complexité.

La complexité obtenue pour calculer Q^* est satisfaisante au sens où elle n'excède pas la complexité de la méthode d'inférence des r4-arbres, généralement en $O(n^4)$. Ainsi la complexité de l'ensemble de la méthode de R.P., conduisant des données

5. La première somme correspond aux différentes étapes i, la deuxième aux composantes de la phylogénie T_i^* et le produit mesure le nombre d'opérations maximum nécessaires pour affecter les i cases d'un L_σ qui sont pertinentes à l'étape i (maximum atteint quand il possède $e-1$ composantes voisines).

moléculaires à la phylogénie inférée, demande $O(n^4)$. Notons de plus que la complexité montrée pour résoudre le problème d'optimisation SMA est la même que celle du meilleur algorithme connu pour résoudre le problème de décision associé, AR (cf. chap. précédent).

2.3 Propriétés statistiques de Q^* en R.P.

L'algorithme IQ^* accepte en entrée un ensemble de r4-arbres Q, et en cela peut être appliqué indépendamment de tout contexte. Ce qui le rattache au contexte, c'est la méthode utilisée pour inférer les r4-arbres depuis les données disponibles, propres au domaine de l'étude. En ce qui concerne la R.P., nous avons vu précédemment qu'il existe plusieurs méthodes possibles pour inférer des r4-arbres. De la consistance de la méthode choisie sur les quadruplets d'espèces dépend invariablement la consistance de IQ^* sur l'ensemble des espèces. Pour montrer des résultats de consistance pour IQ^*, nous nous basons sur un principe de distances. Plus précisément, nous optons pour la méthode FPM (cf. chap. précédent) pour inférer les r4-arbres. Nous appellerons désormais "**méthode** Q^*", l'enchaînement de FPM et de IQ^* : depuis une matrice de distances $\hat{D}$, on applique FPM pour inférer un ensemble Q de r4-arbres qui est ensuite utilisé par IQ^* pour inférer la phylogénie T^*. Remarquons que la méthode Q^* peut être considérée comme une *méthode de distances* car elle accepte en entrée une matrice $\hat{D}$ de distances entre espèces et fournit une estimation $\hat{t}=T^*$ de la phylogénie t des espèces. Rappelons que la méthode FPM appliquée à une matrice de distances $\hat{D}$ choisit le r4-arbre correspondant à la plus petite des trois doubles sommes, p. ex. $xy|zt$ ssi $\hat{D}_{xy}+\hat{D}_{zt} < min(\hat{D}_{xz}+\hat{D}_{yt}, \hat{D}_{xt}+\hat{D}_{yz})$.

Ci-dessous nous montrons des conditions de consistance et un taux de convergence pour la méthode Q^*. Comme Atteson [Att97c], nous considérons à la fois la consistance totale, *i.e.*, lorsque l'ensemble de la phylogénie correcte est retrouvée, et la consistance partielle, *i.e.*, lorsque seulement certaines de ses arêtes sont retrouvées. Nous présentons d'abord les résultats de consistance, basés sur la norme L_∞, puis en déduisons ensuite des vitesses de convergence en nous appuyant sur un modèle de l'évolution relativement simple.

2.3.1 Consistance de la méthode Q^*

Comme le principe FPM d'inférence des r4-arbres est basé sur les distances $\hat{D}_{ij}$, la méthode Q^* est consistante dès que le modèle de l'évolution choisi pour corriger les distances observées est juste : si FPM infère correctement tous les r4-arbres, alors Q est arboré et l'algorithme Q^* retrouve la phylogénie correcte.

Ci-dessous nous montrons des conditions pour lesquelles la méthode Q^* retrouve la phylogénie correcte avec une probabilité de 1 depuis les données dont elle dispose (la distance $\hat{D}$), indépendamment du modèle de l'évolution retenu. Autrement dit, nous caractérisons la *boule d'exactitude garantie* de la méthode.

Rappelons d'abord la définition de la norme L_∞ entre deux distances quelconques D et D' définies sur l'ensemble E des espèces étudiées :

$$L_\infty(D, D') = \max_{xy} \left| D_{xy} - D'_{xy} \right| \ .$$

Rappelons aussi que toute phylogénie P valuée induit une distance, notée ici D^P et représentée habituellement par une matrice (D^P_{xy}) de taille $E \times E$, en fonction de la longueur de ses arêtes. La distance D^P_{xy} induite entre deux espèces x et y est obtenue depuis la longueur $l(e)$ des arêtes e du chemin $[x, y]$ dans P :

$$D^P_{xy} \quad = \quad \sum_{e \in [i,j]} l(e) \ . \tag{2.1}$$

Rappelons aussi qu'une distance D est dite *d'arbre* s'il existe une phylogénie P t.q. $D^P = D$, et que toute distance d'arbre vérifie la condition des quatre points (section 1.5.6 chap. 1, part. I).

Une propriété que l'on peut exiger d'une méthode de reconstruction raisonnable est qu'elle retrouve la phylogénie P correspondant à une distance d'arbre D^P. Cette propriété est vérifiée par les méthodes NJ et *Addtree* [Att97b], SQM [ESSW97b], DCM [ESSW97a], ainsi que par la méthode Q^*, comme le montre la propriété suivante :

Propriété 6 *Soient P une phylogénie et D^P la distance qu'elle définit. La méthode Q^* appliquée à la distance D^P retrouve la phylogénie P.*

Preuve : Si D^P est une distance d'arbre, $\forall q = x, y, z, t$ t.q. $xy|zt \in Q_P$, on a $D^P_{xy} + D^P_{zt} < D^P_{xz} + D^P_{yt} = D^P_{xt} + D^P_{yz}$ (de par la condition des quatre points), donc FPM infère le r4-arbre correct $xy|zt$ pour q. Ceci signifie que l'ensemble Q donné en entrée à l'algorithme IQ^* n'est autre que l'ensemble Q_P des r4-arbres induits par la phylogénie P. Cet ensemble est arboré, si bien que $Q^* = Q_P$, impliquant que l'algorithme IQ^* infère la phylogénie $T^* = P$. □

De même, étant donné une distance $\hat{D}$ suffisamment proche de D^P, *i.e.*, t.q. $L_\infty(D^P, \hat{D}) < \epsilon$, toute méthode raisonnable devrait aussi retrouver P. À partir de quel ϵ ne peut-on plus garantir qu'une méthode raisonnable retrouve P ? Erdös *et* *al* [ESSW97b] ainsi que Atteson [Att97c] répondent à cette question en montrant que pour toute distance d'arbre $D = D^P$, il existe une distance quelconque $\hat{D}$ et une autre distance d'arbre $D' = D^{P'}$, avec $B(P') \neq B(P)$, t.q. :

$$max_{xy}\left|D_{xy} - \hat{D}_{xy}\right| = min_{e \in A_P}\frac{l(e)}{2}$$

$$max_{xy}\left|D'_{xy} - \hat{D}_{xy}\right| = min_{e \in A_{P'}}\frac{l(e)}{2}$$

où A_P est l'ensemble des arêtes internes de la phylogénie P. On peut même choisir D' de façon à ce que e soit aussi la plus courte arête de P' (cf. Fig. 2.7 pour une représentation de ce cas de figure).

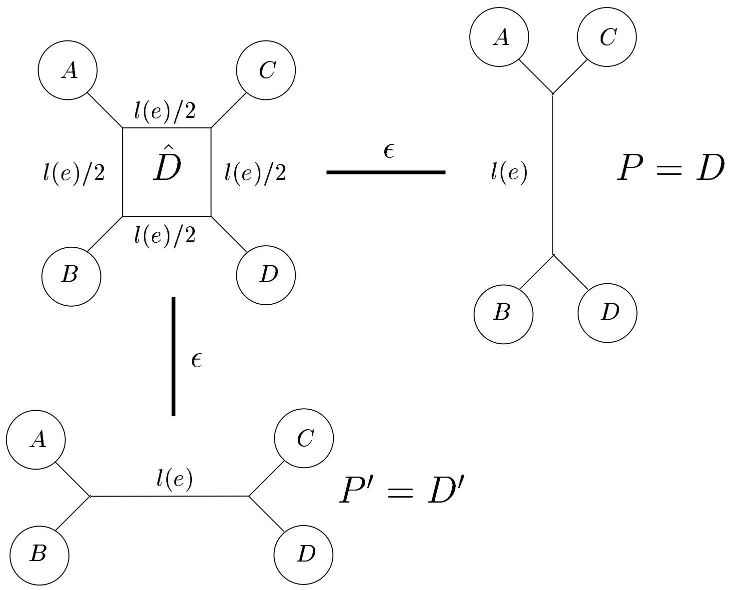


FIG. 2.7 – *(d'après Atteson [Att97c]) Cas d'une distance quelconque $\hat{D}$ définie sur l'ensemble d'espèces $E = A \cup B \cup C \cup D$, située à égale distance (au sens de L_∞), $min_{e \in A_P}\frac{l(e)}{2}$, de deux distances d'arbre D et D' définies sur les mêmes espèces. $l(e)$ est la longueur de la plus petite arête de chacune des deux phylogénies P et P', associées resp. à D et D'.*

Ainsi, on ne peut pas garantir qu'une méthode retrouve systématiquement la phylogénie correcte, P, à partir du moment où $L_\infty(D, \hat{D}) \geq min_{e \in A_P}\frac{l(e)}{2}$, puisque P n'est pas forcément l'unique phylogénie la plus proche des données (ou pire, peut ne pas être la phylogénie la plus proche des données). Cependant, on peut montrer qu'il existe des méthodes qui font aussi bien que possible, *i.e.*, qui retrouvent P dès que $L_\infty(D,\hat{D}) < min_{e\in A_P}\frac{l(e)}{2}$. Par exemple, les méthodes NJ et *Addtree* vérifient cette propriété [Att97b]. Nous montrons ci-dessous que c'est aussi le cas pour la méthode Q^*. Pour ceci, nous avons besoin du lemme suivant (proche du lemme 5 de [ESSW97a]) :

Lemme 5 *(Erdös et al [ESSW97a] modifié) Soit $P=(S_P,A_P)$ une phylogénie sur les espèces E, soient x,y,z,t quatre espèces de E, soit D la distance d'arbre associée à P, et $\hat{D}$ une distance quelconque sur E. Si $L_\infty(D,\hat{D}) < min_{e\in A_P}\frac{l(e)}{2}$ et $xy|zt \in Q_P$ alors FPM retrouve le r4-arbre $xy|zt$.*

Preuve : Pour simplifier les notations dans la preuve, posons $l_m = min_{e\in S_P} l(e)$. $xy|zt \in Q_P$ implique que les chemins $[x,y]$ et $[z,t]$ ont une intersection vide dans P et sont séparés par un chemin $[u_{xy},u_{zt}]$ contenant au moins une arête. Soit $l=\sum_{e\in[u_{xy},u_{zt}]} l(e)$ la longueur de ce chemin, par la condition des quatre points, on sait que :

$$D_{xy}+D_{zt}+2l \;=D_{xz}+D_{yt} \;= D_{xt}+D_{yz}\ . \tag{2.2}$$

Puisque $L_\infty(\hat{D},D) < \frac{l_m}{2}$, on a :

$$\hat{D}_{xy}+\hat{D}_{zt} < D_{xy}+D_{zt}+l_m\ , \tag{2.3}$$
$$\hat{D}_{xz}+\hat{D}_{yt} > D_{xz}+D_{yt}-l_m\ , \tag{2.4}$$
$$\hat{D}_{xt}+\hat{D}_{yz} > D_{xt}+D_{yz}-l_m\ . \tag{2.5}$$

Depuis (2.2), (2.3), (2.4), (2.5), et le fait que $l \geq l_m$, on déduit que $\hat{D}_{xy}+\hat{D}_{zt} < \hat{D}_{xz}+\hat{D}_{yt}$ et $\hat{D}_{xy}+\hat{D}_{zt} < \hat{D}_{xt}+\hat{D}_{yz}$. D'où par définition, FPM appliquée à $\hat{D}$ induit le r4-arbre $xy|zt$. □

On en déduit un résultat pour l'ensemble de la phylogénie :

Théorème 4 *Soient $P=(S_P,A_P)$ une phylogénie et D la distance d'arbre qu'elle induit sur les espèces selon* (2.1)*. La méthode Q^* appliquée à toute distance $\hat{D}$ t.q. $L_\infty(D,\hat{D}) < min_{e\in A_P}\frac{l(e)}{2}$ retrouve la phylogénie P.*

Preuve : D'après le lemme précédent, l'ensemble Q inferé par FPM depuis $\hat{D}$ est exactement l'ensemble Q_P des r4-arbres de P. Et IQ^* appliqué à l'ensemble arboré Q_P retrouve la phylogénie $T^*=P$. □

Ainsi, du point de vue de la norme L_∞, la méthode Q^* a une boule d'exactitude garantie aussi large que possible, au sens où la méthode retrouve la phylogénie des espèces dans toute région de l'espace où une telle garantie peut être formulée. Paradoxalement, comme nous l'avons déjà évoqué dans la partie introductive de cette thèse, aucune preuve n'a encore été trouvée pour montrer qu'une hypothétique méthode exacte (*i.e.*, inférant toujours la phylogénie la plus proche en terme de L_∞ de la matrice de données $\hat{D}$) retrouve la phylogénie correcte P dès que $L_\infty(D^P, \hat{D}) < min_{e \in A_P} \frac{l(e)}{2}$. La seule preuve, dont on dispose pour l'instant [ESSW97b], montre qu'une hypothétique méthode exacte retrouve la phylogénie correcte quand $L_\infty(D^P, \hat{D}) < min_{e \in A_P} \frac{l(e)}{4}$ [Att97a]. La Fig. 2.8-(A) résume les boules d'exactitude garantie connues au sens de la norme L_∞ pour les diverses méthodes mentionnées ci-dessus.

Jusqu'à présent, nous avons considéré l'aptitude des méthodes à retrouver la phylogénie inconnue dans sa totalité. Que se passe-t-il dans le cas où la distance entre D^P et $\hat{D}$ est plus importante que $min_{e \in A_P} \frac{l(e)}{2}$? Peut-on toutefois retrouver certaines arêtes de la phylogénie des espèces ? Ces questions ont un réel intérêt pratique, car il est connu que les taux d'évolution sont parfois relativement différents d'une arête de la phylogénie à l'autre, donnant des longueurs d'arêtes disparates. Dans un tel cas, il est vraisemblable que le bruit évolutif affectant les données fasse que la distance évolutive estimée entre deux espèces éloignées dans la phylogénie soit faussée, par rapport à la réalité, d'une quantité supérieure à $min_{e \in A_P} \frac{l(e)}{2}$. On peut toutefois espérer retrouver les arêtes les plus longues, à condition bien-sûr que le bruit ne soit pas trop important. On dit qu'une arête $e \in A_P$ de la phylogénie P est **reconstruite correctement** par une méthode si cette méthode infère une phylogénie ayant une arête qui induit la même bipartition que e sur les espèces.

Atteson [Att97c] montre qu'on peut garantir que l'algorithme *Addtree* (légèrement modifié) reconstruit correctement toutes les arêtes e t.q. $l(e) > 2\epsilon$, où $\epsilon = L_\infty(D^P, \hat{D})$. En comparaison, il existe un contre-exemple montrant que la méthode NJ ne retrouve pas forcément les arêtes e t.q. $l(e) \leq 4\epsilon$ [Att97c]. De plus, on ne connait pas encore de preuve indiquant qu'elle retrouve les arêtes e t.q. $l(e) > 4\epsilon$. Pour la méthode Q^*, le fait que certaines arêtes ne soient éventuellement pas retrouvées n'a aucune influence sur l'inférence des autres arêtes, puisque les arêtes sont inférées indépendamment les unes des autres (ce qui n'est pas le cas pour *Addtree* et NJ). En conséquence, on peut montrer facilement le résultat suivant :

Théorème 5 *Soit $P = (S_P, A_P)$ une phylogénie, la méthode Q^* reconstruit correctement toute arête $e \in A_P$ t.q. $l(e) > 2L_\infty(D^P, \hat{D})$.*

Preuve : Soient $e \in A_P$ une arête de longueur $l(e)$ et $\sigma|\overline{\sigma}$ la bipartition qu'elle induit sur les espèces. Une preuve en tout point similaire à celle du lemme 5 montre que si $L_\infty(D^P, \hat{D}) < \frac{l(e)}{2}$, alors FPM infère le r4-arbre correct, $xy|zt$, pour tout quadruplet x, y, z, t t.q. $x, y \in \sigma$ et $z, t \in \overline{\sigma}$. Or si tous les r4-arbres induits par l'arête $e = \sigma|\overline{\sigma}$ sont dans l'ensemble Q donné en entrée à l'algorithme IQ^*, alors $\{\sigma|\overline{\sigma}\} \in B^*$ et la phylogénie $T^* \equiv B^*$ inférée par IQ^* contient une arête induisant la bipartition $\sigma|\overline{\sigma}$, *i.e.*, reconstruit correctement l'arête e de P. □

Ainsi, la méthode Q^* (comme la méthode *Addtree* modifiée [Att97c]) retrouve toutes les arêtes pour lesquelles une méthode de reconstruction, aussi sophistiquée soit-elle, peut avoir une garantie, au sens de la norme L_∞. La Fig. 2.8-(B) résume les conditions pour lesquelles une arête est correctement reconstruite en fonction de sa longueur et selon la méthode de reconstruction utilisée.

Les méthodes SQM et DCM n'apparaissent pas dans la Fig. 2.8-(B), car on ne dispose pour l'instant d'aucune garantie qu'elles puissent reconstruire certaines arêtes, même les plus longues. On peut même douter qu'on puisse montrer une quelconque garantie de cette sorte pour ces méthodes, car elles souffrent de l'important défaut algorithmique suivant : si le moindre r4-arbre "court" manque pour inférer une seule arête, alors la méthode conclut à l'inconsistance et ne fournit aucune phylogénie en sortie. Ainsi, dès lors que la plus petite arête e de P ne peut être retrouvée correctement, pour que la méthode SQM (ou DCM) infère une phylogénie, il faudrait qu'il existe une arête e' incorrecte, compatible avec les arêtes de $A_P - \{e\}$, pour laquelle tous les r4-arbres "courts" soient inférés incorrectement vis-à-vis de P, et de plus tel que cette arête ne partage pas de r4-arbres court avec une arête de $A_P - \{e\}$.

Notons que les bornes obtenues dans cette section ne sont que des garanties pour le pire des cas. En pratique les résultats des méthodes sont bien meilleurs, p. ex. , quand $L_\infty(D^P, \hat{D}) < min_{e \in A_P} 2l(e)$, les méthodes retrouvent la phylogénie correcte dans la plupart des cas, même si $L_\infty(D^P, \hat{D}) \geq min_{e \in A_P} \frac{l(e)}{2}$ [Gas97a].

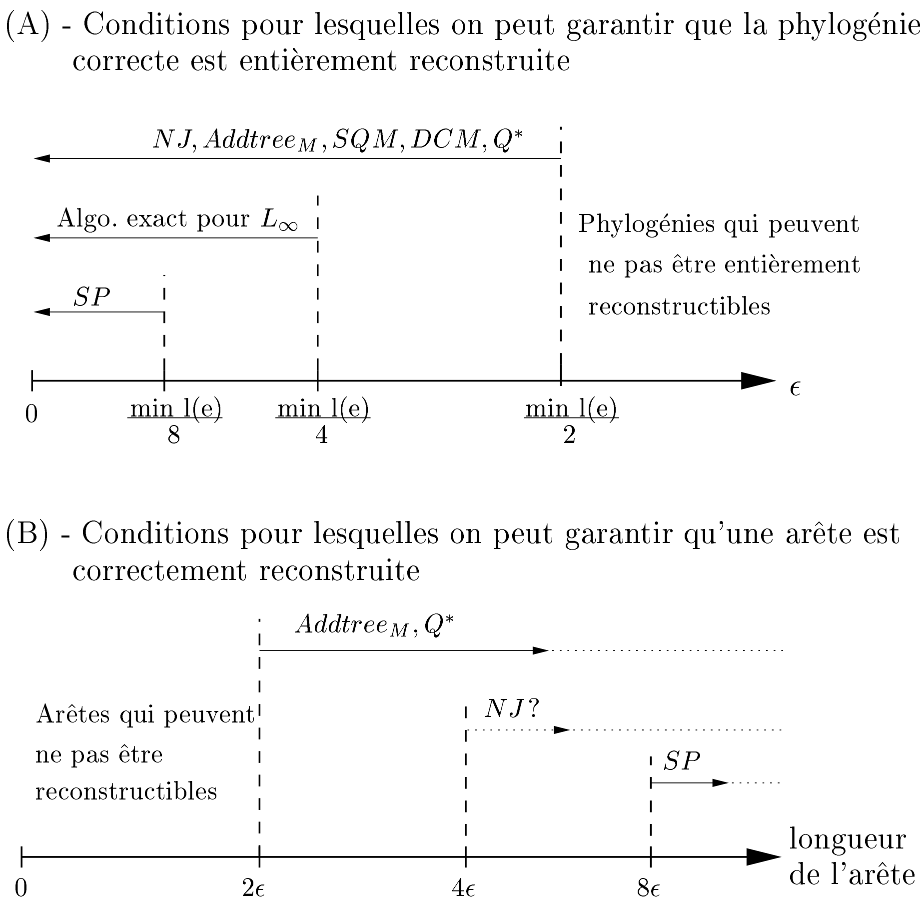


FIG. 2.8 – *Conditions de consistance totale et partielle au sens de la norme L_∞ pour les méthodes NJ [SN87], SP [ABF+96], Addtree (modifié) [ST77, Att97c], SQM [ESSW97b], DCM [ESSW97a], Q^* et pour une hypothétique* méthode exacte *(i.e., trouvant la phylogénie la plus proche de $\hat{D}$ au sens de L_∞). $l(e)$ dénote la longueur d'une arête de P et ϵ dénote $L_\infty(D^P, \hat{D})$, où D^P est la matrice de distances associée à la phylogénie P. (A) Condition pour que la phylogénie correcte P soit complètement retrouvée depuis une matrice de données $\hat{D}$. (B) Condition pour qu'une arête de P soit retrouvée en fonction de sa longueur (pour NJ, la borne est conjecturée et non prouvée, on sait, quoiqu'il en soit, qu'elle ne peut pas être meilleure que 4ϵ).*

2.3.2 Vitesse de convergence de la méthode Q^*

Nous montrons ici comment depuis certains résultats de la section précédente, nous pouvons déduire la complexité d'échantillonnage de la méthode Q^*, *i.e.*, le nombre k de caractères qu'elle requiert pour reconstruire correctement l'intégralité, ou une partie, de la phylogénie P avec une probabilité élevée. Pour ceci, nous supposerons que les séquences évoluent selon le modèle NCF (cf. section 1.3 chap. 1) le long d'une phylogénie P.

Pour bien comprendre l'enchaînement des idées développées ici, rappelons qu'au cours de l'évolution d'un ensemble de n séquences de k caractères le long d'une phylogénie P, plusieurs substitutions se produisent qui ne peuvent être retrouvées depuis les séquences observées aux feuilles de P. En conséquence, les distances évolutives observées entre les différentes espèces sont sous-estimées. On a donc généralement recours à un certain modèle de l'évolution suivant lequel les distances de $\hat{D}$ sont réhaussées. Toutefois, même si on emploie le modèle exact pour corriger $\hat{D}$, on a la garantie de retrouver la matrice D^P des vraies distances évolutives uniquement si l'on dispose d'un nombre k infini de caractères. Lorsque k est fini, on ne peut qu'espérer se rapprocher suffisamment des distances correctes pour être dans la boule d'exactitude garantie de la méthode de reconstruction utilisée, auquel cas le faible écart entre $\hat{D}$ et D^P n'empêche pas la méthode de retrouver la phylogénie P. Par exemple, nous avons montré dans la section précédente que si $L_\infty(D^P, \hat{D}) < min_{e \in A_P} \frac{l(e)}{2}$, alors la méthode Q^* retrouve la phylogénie P. En supposant que les espèces étudiées aient évolué selon le modèle NCF le long de la phylogénie P, nous pouvons évaluer la probabilité que la matrice $\hat{D}$, obtenue depuis des séquences de k caractères et corrigée selon le modèle NCF, soit à moins d'une certaine distance L_∞ de la matrice correcte D^P. De ce résultat, nous pourrons déduire en fonction de k la probabilité que la méthode Q^* retrouve la phylogénie correcte, et donc connaître le nombre k de caractères suffisants pour que la méthode Q^* retrouve la phylogénie correcte avec une probabilité fixée.

Résultat de convergence

Nous pouvons caractériser la vitesse de convergence de la méthode Q^* grâce à l'inégalité de Azuma-Hoeffding ou à celle de Chernoff [AS92]. Plusieurs travaux récents ont utilisé ces inégalités pour montrer la vitesse de convergence de différentes méthodes de R.P. [FK96, Att97b, ESSW97b, AS98]. Pour la méthode Q^*, nous utilisons le lemme suivant, obtenu depuis un résultat similaire dû à Farach et

Kannan[6] (aussi utilisé par Atteson [Att97b]), sur la base de l'inégalité de Azuma-Hoeffding :

Lemme 6 *(Farach, Kannan et Atteson) Supposons qu'un ensemble d'espèces évolue le long d'une phylogénie suivant le modèle de l'évolution de NCF, supposons que l'on dispose de k caractères décrivant ces espèces, et soit $d = max(D_{xy})$ la distance évolutive maximum entre deux espèces x et y induite par la phylogénie. Pour deux espèces x et y quelconques :*

$$P(|D_{xy} - \hat{D}_{xy}| < \epsilon) \geq 1 - 2e^{-(e^{-2\epsilon}-1)^2 e^{-4d} k/8} \ ,$$

où $\hat{D}_{xy}$ est la distance observée entre x et y sur la base des k caractères et corrigée selon le modèle NCF.

Preuve : La preuve est donnée en annexe B. □

Nous avons vu précédemment que la méthode Q^* retrouvait la phylogénie correcte P avec une probablité de 1 si $L_\infty(D, \hat{D}) < min_{e \in A_P} \frac{l(e)}{2}$. En choisissant $\epsilon = min_{e \in A_P} \frac{l(e)}{2}$, le lemme précédent nous permet facilement d'obtenir la probabilité que la méthode Q^* reconstruise correctement une phylogénie P :

Théorème 6 *Soit $P = (S_P, A_P)$ une phylogénie dont les espèces évoluent sous le modèle NCF et dont les les arêtes $e \in A_P$ ont une probabilité de substitution, $p(e)$, bornée inférieurement par $f = min_{e \in A_P} p(e)$. La probabilité que la méthode Q^* retrouve totalement la phylogénie P est au moins :*

$$1 - n^2 e^{(-f^2 e^{-4d} k/8)} \quad .$$

Preuve : Soient

- N l'ensemble des $\binom{n}{2}$ paires (x, y) de deux espèces différentes choisies parmi les n espèces étudiées,
- $A_{(x,y)}$ l'événement $|D_{xy} - \hat{D}_{xy}| < min_{e \in A_P} \frac{l(e)}{2}$,
- B l'événement correspondant à l'intersection de tous les événements $A_{(x,y)}$ (*i.e.*, pour toutes les paires (x, y), $\hat{D}_{xy}$ est correcte à $min_{e \in A_P} \frac{l(e)}{2}$ près). En fait, la matrice $\hat{D}$ n'est pas tout-à-fait quelconque, *i.e.*, $\hat{D}_{xy} = \hat{D}_{yx}$ et $\hat{D}_{xx} = 0$,

6. Originellement [FK96], ces auteurs se sont intéressés à la convergence de la méthode SP au sens d'une *distance variationnelle* ; toutefois, à l'occasion de l'article de Erdös *et al* [ESSW97b], ils ont aussi montré une borne sur la vitesse de convergence de SP pour la distance L_∞, en s'appuyant sur le lemme repris ici.

si bien que seulement $N = \binom{n}{2}$ distances différentes interviennent dans $\hat{D}$, *i.e.*, l'énoncé de l'événement B peut être simplifié : il s'agit de l'intersection des événements $A_{(x,y)}$ pour lesquels $x < y$.

Rappelons aussi l'inégalité de Bonferonni : soit $A_1, \ldots, A_m$ un ensemble d'événements,

$$P\left(\bigcap_{1\leq i\leq m} A_i\right) \;\geq\; 1 - \sum_{1\leq i\leq m} P(\overline{A}_i)\ . \tag{2.6}$$

La méthode Q^* retrouve l'ensemble de la phylogénie P ssi $Q = Q_P$, *i.e.*, ssi *tous* les quadruplets sont reconstruits correctement par FPM. Par le lemme 5, ceci nécessite que *toutes* les distances $\hat{D}_{xy}$ soient correctes à $min_{e\in A_P}\frac{l(e)}{2}$ près, ce qui correspond à l'événement B. On obtient une borne sur la probabilité de cet événement en utilisant (2.6) :

$$P(B) = P\left(\bigcap_{(x<y)\in N} A_{(x,y)}\right) \geq 1 - \sum_{(x,y)\in N} P(\overline{A}_{(x,y)})\ . \tag{2.7}$$

Par le lemme 6, en choisissant $\epsilon = min_{e\in A_P}\frac{l(e)}{2}$, on sait :

$$\forall (x,y) \in N,\ P(\overline{A}_{(x,y)}) \;\leq\; 2e^{(-f^2e^{-4d}k/8)}\ , \tag{2.8}$$

en remarquant que, pour le modèle NCF, $min_{e\in A_P} l(e) = -\frac{1}{2}\ln(1-2f)$, et en remplaçant raisonnablement $(1-\sqrt{1-2f}\,)$ par f, son approximation du 1er ordre. De (2.7) et (2.8) on déduit :

$$P(B) \geq 1 - \binom{n}{2} 2e^{(-f^2e^{-4d}k/8)} \geq 1 - n^2e^{(-f^2e^{-4d}k/8)}\ .$$

□

Suivant le même raisonnement, il est tout aussi simple d'obtenir une borne inférieure sur la probabilité qu'une arête de la phylogénie P soit retrouvée. Cette borne dépend de la longueur de l'arête et non plus de la plus petite arête de la phylogénie :

Théorème 7 *Soit $P = (S_P, A_P)$ une phylogénie, toute arête $e = \sigma|\overline{\sigma} \in A_P$, à laquelle est associée une probabilité de substitution $p(e) > f$, a une probabilité* *d'être inférée par la méthode Q^*, supérieure à :*

$$1 - n^2 e^{(-p(e)^2 e^{-4d} k/8)} \ . \tag{2.9}$$

Preuve : Pour qu'une arête e soit correctement reconstruite, il faut que tous les r4-arbres qu'elle induit soient correctement retrouvés par FPM, or chacun de ces r4-arbres est correctement retrouvé si les 6 distances entre ses 4 espèces comportent une erreur inférieure à $\frac{l(e)}{2}$ (la preuve est identique à celle du lemme 5). Comme toutes les distances $\hat{D}_{xy}$ sont impliquées dans au moins un des r4-arbres induits par l'arête e, demander que tous les r4-arbres induits par cette arête soient correctement reconstruits, revient à demander que les $\binom{n}{2}$ distances $\hat{D}_{xy}$ entre espèces comportent une erreur inférieure à $\frac{l(e)}{2}$. Le reste de la preuve est donc identique à la preuve du théorème précédent, *i.e.*, repose sur l'inégalité de Bonferonni et sur le lemme 6, en fixant, cette fois, $\epsilon = \frac{l(e)}{2}$. □

Du théorème 6, on peut établir facilement le taux de convergence de la méthode Q^*, *i.e.*, le nombre k de caractères qu'elle requiert pour retrouver totalement la phylogénie P :

Corollaire 2 *Supposons que k caractères évoluent selon le modèle NCF sur une phylogénie P, la méthode Q^* reconstruit correctement la phylogénie P avec une probabilité au moins $1-\alpha$ ($\alpha>0$) si :*

$$k > \frac{8\ln(\frac{n^2}{\alpha})e^{4d}}{f^2} \ . \tag{2.10}$$

Comme souvent, le nombre de caractères requis augmente logarithmiquement selon l'inverse du niveau de confiance souhaité.

2.3.3 Comparaison avec les autres méthodes

On obtient ci-dessus la même borne pour la vitesse de convergence de la méthode Q^* que celle obtenue par Atteson [Att97b] pour les méthodes NJ et *Addtree*, ce qui tient au fait que ces vitesses de convergence sont obtenues depuis les boules d'exactitude garantie des méthodes, identiques pour les trois méthodes (cf. section 2.3.1). Notons qu'en adoptant la démarche de Erdös *et al* [ESSW97b], on obtient une borne légèrement moins bonne pour la méthode Q^*, car ces auteurs raisonnent sur les $\binom{n}{4}$ r4-arbres, plutôt que sur les $\binom{n}{2}$ distances entre espèces[7].

La vitesse de convergence (2.10) que nous obtenons pour la méthode Q^* est du même ordre que (ou identique à) celles des méthodes SP [ABF+96, FK96], NJ et *Addtree* [Att97b]. La vitesse de convergence de ce groupe de méthodes est proportionnelle au *diamètre* de la phylogénie P (nombre d'arêtes maximum entre deux feuilles), en raison du paramètre d (2.10). La vitesse de convergence des méthodes SQM [ESSW97b] et DCM [ESSW97a] a une expression très similaire à celle des méthodes précédentes, à la différence près qu'elle dépend non pas du diamètre mais de la *profondeur* de la phylogénie (nombre maximum d'arêtes entre un noeud interne et la feuille la plus proche).

Sur la remarque que le diamètre d'une phylogénie peut être au pire en $O(n)$ et la profondeur au pire en $O(\log n)$, Erdös *et al* [ESSW97a, ESSW97b] concluent que les méthodes SQM et DCM ont une vitesse de convergence bien meilleure que les autres méthodes. Toutefois, comme le font remarquer Ambainis *et al* [ADFK98], les vitesses de convergence des méthodes dépendent du diamètre (ou de la profondeur) *valué*, *i.e.*, relatif à la somme des longueurs de ces arêtes (et non juste à leur nombre). Or en R.P., ces deux quantités sont des constantes proches, car les jeux de données étudiés sont tels que la plus grande distance évolutive considérée entre deux espèces est bornée par une petite constante (p. ex. $d < 1$ [Nei91] ; pour $d \geq 1$, on sort du cadre des modèles de l'évolution utilisés pour corriger les données). Ainsi, la distinction entre diamètre et profondeur n'a donc pas lieu d'être, en revanche, on ne dispose généralement pas de garantie sur la longueur de la plus petite arête de la phylogénie reconstruite, donc sur le paramètre f, devant lequel toutes les méthodes sont à égalité. Plus n augmente, et plus f diminue, car toute espèce supplémentaire vient couper une arête en deux arêtes de longueur inférieure. Ceci rejoint une règle expérimentalement constatée par les biologistes, à savoir que plus le nombre d'espèces considérées est important et plus il faut disposer d'un nombre important de caractères. Les taux de convergence des méthodes sont quadratiques en $O(1/f)$. Or, nous savons que $f \leq O(1/\log n)$[8]. Ainsi dans le meilleur des cas, $f \in O(1/\log n)$, et les taux de convergences sont logarithmiques. Toutefois ce cas est très peu probable, puisqu'il correspond à une phylogénie équilibrée dont toutes les arêtes ont même longueur. En revanche, on peut raisonnablement poser $f = O(1/n)$, ce qui indique une complexité d'échantillonnage en $O(n^2 \log n)$. Il ne faut toutefois pas perdre de vue que f peut être arbitrairement faible, conduisant

7. Cette façon de faire résulte du fait qu'ils doivent introduire une nuance entre les quadruplets *courts* et les autres quadruplets.

8. cet abus de notation traduit que dans le meilleur des cas la phylogénie est de forme équilibrée (cf. Fig. 2.1-(20) et toutes les arêtes ont la même longueur, ainsi la distance entre deux espèces quelconques est répartie en $\log n$ arêtes et $f \approx \frac{d}{\log n}$).

les méthodes mentionnées ci-dessus à requérir un nombre irréaliste de caractères pour reconstruire la phylogénie correcte avec une probabilité élevée. Ce problème est bien connu des biologistes, et il est probable qu'aucune méthode n'existe jamais qui puisse produire une estimation fiable d'une phylogénie possédant des arêtes très courtes.

Notons que la formulation du lemme 6 est spécifique au modèle NCF. Toutefois ce lemme s'étend à d'autres modèles stochastiques de l'évolution (selon le principe employé par [ESSW97b, ADFK98], sur la base de la méthode du *logdet* [Ste94a]). De même, tous les résultats de cette section, déduits pour des caractères à deux états, peuvent être étendus pour des caractères à quatre états (nucléotides), à vingt états (protéines), ou plus.

2.4 Résultats expérimentaux

Après avoir donné un algorithme polynômial pour la méthode Q^* et montré qu'elle possédait de bonnes garanties théoriques, nous étudions maintenant ses performances pratiques. Par son principe même[9], la méthode Q^* doit inférer une très faible proportion d'arêtes incorrectes (*i.e.*, d'*erreur de 1ère espèce*, cf. chap. 1, part. II). En revanche, on s'attend à ce que la méthode infère souvent des phylogénies partiellement irrésolues.

Pour l'importance de ces phénomènes, nous allons examiner les résultats d'expériences menées en simulant l'évolution d'un ensemble d'espèces. Un schéma de simulation similaire à celui décrit dans les chapitres 1 et 2 de la partie précédente est adopté, et nous permet de comparer la méthode Q^* à la célèbre méthode de distances NJ [SN87, SK88]. Nous analyserons les résultats des deux méthodes sur la base de leur erreur structurelle, décomposée soit en termes d'erreur de 1ère et de 2ème espèces, soit en termes de biais et de variance (cf. chap. 1, part. II). Nous terminerons cette étude expérimentale en appliquant la méthode Q^* au jeu de données sur les mammifères.

2.4.1 Schéma de simulation

Pour une description de la façon dont sont générés les jeux de données, nous renvoyons le lecteur à la section 1.4.1, chap. 1, part. II.

Nous ajoutons toutefois une nouvelle mesure d'erreur à celles définies dans le chap. 1, part. II. Selon les conditions d'évolution et la longueur des séquences,

9. Ce principe nécessite la concordance de $\Omega(n^2)$ à $O(n^4)$ r4-arbres de données pour chaque arête de la phylogénie inférée.

chaque arête de la phylogénie modèle t ne génère pas forcément de substitution lors de la production du jeu de données (surtout pour des taux d'évolution faibles). En conséquence, les jeux de données ne contiennent pas toujours de l'information pour chaque arête de t (ce qui rend compte d'un problème souvent rencontré en R.P.). Dans un tel cas, les méthodes de reconstruction n'ont aucun support leur permettant de retrouver de telles arêtes "*muettes*". Sur la base de cette réflexion, nous étudierons aussi les performances des méthodes de reconstruction en terme d'erreur structurelle à la phylogénie "*réalisée*", notée t_R. Cette phylogénie est obtenue depuis t en contractant les arêtes qui n'ont été le support d'aucune substitution pour le jeu de données analysé (on a donc $B(t_R) \subseteq B(t)$). Nous noterons e_1^R et e_2^R les erreurs de 1ère et de 2ème espèces relatives à t_R. Ces erreurs sont calculées depuis des arbres t_R potentiellement différents d'un jeu de données à l'autre. Pour prendre autant que possible la mesure des valeurs obtenues pour e_1^R et e_2^R, nous indiquerons aussi dans les résultats la taille moyenne de la phylogénie t_R en nombre d'arêtes internes ($\in [0, 7]$).

2.4.2 Résultats

Les résultats des simulations sont décrits en détail dans l'annexe E. Nous résumons ici les traits marquants de ces résultats.

Comme prévu, la méthode Q^* infère très peu d'arêtes incorrectes, et elle mène à une phylogénie bien plus stable que celle proposée par exemple par la très connue méthode NJ [SN87]. En contrepartie, la méthode Q^* infère des phylogénies partiellement résolues. Toutefois, dans les cas les plus favorables, cette méthode infère des phylogénies dont la taille est plus proche de la phylogénie réalisée, t_R, que la phylogénie produite par NJ. Ceci dénote une capacité à reconnaître plus ou moins la quantité d'information présente dans les donnéees, et ainsi à éviter de produire une phylogénie totalement résolue quand le jeu de données ne contient pas assez d'information pour cela.

L'importante contrainte, imposée sur chaque arête de la phylogénie inférée par la méthode Q^*, conduit cette méthode a être parfois trop conservatrice. Toutefois, comme la phylogénie qu'elle infère ne comporte presque que des arêtes fiables, cette phylogénie peut être considérée comme une base saine, à laquelle on peut ajouter d'autres arêtes par un quelconque autre principe. Ceci a été vérifié par Rice et Warnow [RSWY97].

2.4.3 Application sur un jeu de données réel

La Fig. 2.9 montre la phylogénie T^* inférée par la méthode Q^* sur le jeu de données regroupant 11 séquences de mammifères de 166 caractères.

La méthode Q^* retrouve correctement le groupe des primates ((homme, chimpanzé), singe) et celui des ongulés (mouton, vache, cochon, cheval). Le groupement (chien, kangourou), aussi inféré par NJ, MP et l'arbre bootstrap réduit au seuil optimum (cf. chap. 1, part. II), est jugé incorrect par une majorité de biologistes, toutefois aucune conclusion ferme n'a été retenue [PHS91] pour l'instant. La méthode Q^* ne prend pas de décision quant à la place du rongeur et du lapin vis-à-vis des autres animaux. La question du positionnement du rongeur dans la phylogénie est actuellement source d'un débat important, et aucun consensus n'apparaît pour l'instant. La dernière hypothèse émise est que le rongeur se situe à la racine de l'arbre d'évolution des mammifères [dGP$^+$96, GHL91].

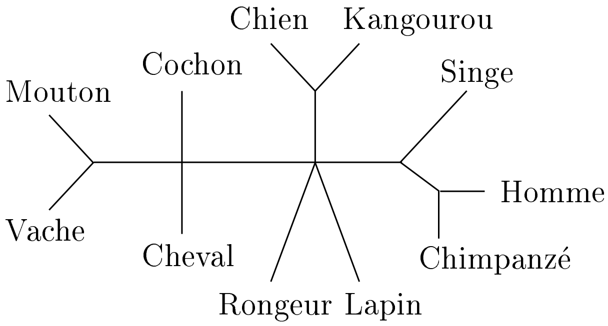


FIG. 2.9 – *La méthode Q^* appliquée au jeu de données sur 11 mammifères*

2.5 Conclusion

La méthode Q^*, objet de ce chapitre, propose une estimation de la phylogénie inconnue, en assurant que chaque arête proposée est combinatoirement sûre. Cette méthode possède une complexité polynômiale ainsi que des garanties de consistance et de vitesse de convergence :

- nous avons caractérisé au sens de la norme L_∞ les conditions dans lesquelles elle retrouve de façon sûre la phylogénie correcte (ou certaines de ses arêtes), indépendamment du nombre de caractères disponibles et de tout modèle de l'évolution ;

– en faisant l'hypothèse d'un modèle de l'évolution, nous avons montré qu'on peut évaluer le nombre de caractères requis par la méthode Q^* pour reconstruire la phylogénie inconnue (ou certaines de ses arêtes) avec une probabilité élevée.

Ce chapitre apporte donc une nouvelle solution à notre objectif général, qui est de proposer des méthodes inférant des estimations structurelles fiables de phylogénies. Cette solution repose sur une approche à résolution croissante, *i.e.*, proposer d'emblée uniquement des arêtes fiables (on évite ainsi la phase d'élagage des arêtes non-sûres, propre à l'approche à résolution décroissante).

Enfin, les résultats expérimentaux indiquent que la méthode Q^* a les qualités requises pour pouvoir servir de *base saine* pour la reconstruction, indiquant un socle fiable auquel on peut éventuellement ajouter de nouvelles arêtes par l'intermédiaire d'autres méthodes de reconstruction.

Chapitre 3

Un schéma générique pour les heuristiques agglomératives basées sur les quadruplets

3.1 Introduction

Dans le chapitre précédent, nous avons montré comment on pouvait inférer une phylogénie, T^*, dont chaque arête est combinatoirement sûre (*i.e.*, a une forte probabilité d'être correcte) car elle correspond à un nombre de r4-arbres des données compris entre $\Omega(n^2)$ et $O(n^4)$. Cependant, cette importante contrainte imposée sur les arêtes de T^*, peut parfois être trop restrictive, impliquant un faible degré de résolution pour cette phylogénie. Les simulations du chapitre précédent ont en effet montré que T^* possède, certes, très peu d'arêtes incorrectes en moyenne, mais oublie parfois un pourcentage non-négligeable d'arêtes de t. Dans un tel cas, on peut souhaiter compléter T^* par des arêtes supplémentaires (vérifiant des contraintes moins fortes que les arêtes initiales) afin d'obtenir une estimation $\hat{t}$ plus résolue de la phylogénie inconnue t.

Toutefois, quel algorithme employer pour compléter T^* ? Il faut un algorithme qui infère des arêtes sur la base des r4-arbres de données non employés pour construire T^* (*i.e.*, $Q - Q^*$) : il est donc préférable que ce soit un algorithme basé sur un principe de quadruplets. De plus, il faut que cet algorithme puisse accepter en entrée une phylogénie déjà partiellement résolue : aucun algorithme de ce style ne semble pour l'instant exister. Qu'à cela ne tienne ! Nous allons en concevoir un qui soit adapté à la situation.

Examinons d'abord rapidement les algorithmes existants, afin de décider sur quelle base sera conçu notre algorithme. Disposant d'un ensemble de r4-arbres Q, on peut utiliser un certain nombre d'algorithmes basés sur les quadruplets pour inférer une phylogénie sur l'ensemble des espèces :

- L'algorithme *Addtree* [ST77, BL84].
  Cet algorithme construit une phylogénie suivant le principe agglomératif dit de *neighbor joining*, depuis une distance $\hat{D}$ (cf. chap. 2, part. II). Pour utiliser cet algorithme il nous faut donc transformer Q en matrice de distances $\hat{D}$. On peut pour cela employer la transformation proposée par Colonius et Schulze [CS79] (cf. aussi [Fit81]) :

$$\hat{D}_{x,y} = \begin{cases} \binom{n-2}{2} + 1 - s(x,y) & \text{si } x \neq y \\ 0 & \text{si } x = y \end{cases},$$

  où $s(x,y)$ désigne le score de la paire x,y, *i.e.*, le nombre de r4-arbres où ils sont voisins : $s(x,y) = \Big|\{z,t\} \text{ t.q. } xy|zt \in Q,\ z \neq t\Big|$. Cette transformation garantit que si Q est arboré, alors $\hat{D}$ est une distance d'arbre [CS79], or disposant d'une telle matrice, il est prouvé que *Addtree* retrouve la phylogénie t t.q. $Q_t = Q$ [BL84, BD86]. Le problème de cette approche est que la transformation $Q \rightarrow \hat{D}$ ci-dessus n'est pas adaptée au cas d'ensembles Q incomplets.

- L'algorithme *SQM* [ESSW97b].
  Cet algorithme est applicable si l'on dispose d'un ordre sur les r4-arbres de Q, soit $r_1, \ldots, r_{|Q|}$, les classant selon la confiance qu'on leur attribue. Dans la plupart des cas, on pourra établir un tel ordre entre les r4-arbres de Q, en accord avec la méthode qui a permis de les inférer. Le principal défaut de l'algorithme *SQM*, de l'aveu même de ses auteurs [Ste97], est d'avoir avant tout été conçu pour que sa convergence puisse être prouvée ; l'applicabilité de l'algorithme s'en ressent d'autant. Cet algorithme ne peut en effet produire que les réponses suivantes : "inconsistant", "insuffisant" ou proposer une phylogénie totalement résolue, $\hat{t}$. De plus, pour obtenir $\hat{t}$ comme réponse, faut-il encore que les r4-arbres du début de l'ordre soient arbre-consistants, *i.e.*, $\exists i \leq |Q|$ t.q. $\{r_1, \ldots, r_i\} \subseteq Q_{\hat{t}}$ et qu'ils définissent exactement une phylogénie binaire. Il est vraisemblable qu'une telle condition ne soit pas souvent vérifiée.

- L'algorithme de recuit simulé proposé par [DvHK86], noté ici *RS*.
  Cet algorithme partage les défauts des méthodes probabilistes : depuis le

même jeu de données Q, il ne donne pas toujours la même phylogénie. De plus, on ne peut pas prouver de complexité pour cet algorithme, et étant donnée la description qu'en donnent Dress *et al* [DvHK86], on peut facilement imaginer qu'il est plus lent que tout algorithme déterministe raisonnable.

– L'algorithme de $QuartetPuzzle$ [SvH96] (noté ici QP).
  Initialement proposé pour construire une phylogénie depuis des r4-arbres inférés par le principe du Maximum de Vraisemblance, cet algorithme glouton peut être employé depuis tout ensemble de r4-arbres Q (complet ou incomplet). Il fonctionne selon un principe incrémental (comme Q^*) : les espèces de E sont successivement greffées sur une phylogénie comportant initialement 4 espèces. À chaque étape, on greffe une espèce i sur une des arêtes de la phylogénie courante. Pour cela on choisit l'arête e telle que l'insertion de i sur e contredise un minimum de r4-arbres de Q. Notons qu'on peut tout à fait envisager une variante agglomérative de cet algorithme (notée QP_a) : partant de la phylogénie étoile, à chaque étape, on décide de regrouper les deux arêtes tel qu'un nombre minimum de r4-arbres de Q soient contredits. La complexité de cet algorithme (version QP ou QP_a) est son point faible, puisqu'elle est de $O(n^5)$ en première analyse, ce qui le classe derrière $Addtree$ pour lequel on peut montrer une complexité $O(n^4)$ [Att97a].

De ces heuristiques, nous ne retiendrons pas SQM et RS en raison de leur principe même. Nous ne pouvons pas non plus retenir la version de QP proposée par Strimmer et von Haeseler [SvH96], car la phylogénie T^* que nous cherchons à compléter contient déjà toutes les espèces de E. De cet examen des algorithmes *de quadruplets* actuels, nous conservons donc uniquement l'idée d'un algorithme agglomératif, tel $Addtree$ ou QP_a.

Dans ce chapitre nous proposons un cadre général, nommé HAQ, pour décrire les *H*euristiques *A*gglomératives basées sur les *Q*uadruplets. Nous verrons que les heuristiques adoptant ce schéma peuvent accepter comme point de départ soit la phylogénie étoile (de façon similaire à $Addtree$ et QP_a), soit une phylogénie déjà partiellement résolue (p. ex. la phylogénie T^*). Le schéma HAQ peut être utilisé en conjonction avec différents critères de reconstruction, définis sur la base de trois facteurs : les r4-arbres de Q satisfaits par la phylogénie construite, les r4-arbres de Q contredits par la phylogénie, et ceux induits par la phylogénie alors que Q ne contient aucune résolution pour leur quadruplet. Nous proposons une instanciation de ce schéma propre à la R.P., en détaillant des critères de reconstruction qui lui sont spécifiques. Le schéma HAQ peut être implémenté en $O(n^4)$ et convient aux éventuels ensembles de r4-arbres incomplets, ce qui répond aux points faibles

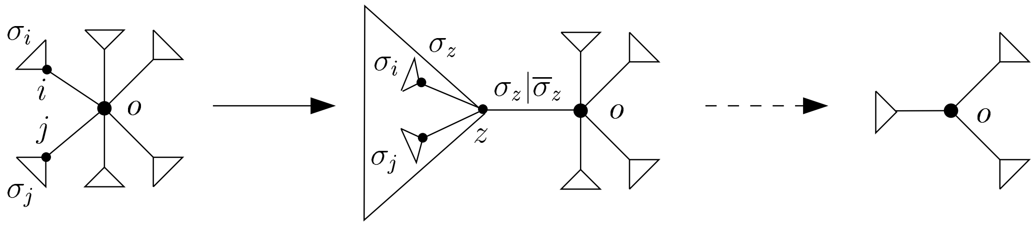


FIG. 3.1 – *Résolution d'un noeud o de degré > 3 suivant le principe agglomératif. À chaque étape deux sous-arbres σ_i et σ_j connectés à o sont réunis en un nouveau sous-arbre σ_z. Ceci a pour effet d'ajouter une arête (oz) à la phylogénie, induisant la bipartition $\sigma_z|\overline{\sigma}_z$ sur les espèces.*

que nous avons précédemment soulevés respectivement pour QP_a et *Addtree*– employé suite à la transformation de Colonius et Schulze [CS79]. Enfin, dans le cadre de notre problématique de fond, nous montrons comment on peut facilement obtenir une mesure de confiance pour les arêtes ajoutées à la phylogénie T^* par un algorithme de type *HAQ*.

3.2 *HAQ*: principe général et algorithme

Nous rappelons ici le principe agglomératif en introduisant les notations qui nous serviront tout au long de ce chapitre. Le principe agglomératif agit par modifications successives d'une phylogénie initiale, notée ici $\hat{t}_0$, en agglomérant à chaque étape deux sous-arbres connectés à un noeud o irrésolu (cf. Fig. 3.1). La phylogénie $\hat{t}_0$ peut en fait comporter plusieurs noeuds de ce type, mais nous nous limitons pour l'instant à envisager le cas d'un seul de ces noeuds. Chaque agglomération de sous-arbres diminue le degré du noeud o; on peut donc répéter une telle opération jusqu'à ce que o soit de degré 3 (*i.e.*, de degré minimum pour un noeud interne). À chaque étape, on considère toutes les agglomérations possibles (σ_i, σ_j) de deux sous-arbres σ_i et σ_j et on en choisit une selon le critère de reconstruction qu'on veut optimiser. Cette agglomération entraîne la création d'un nouveau noeud interne $z = (\sigma_i, \sigma_j)$ et d'une nouvelle arête interne $e = (oz)$ dans la phylogénie. Cette arête induit une bipartition b_{ij} sur les espèces étudiées et si on note $\hat{t}_m$ la phylogénie obtenue après l'étape m, on a $B(\hat{t}_{m-1}) \subseteq B(\hat{t}_m)$. La bipartition b_{ij} modifie (selon les quadruplets qu'elle résout) le nombre de r4-arbres de Q satisfaits (soit s) ou contredits (soit c) par la phylogénie, ainsi que le nombre de quadruplets résolus par la phylogénie sans que Q ne contienne de r4-arbre pour eux (soit u). À l'étape m, chacune des agglomérations possibles entraîne l'ajout dans $B(\hat{t}_m)$ d'une bipartition différente de celle induite par les autres agglomérations, donc une modification particulière des quantités s, c et u[1]. En fait, la modification apportée à ces quantités par une agglomération est aisément caractérisée. Au début de chaque étape m, certains quadruplets sont déjà résolus par la phylogénie $\hat{t}_{m-1}$ résultant de l'étape précédente, tandis que d'autres quadruplets sont encore irrésolus, *i.e.*, aucun de leurs r4-arbres n'appartient à $Q_{\hat{t}_{m-1}}$. À l'étape m, l'ajout d'une arête $(oz) \equiv b_{ij}$ dans la phylogénie ne peut contribuer qu'à résoudre des quadruplets précédemment irrésolus ; tout quadruplet précédemment résolu par la phylogénie reste résolu de la même façon (une arête de $\hat{t}_{m-1}$ sépare ses quatre espèces en deux paires et cette arête est toujours présente dans $\hat{t}_m$). Ainsi **un quadruplet ne peut *influer* sur le score des agglomérations (et donc sur leur choix) que tant qu'il est irrésolu dans la phylogénie**. À chaque étape m, le score d'une agglomération (σ_i, σ_j) dépend exactement des quadruplets irrésolus dans $\hat{t}_{m-1}$ qu'elle résout dans $\hat{t}_m$: les r4-arbres de Q qu'elle satisfait, soit s_{ij}, les r4-arbres de Q qu'elle contredit, soit c_{ij} et les quadruplets qu'elle résout et que Q ne résout pas, soit u_{ij}. Si (σ_i, σ_j) est l'agglomération choisie à l'étape m, nous avons

$$s_m - s_{m-1} = s_{ij} \tag{3.1}$$
$$c_m - c_{m-1} = c_{ij} \tag{3.2}$$
$$u_m - u_{m-1} = u_{ij} \tag{3.3}$$

où s_m, resp. c_m et u_m, dénote le terme s, resp. c et u, associés à la phylogénie $\hat{t}_m$ résultant de l'étape m.

À chaque étape, le score de chaque agglomération (σ_i, σ_j) est évalué selon un certain critère, que nous notons l (*local*), dont l'expression analytique dépend des trois valeurs s_{ij}, c_{ij} et u_{ij}. L'expression exacte de l dépend du contexte d'application de *HAQ* mais doit tendre à mesurer la qualité de l'arête (oz) (résultant de l'agglomération (σ_i, σ_j)) pour représenter Q. À chaque étape on choisit bien-sûr l'agglomération qui induit l'arête de meilleure qualité au sens de l.

Dans le schéma *HAQ*, on ne cherche pas à forcer la résolution complète du noeud o : on ajoute des arêtes à la phylogénie tant que la qualité de celui-ci pour représenter Q au sens d'un certain critère g (*global*) s'en trouve améliorée. Pour pouvoir évaluer la qualité de la phylogénie $\hat{t}_m$ pouvant résulter de l'étape m, le critère g doit se baser sur les termes s_{ij}, c_{ij}, u_{ij} propres à chaque aggloméra-

1. Ceci résulte du fait que $Q_{b_{ij}}$ dépend de la composition des deux sous-arbres agglomérés, σ_i et σ_j.

tion candidate, mais aussi sur s_{m-1}, c_{m-1} et u_{m-1} (en accord avec les équations (3.1)-(3.3)). Par exemple, lorsque toute agglomération encore possible conduit la phylogénie à contredire plus d'un certain pourcentage des r4-arbres initiaux (ou fait passer le rapport r4-arbres satisfaits sur r4-arbres contredits au dessous d'un certain seuil), on peut raisonnablement penser qu'aucune de ces agglomérations ne doit être effectuée, et qu'il faut arrêter là le processus de résolution de la phylogénie.

Le critère g est le critère à optimiser par la phylogénie inférée, aussi le critère l est choisi en fonction de g, mais n'a pas forcément la même expression analytique. Le choix de ces deux critères dépend du contexte dans lequel on emploie *HAQ*. Dans la plupart des cas, en raison même du principe glouton de ce schéma heuristique, les critères l et g ne possèdent pas la même expression analytique (nous renvoyons le lecteur à la section 3.5 de ce chapitre pour un tel exemple). Le schéma *HAQ* opérant sur un noeud irrésolu o est résumé dans l'algorithme 6 donné ci-dessous.

Algorithme 6: résolution d'un noeud irrésolu o selon le principe *HAQ*.

---

donnée : un ensemble de r4-arbres Q à résoudre et une phylogénie $\hat{t}_0$ possédant un noeud o de degré $d(o) > 3$ connecté aux sous-arbres $S = \{\sigma_1, \ldots, \sigma_i, \sigma_j, \ldots, \sigma_d\}$ par les arêtes $e_1, \ldots, e_i, e_j, \ldots, e_d$.

résultat: une phylogénie $\hat{t}_m$ t.q. $B(\hat{t}_0) \subseteq B(\hat{t}_m)$

$m \leftarrow 1$

tant que $d(o) > 3$ *et* $\exists\, (\sigma_i, \sigma_j)$ *t.q.* $g\,(s_{ij}, c_{ij}, u_{ij}, s_{m-1}, c_{m-1}, u_{m-1}) > 0$ **faire**

- Choisir $\{\sigma_i, \sigma_j\} \in S$ qui optimise $l(\sigma_i, \sigma_j)$
- /* *Modification résultante de la phylogénie* */
- $\hat{t}_m \leftarrow \hat{t}_{m-1} \cup \{(oz), (zi), (zj)\} - \{(oi), (oj)\}$
- $S \leftarrow S \cup \{\sigma_z\} - \{\sigma_i, \sigma_j\}$
- /* *Mise à jour des divers compteurs* */
- $s_m \leftarrow s_{m-1} + s_{ij}$
- $c_m \leftarrow c_{m-1} + c_{ij}$
- $u_m \leftarrow u_{m-1} + u_{ij}$
- $m \leftarrow m + 1$

3.3 Complexité de *HAQ*

Nous montrons maintenant comment le schéma générique *HAQ* peut être efficacement implémenté grâce à un certain nombre de propriétés. Pour les premières de ces propriétés, nous ne nous limitons plus à décrire la situation du point de vue d'un seul noeud irrésolu o, et nous reconsidérons momentanément l'ensemble de la phylogénie à résoudre.

Chaque agglomération effectuée par le schéma HAQ autour d'un noeud irrésolu résout un certain nombre de quadruplets. Inversement, chaque quadruplet peut être résolu par plusieurs agglomérations candidates à un moment donné. Une question naturelle est de savoir comment sont réparties dans la phylogénie les agglomérations susceptibles de résoudre un quadruplet q, *i.e.*, est-il possible de caractériser les noeuds irrésolus et les sous-arbres σ_i, σ_j qui leur sont connectés tel que l'agglomération (σ_i, σ_j) résout le quadruplet q?

Observation 1 *Soit q un quadruplet non-résolu par la phylogénie initiale $\hat{t}_0$. Il existe un et un seul noeud irrésolu o dans $\hat{t}_0$ tel que l'agglomération de deux sous-arbres connectés à o conduit à résoudre q.*

Preuve : Tout quadruplet q n'est pas résolu par $\hat{t}_0$ ssi $\hat{t}_0$ ne possède pas d'arête induisant une bipartition $b = \sigma|\overline{\sigma}$ t.q. deux des quatre espèces de q appartiennent à σ et les deux autres à $\overline{\sigma}$. En conséquence, si q n'est pas résolu par $\hat{t}_0$, alors $\hat{t}_0$ possède forcément un noeud o t.q. chacune des espèces de q appartient à un sous-arbre différent vis-à-vis du noeud o (cf. Fig. 3.2). Le noeud o est donc *irrésolu* car de degré au moins 4. Le quadruplet q est résolu par HAQ au moment où celui-ci effectue une agglomération de deux sous-arbres connectés à o et contenant chacun une espèce de q.

De plus, aucun autre noeud $o' \neq o$ n'est connecté à quatre sous-arbres différents contenant chacun une des espèces de q (*i.e.*, la résolution d'aucun noeud o' autre que o ne conduit à résoudre q). En effet, $\forall o' \neq o$, o' appartient nécessairement à un sous-arbre $\sigma_{o'}$ connecté à o, et donc au moins trois des quatre espèces de q appartiennent au même sous-arbre que o, soit σ_o, connecté à o'. □

L'observation ci-dessus a l'implication suivante :

Lemme 7 *Les différents noeuds irrésolus $o, o', \ldots$ de la phylogénie $\hat{t}_0 = (S_0, A_0)$* concernent *chacun des ensembles de quadruplets disjoints : $\mathcal{Q}_o, \mathcal{Q}_{o'}, \ldots$, et ces ensembles sont t.q. $\mathcal{Q}_o \cup \mathcal{Q}_{o'} \cup \ldots \supseteq \mathcal{Q}(Q)$.*

On sait déjà que $\forall o, o' \in S_0,\ \mathcal{Q}_o \cap \mathcal{Q}_{o'} = \emptyset$ par l'observation précédente. De plus, tout quadruplet pour lequel Q contient un r4-arbre appartient forcément à un des ensembles $\mathcal{Q}_o, \mathcal{Q}_{o'}, \ldots$ car l'union de ceux-ci contient tout quadruplet encore non-résolu par $\hat{t}_0$ (ceci découle du fait que lorsque tous les noeuds $o, o', \ldots$ de la phylogénie ont été résolus, la phylogénie inférée, notée $\hat{t}$, est totalement résolue,

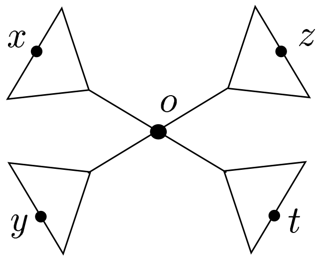


FIG. 3.2 – *La résolution du quadruplet $q = (x, y, z, t)$ dépend uniquement d'agglomérations de sous-arbres connectés au noeud o. Toute agglomération de sous-arbres connectés à un noeud $o' \neq o$ ne contribue pas à regrouper deux des quatre espèces de q dans le même sous-arbre.*

i.e., $Q_{\hat{t}}$ est complet).

Le score d'une agglomération de sous-arbres connectés à un noeud o ne dépend que des quadruplets de $\mathcal{Q}_o$ qu'elle résout, aussi le lemme précédent a pour implication qu'une agglomération effectuée autour d'un noeud o de la phylogénie ne change pas le score (et le choix) des agglomérations liées à tout autre noeud irrésolu o' de la phylogénie. En conséquence, **nous pouvons résoudre séparément les différents noeuds irrésolus de $\hat{t}_0$.**

Ceci implique que le schéma HAQ peut accepter comme phylogénie initiale $\hat{t}_0$, soit la phylogénie étoile, comme c'est le cas habituellement (cf. *Addtree*, NJ, QP_a,...), soit une phylogénie déjà partiellement résolue (contenant éventuellement plusieurs noeuds irrésolus), comme la phylogénie T^* inférée par la méthode Q^* du chapitre précédent.

Nous décrivons maintenant le traitement appliqué à chaque noeud o irrésolu de $\hat{t}_0$.

Lemme 8 *À chaque étape m du schéma HAQ, tout quadruplet q encore irrésolu dans la phylogénie influence le score de 6 agglomérations exactement.*

Preuve : Tout quadruplet irrésolu influence uniquement le score des agglomérations qui le résolvent. Comme nous l'avons vu précédemment, si $q = x, y, z, t$ est irrésolu dans la phylogénie $\hat{t}_0$, alors ses quatre espèces sont réparties dans quatre sous-arbres distincts connectés à un noeud o irrésolu, soit $\sigma_x, \sigma_y, \sigma_z, \sigma_t$ (cf. Fig. 3.2). À chaque étape m, seules les 6 agglomérations (σ_x, σ_y), (σ_x, σ_z), (σ_x, σ_t), (σ_y, σ_z), (σ_y, σ_t) et (σ_z, σ_t) regroupent 2 des 4 espèces de q, *i.e.*, résolvent q. $\square$

Comme nous l'avons dit plus haut, le score de toute agglomération (σ_i, σ_j) dépend des trois types de quadruplets que la nouvelle arête (oz) résout alors qu'ils étaient précédemment irrésolus : l'ensemble des quadruplets résolus de la même façon que dans Q, soit S_{ij}, l'ensemble des quadruplets résolus d'une façon différente que dans Q, soit C_{ij}, et l'ensemble de quadruplets résolus et pour lesquels Q ne contient aucun r4-arbre, soit U_{ij}. Dorénavant, ces ensembles seront appelés les **ensembles supports** de l'agglomération, puisque leurs cardinaux, $s_{ij} = |S_{ij}|$, $c_{ij} = |C_{ij}|$ et $u_{ij} = |U_{ij}|$, sont utilisés pour évaluer le score de (σ_i, σ_j) selon le critère l.

Soit (σ_i, σ_j) l'agglomération choisie à l'étape m, les agglomérations possibles à l'étape $m+1$, ainsi que leurs ensembles supports, et par là leur score, peuvent être obtenus facilement depuis ceux de l'étape m :

- Les agglomérations (σ_k, σ_l), t.q. $\{k, l\} \cap \{i, j\} = \emptyset$, sont toujours possibles à l'étape $m + 1$ (si o est toujours de degré > 3). Leurs ensembles supports sont exactement ceux de l'étape courante, desquels les quadruplets résolus par (σ_i, σ_j) ont été retirés.

- De plus, de nouvelles agglomérations sont possibles à l'étape $m+1$, ce sont les agglomérations (σ_z, σ_k) t.q. $k \neq \{i, j\}$, où σ_z dénote le sous-arbre résultant de l'agglomération de σ_i et σ_j. Une fois que les quadruplets résolus par (σ_i, σ_j) ont été retirés de tous les ensembles supports, les agglomérations (σ_z, σ_k) nouvellement possibles ont pour ensembles supports $S_{z,k} = S_{i,k} \cup S_{j,k}$, $C_{z,k} = C_{i,k} \cup C_{j,k}$ et $U_{z,k} = U_{i,k} \cup U_{j,k}$.

L'algorithme 7 détaille le fonctionnement du schéma *HAQ* en appliquant les propriétés précédentes. Cet algorithme prend en entrée une phylogénie $\hat{t}_0$ et un ensemble de r4-arbres Q permettant d'évaluer les différentes agglomérations candidates. Quand la phylogénie $\hat{t}_0$ est la phylogénie T^* inférée par la méthode Q^*, l'ensemble Q ne contient plus qu'une partie des r4-arbres donnés en entrée à l'algorithme IQ^*, *i.e.*, les r4-arbres non encore résolus par la phylogénie $\hat{t}_0 = T^*$.

Algorithme 7: Schéma générique *HAQ*

entrée: Une phylogénie $\hat{t}_0$ et un ensemble Q de r4-arbres t.q. $Q \cap Q_{\hat{t}_0} = \emptyset$

sortie : une phylogénie $\hat{t}$

$\hat{t} \longleftarrow \hat{t}_0$

pour chaque *noeud o irrésolu de* $\hat{t}$ **faire**

  Initialiser SA par un parcours de $\hat{t}$; Vider(M)

  /* *Initialisation des ensembles supports* S, C *et* U */

  pour chaque *quadruplet q concerné par la résolution de o* **faire**

    pour chaque *agglomération* (σ_i, σ_j) *résolvant q* **faire**

      si (σ_i, σ_j) *résout q de la même façon que Q* **alors**

        $S_{ij} \leftarrow S_{ij} \cup \{q\}$; $s_{ij} \leftarrow s_{ij} + 1$

      si (σ_i, σ_j) *résout q différemment de Q* **alors**

        $C_{ij} \leftarrow C_{ij} \cup \{q\}$; $c_{ij} \leftarrow c_{ij} + 1$

      si *Q ne résout pas q* **alors**

        $U_{ij} \leftarrow U_{ij} \cup \{q\}$; $u_{ij} \leftarrow u_{ij} + 1$

  /* *Processus agglomératif* */

  tant que $\exists$ (σ_i, σ_j) *t.q.* $g(c, u, s) > 0$ *et* $d(o) > 3$ **faire**

    Choisir (σ_i, σ_j) qui optimise $l(c_{ij}, u_{ij}, s_{ij})$

    Mettre à jour la structure de $\hat{t}$

    pour chaque *quadruplet* $q \in S_{ij} \cup C_{ij} \cup U_{ij}$ **faire**

      pour chaque (σ_k, σ_l) *influencée par q* **faire**

        Enlever q des ensembles S_{kl}, C_{kl}, U_{kl}

        et mettre à jour les compteurs s_{kl}, c_{kl}, u_{kl}

    Soit σ_z le sous-arbre résultant de (σ_i, σ_j)

    pour chaque σ_k, $k \neq i, j$ **faire**

      $S_{zk} \leftarrow S_{ik} \cup S_{jk}$, $C_{zk} \leftarrow C_{ik} \cup C_{jk}$, $U_{zk} \leftarrow U_{ik} \cup U_{jk}$

    Mettre à jour SA et $\hat{t}$ en agglomérant les sous-arbres σ_i et σ_j

Comme dans les précédents chapitres, nous supposerons que l'ensemble Q est codé sous la forme d'une matrice à quatre dimensions dont chaque entrée Q_{xyzt} t.q. $x > y, z > t, x > z$ correspond à un quadruplet $q = x, y, z, t$. Cette matrice n'est examinée que lorsque HAQ considère un nouveau noeud o de $\hat{t}_0$: les r4-arbres de Q qui peuvent être résolus par agglomération de sous-arbres connectés à o sont répartis dans des ensembles supports associés aux agglomérations concernées. Ainsi, lorsqu'une agglomération est choisie on a accès rapidement aux r4-arbres qu'elle résout [2].

Les deux structures de données principalement utilisées par HAQ sont un tableau unidimensionnel, SA, et une matrice bidimensionnelle, M (cf. Fig. 3.3).

- Le tableau SA indique pour chaque espèce le sous-arbre auquel elle appartient vis à vis du noeud o. Ce tableau est facilement mis à jour suite à une agglomération de deux sous-arbres et nous permet de connaître en $O(1)$ les sous-arbres concernés par un quadruplet en fonction des espèces x, y, z, t qu'il référence. Ainsi, suite à une agglomération, quand on enlève d'un ensemble support un r4-arbre associé au quadruplet x, y, z, t, on sait en $O(1)$ quels compteurs s_{ij}, c_{ij}, u_{ij} doivent être décrémentés.

- Dans la matrice M, chaque sous-arbre σ_i adjacent à o est associé à une colonne (soit i) et à une ligne (soit i aussi). Les entrées de M, dénotées M_{ij} correspondent respectivement aux agglomérations candidates (σ_i, σ_j) de l'étape courante. On n'utilise que la partie inférieure de M (*i.e.*, les entrées M_{ij} t.q. $i > j$) puisque $(\sigma_i, \sigma_j) \equiv (\sigma_j, \sigma_i)$. À chaque entrée M_{ij} sont associés les ensembles supports S_{ij}, C_{ij} et U_{ij}, implémentés comme des listes circulairement chaînées, pour que l'union de deux de ces ensembles (p. ex. $S_{zk} = S_{ik} \cup S_{jk}$) soit en $O(1)$. Par le lemme 8, 6 agglomérations sont influencées par le même quadruplet, aussi 6 occurrences de chaque quadruplet sont réparties dans les divers ensembles supports associés aux entrées de M. Par exemple pour le quadruplet $q = x, y, z, t$, si $xy|zt \in Q$, $SA_x = \sigma_x$, $SA_y = \sigma_y$, $SA_z = \sigma_z$, et $SA_t = \sigma_t$, alors les 6 occurrences de q sont dans les ensembles supports S_{xy}, S_{zt}, C_{xz}, C_{yt}, C_{xt} et C_{yz}. Notons que les ensembles U_{ij} ne contiennent que des occurrences de quadruplets non-résolus dans Q, *i.e.*, q t.q. $\mathcal{R}(q) \cap Q = \emptyset$. Lorsque nous constituons les ensembles supports, nous assurons aussi que les 6 occurrences d'un quadruplet sont reliées entre elles par une liste circulairement chaînée. De cette façon, quand un quadruplet est résolu par l'agglomération (σ_i, σ_j) choisie à l'étape courante, enlever des divers ensembles supports les 6 occurrences d'un même quadruplet ne demande que $O(1)$ depuis l'occurrence de ce quadruplet appartenant à un ensemble support de (σ_i, σ_j). On associe aussi à chaque entrée M_{ij} les cardinaux s_{ij}, c_{ij} et u_{ij}

2. les ensembles supports ne retiennent comme information pour un r4-arbre $xy|zt$, que le quadruplet x, y, z, t auquel il correspond. La résolution de ce quadruplet peut en effet être retrouvée en $O(1)$ par l'intermédiaire des autres structures de données.

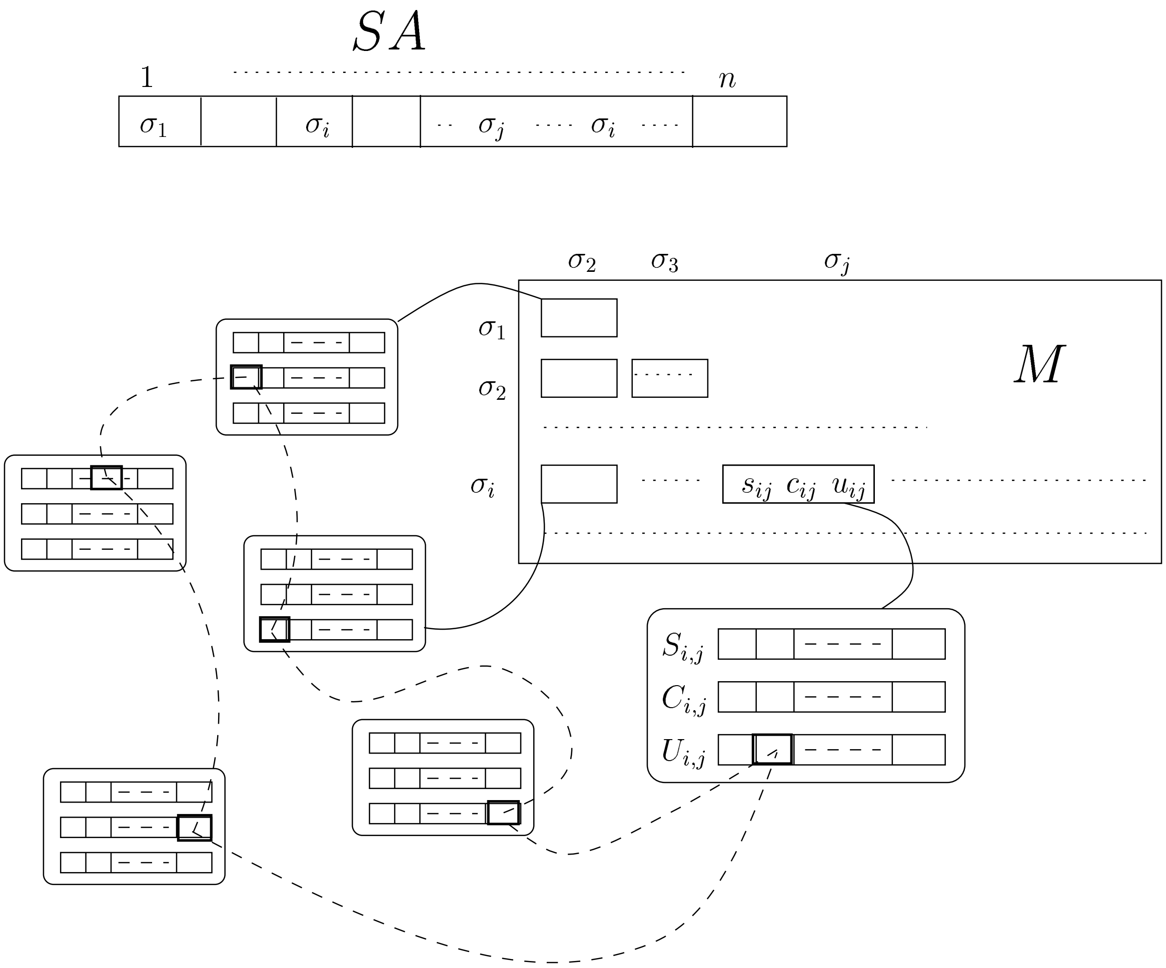


Fig. 3.3 – *Structures de données pour le schéma* HAQ *: le tableau SA indique pour chaque espèce le sous-arbre dans lequel elle est contenue vis-à-vis du noeud o ; la matrice M recense les informations propres aux agglomérations de sous-arbres possibles autour de o : à chaque entrée $M_{ij}, i < j$, correspondant à l'agglomération (σ_i, σ_j), sont associés les ensembles supports S_{ij}, C_{ij}, U_{ij} et leurs cardinaux s_{ij}, c_{ij}, u_{ij}. Les 6 occurrences d'un même quadruplet réparties dans divers ensembles supports sont reliées entre elles (cf. texte).*

des ensembles supports, afin de connaître le nombre d'éléments d'un ensemble support résultant de l'union de deux autres sans avoir à le parcourir (p. ex. $s_{zk} = s_{ik} + s_{jk}$).

En utilisant ces structures de données on peut montrer le résultat suivant :

Théorème 8 HAQ *peut être exécuté en temps* $O(n^4)$.

Preuve : On étudie en détail les opérations effectuées par l'algorithme 7, et on constate que chacune demande au plus $O(n^4)$ opérations sur l'ensemble de l'exécution :

- **ligne** `15` **:** parcourir la phylogénie $\hat{t}$ ne demande qu'un temps $O(n)$ proportionnel au nombre de ses noeuds. Au cours de ce parcours, chaque entrée du tableau SA est initialisée une et une seule fois. Connaissant *de facto* la liste des sous-arbres connectés à o, on associe chacun d'eux à une ligne et une colonne de M. Les entrées M_{ij} sont ensuite réinitialisées ($(S_{ij}, C_{ij}, U_{ij}) \leftarrow \emptyset$; $(s_{ij}, c_{ij}, u_{ij}) \leftarrow 0$). Cette étape coûte donc $O(n^2)$ à chaque fois qu'un noeud irrésolu o est considéré, et $\hat{t}_0$ peut en contenir $O(n)$. On obtient ainsi une complexité globale de $O(n^3)$ pour cette ligne.

- **ligne** `16` **:** pour obtenir chaque quadruplet q concerné par la résolution du noeud o on procède ainsi :

  - En un parcours de la phylogénie on établit l'ensemble E_i des espèces que contient chaque sous-arbre σ_i connecté au noeud o.
  - Puis on considère toute combinaison $\sigma_x, \sigma_y, \sigma_z, \sigma_t$ de quatre sous-arbres connectés à o et pour chacune on énumère toutes les combinaisons de quatre espèces (x, y, z, t) t.q. $x \in E_x$, $y \in E_y$, $z \in E_z$, $t \in E_t$. On obtient ainsi directement tous les quadruplets $q = x, y, z, t$ qui peuvent être résolus lors d'agglomérations de sous-arbres autour du noeud o.

  L'observation 1 nous assure qu'un seul noeud irrésolu o permet de résoudre tout quadruplet q, en conséquence le mécanisme précédent n'envisagera chaque quadruplet q qu'une seule fois au cours de l'algorithme. De plus, seuls des quadruplets q irrésolus sont considérés par ce mécanisme, *i.e.*, exactement $|Q| < \binom{n}{4} \in O(n^4)$ quadruplets, chacun étant obtenu en $O(1)$ depuis les ensembles E_i.

  Cette ligne coûte donc globalement $O(n^4)$ opérations, en raison du nombre $O(n^4)$ de fois où elle est exécutée.

- **boucle de la ligne** `17` **:** On sait par le lemme 8 que tout quadruplet $q = x, y, z, t$ est résolu par 6 agglomérations candidates et les entrées correspondantes dans la matrice M sont obtenues en $O(1)$ par SA_x, SA_y, SA_z, SA_t. Savoir si Q résout q ou non demande $O(1)$, de même que de savoir à quel ensemble support ajouter q. Chaque ajout se fait en $O(1)$ puisque les ensembles supports sont codés comme des listes chaînées. Les compteurs s_{ij}, c_{ij}, u_{ij} sont mis à jour en $O(1)$ conjointement aux 6 ensembles supports auxquels sont rajoutés les occurrences de q. La boucle `17` s'exécute donc en temps constant, sur l'ensemble de l'algorithme elle a donc une complexité de $O(n^4)$, proportionnelle au nombre de fois où elle est exécutée (cf. ligne précédente).

- **ligne** `18-19` **:** Ces lignes examinent le score des $O(n^2)$ agglomérations candidates. Obtenir le score de chacune pour les critères l et c demande $O(1)$. De plus, ces lignes sont exécutées un nombre proportionnel de fois au nombre d'agglomérations réalisées dans la phylogénie, *i.e.*, $O(n)$ fois. Sur l'ensemble de l'algorithme, ces lignes sont donc en $O(n^3)$.

- **ligne** `20` **:** La structure de la phylogénie est mise à jour comme décrit dans l'algorithme 6. Ceci demande un temps constant, si on connaît, par la structure de données codant la phylogénie, le noeud que référence chaque sous-arbre.

- **ligne** `21` **:** Les listes de quadruplets codant les ensembles supports S_{ij}, C_{ij}, et U_{ij} sont parcourues d'un bout à l'autre, donnant accès en $O(1)$ à chaque quadruplet qu'elles contiennent. Tout quadruplet possible est présent dans un nombre constant de ces listes supports tout au long de l'algorithme (cf. ligne `16`), et une fois éliminé, il n'est plus reconsidéré par la ligne `21`. Cette ligne demande donc $O(1)$ par quadruplet considéré, *i.e.*, $O(n^4)$ en tout, en regard du nombre de quadruplets existants.

- **ligne** `22-24` **:** Les 6 agglomérations (σ_k, σ_l) influencées par un quadruplet q sont obtenues en temps constant depuis le tableau SA ; de même on sait en temps constant comment chacune résout q et donc quel compteur s_{kl}, c_{kl} ou u_{kl} de chaque agglomération doit être décrémenté. Quand on considère une occurrence d'un quadruplet q lors du parcours des listes S_{ij}, C_{ij}, U_{ij}, enlever les 5 autres occurrences de q des divers ensembles supports auxquelles elles appartiennent ne demande que $O(1)$ car les occurrences de q sont chaînées entre elles. Donc, enlever aux diverses agglomérations l'influence d'un quadruplet q résolu par l'agglomération choisie demande un temps constant.

Les lignes 22–24 ont donc une complexité de $O(n^4)$ sur l'ensemble de l'algorithme, en raison du nombre de quadruplets que HAQ peut résoudre.

– **ligne** 25 **:** Chaque union de deux ensembles supports demande $O(1)$ car ceux-ci sont codés par des listes circulairement chaînées.

□

Signalons qu'une meilleure complexité ($O(n^3 + |Q|)$) peut être obtenue pour cet algorithme si la phylogénie $\hat{t}_0$ est la phylogénie étoile et si les quadruplets de Q sont initialement disponibles sous la forme d'une liste chaînée, et non d'une matrice (pour plus de détails, cf. [BG96b]).

L'algorithme HAQ a une moins bonne complexité que l'algorithme aggloméràtif basé sur les distances (nommé *Neighbor-Joining* [SN87]) s'exécutant lui en $O(n^3)$ [SK88], mais il possède cependant la meilleure complexité existante pour un algorithme basé sur les quadruplets, à égalité avec les algorithmes IQ^* et *Addtree* ([Att97a]).

La complexité obtenue pour HAQ fait que, si nous l'employons à la suite de l'algorithme IQ^* pour compléter la phylogénie T^*, nous n'augmentons pas la borne de complexité de la méthode de reconstruction, toujours en $O(n^4)$.

3.3.1 Cas particuliers : *Addtree* et QP_a

La variante QP_a de l'algorithme $QuartetPuzzle$ [SvH96] définie en introduction correspond à l'instanciation de HAQ pour laquelle $l = min(c_{ij})$ et g est choisi de façon à proposer une phylogénie complètement résolue (p. ex. $g = cte \geq 0$).

Addtree n'est pas un cas particulier de HAQ. C'est en fait un algorithme mixte : même s'il optimise le critère de quadruplets $l = max(s_{ij})$ sur l'ensemble Q à chaque étape, le critère global qu'il cherche à optimiser est le critère de distances ME sur une matrice de données $\hat{D}$. De plus, à chaque étape, un nouvel ensemble de r4-arbres Q est obtenu par le principe FPM depuis la matrice $\hat{D}$. Cependant, $\hat{D}$ est transformée d'étape en étape selon un principe de distance. Ceci induit des transformations non-uniformes sur l'ensemble Q de r4-arbres considérés d'une étape sur l'autre, d'autant plus que chaque sous-arbre connecté au noeud irrésolu o est considéré comme une espèce *virtuelle*, indépendamment de l'identité et du nombre des espèces qu'il contient. Les r4-arbres inférés dans Q et faisant référence à une telle espèce *virtuelle* sont en quelque sorte une moyenne des r4-arbres qui seraient inférés si on considérait les espèces *réelles* contenues dans ce sous-arbre. Dans la version initiale de *Addtree*, dite "*weighted*", chaque espèce (virtuelle aussi bien que réelle) connectée à o se voit attribuer la même importance, indépendamment du nombre d'espèces qu'elle représente effectivement. Les r espèces réelles appartenant à un sous-arbre connecté à o par une espèce virtuelle se voient donc attribuer un poids implicite de $1/r$. Pour remédier à ce phénomène attribuant des poids inégaux et aléatoires (car dépendant de l'ordre des agglomérations effectuées précédemment) aux espèces réelles pour l'inférence de Q, une version "*unweighted*" de *Addtree* ($Addtree_u$) a été proposée [BL84, BD86] : elle attribue à chaque espèce virtuelle un poids proportionnel au nombre d'espèces réelles qu'elle représente. Toutefois, si l'ensemble Q initialement considéré est incomplet, *i.e.*, si chaque espèce réelle ne participe pas au même nombre de r4-arbres, les r4-arbres inférés à chaque étape par $Addtree_u$ ne représentent qu'une moyenne des r4-arbres entre les espèces réelles ; ils sont par là une approximation de l'ensemble de r4-arbres initiaux et à chaque étape un ensemble Q, de plus en plus déformé par rapport à l'ensemble Q initial, est considéré pour choisir les agglomérations.

Tout ceci fait que *Addtree* (ou $Addtree_u$) n'est pas une variante de HAQ. En fait, HAQ, avec $l = max(s_{ij})$, peut être vu en quelque sorte comme une version corrigée de *Addtree*, puisqu'il travaille toujours depuis l'ensemble Q originel (toutefois, rappelons que *Addtree* a été initialement proposé pour construire un arbre depuis une matrice de distances, et non depuis un ensemble de r4-arbres). Paradoxalement, *Addtree* retrouve la phylogénie correcte quand l'ensemble Q de départ est arboré (et la transformation de Colonius et Schulze [CS79] est employée) [BD86], alors que sa version corrigée, HAQ avec $l = max(s_{ij})$, ne possède pas cette propriété (sauf si Q est complet) ! Pour que HAQ retrouve la phylogénie correcte depuis des données arborées, il faut employer pour l un certain type de critères, dont $max(s_{ij})$ ne fait pas partie. Nous montrons ci-dessous quelle gamme de critères il faut employer.

3.3.2 Consistance et vitesse de convergence de *HAQ*

Cas de données arborées

Les critère g et l, surtout, sont particulièrement importants pour la consistance de HAQ. De leur choix dépend le fait que HAQ puisse retrouver ou non la phylogénie correspondant aux données quand celles-ci sont arborées. C'est une propriété importante pour une méthode de reconstruction, non pas parce qu'on traite souvent des ensembles de r4-arbres arborés, mais parce que ceci montre le bien fondé de la méthode. Pour que HAQ respecte cette propriété, il faut porter un soin tout particulier aux quadruplets contredits. Dans le cas où les données désignent une phylogénie t totalement résolue, on a la propriété suivante :

Propriété 7 *Soit t une phylogénie totalement résolue, et $\hat{t}_0$ toute phylogénie t.q. $B(\hat{t}_0) \subseteq B(t)$,* HAQ *retrouve la phylogénie t depuis $Q = Q_t - Q_{\hat{t}_0}$ si l est n'importe quel critère choisissant toujours une agglomération (σ_i, σ_j) ne contredisant aucun r4-arbre (i.e., t.q. $c_{ij} = 0$) et g n'interrompt pas la construction de $\hat{t}$ tant que $\exists i, j$ t.q. $s_{ij} > 0$.*

Preuve : Remarquons que Q_t est complet puisque t est totalement résolu. Ceci signifie qu'à chaque étape, tout quadruplet q encore irrésolu par la phylogénie $\hat{t}$ est t.q. $\exists r \in \mathcal{R}(q) \cap Q$, donc $\forall (\sigma_i, \sigma_j)$ on a $u_{ij} = 0$.

Toute phylogénie est désignée de façon unique par l'ensemble de bipartitions induites sur les espèces par ses arêtes [Bun71], ainsi pour avoir $\hat{t} = t$ après la dernière étape de *HAQ*, il faut que $B(\hat{t}) = B(t)$. Comme initialement $B(\hat{t}_0) \subseteq B(t)$, il faut donc montrer qu'à chaque étape, on choisit une agglomération (σ_i, σ_j) ajoutant à $\hat{t}$ une arête $e_{ij} = \sigma | \overline{\sigma}$ t.q. $\{\sigma | \overline{\sigma}\} \in B(t)$.

À chaque étape i, on peut classer les agglomérations (σ_i, σ_j) candidates en deux types, selon l'arête e_{ij} qu'elles induisent respectivement :

(1) $e_{ij} = \sigma | \overline{\sigma}$ t.q. $\{\sigma | \overline{\sigma}\} \in B(t)$, ce qui signifie, de par la composition de Q, que $s_{ij} > 0$ et $c_{ij} = 0$;

(2) $e_{ij} = \sigma | \overline{\sigma}$ t.q. $\{\sigma | \overline{\sigma}\} \notin B(t)$, ce qui signifie que e_{ij} résout au moins un r4-arbre différemment de $Q \subseteq Q_t$, donc $c_{ij} > 0$.

Ainsi :

> (a) à l'étape i, tout critère l, qui choisit une agglomération (σ_i, σ_j) t.q. $c_{ij} = 0$, induit dans la phylogénie $\hat{t}$ une arête de type (1) (*i.e.*, ajoute à $B(\hat{t})$ une bipartition de $B(t)$), s'il existe une agglomération correspondant à une telle arête ;
>
> (b) or tant que $B(\hat{t}) \neq B(t)$, il existe une telle agglomération si $B(\hat{t}) \subset B(t)$, ce qui est vérifié à l'étape initiale et lors des étapes suivantes en raison du point (a).

Les preuves des points (a) et (b) sont interdépendantes et sont résolues par induction croisée. □

Si t est partiellement résolue (ce qui n'est toutefois jamais supposé en R.P.), *HAQ* ne le retrouve pas forcément depuis $Q = Q_t$, même avec un critère l respectant la règle ci-dessus. Par contre, on peut concevoir une version légèrement différente de *HAQ*, de façon similaire à la variante de *Addtree* qui permet de s'accomoder de données arborées mais correspondant à une phylogénie partiellement résolue [BD86, BL84]. Dans cette version, on peut agglomérer ensemble plus de deux sous-arbres à la fois. À l'étape m, si (σ_i, σ_j) est une agglomération qui possède le meilleur score pour l, on note $\sigma_i \sim \sigma_j$. Et à chaque étape on choisit d'agglomérer ensemble tous les sous-arbres appartenant à une même composante connexe de la fermeture transitive de la relation $\sim$. On peut noter au passage, qu'agglomérer plus de deux sous-arbres, crée un nouveau noeud irrésolu dans la phylogénie, dont la résolution devra être considérée postérieurement par l'algorithme. La création de nouveaux noeuds irrésolus n'empêche pas la terminaison de l'algorithme car la somme S_d des degrés des noeuds de la phylogénie $\hat{t}$ est augmentée à chaque étape, et toute phylogénie sur n espèces peut avoir au maximum une somme $S_d = 4n - 9$, signifiant que $\hat{t}$ est totalement résolue, auquel cas l'algorithme s'arrête. La preuve que cette variante de HAQ permet de retrouver la phylogénie correcte, quand les données sont arborées et correspondent à une phylogénie partiellement résolue, est similaire à celle donnée par Bandelt et Dress [BD86] pour la variante correspondante de $Addtree$.

Cas de données bruitées

On peut conduire la même analyse que celle menée au chapitre précédent pour la méthode Q^* : si les r4-arbres sont inférés par le principe FPM et que celui-ci ne se trompe sur aucun r4-arbre, alors Q est complet et arboré, si bien que par la propriété 7, le schéma HAQ retrouve la phylogénie correcte. Ainsi, le schéma HAQ possède-t-il une vitesse de convergence au moins aussi bonne que celle de la méthode Q^*, ainsi qu'une boule d'exactitude garantie aussi large que possible au sens de la norme L_∞, pour retrouver la totalité de la phylogénie correcte. En revanche, dans le cas où certains r4-arbres ne sont pas reconstruits correctement, HAQ ne possède aucune garantie démontrée pour retrouver certaines arêtes (même les plus longues). Ceci résulte du fait que, contrairement à la méthode Q^*, l'ordre des modifications effectuées au cours des étapes sur la phylogénie (ici par l'intermédiaire des agglomérations) a une importance sur la phylogénie finalement proposée.

3.4 Une mesure de confiance sur les arêtes inférées par HAQ

On peut très facilement classer les arêtes inférées par HAQ par ordre de confiance en se basant sur les quadruplets résolus par chacune. Cet ordre n'est pas forcément celui dans lequel elles ont été inférées. HAQ ne regroupe que deux sous-arbres à chaque étape, aussi certaines arêtes (*i.e.*, bipartitions) n'ont pu être inférées qu'à partir du moment où un certain nombre d'agglomérations avait déjà été fait.

Par exemple, une indication qui peut sembler pertinente est la valeur du rapport nombre de r4-arbres de Q satisfaits sur nombre de r4-arbres contredits. Contrairement à l'algorithme glouton, évaluant les agglomérations candidates sur le nombre de r4-arbres de Q non-encore résolus, afin qu'on obtienne la meilleure augmentation du score de la phylogénie au sens de g à chaque étape, chaque arête est ici évaluée sur le nombre total de r4-arbres qu'elle résout, indépendamment du fait que d'autres arêtes résolvent les mêmes r4-arbres. On s'intéresse ici à la valeur intrinsèque d'une arête, non à la valeur ajoutée qu'elle représente pour la phylogénie.

Une deuxième indication de confiance pour une arête est le pourcentage de r4-arbres qu'elle résout et qui ne sont pas dans Q. On obtient ainsi un *écart* aux conditions pour que l'arête appartienne à la phylogénie T^*, pour laquelle on connaît quelques propriétés statistiques (cf. chap. précédent).

La Fig. 3.4 donne un exemple de ces mesures sur un jeu de données décrivant des bactéries [HPT97]. Sur la base des mesures indiquées pour les deux arêtes de la phylogénie dues à HAQ, on peut raisonnablement penser que l'arête regroupant (Viridoch CS, Methyloc Ca, Azosp Bra) est correcte (elle satisfait 16 fois plus de r4-arbres de Q qu'elle n'en contredit), tandis que le regroupement (Methyloc Ca, Azosp Bras) semble incertain (un tiers des r4-arbres induits par cette arête ne sont pas dans Q).

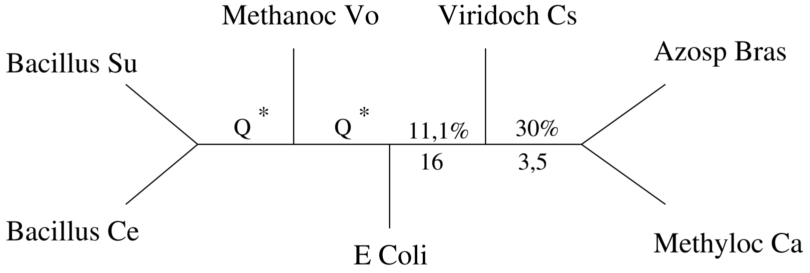


FIG. 3.4 – *Phylogénie obtenue par l'algorithme* HAQ*, exécuté sur la base de la phylogénie T^* (arêtes étiquetées Q^*), inférée par la méthode Q^*. Pour chacune des arêtes inférées par* HAQ*, sont indiqués, au dessus de l'arête, l'écart aux conditions pour que l'arête soit inférée par la méthode Q^*, et, sous l'arête, le rapport entre le nombre de r4-arbre qu'elle satisfait et le nombre de r4-arbres qu'elle contredit.*

3.5 *Addquad*: une instanciation de ***HAQ*** spécifique à la R.P.

Nous rejoignons maintenant les préoccupations de la R.P.: on suppose qu'il existe une phylogénie t ayant servi de support à l'évolution des espèces étudiées, et on cherche à proposer un estimateur le plus proche de cette phylogénie, depuis des données dont la composition dépend de t. La phylogénie inférée par le schéma de HAQ sur la base des données est vue comme une estimation $\hat{t}$ possible de la phylogénie inconnue t. Cependant, comment mesurer l'erreur d'estimation? On ne peut obtenir une expression analytique de l'erreur d'estimation, car on ne connaît pas Q_t. En revanche, on peut se baser sur Q qui peut être vu comme une estimation de Q_t et obtenir une estimation de l'erreur de $\hat{t}$, notée $\hat{e}(\hat{t})$, en comparant $Q_{\hat{t}}$ à $Q = \hat{Q}_t$. Comme nous l'avons fait dans la partie II de cette thèse, nous pouvons ici aussi décomposer $\hat{e}(\hat{t})$ en estimation de l'erreur de 1ère espèce, notée $\hat{e}_1(\hat{t})$, et d'erreur de 2ème espèce, notée $\hat{e}_2(\hat{t})$. L'erreur de 1ère espèce est estimée par le nombre de r4-arbres induits par la phylogénie mais non présents dans l'ensemble Q:

$$\hat{e}_1(\hat{t}) = \left|Q_{\hat{t}} - Q\right| = c + u \ .$$

De façon symétrique, l'erreur de 2ème espèce est estimée par le nombre de r4-arbres de Q que la phylogénie n'induit pas:

$$\hat{e}_2(\hat{t}) = \left|Q - Q_{\hat{t}}\right| = |Q| - s \ .$$

Notons que durant la construction de la phylogénie $\hat{t}$, chaque arête ajoutée à la phylogénie ne peut qu'augmenter (ou laisser inchangé) $\hat{e}_1(\hat{t})$, et ne peut que diminuer (ou laisser inchangé) $\hat{e}_2(\hat{t})$. Dans le cadre du schéma HAQ, nous attribuerons comme précédemment un poids identique à ces deux termes d'erreur. Nous essayerons donc de minimiser le critère global:

$$g \ = \ \hat{e}_1(\hat{t}) + \hat{e}_2(\hat{t}) \equiv c + u - s \ . \tag{3.4}$$

Si on note $\hat{e}_1^m$, resp. $\hat{e}_2^m$, les valeurs de $\hat{e}_1(\hat{t})$, resp. $\hat{e}_2(\hat{t})$ à la fin de l'étape m du processus agglomératif, nous avons:

$$(\hat{e}_1^m + \hat{e}_2^m) - (\hat{e}_1^{m-1} + \hat{e}_2^{m-1}) \ = \ c_{ij}^m + u_{ij}^m - s_{ij}^m \ , \tag{3.5}$$

où c_{ij}^m, u_{ij}^m et s_{ij}^m dénotent les valeurs associées à l'agglomération (σ_i, σ_j) choisie à l'étape m. Le schéma HAQ effectue des agglomérations tant que la valeur de la phylogénie $\hat{t}$ pour g peut augmenter. D'après (3.4) et (3.5), ceci est vérifié tant que

$\exists i, j$ t.q. $c_{ij}^m + u_{ij}^m - s_{ij}^m > 0$.

Pour choisir un critère local l, on peut aussi s'inspirer de l'équation (3.5), indiquant que minimiser $c_{ij}^m + u_{ij}^m - s_{ij}^m$ revient à minimiser $\hat{e}_1^m + \hat{e}_2^m$. Ainsi nous pourrions fixer $l = c_{ij} + u_{ij} - s_{ij}$ pour évaluer chaque agglomération. Toutefois nous n'emploierons pas ce critère car il ne garantit pas à HAQ de retrouver la phylogénie correcte, quand les données sont arborées ($Q = Q_t$). Nous avons vu précédemment que pour que HAQ retrouve la phylogénie correcte, l doit avant tout éviter de contredire des r4-arbres de Q (propriété 7). Divers critères l peuvent alors être choisis en fonction de ces remarques. Nous choisissons ici l'ordre lexicographique sur $\hat{e}_1^m$ puis $\hat{e}_2^m$, noté $lex(\hat{e}_1^m, \hat{e}_2^m)$, *i.e.*,

$$l \;=\; lex(c_{ij} + u_{ij}\,,\, -s_{ij})\;, \tag{3.6}$$

l'agglomération minimisant l étant choisie à chaque étape.

Exemple 3 *À l'étape m si le schéma* HAQ *a le choix entre les agglomérations (σ_x, σ_y), (σ_x, σ_z) et (σ_y, σ_z), sachant que $(s_{xy} = 20, c_{xy} = 3, u_{xy} = 1)$, $(s_{xz} = 2, c_{xz} = 0, u_{xz} = 1)$, et $(s_{yz} = 1, c_{yz} = 1, u_{yz} = 0)$, nous avons :*

$$(\sigma_x, \sigma_z) <_l (\sigma_y, \sigma_z) <_l (\sigma_x, \sigma_y)\ .$$

L'agglomération choisie par le critère l est donc (σ_x, σ_z).

Le critère l décrit ci-dessus place les r4-arbres c_{ij} et u_{ij} au même niveau et on peut se demander si cela ne contribue pas à le faire sortir du champ d'application de la propriété 7. Toutefois, on peut facilement montrer que ce n'est pas le cas, puisque quand $Q = Q_t$ et t est totalement résolue (comme supposé pour la phylogénie des espèces), alors $\forall i, j$, on a $u_{ij} = 0$. Dans un tel cas, l est donc réduit à l'expression $lex(c_{ij}, -s_{ij})$ et HAQ retrouve ainsi t par la propriété 7.

Remarquons que les critères décrits ci-dessus sont des extensions naturelles des critères implicitement optimisés par la méthode Q^* du chap. 2. Celle-ci produit en effet la phylogénie T^* tel que $\hat{e}_2(T^*)$ est minimisé tout en assurant que $\hat{e}_1(T^*) = 0$. En ce sens, le schéma HAQ constitue une extension naturelle de la méthode Q^*.

3.6 Résultats expérimentaux

Nous utilisons le schéma HAQ pour étudier différents problèmes de reconstruction de phylogénies. Dans un premier temps, nous étudions empiriquement la difficulté des problèmes lié à l'arbre-consistance d'un ensemble de r4-arbres.

Puis nous rejoignons le cadre de la R.P. où nous mesurons l'aptitude de la version *Addquad* du schéma HAQ pour compléter la phylogénie T^*, inférée par la méthode Q^*. Nous terminons en testant *Addquad* sur l'exemple des 11 mammifères.

3.6.1 Problèmes AC et $SMAC$

La propriété 7 indique que pour une certaine gamme de critères l, l'algorithme HAQ retrouve la phylogénie correcte depuis des données arborées. Toutefois, si à un ensemble arboré on enlève certains r4-arbres, on obtient un ensemble arbre-consistant, et le problème de savoir si un ensemble de r4-arbres est arbre-consistant est NP-complet (cf. chap. 1). On peut donc naturellement penser qu'il est difficile pour un algorithme de reconstruction d'inférer une phylogénie satisfaisant tous les r4-arbres de Q (même si une telle phylogénie existe), et ce d'autant plus que l'ensemble Q s'éloigne d'un ensemble arboré, *i.e.*, d'autant plus que Q est incomplet.

Pour tester la difficulté pratique du problème d'arbre-consistance, noté ACR au chap. 1, on génère un certain nombre d'instances positives, *i.e.*, arbre-consistantes, en enlevant de plus en plus de r4-arbres d'un ensemble initialement arboré. On s'intéresse alors au pourcentage de fois où HAQ retrouve une phylogénie consistante avec les données, *i.e.*, donne la bonne réponse au problème ACR[3].

Plus précisément, pour chaque jeu de données, on part d'une phylogénie t totalement résolue sur 15 espèces. Soit Q_t l'ensemble de r4-arbres qu'il induit, l'ensemble Q, donné en entrée aux algorithmes, est obtenu depuis Q_t duquel a été enlevée une proportion p_e $(0 \leq p_e \leq 1)$ de r4-arbres. On s'intéresse alors au comportement de HAQ en fonction de la proportion p_e de r4-arbres de l'ensemble initial et qui ont été "perdus" (différentes valeurs entre 5% et 95% sont essayées pour p_e). Le critère l de HAQ est choisi de façon à ne contredire aucun r4-arbre de Q, p. ex. $l = lex(c_{ij} + u_{ij}, -s_{ij})$, qu'on cherche à minimiser. On choisit un critère global $g = max(s)$, poussant HAQ à résoudre la phylogénie construite tant qu'on peut lui ajouter des arêtes ne contredisant aucun r4-arbre de Q. La table 3.1 présente les résultats moyennés sur 200 jeux de données).

Ces résultats montrent que même si la théorie classe le problème ACR dans les problèmes NP-complets, en pratique il semble qu'un algorithme raisonnable soit capable de résoudre ce problème avec un bon degré de confiance, même pour des valeurs de p_e très élevées.

---

3. Les algorithmes de reconstruction ne peuvent se tromper que sur les instances positives de ACR. Pour les instances négatives, *i.e.*, pour lesquelles il n'existe aucune phylogénie satisfaisant tous les r4-arbres de Q, les algorithmes ne peuvent trouver qu'une phylogénie contredisant au moins certains r4-arbres, et répondre que l'ensemble n'est pas arbre-consistant, ce qui est correct.

p_e	5%	10%	50%	70%	80%	90%	95%
HAQ	100%	100%	100%	100%	100%	92.5%	84%

TAB. 3.1 – *Proportion des ensembles Q détectés arbre-consistants en fonction du pourcentage p_e de r4-arbres manquant vis-à-vis d'un ensemble arboré complet, Q_t.*

De ces bons résultats pour le problème ACR, on peut espérer que les algorithmes de reconstruction étudiés ci-dessus soient efficaces pour le problème d'optimisation associé, noté $SMAC$ au chap. 1, consistant à trouver le sous-ensemble arbre-consistant maximum d'un ensemble quelconque Q de r4-arbres.

Pour étudier le problème $SMAC$, on peut générer des jeux de données de la même façon que ci-dessus pour le problème ACR, à la différence qu'une composante de bruit est maintenant ajoutée aux données : partant d'un ensemble Q_t complet arboré, on élimine une proportion p_e des r4-arbres, puis on *inverse* une proportion p_i des r4-arbres restants. *Inverser* un r4-arbre signifie le remplacer par une des deux autres résolutions possible pour le quadruplet concerné. Par exemple, $xz|yt, xt|yz$ sont les inversions possibles du r4-arbre $xy|zt$. Quand on inverse un r4-arbre, on choisit au hasard des deux r4-arbres possibles celui qui le remplace. Si p_i reste suffisamment faible, il est vraisemblable que la phylogénie t reste la phylogénie qui satisfait un maximum de r4-arbres de Q. Sur cette idée, on obtient une estimation $\hat{n}_{opt}$ du nombre maximum, n_{opt}, de r4-arbres de Q pouvant être satisfaits : $\hat{n}_{opt} = (1 - p_e)(1 - p_i)\binom{n}{4}$, correspondant au nombre de r4-arbres de Q_t conservés dans Q. Ainsi quand p_i est faible, $\hat{n}_{opt}$ est une borne inférieure précise de n_{opt}, et quand p_i est importante, la borne devient lâche car il est possible qu'un plus grand sous-ensemble de Q soit satisfiable en une phylogénie autre que t. Une telle situation est susceptible d'être obtenue quand une bonne proportion des quadruplets inversés sont mutuellement *compatibles* (arbre-consistants). La probabilité d'un tel événement reste toutefois faible car les r4-arbres inversés sont choisis totalement au hasard, ainsi que la résolution alternative choisie pour leur quadruplet.

La table 3.2 montre les résultats obtenus pour différentes valeurs de p_e et p_i, en terme de $n_s/\hat{n}_{opt}$, où n_s est le nombre de r4-arbres de Q satisfaits par la phylogénie inférée. Cette valeur est calculée indépendamment sur chaque jeu de données, les résultats de la table 3.2 étant moyennés sur 300 jeux de données.

	$p_e = 50\%$		$p_e = 75\%$	
	$p_i = 10\%$	$p_i = 33\%$	$p_i = 10\%$	$p_i = 33\%$
HAQ	100%	99%	99.8%	98.3%

TAB. 3.2 – *Efficacité d'un algorithme polynômial pour résoudre le problème SMAC, depuis des jeux de données Q complets arborés dont une proportion p_e de r4-arbres a été enlevée et une proportion p_i des r4-arbres restants a été inversée.*

Si on admet que $\hat{n}_{opt}$ est une borne inférieure proche de n_{opt}, les résultats précédents montrent que le problème $SMAC$ peut être résolu en pratique de façon quasi-optimale par des algorithmes polynômiaux, même pour des instances relativement difficiles. Notons que pour un niveau de bruit élevé ($p_i = 0.5$), dans la plupart des cas, une faible proportion ($< 1\%$) des r4-arbres satisfaits par les phylogénies inférées sont des r4-arbres inversés. Comme nous l'avons supposé plus tôt, la composante de bruit conduit parfois les algorithmes de reconstruction à inférer des arêtes différentes de celles de la phylogénie t depuis laquelle Q a été généré.

Des expériences complémentaires montrent que le nombre d'espèces considérées (ici 15) n'a pas d'influence sur les résultats obtenus.

3.6.2 Application à la R.P.

Nous en venons ici à la motivation principale nous ayant conduit au schéma HAQ: compléter la phylogénie T^* dans le contexte de la R.P., pour obtenir de meilleurs estimateurs de la phylogénie inconnue. On s'intéresse donc ici à l'instanciation de HAQ dédiée à la R.P. : *Addquad*.

Au cours de simulations suivant en tout point le schéma décrit dans les chapitres précédents (p. ex. cf. section 1.4 chap. 1 part. II et Annexes C,E), les performances de *Addquad* sont évaluées selon que l'algorithme utilise pour point de départ la phylogénie étoile (on note alors $Addquad(\emptyset)$) ou la phylogénie T^* (notée $Addquad(T^*)$). Dans un deuxième temps nous comparons $Addquad(T^*)$ à la méthode NJ [SN87, SK88]. Les simulations effectuées incluaient aussi *Addtree* [ST77] et une heuristique pour le problème MP [Fel93], toutefois les résultats de ces deux méthodes ne sont pas détaillés car ils sont relativement proches des résultats présentés pour NJ.

Une analyse détaillée des résultats est présente en Annexe F, dont nous nous contentons ici de résumer rapidement les traits marquants.

Commencer la reconstruction depuis la base saine inférée par la méthode Q^*, et non depuis la phylogénie étoile, améliore dans tous les cas les résultats de *AddQuad*, *i.e.*, $Addquad(T^*)$ est toujours meilleur en moyenne que $Addquad(\emptyset)$.

De plus, dans tous les cas, compléterla phylogénie T^* inférée par la méthode Q^* par l'algorithme *AddQuad* donne un meilleur estimateur de la phylogénie t des espèces, et dans tous les cas sauf un, on obtient aussi un meilleur estimateur de la phylogénie *réalisée* (cf. chap. précédent).

Le degré de résolution de la phylogénie inférée par $Addquad(T^*)$ se situe toujours entre celui des phylogénies inférées par Q^* et $Addquad(\emptyset)$. La méthode $Addquad(T^*)$ a un comportement très sain vis-à-vis de la quantité d'information présente dans les données, au sens où elle propose toujours a peu près autant d'arêtes que le nombre d'arêtes pour lesquelles le jeu de données contient de l'information (*i.e.*, $|t_R|$). En conséquence, $Addquad(T^*)$ propose, dans presque tous les cas, des estimations plus proches de la phylogénie des espèces (t_R ou t) que les estimations proposées par la méthode NJ.

3.6.3 Application au jeu de données sur les mammifères

La Fig. 3.5 montre la phylogénie inférée par $Addquad(Q^*)$ sur le jeu de données regroupant 11 séquences de mammifères de 166 caractères.

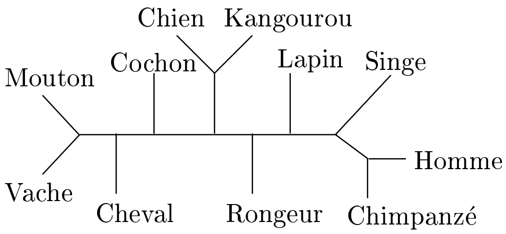


FIG. 3.5 – *La méthode $Addquad(Q^*)$ appliquée au jeu de données sur 11 mammifères.*

Exécuter l'algorithme *Addquad* à la suite de la méthode Q^* ajoute les groupes (mouton, vache, cheval), (lapin, rongeur, primates) et (lapin, primates). Selon [PHS91], le premier groupe est incorrect et les deux autres corrects. Notons que le dernier groupe était incorrectement résolu par l'arbre bootstrap réduit au seuil optimum (cf. chap. 1, part. *II*)

3.7 Conclusion

Nous avons présenté dans ce chapitre un schéma agglomératif basé sur les quadruplets. Ce schéma possède plusieurs qualités :

- *HAQ* semble être le premier algorithme à avoir été décrit de façon à pouvoir accepter une phylogénie déjà résolue en entrée, même si cette propriété doit pouvoir aisément être étendue à divers autres algorithmes usuels (cf. p. ex. le programme QR^2 de Gascuel et Levy [GL96] doté d'un mécanisme additionnel permettant d'indiquer certaines contraintes structurelles). Cette propriété permet de répondre à certaines attentes des biologistes [Roo95] : certaines parties des phylogénies qu'on cherche à reconstruire sont déjà connues de façon sûre. Aussi quand on reconstruit la phylogénie, on ne veut pas que la méthode de reconstruction remette en question ces connaissances partielles, on souhaite juste qu'elle place au mieux les autres espèces dans la phylogénie. Les quadruplets sont le formalisme idéal pour imposer des contraintes sur des parties de la phylogénie (puisque les r4-arbres peuvent être vus comme des contraintes structurelles) [GL96]. Le schéma *HAQ* permet en plus d'imposer initialement des bipartitions entre différents groupes d'espèces (par la phylogénie $\hat{t}_0$).

- Le schéma *HAQ* est adapté au traitement d'ensembles de quadruplets incomplets, au sens où il tient compte indépendamment de chaque r4-arbre de Q, contrairement à *Addtree*.

- La complexité $O(n^4)$ de *HAQ* semble raisonnable : les algorithmes agglomératifs basés sur les distances sont en $O(n^3)$, cependant ici on doit considérer $O(n^4)$ quadruplets dans le pire des cas.

- Le schéma *HAQ* constitue une façon efficace de compléter la phylogénie inférée par la méthode Q^* du chapitre précédent. Combiner *HAQ* et Q^* permet d'obtenir de meilleurs estimations structurelles de la phylogénie inconnue que celles obtenues par des méthodes de reconstruction usuelles.

Conclusion

Au cours de cette thèse, nous avons proposé plusieurs nouvelles méthodes de reconstruction phylogénétique, en mettant systématiquement l'accent sur la stabilité des phylogénies proposés pour estimer les phylogénies étudiées, c'est-à-dire sur le fait que les arêtes inférées ne soient pas spécifiques du jeu de données dont on dispose, mais traduisent effectivement l'histoire évolutive des espèces.

Les méthodes présentées procèdent soit d'une approche à résolution décroissante (partant d'une phylogénie complète dont sont enlevées les arêtes non-fiables), soit d'une approche à résolution croissante (inférant d'emblée une phylogénie ne comportant que des arêtes fiables). Les simulations réalisées ont permis de valider ces deux approches, en montrant notamment que les méthodes proposées ici produisent en moyenne de meilleures estimations d'une phylogénie que les méthodes traditionnelles.

En mettant en oeuvre ces méthodes, nous avons pris garde à ce qu'elles soient réalisables en un temps polynômial, aussi faible que possible, afin de pouvoir être utilisées en pratique par les biologistes. Dans le cas des méthodes à résolution croissante, nous avons aussi pu caractériser les conditions où ces méthodes retrouvent avec une probablité de 1 la phylogénie correcte, ainsi que leur vitesse de convergence (*i.e.*, la taille des données requise pour que la méthode retrouve la phylogénie correcte avec la probablité souhaitée). Suite à ces bons résultats, un package regroupant la plupart des méthodes présentées ici a été réalisé. Ce package est accessible via *internet*, comme il est de coutume dans la communauté R.P.

Les travaux menés dans cette thèse peuvent être poursuivis dans plusieurs directions possibles.

Premièrement, on peut étendre la méthode du bootstrap du chapitre 2 de la partie II, afin d'estimer d'autres paramètres que le nombre d'arêtes que doit posséder la phylogénie inférée. Des paramètres plus fins, comme la valeur d'une arête pour le critère de reconstruction, peuvent sûrement conduire à des résultats encore meilleurs.

En ce qui concerne les méthodes de quadruplets, la méthode Q^* du chapitre 2 de la partie III, peut être étendue en allégeant la contrainte imposée pour chaque arête de la phylogénie T^* (*i.e.*, pour chaque bipartition de B^*), dans le but d'obtenir une phylogénie plus résolue. Un premier pas dans ce sens a été effectué par Moulton et Steel [MS96] qui obtiennent un raffinement de la phylogénie T^* en ne s'intéressant qu'à $n-3$ des r4-arbres induits par une bipartition. Cependant leur méthode ne dispose pour l'instant que d'un algorithme exponentiel. Il reste donc à montrer que cette méthode est NP-difficile ou bien à exhiber un algorithme polynômial.

On peut aussi s'intéresser au cas d'ensembles de r4-arbres incomplets. Pour de tels ensembles, il existe des règles d'inférence permettant depuis plusieurs r4-arbres d'inférer certains r4-arbres manquants [Ste92, ESSW97a]. De telles règles s'apparentent fortement aux algorithmes d'arc-consistance et de chemin-consistance rencontrés dans le domaine des CSP (problèmes de satisfaction de contraintes). Des connections effectuées avec ce domaine peut sûrement émerger un algorithme exact de recherche du sous-ensemble arbre-consistant maximum d'un ensemble de r4-arbres (problème NP-difficile), selon le principe d'essais-erreurs bien connu dans ce domaine (aussi connu sous le nom d'algorithme de *backtrack* ou de *branch-and-bound*).

Enfin le schéma d'heuristiques agglomératives HAQ, présenté au chapitre 3 de la partie III, est idéal pour imposer à la phylogénie inférée un certain nombre de contraintes structurelles, issues de la connaissance du biologiste sur la phylogénie considérée. Poursuivant cette voie, on peut encore améliorer l'applicabilité de cet algorithme en lui permettant d'accepter des contraintes structurelles plus complexes (clauses exclusives, disjonctives, négatives), ou même permettre que chaque quadruplet ait un poids différent. On peut aussi utiliser le schéma HAQ comme point de départ pour la conception d'algorithmes ayant une borne d'approximation garantie pour le problème du sous-ensemble maximum de r4-arbres arbre-consistant. Ce type de résultat, pourtant intéressants, a été relativement peu étudié pour le momenten reconstruction phylogénétique (à l'exception de [ABF$^+$96, War93]).

Comme on peut s'en rendre compte à l'issue de cette thèse, les pistes d'investigation possibles fleurissent à merveille dans le domaine de la reconstruction phylogénétique, à l'image de la bioinformatique en général. Actuellement, ce domaine de recherche est le théâtre d'un foisonnement intense de travaux de recherche, mêlant biologistes, mathématiciens, et informaticiens. En informatique, cet engouement est nettement visible par l'élargissement croissant de la partie consacrée aux problèmes liés à la bioinformatique dans les conférences et les revues internationales. Mais quelle est la véritable raison de cet engouement ? Est-ce le fait, reconnu, que la bioinformatique propose de nouveaux problèmes aux informaticiens ? Ou est-ce le fait que ceux-ci retrouvent dans ce domaine une quête qui est chère à l'homme depuis bien des siècles : la quête de ses origines ?

Quatrième partie

Annexes

Annexe A

Publications dans le cadre de la thèse

- *On the interpretation of bootstrap trees: appropriate threshold of clade selection and induced gain*, publié dans la revue *Molecular Biology and Evolution* [BG96a]. Les résultats de cet article sont ceux exposés dans le chapitre 1 de la partie *II*. Suite cette étude, j'ai aussi été invité (*young investigator NSF grant*) à la *3rd Albany Conference on Computational Biology*, (Etats-Unis, 1995).

- *Choosing the tree which best explains the data: another look at the bootstrap in phylogenetic reconstruction*, présenté à la *2nd World Conf. of the International Association for Statistical Computing*, dont les actes ont été publiés dans *Computing Science and Statistics*, 1997, (Pasadena, Etats-Unis). Cet article reprenant les résultats du chap. 2, part. *II*, est aussi soumis à la revue *Computational Statistics and Data Analysis.*

- *Inferring evolutionary trees with strong combinatorial evidence*, présenté à la *3rd Annual International Computing and Combinatorics Conference (COCOON)*, Shanghaï (Chine), 1997. Cet article reprend les résultats présentés au chap. 2, part. *III*. Il a été soumis à la revue *Theoretical Computer Science*, éditant un numéro spécial pour la conférence citée ci-dessus.

- Le chapitre 3, part. *III*, est inclus dans un article actuellement soumis à la revue *Algorithmica.* Cet article reprend aussi certains résultats du chap. 2, part. *II*.

Annexe B

Preuve de la convergence des algorithmes de quadruplets

Au chapitre 2 et 3, de la partie III nous avons exhibé une borne sur la vitesse de convergence des méthodes Q^* et HAQ. Ce résultat reposait sur le lemme suivant (inspiré des résultats de Farach, Kannan et Atteson) :

Lemme : *Supposons qu'un ensemble d'espèces évolue le long d'un arbre suivant le modèle de l'évolution de NCF, supposons que l'on dispose de k caractères décrivant ces espèces, et soit $d = max(D_{xy})$ la distance évolutive maximum entre deux espèces x et y induite par l'arbre. Pour deux espèces x et y quelconques, et pour tout $\epsilon > 0$:*

$$P(|D_{xy} - \hat{D}_{xy}| < \epsilon) \ > \ 1 - 2\ e^{-(e^{-2\epsilon}-1)^2 e^{-4d} k/8}\ ,$$

où $\hat{D}_{xy}$ est la distance observée entre x et y sur la base des k caractères et corrigée selon le modèle NCF.

Preuve : Pour plus de clarté dans la preuve nous allégeons les notations en notant D pour D_{xy}, $\hat{D}$ pour $\hat{D}_{xy}$, p pour $p_{xy}(t)$ et f pour f_{xy}. Rappelons que sous le modèle NCF, $D = -\frac{1}{2}\ln(1-2p)$, formule que nous utilisons aussi pour corriger les distances observées : $\hat{D} = -\frac{1}{2}\ln(1-2f)$ (cf chap. 1 part. I). Pour prouver le lemme, nous nous intéressons à l'évènement inverse de $|D - \hat{D}| < \epsilon$:

$$|D - \hat{D}| \ \ \geq \ \ \epsilon\ . \tag{B.1}$$

Analysons séparément les deux cas possibles :

1. $\underline{D > \hat{D}}$:

$$
\begin{aligned}
|D - \hat{D}| \geq \epsilon &\equiv D - \hat{D} \geq \epsilon \\
&\equiv -\frac{1}{2}\ln(1-2p) + \frac{1}{2}\ln(1-2f) \geq \epsilon \\
&\equiv \frac{1}{2}\ln\left(\frac{1-2p}{1-2f}\right) \leq -\epsilon \\
&\equiv 1 - \frac{2\,(p-f)}{1-2f} \leq e^{-2\epsilon} \\
&\equiv p - f \geq \frac{1}{2}(1-e^{-2\epsilon})(1-2f) \\
&\Rightarrow p - f \geq \frac{1}{2}(1-e^{-2\epsilon})(1-2p)
\end{aligned}
\tag{B.2}
$$

(la dernière implication résulte du fait que $D > \hat{D}$).

2. $\underline{D < \hat{D}}$:

$$
\begin{aligned}
|D - \hat{D}| \geq \epsilon &\equiv \hat{D} - D \geq \epsilon \\
&\equiv -\frac{1}{2}\ln(1-2f) + \frac{1}{2}\ln(1-2p) \geq \epsilon \\
&\equiv \frac{1}{2}\ln\left(\frac{1-2f}{1-2p}\right) \leq -\epsilon \\
&\equiv 1 + \frac{2(p-f)}{1-2p} \leq e^{-2\epsilon} \\
&\equiv p - f \leq \frac{1}{2}(e^{-2\epsilon}-1)(1-2p) \\
&\Rightarrow p - f \leq \frac{1}{2}(1-e^{-2\epsilon})(1-2p)
\end{aligned}
\tag{B.3}
$$

(la dernière implication résulte du fait que $\epsilon > 0$).

De (B.2) et (B.3) on déduit :

$$\begin{aligned} |D - \hat{D}| \geq \epsilon \quad &\Rightarrow \quad |p - f| \geq \frac{1}{2}(1 - e^{-2\epsilon})(1 - 2p) \\ &\Rightarrow \quad |p - f| \geq \frac{1}{2}(1 - e^{-2\epsilon})e^{-2D} \\ &\Rightarrow \quad |p - f| \geq \frac{1}{2}(1 - e^{-2\epsilon})e^{-2d} \ , \end{aligned} \tag{B.4}$$

(puisque $d = max_{xy} D_{xy}$). L'inégalité de Azuma-Hoeffding [AS92] peut être appliquée dans ce contexte [Att97b, ESSW97b] et garantit que

$$P\left(|\mathrm{E}(f) - f| \geq g\right) \ \leq \ 2\ e^{-kg^2/2} \ ,$$

i.e., que

$$P\left(|p - f| \geq \frac{1}{2}(1 - e^{-2\epsilon})e^{-2d}\right) \ \leq \ 2\ e^{-(e^{-2\epsilon}-1)^2 e^{-4d} k/8} \ . \tag{B.5}$$

De (B.4) et (B.5) on obtient

$$P(|D - \hat{D}| \geq \epsilon) \ \leq \ 2\ e^{-(e^{-2\epsilon}-1)^2 e^{-4d} k/8} \ ,$$

d'où le résultat :

$$P(|D - \hat{D}| < \epsilon) \ > \ 1 - 2\ e^{-(e^{-2\epsilon}-1)^2 e^{-4d} k/8} \ .$$

□

Annexe C

Simulations sur la technique du bootstrap décrite au chapitre 1 de la partie II

Les résultats des simulations menées sur la technique du bootstrap (chap. 1, part. II) sont présentés ci-dessous.

Seuil optimum de sélection des arêtes

Nous avons proposé deux estimations analytiques simples du seuil optimal de sélection des arêtes ($S_1(\lambda)$ et $S_2(\lambda)$) et expliqué aussi comment calculer de façon empirique une troisième estimation $S_e(\lambda)$. Cette dernière estimation a été obtenue selon l'équation (1.8) (cf. chap. 1, part. II), depuis les $c=16$ conditions d'évolution étudiées : les quatre possibilités de taux d'évolution évoquées dans la section 1.4 du chap. 1, part. II, combinées avec les quatres longueurs de séquences : 100, 300, 1000, 3000 sites.

Nous évaluons la qualité des trois approximations $S_1(\lambda)$, $S_2(\lambda)$, et $S_e(\lambda)$ au moyen des mesures L_A (1.9) et L_R (1.10) (cf. chap. 1, part II), où $v=10$ valeurs de λ ont été employées ($\lambda \in \{1; 1,5; 2; 2,5; 3; 3,5; 4; 5; 6; 7; 8; 9; 10\}$). Les calculs sont effectués depuis 1600 phylogénies différentes et 160 000 jeux de données (*i.e.*, les valeurs relevées sont les valeurs maximales constatées sur 160 000 jeux de données différrents). La table C.1 détaille les résultats obtenus. Pour chaque méthode de reconstruction (ME, resp. MP), une des approximations simples calculées *a priori* ($S_1(\lambda)$, resp. $S_2(\lambda)$) est clairement plus efficace que l'autre, et, plus important, presque aussi efficace que l'approximation calculée *a posteriori* ($S_e(\lambda)$). Les $L_A(x)$ représentent des variation d'erreur comprises dans $[0; 14]$. En regard de cette

	L_A		L_R	
threshold	MP	ME	MP	ME
$S_1(\lambda)$	0,46	**0,16**	24%	**8%**
$S_2(\lambda)$	**0,20**	0,20	**13%**	18%
$S_e(\lambda)$	0,18	0,10	13%	6%

TAB. C.1 – *Pertes de qualité absolues (L_A) et relatives (L_R) observées pour les deux approximations analytiques ($S_1(\lambda)$ et $S_2(\lambda)$) et l'approximation empirique ($S_e(\lambda)$) du seuil optimal de sélection des arêtes. Il s'agit du maximum obtenu sur 16 conditions d'évolution et pour* $\lambda \in \{1; 1,5; 2; 2,5; 3; 3,5; 4,\ 5; 6; 7; 8; 9; 10\}$. *$L_A$ est mesuré en nombre d'arêtes et prend ses valeurs dans* $[0; 14]$. *Les résultats sont détaillés séparément pour la méthode basée sur les caractères, MP, et la méthode de distances, ME.*

échelle, les valeurs observées pour les trois approximations sont très faibles. Par exemple, considérons la perte de qualité maximum induite par $S_2(\lambda)$ pour MP, qui est de $0,20$ arêtes (table C.1). Cette perte est obtenue pour $\lambda = 3,5$, pour un taux de substitution faible et pour des séquences de 300 sites. Ceci signifie que sous ces conditions, il existe un jeu de données pour lequel l'arbre bootstrap réduit au seuil $S_2(\lambda)$ induit une erreur de $0,20$ arête supplémentaire à l'arbre bootstrap réduit selon le seuil optimal pour ces conditions, soit $S^i(\lambda)$. Plus précisément dans ce cas, pour $\lambda = 3,5$, $\bar{e}^{\lambda}_{RF}(\bar{t}^*_{S^i(\lambda)}) = 1,61$ arêtes et $\bar{e}^{\lambda}_{RF}(\bar{t}^*_{S_2(\lambda)}) = 1,81$ arêtes, si bien que même dans le pire des cas, on peut juger que les arbres bootstrap réduits $\bar{t}^*_{S^i(\lambda)}$ et $\bar{t}^*_{S_2(\lambda)}$ sont approximativement aussi proches de la phylogénie correcte t. Des valeurs relativement faibles sont aussi observées pour la mesure L_R : $L_R(S_1) = 8\%$ pour ME et $L_R(S_2) = 13\%$ pour MP. Notons que les conditions les moins avantageuses pour les diverses approximations ne sont pas forcément les mêmes pour les mesures L_A et L_R. Ainsi, la condition la plus défavorable à $S_2(\lambda)$ pour MP en terme de L_A (cf. ci-dessus) n'induit-elle qu'une perte relative de 11% (le maximum étant 13%).

Pour expliquer ces bons résultats, nous avons tracé la fonction empirique reliant la valeur de bootstrap (*bp*) et la probabilité que l'arête correspondante soit correcte (Fig. 1.2 chap. 1, part. II). Une telle courbe a déjà été étudiée par Hillis et Bull [HB93] pour la méthode MP, sur la base de simulations de phylogénies de 4 et 9 espèces. La figure 1.2 représente aussi les fonctions f_1 et f_2 depuis lesquelles les approximations $S_1(\lambda)$ et $S_2(\lambda)$ sont obtenues. La fonction sinusoïdale est constatée relativement proche de la courbe empirique associée à MP (f_e^{MP}), et la fonction identité joue le même rôle pour la courbe empirique associée à la méthode ME

(f_e^{ME}), pour $bp > 60\%$. Cette figure explique déjà les performances similaires de $S_1(\lambda)$ et $S_e(\lambda)$ pour ME et de $S_2(\lambda)$ et $S_e(\lambda)$ pour MP.

La grande qualité des approximations est aussi liée à la forme de la courbe représentant l'erreur moyenne induite par l'arbre bootstrap réduit en fonction du seuil de sélection des arêtes. La figure 1.4, chap. 1, part. II, détaille cette courbe pour le pire des cas précédemment mentionné (méthode MP, $\lambda = 3,5$, 300 sites évoluant avec un taux faible). Dans ce cas, nous avons $S^i(\lambda) \approx 59\%$, $S_2(\lambda) \approx 69\%$ et $S_1(\lambda) \approx 78\%$, *i.e.*, on constate que $S_2(\lambda)$ est plus efficace que $S_1(\lambda)$ pour approximer $S^i(\lambda)$, comme précédemment constaté pour MP (table C.1 et Fig. 1.2). La platitude de la courbe de la Fig. 1.4 autour du seuil optimal $S^i(\lambda)$ explique pourquoi l'erreur de l'arbre bootstrap réduit est modérément sensible au choix exact d'un seuil de sélection des arêtes dans cette région : la dérivée seconde de la courbe est très faible (*i.e.*, son rayon de courbure est important), si bien que la perte de qualité induite par l'utilisation de $S_2(\lambda)$ plutôt que $S^i(\lambda)$ est seulement du second ordre relativement à la différence entre ces deux seuils.

Réduction d'erreur dûe à la suppression des arêtes de faible *bp*

Nous examinons maintenant l'étendue de la réduction d'erreur, obtenue en estimant la phylogénie correcte t par l'arbre bootstrap réduit $t^*_{S(\lambda)}$ plutôt que par la phylogénie $\hat{t}$ inférée directement depuis le jeu de données originel (aussi appelée *arbre direct*).

Erreur généralisée

La table C.2 donne la réduction d'erreur absolue et relative obtenue par $t^*_{S(\lambda)}$ vis-à-vis de $\hat{t}$, suivant les différentes conditions d'évolution et les différents poids accordés aux erreurs de 1ère et de 2ème espèces. En accord avec les précédents résultats, $S_1(\lambda)$, resp. $S_2(\lambda)$, a été utilisé pour approximer le seuil optimum $S(\lambda)$ dans le cas de ME, resp. MP. La réduction d'erreur obtenue est détaillée pour ces deux méthodes de reconstruction pour les 16 conditions d'évolution retenues et pour différentes valeurs de λ (1, 2 et 5). L'étendue de la réduction d'erreur dépend grandement des conditions d'évolution. Le procédé bootstrap apparaît plus efficace quand on dispose de séquences courtes. De plus, la réduction d'erreur est constatée plus importante avec MP qu'avec ME. Augmenter λ conduit aussi à des réductions d'erreur plus importantes. Comme $e^{\lambda}_{RF}(\hat{t})$ est indépendant de λ, ceci signifie que $e^{\lambda}_{RF}(\bar{t}^*_{S(\lambda)})$ diminue quand λ augmente. En fait, pour des valeurs extrêmement élevées de λ (et pour un nombre élevé de réplications du bootstrap), on s'attend à ce que $\bar{t}^*_{S(\lambda)} \approx \bar{t}^*_{100}$ soit la phylogénie étoile. Pour cette phylogénie extrême, l'erreur de 1ère espèce est nulle et l'influence de l'erreur de 2ème espèce

A. Méthode MP

		$\lambda = 1$		$\lambda = 2$		$\lambda = 5$	
taux	sites	rel	abs	rel	abs	rel	abs
faibles	100	39%	3,26	56%	4,74	75%	6,29
	300	33%	1,49	51%	2,37	69%	3,19
	1000	24%	0,51	45%	0,99	65%	1,42
	3000	20%	0,14	39%	0,31	59%	0,47
élevés	100	18%	0,51	36%	1,02	61%	1,71
	300	9%	0,11	25%	0,33	51%	0,68
	1000	4%	0,03	21%	0,16	48%	0,36
	3000	4%	0,01	21%	0,06	46%	0,13
inégaux selon arêtes	100	27%	1,45	46%	2,49	69%	3,75
	300	21%	0,76	40%	1,43	64%	2,31
	1000	14%	0,27	31%	0,59	55%	1,04
	3000	4%	0,03	14%	0,15	31%	0,34
inégaux selon sites	100	24%	1,07	42%	1,90	66%	2,97
	300	16%	0,32	33%	0,64	58%	1,13
	1000	8%	0,06	23%	0,18	49%	0,39
	3000	6%	0,01	17%	0,05	42%	0,13

B. Méthode ME

		$\lambda = 1$		$\lambda = 2$		$\lambda = 5$	
taux	sites	rel	abs	rel	abs	rel	abs
faibles	100	33%	2,63	49%	4,00	72%	5,92
	300	24%	1,17	38%	1,93	64%	3,21
	1000	11%	0,15	22%	0,36	51%	0,87
	3000	5%	0,03	19%	0,15	49%	0,40
élevés	100	9%	0,24	25%	0,67	53%	1,41
	300	6%	0,09	22%	0,36	49%	0,79
	1000	3%	0,02	18%	0,10	46%	0,27
	3000	3%	0,01	23%	0,09	54%	0,20
inégaux selon arêtes	100	17%	0,86	37%	1,93	65%	3,37
	300	9%	0,28	29%	0,87	59%	1,77
	1000	9%	0,20	28%	0,60	58%	1,22
	3000	7%	0,08	25%	0,28	53%	0,61
inégaux selon sites	100	14%	0,59	32%	1,37	61%	2,57
	300	6%	0,13	23%	0,47	51%	1,05
	1000	3%	0,03	21%	0,20	49%	0,48
	3000	3%	0,02	23%	0,14	55%	0,34

TAB. C.2 – *Réduction relative (rel) et absolue (abs) de la distance à la phylogénie correcte t obtenue en estimant t par l'arbre bootstrap réduit au seuil adéquat plutôt que par la phylogénie directement inférée depuis le jeu de données initial.* λ *est l'importance attribuée à l'erreur de 1ère espèce vis-à-vis de l'erreur de 2ème espèce. Le seuil optimum de sélection des arêtes utilisé par l'arbre bootstrap réduit a été approximé par* $S_1(\lambda)$ *pour* D *et par* $S_2(\lambda)$ *pour* MP.

est négligeable, si bien qu'on s'attend à ce que $e^{\lambda}_{RF}(\bar{t}^*_{S(\lambda)}) \approx 0$. Cependant, quand on augmente λ sur des valeurs plus raisonnables, deux tendances s'opposent dans chaque composante de l'erreur e^{λ}_{RF} : l'erreur de 1ère espèce diminue mais son poids $(\frac{2\lambda}{\lambda+1})$ augmente, alors que l'erreur de 2ème espèce augmente mais son influence diminue $(\frac{2}{\lambda+1})$. Ainsi on ne pouvait pas prédire les résultats observés dans la table C.2, résultats qui semblent indiquer que la réduction d'erreur obtenue par l'arbre bootstrap réduit avec un seuil adéquat augmente de façon monotone avec λ.

Erreur standard de Robinson et Foulds

Nous nous concentrons maintenant sur la mesure d'erreur obtenue en donnant un même poids aux erreurs de 1ère et de 2ème espèces ($\lambda = 1$). Les tables C.3-C.6 quantifient la réduction d'erreur dûe à l'utilisation de $\bar{t}^*_{50\%}$, l'arbre bootstrap réduit au seuil approximativement optimal, plutôt que de $\hat{t}$ pour estimer t. Notons que nous sommes déjà assurés d'un gain (*i.e.*, d'une réduction d'erreur), puisque $\hat{t}$ peut être vu comme $\hat{t}^*_0$, *i.e.*, comme un arbre bootstrap réduit à un seuil non-optimal. Ces tables détaillent aussi l'erreur moyenne $\bar{e}_{RF}$ (décomposée en $\bar{e}_1$ et $\bar{e}_2$) des estimations de t. Rappelons que cette mesure, resp. chacune de ses composantes, prend ses valeurs dans $[0;14]$, resp. $[0;7]$. Notons aussi, comme point de repère pour les valeurs observées, que la phylogénie étoile induit sous toutes les conditions une erreur $\bar{e}_{RF} = 7$ ($e_1 = 0$, $e_2 = 7$).

En écartant des arêtes de la phylogénie $\hat{t}$, on s'attend bien-sûr à ce que $\bar{t}^*_{50\%}$ induise une plus faible erreur de 1ère espèce et une plus forte erreur de 2ème espèce. Les tables confirment ce phénomène mais montrent toutefois que ces deux tendances n'ont pas la même importance, *i.e.*, l'erreur de 1ère espèce est grandement réduite alors que l'erreur de 2ème espèce n'est que faiblement augmentée. Comme exemple extrême, on peut considérer le cas du taux d'évolution faible, où $\bar{e}_1(\hat{t}) = \bar{e}_2(\hat{t}) \approx \bar{e}_2(\bar{t}^*_{50\%})$, tandis que $\bar{e}_1(\bar{t}^*_{50\%}) \approx 0$, si bien que l'arbre bootstrap réduit induit une erreur deux fois plus faible que $\hat{t}$.

Finalement, on retiendra que la réduction relative d'erreur d'estimation dûe à l'utilisation de la technique du bootstrap est observée entre 3% et 39%, selon les conditions d'évolution.

On peut comparer les résultats de ces simulations avec celles de Kuhner et Felsenstein [KF94], qui examinent le comportement de plusieurs méthodes de R.P. (FM [FM67],MP, ME, MV) au sens de l'erreur standard de Robinson et Foulds, $\bar{e}_{RF}$:

- Les erreurs moyennes observées ici pour l'arbre direct $\hat{t}$ sont semblables à celles indiquées par Kuhner et Felsenstein pour des conditions similaires,

A. Méthode MP

	Réduction d'erreur			$\hat{t}$		$\bar{t}^*_{50\%}$		
#Sites	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1 = \bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$
100	**39%**	47%	-8%	**8,42**	4,21	**5,16**	0,29	4,87
300	**33%**	45%	-12%	**4,66**	2,33	**3,13**	0,23	2,90
1000	**24%**	39%	-15%	**2,11**	1,06	**1,61**	0,23	1,37
3000	**20%**	39%	-19%	**0,79**	0,40	**0,64**	0,09	0,55

B. Méthode ME

	Réduction d'erreur			$\hat{t}$		$\bar{t}^*_{50\%}$		
#Sites	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1 = \bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$
100	**33%**	37%	-4%	**8,22**	4,11	**5,48**	1,03	4,44
300	**24%**	31%	-7%	**5,02**	2,51	**3,84**	0,96	2,88
1000	**11%**	18%	-7%	**1,69**	0,85	**1,52**	0,55	0,97
3000	**5%**	16%	-11%	**0,79**	0,40	**0,76**	0,27	0,49

TAB. C.3 – *Erreur standard de Robinson et Foulds, $\bar{e}_{RF}$, induite en moyenne par la phylogénie initialement inférée, $\hat{t}$, et par l'arbre bootstrap réduit, $\bar{t}^*_{50\%}$. Cette erreur est décomposée en erreur de 1ère espèce, $\bar{e}_1$, et erreur de 2ème espèce, $\bar{e}_2$. La réduction relative en erreur obtenue en utilisant $\bar{t}^*_{50\%}$ plutôt que $\hat{t}$ pour estimer t, est aussi détaillée, ainsi que sa répartition en erreur de 1ère espèce et de 2ème espèce. Les résultats de la table ont été obtenus pour des* **taux d'évolution élevés**.

sauf peut-être dans le cas de taux inégaux entre les sites où nous observons une vitesse de convergence plus importante ici.

– Pour des séquences de 100 nucléotides, l'arbre bootstrap réduit $\bar{t}^*_{50\%}$, qu'il soit inféré avec MP ou ME, est plus proche en moyenne de la phylogénie correcte t, que l'arbre direct inféré par les différentes méthodes de reconstruction étudiées par Kuhner et Felsenstein.

– Pour des séquences de 300 nucléotides et plus, $\bar{t}^*_{50\%}$ obtient en moyenne (sur l'ensemble des conditions d'évolution testées) des résultats similaires à ceux de l'arbre direct inféré par la méthode MV (demandant un temps calcul bien plus important que pour obtenir l'arbre bootstrap réduit), *i.e.*, que la méthode qui obtient de loin les meilleurs résultats dans l'étude de Kuhner et Felsenstein.

A. Méthode MP

	Réduction d'erreur			$\hat{t}$		$\bar{t}^*_{50\%}$		
#Sites	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1 = \bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$
100	**18%**	34%	-16%	**2,82**	1,41	**2,32**	0,46	1,85
300	**9%**	26%	-17%	**1,32**	0,66	**1,20**	0,32	0,88
1000	**4%**	16%	-12%	**0,75**	0,38	**0,72**	0,26	0,47
3000	**4%**	14%	-10%	**0,28**	0,14	**0,27**	0,10	0,17

B. Méthode ME

	Réduction d'erreur			$\hat{t}$		$\bar{t}^*_{50\%}$		
#Sites	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1 = \bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$
100	**9%**	22%	-13%	**2,69**	1,35	**2,44**	0,75	1,70
300	**6%**	18%	-12%	**1,53**	0,76	**1,44**	0,48	0,95
1000	**3%**	15%	-12%	**0,58**	0,29	**0,56**	0,20	0,36
3000	**3%**	12%	-9%	**0,32**	0,16	**0,31**	0,12	0,19

TAB. C.4 – *cf. légende de la table C.3. Les résultats de la table ont été obtenus pour des* **taux d'évolution faibles**.

A. Méthode MP

	Réduction d'erreur			$\hat{t}$		$\bar{t}^*_{50\%}$		
#Sites	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1 = \bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$
100	**27%**	38%	-11%	**5,38**	2,69	**3,92**	0,63	3,29
300	**21%**	36%	-15%	**3,59**	1,80	**2,82**	0,49	2,33
1000	**14%**	27%	-13%	**1,91**	0,95	**1,64**	0,44	1,20
3000	**4%**	10%	-6%	**1,07**	0,54	**1,04**	0,43	0,60

B. Méthode ME

	Réduction d'erreur			$\hat{t}$		$\bar{t}^*_{50\%}$		
#Sites	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1 = \bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$
100	**17%**	26%	-9%	**5,21**	2,60	**4,29**	1,24	3,05
300	**10%**	20%	-10%	**3,04**	1,52	**2,74**	0,91	1,83
1000	**9%**	20%	-11%	**2,11**	1,06	**1,91**	0,62	1,29
3000	**7%**	18%	-11%	**1,13**	0,56	**1,04**	0,36	0,68

TAB. C.5 – *cf. légende de la table C.3. Les résultats de la table ont été obtenus pour des* **taux d'évolution faibles sur la moitié des arêtes et élevés sur l'autre moitié**.

A. Méthode MP

	Réduction d'erreur			$\hat{t}$		$\bar{t}^*_{50\%}$		
#Sites	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1 = \bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$
100	**24%**	38%	-14%	**4,51**	2,25	**3,43**	0,53	2,89
300	**16%**	31%	-15%	**1,94**	0,97	**1,64**	0,37	1,27
1000	**8%**	20%	-12%	**0,75**	0,38	**0,69**	0,22	0,47
3000	**6%**	19%	-13%	**0,32**	0,16	**0,29**	0,10	0,20

B. Méthode ME

	Réduction d'erreur			$\hat{t}$		$\bar{t}^*_{50\%}$		
#Sites	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1 = \bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$
100	**14%**	26%	-12%	**4,19**	2,10	**3,62**	1,03	2,59
300	**7%**	20%	-13%	**2,04**	1,02	**1,91**	0,61	1,29
1000	**3%**	13%	-10%	**0,97**	0,49	**0,94**	0,36	0,58
3000	**3%**	11%	-8%	**0,62**	0,31	**0,60**	0,24	0,36

TAB. C.6 – *cf. légende de la table C.3. Les résultats de la table ont été obtenus pour des* **taux d'évolution faibles pour la moitié des sites et élevés pour l'autre moitié**.

Erreur en termes de biais et de variance

En s'intéressant toujours à l'erreur standard de Robinson et Foulds (donnant un même poids à e_1 et e_2), nous abordons maintenant les résultats selon les termes de biais et de variance définis dans la section 1.2.2, chap. 1, part. *II*.

La table C.7 montre les résultats obtenus par $\hat{t}$ et $\bar{t}^*_{50\%}$ au sens de la mesure de biais dûe à Kuhner et Felsenstein [KF94], pour des taux d'évolution inégaux suivant les arêtes. Ces résultats suffisent à illustrer les tendances générales constatées sur l'ensemble des conditions. Les résultats obtenus pour l'arbre direct, $\hat{t}$, sont similaires à ceux de l'étude de Kuhner et Felsenstein [KF94].

Quand nous avons présenté cette mesure (section 1.2, chap. 1, part. *II*), nous avons montré qu'elle était sensible au biais structurel des phylogénies inférées, *i.e.*, qu'elle indique systématiquement un biais élevé pour une méthode proposant des phylogénies partiellement résolues, sans que cela indique la présence d'un biais fondamental (son but initial). Les résultats de la table C.7 semblent confirmer cette remarque, puisque la phylogénie $\hat{t}^*_{50\%}$, partiellement résolue, apparaît très biaisé au sens de cette mesure. Le biais est d'autant plus important que les séquences sont courtes, *i.e.*, que l'arbre bootstrap réduit est peu résolu.

On a confirmation du fait que le biais observé pour $\hat{t}^*_{50\%}$ par la mesure précé-

	Parcimonie		Distance	
#Sites	$\hat{t}$	$\bar{t}^*_{50\%}$	$\hat{t}$	$\bar{t}^*_{50\%}$
100	0	41	1	20
300	3	22	0	6
1000	5	15	2	6
3000	9	15	1	3

TAB. C.7 – *Nombre de phylogénies modèles (sur 50) pour lesquelles les phylogénies inférées ($\hat{t}$ et $\bar{t}^*_{50\%}$) apparaissent biaisées au sens de la mesure de Kuhner et Felsenstein [KF94]. Les distances entre phylogénies sont mesurées suivant la distance standard de Robinson et Foulds. Les valeurs affichées ont été obtenues pour des taux inégaux de substitution suivant les arêtes.*

	Parcimonie		Distance	
	$\hat{t}$	$\bar{t}^*_{50\%}$	$\hat{t}$	$\bar{t}^*_{50\%}$
$2\overline{V}$	5,20	1,06	5,04	2,28
$\overline{BS}$	0	2,66	0	1,81
$2\overline{A}$	0,18	0,20	0,17	0,20
$\bar{e}_{RF}$	5,38	3,92	5,21	4,29

TAB. C.8 – *Erreur standard de Robinson et Foulds moyenne pour des séquences de 100 sites et des taux d'évolution inégaux selon les arêtes. L'erreur est décomposée en un terme de variance ($2\overline{V}$), un terme de biais structurel ($\overline{BS}$) et un terme d'acentralité ($2\overline{A}$), pour l'arbre direct ($\hat{t}$), et pour l'arbre bootstrap réduit au seuil adéquat ($\bar{t}^*_{50\%}$).*

dente est essentiellement structurel en analysant les résultats au travers des trois mesures de biais et de variance issues de la décomposition de l'erreur que nous avons aussi introduite en section 1.2, chap. 1, part. *II*. Par exemple, la table C.8 montre les valeurs moyennes observées pour ces trois termes dans le cas de taux d'évolution inégaux entre les arêtes et pour des séquences de 100 nucléotides, condition où $\hat{t}^*_{50\%}$ apparaissait clairement biaisé dans la table C.7.

Rappelons que les trois termes de biais et de variance sont basés sur les poids $\mathcal{P}(b)$ associés respectivement aux différentes bipartitions b possibles sur n espèces. Chaque poids $\mathcal{P}(b)$ a été approximé par la fréquence d'apparition de b dans la phylogénie inferée ($\hat{t}$ ou $\hat{t}^*_{50\%}$) sur l'ensemble des jeux de données associés à une même phylogénie modèle. Les valeurs des termes de biais et de variance ont été moyennées sur les 50 phylogénies t générées, expliquant la concordance des valeurs $\bar{e}_{RF}$ avec celles de la table C.5. Les résultats de la table C.8 confirment que pour la condition d'évolution associée à la table, la phylogénie $\hat{t}^*_{50\%}$ est faiblement résolu en moyenne (comme indiqué par $7 - \overline{BS}$). Le biais précédemment détecté pour cette phylogénie est donc presque entièrement structurel, et non fondamental (faible valeur de $\bar{A}$). Ainsi la phylogénie $\hat{t}^*_{50\%}$, souvent partiellement résolue, (≈ 5 arêtes en moyenne pour la condition d'évolution liée à la table C.8) montre-t-elle un biais structurel plus fort que la phylogénie $\hat{t}$, qui elle est complètement résolue. En contrepartie, $\hat{t}^*_{50\%}$ présente une variance bien plus faible que $\hat{t}$, qu'il soit inféré par la méthode ME ou MP. Ceci vient du fait que $\hat{t}^*_{50\%}$ contient uniquement des arêtes ayant une *bp* relativement élevée, donc des arêtes qui sont fortement susceptibles d'être retrouvés d'un jeu de données sur l'autre, ce qui n'est pas le cas de $\hat{t}$ dont certaines arêtes dépendent de la composition exacte du jeu de données analysé. Des valeurs faibles sont observées en moyenne pour le terme d'acentralité, exprimant l'absence de biais fondamental pour les deux estimateurs de t ($\hat{t}$ et $\hat{t}^*_{50\%}$). C'est la tendance générale que nous avons observée pour ce terme sur l'ensemble des simulations, un maximum de $1,75$ étant constaté pour une phylogénie déclarée aussi très clairement biaisée par la mesure de Kuhner et Felsenstein. Ces bons résultats pour le terme d'acentralité sont expliqués par le fait que les méthodes de reconstruction ME et MP sont globalement consistantes pour les conditions d'évolution examinées. Des conditions plus extrêmes conduiraient sûrement l'acentralité à être une composante de l'erreur bien plus importante.

Autres remarques

Seuil de 95% et test d'hypothèses

La table C.9 montre l'erreur de 1ère et de 2ème espèce lié à l'arbre bootstrap réduit au seuil de 95%. Rappelons que pour l'erreur standard, e_{RF} (correspondant à la distance de Robinson et Foulds), on s'attend à ce que les performances de cette phylogénie soient faibles, puisque 95% est loin d'être le seuil optimal ($\approx$ 50% dans ce cas). Les tables C.3-C.6 et C.9 montrent que c'est effectivement le cas. Cependant, le seuil 95% est généralement considéré pour effectuer un test d'hypothèse. Dans ce cas, la proportion bootstrap joue le rôle de la variable de décision du test, et on espère que la probabilité d'accepter une arête incorrecte soit inférieure à 5%, tout en étant proche de cette valeur. Dans le cas de 10 espèces, $\hat{t}$ contient 7 arêtes, si bien qu'on peut estimer cette probabilité par $\frac{\bar{e}_1}{7}$. Cette quantité est bien souvent observée (table C.9) proche de 0% et atteint un maximum de 2% (quand les taux de substitutions sont inégaux entre les arêtes). Ceci indique qu'un test construit sur la base des *bp* est bien fondé (puisque $\frac{\bar{e}_1}{7} < 5\%$) mais très conservatif (puisque $\frac{\bar{e}_1}{7} << 5\%$). Ceci nous amène à questionner la *puissance* de ce test, *i.e.*, la probabilité de conserver une arête correcte. On peut obtenir

		Parcimonie		Distance	
taux	sites	$\bar{e}_1$	$\bar{e}_2$	$\bar{e}_1$	$\bar{e}_2$
faibles	100	0,00	6,77	0,06	6,71
	300	0,00	5,76	0,05	5,66
	1000	0,00	3,21	0,02	2,90
	3000	0,00	1,53	0,00	1,52
élevés	100	0,00	4,48	0,05	4,34
	300	0,00	2,50	0,01	2,83
	1000	0,01	1,29	0,00	1,17
	3000	0,00	0,58	0,00	0,51
inégaux selon arêtes	100	0,00	5,26	0,04	5,35
	300	0,00	4,40	0,03	3,88
	1000	0,03	2,73	0,02	2,86
	3000	0,14	1,43	0,01	1,77
inégaux selon sites	100	0,00	5,66	0,02	5,48
	300	0,00	3,40	0,01	3,64
	1000	0,00	1,64	0,01	1,72
	3000	0,00	0,74	0,00	0,92

TAB. C.9 – *Erreur de 1ère espèce et de 2ème espèce pour l'arbre bootstrap réduit au seuil de* 95%, $\bar{t}^*_{95\%}$.

une estimation de cette probabilité par $1 - \frac{\bar{e}_2(\bar{t}^*_{95\%})}{7}$. La table C.9 indique que $\bar{e}_2$, et donc la puissance du test, varie grandement suivant les conditions d'évolution et la taille des séquences, mais relativement peu selon la méthode de reconstruction. La puissance du test construit sur la base du procédé bootstrap est constatée très faible pour des séquences courtes (de $3,3\%$ à 38% avec 100 sites), mais plus acceptable pour des séquences longues (de 75% à 93% avec 3000 sites). Cette remarque montre l'intérêt des études récentes sur la puissance du bootstrap, essayant de fournir de meilleures approximations du niveau de confiance du test que celui obtenu par simple bootstrap.

Influence de la méthode de reconstruction

Les tables C.3-C.6 montrent que la phylogénie $\hat{t}$, inférée directement depuis le jeu de données initial, a des performances similaires quand il est inféré par la méthode MP ou la méthode ME. Cependant, la réduction d'erreur dûe à l'usage du bootstrap est toujours (sauf une fois) plus importante dans le premier cas. On obersve que le bootstrap conduit à écarter plus d'arêtes avec MP qu'avec ME (cf. différences en e_1 et e_2 dans les tables C.3-C.6 et en BS dans la table C.8). Le choix du seuil de 50% semble être la clef de ce phénomène. On peut en effet constaté sur la Fig. 1.2 que ce seuil convient bien mieux à MP qu'à ME, pour lequel un seuil de 55% aurait été plus adéquat. Ce seuil légèrement plus élevé aurait conduit à écarter plus d'arêtes, donnant alors des résultats similaires à ceux de MP.

Influence des conditions d'évolution

Les méthode de reconstruction utilisées apparaissent comme consistantes sur les conditions d'évolution étudiées, puisqu'on observe (tables C.3-C.6) que plus on dispose de séquences longues, plus l'erreur moyenne des méthodes est réduite. En conséquence, les estimations de t obtenues pour de longues séquences sont de bonne qualité, laissant peu de place aux améliorations. De plus, pour de longues séquences, le bootstrap amène moins de variabilité entre les échantillons. Ces deux effets contribuent à réduire l'efficacité relative de $\bar{t}^*_{50}$, si bien que, pour des séquences de 3000 sites, il montre une erreur similaire à celle de $\hat{t}$.

L'écart constaté entre les deux phylogénies varie aussi selon les taux de substitution, mais est généralement plus faible quand les conditions d'évolution sont favorables à la reconstruction de phylogénies. La condition la plus favorable correspond à des taux d'évolution "élevés" et uniformes (hypothèse de l'horloge moléculaire). En contraste, les plus fortes valeurs de $\bar{e}_{RF}$ sont obtenues pour des taux "faibles" et c'est précisément pour cette condition qu'on observe l'écart le plus important entre $\bar{t}^*_{50}$ et $\hat{t}$ (réduction d'erreur de 39% avec MP et de 33% avec D). Pour cette condition d'évolution, il faut disposer de séquences relativement longues pour obtenir des estimations précises de t, en raison de la très faible proportion de substitutions et donc de sites informatifs.

Annexe D

Simulations concernant la nouvelle utilisation de la technique du bootstrap introduite au chapitre 2 de la partie II

Schéma de simulations

Des jeux de données X sont générée pour 10 espèces suivant des conditions d'évolution identiques à celles détaillées au chapitre précédent : taux d'évolution faibles constant (cf), taux inégaux selon les arêtes (i/a) ou encore taux inégaux selon les sites (i/s). Notons que cette dernière condition viole une hypothèse importante du modèle de l'évolution utilisé (modèle de Kimura à deux paramètres [Kim80]) pour obtenir $\hat{D}$ depuis X, ce qui permet d'éprouver la robustesse des méthodes. Disposant d'un jeu de données X, les trois approches testées sont lancées séparément :

– *Approche directe :* $\hat{D}$ est calculée depuis X suivant le modèle de Kimura [Kim80] à deux paramètres, avec un taux de transitions/transversions de 2.0. Puis $\hat{D}$ est donnée en entrée à la méthode M décrite en section 2.2.2, ce qui aboutit à la production d'une phylogénie $\hat{t}$, estimant la phylogénie t depuis laquelle X a été obtenu.

– *Approche à seuil fixe :* le rééchantillonnage bootstrap de X produit des pseudojeux de données X^*, qui sont donnés en entrée à la méthode M, résultant en autant d'arbres bootstrap $\hat{t}^*$. Nous avons généré $r=100$ pseudojeux bootstrap pour chaque jeu de données initial X (donnant un écart-type de 5% au maximum pour les *bp*). L'arbre bootstrap réduit est obtenu en contractant dans la phylogénie $\hat{t}$, inférée par l'approche directe, les arêtes ayant une valeur de bootstrap inférieure à un certain seuil S. Selon l'analyse menée au chapitre précédent, un seuil $S = 50\%$ sera retenu, puisque δ (éq. (2.1), chap. 2, part II) donne un poids identique aux erreurs de 1ère espèce et 2ème espèce. On notera $\hat{t}_{sf}$ l'estimation de t produite par cette approche à *seuil fixe*.

– *Approche à complexité variable :* elle utilise elle aussi des pseudo-jeux de données X^*, produits par rééchantillonnage bootstrap (on étudiera aussi l'efficacité de cette approche quand le rééchantillonnage est effectué par validation croisée). Pour cette approche, $r' = 10$ réplications bootstrap sont suffisantes puisque le seul paramètre à estimer ici est la complexité de la phylogénie tandis que l'approche à seuil fixe doit estimer les $O(n)$ valeurs *bp* (d'où $r' < r$). On notera $\hat{t}_{cv}$ l'estimation de t produite par cette approche à *complexité variable*.

Ces trois approches sont comparées, sur la base de leurs erreurs moyennes respectives, mesurées par le critère δ (*i.e.*, $\bar{\delta}(D, \hat{t})$, $\bar{\delta}(D, \hat{t}_{sf})$ et $\bar{\delta}(D, \hat{t}_{cv})$).

Résultats

Les résultats sont présentés dans les tables sous forme de pourcentages d'erreur, obtenus en divisant l'erreur de chaque méthode par $|Q_{\hat{D}}|$, le nombre de r4-arbres du jeu de données.

Métodes de rééchantillonnage pour l'approche à complexité variable

Nous avons vu précédemment que l'approche à complexité variable pouvait être appliquée avec diverses techniques de rééchantillonnage. Nous examinons ici l'efficacité relative du bootstrap et de la validation croisée quand ils sont employés dans le cadre de cette approche. La table D.1 montre les résultats moyens obtenus par simulation de 20 phylogénies $\times$ 20 jeux de données, et ce pour diverses conditions d'évolution et différents nombres d'espèces considérées.

Ces résultats montrent que la validation croisée est moins efficace que le bootstrap pour les différents types de jeux de données considérés (nombre d'espèces, taux d'évolution, longueur des séquences). Ceci peut être expliqué par le fait que la validation croisée n'utilise que la moitié de chaque jeu de données initial pour construire les séquences de phylogénies, et l'autre moitié pour les évaluer. Ceci

TAB. D.1 – *Efficacité relative de deux techniques de rééchantillonnage dans l'approche à complexité variable (cv) pour différentes conditions d'évolution simulées : taux constants et faibles (cf), constants élevés (ce), inégaux selon les arêtes (i/a).*

	Const^o	Valid^o Croisée		Bootstrap	
Condition	$\%\delta(t,\hat{t})$	$\%\delta(t,\hat{t}_{cv})$	$\lvert\hat{t}_{cv}\rvert$	$\%\delta(t,\hat{t}_{cv})$	$\lvert\hat{t}_{cv}\rvert$
10 esp. (cf) 300	31,52	21,0	2,03	8,0	3,8
10 esp. (ce) 300	10,97	10,1	5,8	10,0	6,3
10 esp. (i/a) 300	25,61	18,4	4.2	17,8	5,5
20 esp. (i/a) 300	19,27	15,7	10,3	15,0	13,2

mène à une approche trop conservatrice, sous-évaluant la qualité des arêtes proposées. En contraste, le bootstrap est naturellement optimiste puisque les phylogénies sont évaluées par des jeux de données relativement proches de ceux ayant servi à les inférer. On retrouve ces deux tendances dans les résultats où les phylogénies inférées par validation croisée sont en moyenne de complexité plus faible que ceux inférés par bootstrap.

Aux vues de ces résultats, nous considèrons dans la suite uniquement la technique du bootstrap pour représenter la méthode complexité variable.

Approches de rééchantillonnage contre approche directe

Les résultats de table (D.2) sont des moyennes obtenues sur 20 phylogénies $\times$ 20 jeux de données suivant diverses conditions d'évolution. Dans la table (D.2), chaque mesure d'erreur moyenne m est accompagnée de son écart-type $(\sigma(m)/20)$ indiqué entre parenthèses.

Ces résultats confirment le bien-fondé du recours aux techniques de rééchantillonnage puisque les approches (sf) et (cv) amènent à des gains significatifs vis-à-vis de l'approche directe (jusqu'à une réduction d'erreur de 28% et 22% respectivement). L'importance du gain dépend grandement des conditions d'évolution et de la longueur des séquences. Le gain est d'autant plus élevé qu'on dispose de séquences courtes, ce qui est important car c'est le paramètre le plus limitatif pour les jeux de données réels, d'autant plus qu'on considère un nombre important d'espèces. Les plus faibles gains dûs à l'utilisation de techniques de rééchantillonnage sont obtenus pour la condition d'évolution la plus difficile pour la reconstruction, (i/s). Ceci résulte simplement du fait que les approches (sf) et (cv) sont basées, comme l'approche directe, sur la méthode de reconstruction M, et que cette méthode rencontre des problèmes d'inconsistance pour traiter les données associées

TAB. D.2 – *Efficacité des approches* directes *(proposant $\hat{t}$), à* seuil fixe *(proposant $\hat{t}_{sf}$) et à* complexité variable *(proposant $\hat{t}_{cv}$) pour estimer t.*

# sites	condition	% $\delta(D,\hat{t})$	% $\delta(D,\hat{t}_{sf})$	% $\delta(D,\hat{t}_{cv})$
100	(cf)	36 $\pm 0,01$	8 $\pm 0,01$	14 $\pm 0,01$
	(i/a)	36 $\pm 0,02$	22 $\pm 0,01$	25 $\pm 0,01$
	(i/s)	34 $\pm 0,01$	30 $\pm 0,01$	32 $\pm 0,01$
300	(cf)	26 $\pm 0,02$	10 $\pm 0,01$	11 $\pm 0,01$
	(i/a)	23 $\pm 0,01$	17 $\pm 0,01$	17 $\pm 0,01$
	(i/s)	20 $\pm 0,01$	19 $\pm 0,01$	19 $\pm 0,00$
600	(cf)	8 $\pm 0,01$	6 $\pm 0,01$	14 $\pm 0,01$
	(i/a)	6 $\pm 0,01$	6 $\pm 0,01$	7 $\pm 0,01$
	(i/s)	$\pm$	$\pm$	$\pm$

à cette condition d'évolution. En conséquence, pour cette condition, les trois approches donnent des estimations moins bonnes de t, et l'écart entre $\hat{t}$ et $\hat{t}_{cv}$ ou $\hat{t}_{sf}$ est réduit.

Approche à complexité variable contre approche à seuil fixe

Les deux approches basées sur le rééchantillonnage, (sf) et (cv), donnent des résultats d'efficacité comparable (cf. table D.2). Toutefois, pour obtenir ces résultats, nous avons employé uniquement 10 réplications bootstrap pour la méthode à complexité variable, et 100 réplications pour l'approche à seuil fixe. En effet, moins de réplications bootstraps sont nécessaires pour la première méthode puisqu'elle n'a besoin d'estimer qu'un seul paramètre (le niveau de complexité $\hat{\alpha}$) alors que l'approche à seuil fixe doit estimer un nombre de paramètres proportionnel au nombre d'espèces (les *bp* des $O(n)$ arêtes).

Un certain nombre d'expériences complémentaires montrent que les tendances constatées ci-dessus sont retrouvées pour des tailles de séquences différentes et pour des nombres différents d'espèces considérées.

Annexe E

Simulations concernant la méthode Q^*

Nous donnons ici les résultats de simulations effectuées pour la méthode Q^*, décrite au chap. 2, part. III. Les simulations sont réalisées suivant le schéma défini précédemment. La seule nouvelle condition d'évolution testée est une condition où les taux d'évolution ont une vitesse constante de $0,05$ substitution par site et par unité de temps, à mi chemin entre les taux faibles $(0,01)$ et élevés $(0,1)$ caractérisant les autres conditions d'évolution.

La table E.1 montre les valeurs obtenues pour les différents termes d'erreurs définis ci-dessus, pour la méthode Q^* et pour la méthode NJ. Chaque valeur contenue dans cette table est une moyenne obtenue sur 2500 jeux de données.

Erreurs de 1ère espèce et de 2ème espèce

Les simulations confirment que la la méthode Q^* infère peu d'arêtes incorrectes (faibles valeurs de e_1), et ce, de façon intéressante, sur l'ensemble des conditions étudiées. La plupart du temps $e_1(Q^*) < 0,1$ ce qui correspond à l'inférence d'une seule arête incorrecte sur 10 phylogénies inférées. Un maximum de $0,33$ est constaté pour la plus difficile condition analysée (taux d'évolution inégaux selon les sites), sur laquelle FPM peut être partiellement inconsistante. La contrepartie du fait que Q^* n'infère presque uniquement que des arêtes "sûres", est qu'elle infère des phylogénies qui sont en moyenne moins résolues que la phylogénie t (comme l'indique les valeurs élevées du terme BS). Ainsi, une erreur systématique en e_2 vis-à-vis de t ne peut être évitée : on a forcément $e_2(Q^*) \geq t - |T^*|$, ce qui explique les importantes valeurs de la table pour $e_2(Q^*)$.

Taux	#sites	méth.	$\|t_R\|$	e_1^R	e_2^R	e_1	e_2	e_{RF}	$2A$	$2V$	BS
faible	300	NJ	4,10	2,98	0,08	2,46	2,46	4,93	0,14	4,80	0,00
		Q^*	-	0,09	0,36	0,07	3,24	3,31	0,00	0,14	3,17
	1000	NJ	5,77	1,30	0,07	0,98	0,98	1,96	0,04	1,92	0,00
		Q^*	-	0,07	0,46	0,05	1,67	1,72	0,00	0,10	1,62
medium	300	NJ	6,21	1,05	0,26	0,82	0,82	1,65	0,02	1,64	0,00
		Q^*	-	0,06	1,19	0,04	1,97	2,01	0,00	0,08	1,93
	1000	NJ	6,76	0,43	0,20	0,37	0,37	0,73	0,02	0,72	0,00
		Q^*	-	0,02	0,84	0,01	1,08	1,09	0,00	0,02	1,07
élevé	300	NJ	6,65	0,73	0,38	0,62	0,62	1,25	0,00	1,24	0,00
		Q^*	-	0,03	1,42	0,02	1,77	1,79	0,00	0,04	1,74
	1000	NJ	6,91	0,29	0,19	0,25	0,25	0,51	0,00	0,50	0,00
		Q^*	-	0,01	0,87	0,01	0,96	0,96	0,00	0,00	0,95
faible/élevé par arêtes	300	NJ	5,48	1,93	0,41	1,55	1,55	3,10	0,14	2,96	0,00
		Q^*	-	0,13	1,43	0,10	2,91	3,01	0,08	0,12	2,82
	1000	NJ	6,36	1,05	0,42	0,87	0,87	1,74	0,08	1,66	0,00
		Q^*	-	0,11	1,50	0,08	2,11	2,18	0,10	0,06	2,03
faible/élevé par sites	300	NJ	6,66	1,65	1,31	1,56	1,56	3,12	0,96	2,16	0,00
		Q^*	-	0,25	2,75	0,24	3,07	3,31	0,36	0,12	2,84
	1000	NJ	6,89	1,31	1,20	1,28	1,28	2,56	1,38	1,18	0,00
		Q^*	-	0,31	2,25	0,31	2,35	2,66	0,56	0,06	2,04

TAB. E.1 – *Erreur structurelle des méthodes de reconstruction NJ et Q^*. $|t_R|$ dénote le nombre moyen d'arêtes internes de la phylogénie correcte* réalisée, t_R. *e_1^R est le nombre moyen d'arêtes inférées qui sont incorrectes vis-à-vis de t_R, et e_2^R est le nombre moyen d'arêtes de t_R qui ne sont pas retrouvées. De façon similaire, e_1 est l'erreur moyenne de 1ère espèce, relativement à la phylogénie correcte* complete, *t, et e_2 l'erreur moyenne de 2ème espèce relativement à t. Le terme A représente le biais fondamental induit en moyenne par une méthode, V est la variance moyenne, et BS le biais structurel moyen.*

La relation mise en évidence ici, entre la taille de la phylogénie inférée et la répartition de l'erreur entre les composantes e_1 et e_2, exprime bien, encore une fois, le dilemme entre les deux catégories d'estimateurs qu'on peut utiliser pour retrouver une phylogénie inconnue :

- Les méthodes inférant des phylogénies totalement résolues et possédant un nombre non-négligeable d'arêtes non-fiables (la grande majorité des méthodes usuelles entrent dans cette catégorie).
- Les méthodes n'inférant que des arêtes jugées fiables et donc des phylogénies partiellement résolues (c'est le cas des méthodes présentées dans cette thèse).

En accord avec cette remarque, on constate dans la table E.1 que $e_1(NJ) > e_1(Q^*)$ et que $e_2(NJ) < e_2(Q^*)$. Si on accorde la même importance aux termes e_1 et e_2, comme le fait la mesure e_{RF}, la table indique de meilleurs résultats moyens pour NJ que pour Q^* dans la majorité des conditions examinées. Toutefois, donner une importance égale à e_1 et e_2 est arbitraire. Si on donne à e_1 deux fois plus d'importance qu'à e_2 (*i.e.*, pour $\lambda = 2$), la méthode Q^* induit une erreur totale plus faible que la méthode NJ sur *toutes* les conditions testées ici.

Quand on mesure l'erreur structurelle vis-à-vis de la phylogénie réalisée, t_R, plutôt que vis-à-vis de la phylogénie complete t, les différences en terme de e_{RF} entre NJ et Q^* s'atténuent (4 fois sur 10, $e^R_{RF}(Q^*) < e^R_{RF}(NJ)$) et dans les conditions les plus favorables (p. ex. taux faible, 300 sites), Q^* infère une phylogénie qui est plus proche en taille de t_R que ne l'est la phylogénie inférée par NJ. Dans ces cas là, une quantité non-négligeable de e_1 est inévitablement induite pour NJ : $e^R_1(NJ) > |t| - |t_R|$ (ceci est confirmé par les résultats des simulations). De plus, $e_1(NJ) \approx e^R_1(NJ)$, ce qui explique que NJ ne montre pas de très bons résultats quand $|t_R| \ll 7$.

Biais et de variance

La table E.1 montre une claire opposition entre les méthodes Q^* et NJ en termes de biais et de variance. Les résultats expriment encore une fois le dilemme existant entre :

- inférer une phylogénie totalement résolue (de BS nul) possédant des arêtes non-fiables et donc extrêmement variables (valeur de V élevée) ;
- inférer une phylogénie partiellement résolue (valeur de BS élevée) mais ne contenant presque uniquement que des arêtes fiables, retrouvées d'un jeu de données sur l'autre pour la même phylogénie modèle (faible V).

Comme au chap. 1, part. II, on observe des valeurs faibles pour le terme d'acentralité, A, et ce pour les deux méthodes. Ceci indique qu'elles ne souffrent d'aucun biais fondamental (*i.e.*, d'aucun biais au sens statistique du terme). Le terme A est constaté plus faible pour Q^* que pour NJ, toutefois on ne peut pas tirer de conclusion véritable de cette observation puisque cette mesure est corrélée avec la taille de la phylogénie inférée (A augmente avec le degré de résolution de la phylogénie inférée), différence qui oppose clairement les méthodes NJ et Q^*.

Influence de divers paramètres

Comme dans les études menées précédemment (chap. 1 et 2, part. II), la condition correspondant à un taux d'évolution élevé est la plus favorable pour les méthodes de reconstruction, et la condition correspondant à des taux inégaux selon les sites est la plus difficile. Rappelons que cette dernière condition viole une hypothèse du modèle de correction des données, ce qui se traduit par une augmentation sensible des termes A et e_1, pour NJ et Q^*.

Les résultats au sens de t_R sont intéressants dans la mesure où ils indiquent clairement, que lorsque l'on ne dispose pas d'une quantité suffisante de données, il n'est pas raisonnable de chercher à retrouver la phylogénie t dans son intégralité, *i.e.*, des phylogénies totalement résolues ne devraient pas être proposés. Dans la table E.1, c'est particulièrement évident pour les conditions où les taux d'évolution sont faibles, ou quand seulement 300 sites sont disponibles, ce qui est pourtant loin d'être le nombre minimum de sites utilisés dans les jeux de données réels pour reconstruire des phylogénies sur 10 espèces. Dans un tel cas il est bien préférable d'inférer une phylogénie partiellement résolue (voir [Kum96] pour une discussion sur ce point). Les résultats des différentes méthodes s'améliorent très sensiblement quand on passe de 300 à 1000 sites, ce qui résulte du fait qu'on dispose d'un nombre suffisant de nucléotides pour que les méthodes parviennent à extraire des données des informations sur la plupart des arêtes de t. En conséquence, l'écart entre les méthodes Q^* et NJ est plus réduit que pour 300 sites.

Des expériences complémentaires (résultats non montrés ici) indiquent que le nombre d'espèces étudiées n'a qu'une légère influence sur les résultats, au sens où les tendances constatées ici pour la méthode Q^* relativement à la méthode NJ ne sont pas modifiées. Toutefois, l'écart entre les méthodes est proportionnellement réduit quand le nombre d'espèces considérées augmente.

Annexe F

Simulations concernant la variante *AddQuad* du schéma ***HAQ*** décrit au chapitre 3, de la partie *III*

Le schéma de simulation suit celui décrit lors des annexes précédentes (cf. aussi section 1.4 chap. 1 part. *II*).

Efficacité de *AddQuad* pour compléter la phylogénie T^*

Les tables F.2 et F.3 donnent les termes d'erreurs obtenus en moyenne par les méthodes Q^*, $Addquad(\emptyset)$ et $Addquad(T^*)$ pour reconstruire l'ensemble de la phylogénie correcte t (table F.1) et vis-à-vis de la phylogénie correcte réalisée t_R (table F.2). Chaque valeur de ces tables est une moyenne obtenue sur 20 phylogénies $\times$ 50 jeux de données. Chaque jeu de données contient 10 espèces.

Dans tous les cas examinés, l'algorithme $Addquad(T^*)$ améliore les résultats en moyenne de l'algorithme Q^* pour estimer la phylogénie complète t, et dans tous les cas sauf un (taux inégaux selon les arêtes) pour estimer la phylogénie réalisée t_R. Le gain est inégal selon les conditions d'évolution : on constate une gain faible pour des taux d'évolution faibles (gain $< 7\%$ pour estimer t) et un gain important pour des taux d'évolution élevés (gain $> 23\%$ pour estimer t et t_R). Ceci semble indiquer que pour des taux élevés la méthode Q^* est trop prudente.

Les résultats de ces tables donnent aussi confirmation du fait qu'une méthode de reconstruction gagne à débuter la reconstruction depuis la base saine que constitue la phylogénie T^* (ceci a aussi été montré par Rice *et al* [RSWY97]) :

	Q^*			$Addquad(\emptyset)$			$Addquad(T^*)$		
cond^o	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$
cf 100	4,57	0,05	4,52	4,89	0,60	4,29	4,51	0,08	4,43
cf 300	3,16	0,06	3,10	3,17	0,57	2,6	**2,96**	0,12	2,84
i/a 100	3,59	0,18	3,41	3,58	1,01	2,57	**3,40**	0,64	2,76
i/a 300	3,19	0,10	3,09	2,73	0,79	1,94	**2,67**	0,61	2,06
ce 100	3,23	0,06	3,16	2,49	0,72	1,78	2,49	0,55	1,94
ce 300	1,78	0,04	1,75	1,31	0,59	0,72	1,30	0,31	0,99

TAB. F.1 – *Efficacité de Addquad pour compléter la phylogénie T^* dans le but d'estimer la phylogénie complète t, pour différentes conditions d'évolution: taux constants faibles (cf), taux cosntants élevés (ce) ou taux inégaux selon les arêtes (i/a). Les valeurs e_{RF} en caractère gras indiquent quand une méthode est significativement meilleure que les autres suivant cette mesure (α=5%).*

Addquad(T^*) induit une estimation de t et de t_R de moindre erreur (en moyenne) que *Addquad*($\emptyset$), et ce pour toutes les conditions d'évolution examinées. Le gain est très important pour des taux d'évolution faibles (au maximum $\approx$ 79% pour estimer t_R disposant de 100 sites), mais faible ($\approx$ 1%) pour des taux élevés quand on mesure l'erreur au sens de t. Ce gain faible s'explique par le fait que, pour un taux d'évolution élevé, suffisamment d'information sur les arêtes de t est contenue dans les données pour que *Addquad* fasse peu d'erreurs, *i.e.*, les arêtes de la phylogénie T^* sont souvent retrouvées par *Addquad*($\emptyset$). Dans ce cas, le fait que *Addquad* parte de la phylogénie étoile ou de la phylogénie T^* ne change pas grand chose et les phylogénies inférées par *Addquad*(T^*) et *Addquad*($\emptyset$) sont similaires.

On peut voir la méthode *Addquad*(T^*) comme un compromis entre Q^* et *Addquad*($\emptyset$), puisque, pour toutes les conditions testées, on a $e_1(T^*) < e_1(\hat{t}_A) < e_1(\hat{t}_{A^*})$ et, pour toute sauf une (taux faible, 100 sites), on a $e_2(T^*) > e_2(\hat{t}_A) > e_2(\hat{t}_{A^*})$, de même *Addquad*($T^*$) infère une phylogénie dont la taille moyenne est comprise entre celle de la phylogénie inférée par Q^* et celle de la phylogénie inférée par *Addquad*($\emptyset$) (résultats non montrés). Toutefois, *Addquad*(T^*) est un très bon compromis car les résultats en terme de e_{RF} sont meilleurs en moyenne que ceux des autres méthodes, dans tous les cas vis-à-vis de t, et dans tous, sauf un, vis-à-vis de t_R (taux inégaux selon les arêtes pour 100 sites).

	Q^*			$Addquad(\emptyset)$			$Addquad(T^*)$		
cond	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$
cf 100	0,25	0,07	0,18	1,04	0,86	0,18	**0,22**	0,12	0,11
cf 300	0,45	0,08	0,37	1,01	0,87	0,14	**0,32**	0,18	0,14
i/a 100	**1,29**	0,24	1,06	1,84	1,35	0,5	1,39	0,84	0,55
i/a 300	1,68	0,13	1,55	1,63	1,02	0,61	**1,47**	0,79	0,68
ce 100	2,11	0,08	2,03	1,74	0,91	0,83	**1,61**	0,68	0,92
ce 300	1,39	0,05	1,34	1,11	0,70	0,41	**1,02**	0,37	0,65

TAB. F.2 – *Efficacité de Addquad pour compléter la phylogénie T^* dans le but d'estimer la phylogénie réalisée t_R. Les valeurs e_{RF} en caractère gras indiquent quand une méthode est significativement meilleure que les autres suivant cette mesure (α=5%).*

Comparaison avec la méthode NJ

Nous comparons ici le schéma agglomératif basé sur les *quadruplets*, *Addquad*, à la méthode la plus connue en R.P., NJ [SN87], utilisant elle aussi le même schéma (dit aussi de *neighbor joining*), mais sur des *distances*. Les résultats sont détaillés dans la table F.3.

Compléter la phylogénie T^* par la méthode *Addquad* (combinaison d'algorithmes dénotée A^* dans la table) induit des phylogénies plus résolues que T^*, mais qui ne sont toutefois pas totalement résolus (comme l'indique le terme BS non-nul). Ceci indique la capacité de *Addquad* (que ne partagent pas NJ et d'autres algorithmes usuels) à se rendre compte que les arêtes ajoutées à la phylogénie au-delà d'une certaine étape ne sont pas très fiables. La méthode *Addquad*(T^*) a un comportement sain vis-à-vis de t_R, au sens où ses deux termes e_1^R et e_2^R sont presque identiques, indiquant que la phylogénie inférée est de même taille que la phylogénie t_R (ce qui est confirmé par $7 - BS \approx |t_R|$). Comme t_R est précisément la partie de t pour laquelle on dispose d'informations dans les données, ceci traduit que la méthode *Addquad*(T^*) reconnaît, dans une certaine mesure, la quantité d'information utilisable depuis le jeu de données. En conséquence elle donne de meilleures estimations de la phylogénie en termes de e_{RF} et e_{RF}^R que la méthode NJ (cf. table F.3) et que la méthode Q^* (cf. table E.1), pour toutes les conditions d'évolution examinées (sauf une). La différence la plus importante entre *Addquad*(T^*) et NJ est obtenue pour 300 sites, quand $|t_R|$ est peu importante. Plus généralement quand t_R contient peu d'arêtes, les méthodes de quadruplets

Taux	#sites	méth.	$\lvert T_R\rvert$	e_1^R	e_2^R	e_1	e_2	e_{RF}	$2A$	$2V$	BS
faible	300	NJ	4,10	2,98	0,08	2,46	2,46	4,93	0,14	4,80	0,00
		A^*	-	0,22	0,14	0,16	2,99	**3,15**	0,02	0,30	2,83
	1000	NJ	5,77	1,30	0,07	0,98	0,98	1,96	0,04	1,92	0,00
		A^*	-	0,24	0,17	0,18	1,33	**1,51**	0,02	0,34	1,16
medium	300	NJ	6,21	1,05	0,26	0,82	0,82	1,65	0,02	1,64	0,00
		A^*	-	0,42	0,45	0,32	1,14	**1,46**	0,00	0,64	0,82
	1000	NJ	6,76	0,43	0,20	0,37	0,37	0,73	0,02	0,72	0,00
		A^*	-	0,28	0,35	0,17	0,52	0,69	0,02	0,32	0,35
élevé	300	NJ	6,65	0,73	0,38	0,62	0,62	1,25	0,00	1,24	0,00
		A^*	-	0,32	0,59	0,26	0,88	**1,14**	0,00	0,52	0,62
	1000	NJ	6,91	0,29	0,19	0,25	0,25	0,51	0,00	0,50	0,00
		A^*	-	0,11	0,34	0,10	0,41	0,51	0,00	0,20	0,31
faible/élevé selon arêtes	300	NJ	5,48	1,93	0,41	1,55	1,55	3,10	0,14	2,96	0,00
		A^*	-	0,82	0,63	0,63	1,95	**2,58**	0,22	1,02	1,33
	1000	NJ	6,36	1,05	0,42	0,87	0,87	1,74	0,08	1,66	0,00
		A^*	-	0,67	0,74	0,54	1,25	1,79	0,36	0,72	0,71
faible/élevé selon sites	300	NJ	6,66	1,65	1,31	1,56	1,56	3,12	0,96	2,16	0,00
		A^*	-	1,51	1,39	1,42	1,64	3,06	0,90	1,94	0,21
	1000	NJ	6,89	1,31	1,20	1,28	1,28	2,56	1,38	1,18	0,00
		A^*	-	1,24	1,26	1,22	1,34	2,56	1,32	1,10	0,12

TAB. F.3 – *Erreur structurelle des méthodes de reconstruction NJ et Addquad(T^*) (notée ici A^*). Le terme $|t_R|$ dénote le nombre moyen d'arêtes internes de la phylogénie correcte* réalisé, *t_R ; e_1^R est le nombre moyen d'arêtes inférées qui sont incorrectes vis-à-vis de t_R ; e_2^R est le nombre moyen d'arêtes de t_R qui ne sont pas retrouvées. De façon similaire, e_1 est l'erreur moyenne de 1ère espèce, relativement à la phylogénie correcte* au complet, *t ; e_2 est l'erreur moyenne de 2ème espèce relativement à t. Le terme A représente le biais fondamental induit en moyenne par une méthode, V est la variance moyenne, et BS le biais structurel moyen. Les chiffres en caractère gras indiquent les cas où A^* est significativement meilleur que NJ au sens de e_{RF} (α=5%).*

sont plus efficaces que NJ puisque celle-ci produit toujours une phylogénie totalement résolue. Ainsi, pour certaines arêtes non traduites dans le jeu de données, NJ doit deviner aléatoirement comment est résolu t, tandis que les méthodes de quadruplets préfèrent ne pas proposer d'arête dans un tel cas.

Comme précédemment, des expériences complémentaires montrent que le nombre d'espèces considérées (ici 10) a une faible influence sur les performances relatives des méthodes.

Annexe G

Jeu de données sur les 11 mammifères

Données issues des hémoglobines α et β, des fibrinopeptides A et B, du cytochrome c, de la myoglobine, du christallin α. Le mouton et le chimpanzé n'ayant pas de séquence pour le christallin α, un état spécifique (différent de tous les autres états) leur est assigné.

Singe	ctcggggggattacggaaactaggaaaggcagctgcagaggccacaaacacccccatctcccg
Mouton	gaggggggcattacgggcaaaagccggagactgtaacagaaacaggcacgctccgatcaggcg
Cheval	gacgggggattgcgcaaactagccggggacgttaacgaggcgtggcttgcccagctcaaccg
Kangourou	ggggaggggacaacacaagcaaaacggaaaagctgaaggcgccagaaacagcaaaccaaagag
Rongeur	gtcaggaggatcacggaagaaccccggggcatcagccgaggcatggctcggcccgatctcgcg
Lapin	cacagaaggattgaacaaaaaagccgggacaaattcagaagcgaatcacgcccagatctcgaa
Chien	cacaaaaggccaacgcaaaacacccggggcatgaacaggcgccagtctcggcccactcaccca
Cochon	ggcggggcgattgagcgaaaaagcaggggcatctaacaaagcctggcttggcccgatcaagcg
Homme	cacggggcgattacggaaaaaccgaaaggcagctgcagaggccactcacgccccgatcacccg
Vache	ggggggggcattacgggcaaaaaccggggaatctagcgaaaacaggcacggtacgacaacgca
Chimpanzé	cacggggcgcttacggaaaaaccgaaaggcagctgcagaggccactcacgccccgatcacccg

Singe	gcgggaccaaccacaagctgataacagcagagggaggacccgagagagaagcctagaattcat
Mouton	acgatacaaaacggcagcgtgtcagtcatgcgtggggaaccgagggcaactcctgagcagctt
Cheval	gccatgagcaccagcggcgtgtacgaccacagagaggaacagaggacaagtactgagcagcta
Kangourou	ccgtggcaaaaaaaaaaaatataagaacgcaaagagacaacgagggtaaagagggagcagcta
Rongeur	ggacgaccaaccatcgaaatacacgcgctgaggacagaaaacagcaagtgaattaggcaggtt
Lapin	acgcgaacagccaaaggcatataacagcagcgggaaacaaacagcgagagtttcgagcaggtt
Chien	ggaataccaaaaaaaagctgataacagcaataagaggaaacgaggatgtagacggagcagctt
Cochon	gcgataacaaacaaaggagtgtccgtccaggcaaaggaaccgaggccaagtcaggagcagctt
Homme	gcgggaccgcccaacggctgataaccgcagtgggaggacccgtaagagaggtttagaattcat
Vache	acgatacacaacgcaggcgtgccagtcatgcccagggacccgaggccaagtggtgggcagctt
Chimpanzé	gcgggaccgcccaacagctgataaccgcagtgggaggacccgtaagagaggtttagaattcat

Singe	ccctacgtgacaggtacgtacccgatcaatggacagtaaagaactcacacaaaagctaagcc
Mouton	acacgagcgggcggaccaagagagcgcccgaggccggcatnnnnnnnnnnnnnnnnnnnnn
Cheval	acaccagtgagcggaccacccccacgccctggcccagaaaaggctcaaacacatgagcaggc
Kangourou	atccgcataggaggtacatccccgcgcaatgcaacggaaaaagtcatcgaaactaaacctat
Rongeur	acccgcgtgcgaaatagaaacccaagcaatgggccagaaagagctctaacgaatgcgcagcc
Lapin	acccacgtgcgaactacaaacccgagccctgcgccagcaagaactcaaacgaatgcgcagcc
Chien	acacgcatacgaaaaagaaccccgcgcaatgcaacagaaaaggctctaacgcatgggcagcc
Cochon	acacgagtgggaggtacaaccccgagcaatgggccggaataggctcaaacgcatgggcagcc
Homme	ctctacgtgacaggtacgcacccgatgaatggacagtacagaactcacacaaaagctaagac
Vache	acacgagcgggcggaccaagagaacgcccgaggcctgaatagactcaaacacatgggcagcc
Chimpanzé	ctctacgtgataggtacgcacccgatgaatggacagtacaxxxxxxxxxxxxxxxxxxxxx

Liste des Algorithmes

Index

Bibliographie

[ABF$^+$96] R. Agarwala, V. Bafna, M. Farach, B. Narayanan, M. Paterson, and M. Thorup. On the approximability of numerical taxonomy: Fitting distances by tree metrics. In *Proc. of the 7th Annual ACM-SIAM Symposium on Discrete Algorithms*, 1996.

[ADFK98] A. Ambainis, R. Desper, M. Farach, and S. Kannan. Nearly tight bounds on the learnability of evolution. In *FOCS*, 1998.

[Ald95] D.J. Aldous. *Mathematics and its application*, volume 76 of *IMA*, chapter Probability distributions on cladograms, pages 1–18. Springer Verlag, 1995.

[And97] G. Andrieu. *Estimation par intervalle d'une distance évolutive.* PhD thesis, Univ. Montpellier II, décembre 1997.

[AS92] N. Alon and J.H. Spencer. *The Probabilistic Method.* John Wiley and Sons, New York, 1992.

[AS98] ACM-SIGACT, editor. *Ordinal beats additive*, volume Proc. of the 2nd Annual International Conference on Computational Molecular Biology (RECOMB), 1998.

[ASSU81] A. Aho, Y. Sagiv, T.G. Szymanski, and J.D. Ullman. Inferring a tree from the lowest common ancestors with an application to the optimization of relational expressions. *SIAM J. of Computing*, 10(3):405–421, 1981.

[Att97a] K. Atteson. Communication personnelle, 1997.

[Att97b] K. Atteson. On the performance of neighbor-joining. In *Proc. of COCOON 97*, LNCS. Springer Verlag, 1997.

[Att97c] K. Atteson. The performance of neighbor-joining algorithms of phylogeny reconstruction. Submitted to Algorithmica, 1997.

[AU87] A.V. Aho and J.D. Ullman. *Concepts fondamentaux de l'informatique*. InterEditions, 1987.

[Ban96] H-J. Bandelt. communication personnelle, 1996.

[BD86] H-J. Bandelt and A. Dress. Reconstructing the shape of a tree from observed dissimilarity data. *Adv. in appl. math.*, 7:309–343, 1986.

[BD92] H.-J. Bandelt and A. Dress. A canonical decomposition theory for metrics on a finite set. *Advances Math*, 92:47–105, 1992.

[Ber70] C. Berge. *Graphes et hypergraphes*. Dunod, Paris, 1970.

[BFOC84] L. Breiman, J.H. Friedman, A. Olshen, and Stone C.J. *Classification and regression trees*. Wadsworth, 1984.

[BFW92] H.L. Bodlaender, M.R. Fellows, and T. Warnow. Two strikes against perfect phylogeny. In *Actes de ICALP*, LNCS, pages 273–287. Springer, 1992.

[BG88] J.-P. Barthélemy and Alain Guénoche. *Les arbres et les représentations des proximités*. Méthodes + Programmes. Masson, 1988.

[BG91] J.P. Barthélemy and A. Guénoche. *Trees and proximities representations*. Wiley, 1991.

[BG96a] V. Berry and O. Gascuel. On the interpretation of bootstrap trees: appropriate threshold of clades selection and induced gain. *Mol. Biol. Evol.*, 13(7):999–1011, 1996.

[BG96b] V. Berry and O. Gascuel. A quartet-based heuristics for phylogeny reconstruction. Technical Report 96-044, LIRMM, 1996.

[BG97] V. Berry and O. Gascuel. Inferring evolutionary trees with strong combinatorial confidence. In T. Jiang, editor, *Proc. of the 3rd Ann. Int. Computing and Combinatorics Conf.*, LNCS, pages 111–123. Springer Verlag, 1997.

[BL84] J.-P. Barthélemy and N.X. Luong. Représentations arborées de mesures de dissimilarité. In *Colloque ASSU*, La Grande-Motte, 1984.

[Bro94] J.K.M. Brown. Probabilities of evolutionary trees. *Syst. Biol.*, 43(1):78–91, 1994.

[Bun71] P. Buneman. *Mathematics in Archeological and Historical Sciences*, chapter The recovery of trees from measures of dissimilarity, pages 387–395. Edhinburgh University Press, 1971.

[Cav78] J. A. Cavender. Taxonomy with confidence. *Math. Biosci.*, 40:271–280, 1978.

[Cav81] James A. Cavender. Tests of phylogenetic hypotheses under generalized models. *Math. Biosci.*, 54:217–229, 1981.

[CF97] V. Chepoi and B. Fichet. l_∞-approximation via subdominants, 1997. soumis.

[CGAL95] G. Caraux, O. Gascuel, G. Andrieu, and D. Levy. Approches informatiques de la reconstruction d'arbres phylogénétiques à partir de données moléculaires. *Tech. Sci. Inf.*, 14(2):113–139, 1995.

[CGAS88] R. Cedergren, M.W. Gray, Y. Abel, and D. Sankoff. The evolutionary relationships among known life forms. *J. Mol. Evol.*, 28:98–112, 1988.

[CLR94] T. Cormen, C. Leiserson, and R. Rivest. *Introduction à l'algorithmique*. Dunod (pour l'édition française), 1994.

[CR61] R. Courant and H. Robbins. *What is Mathematics?* Oxford Univ. Press, 1961.

[CS65] J.H. Camin and R.R. Sokal. A method for deducing branching sequences in phylogeny. *Evolution*, 19:311–326, 1965.

[CS79] H. Colonius and H.H. Schulze. Repräsentation nichtnümerischer Ähnlichkeitsdaten durch baumstrukturen. *Psych. Beitr.*, pages 98–111, 1979.

[CS81] H. Colonius and H.H. Schulze. Tree structures for proximity data. *British Journal of Math. and Stat. Psychology*, 34:167–180, 1981.

[CSE67] L. L. Cavalli-Sforza and A. W. F. Edwards. phylogenetic analysis models and estimation procedures. *Amer. J. Human Genetics*, 19(3):233–257, 1967.

[Day86] W.H.E. Day. The computational complexity of inferring rooted phylogenies by parsimony. *Math. Biosci.*, 81:33–42, 1986.

[Day87] W. H. E. Day. Computational complexity of inferring phylogenies from dissimilarity matrices. In *Bulletin of Mathematical Biology*, volume 49, pages 461–467. 1987.

[DeB92] R. W. DeBry. The consistency of several phylogeny-inference methods under varying evolutionary rates. In *Mol. Biol. Evol.*, volume 9, pages 537–551. 1992.

[dF36] P. de Fermat. *Oeuvres de Fermat*, chapter Maxima et minima, pages 133–174. Gauthier-Villars, Paris, 1836.

[dGP+96] A.M. d'Erchia, C. Gissi, G. Pesole, C. Saccone, and U. Arnason. The guinea-pig is not a rodent. *Nature*, 381:597–600, 1996.

[dQG94] A. de Queiroz and C. Gauthier. Toward a phylogenetic system of biological nomenclature. *TREE*, 9:27–31, 1994.

[dS83] Geert de Soete. A least squares algorithm for fitting additive trees to proxumity data. *Psychometrika*, 48(4):621–626, 1983.

[DS86] W.H.E. Day and D. Sankoff. Computational complexity of inferring phylogenies by compatibility. *Syst. Zool.*, 35(2):224–229, 1986.

[DT93] P. Darlu and P. Tassy. *Reconstruction phylogénétique*. Concepts et méthodes. Masson, 1993.

[DvHK86] A. Dress, A. von Haeseler, and M. Krueger. Reconstructing phylogenetic trees using variants of the "four-point condition". *Studien zur Klassifikation*, pages 299–305, 1986.

[EC80] N. Eldredge and M. Cracraft. *Phylogenetic patterns and the evolutionary process*. Columbia University Press, 1980.

[ECS63] A.W.F. Edwards and L.L. Cavalli-Sforza. The reconstruction of evolution. *Ann. Hum. Genet.*, 27:104–105, 1963.

[ECS64] A.W.F. Edwards and L.L. Cavalli-Sforza. chapter The reconstruction of evolutionary trees. Systematics Association, London, 1964.

[ED66] R.V. Eck and M.O. Dayhoff. *Altlas of protein sequence and structure.* National Biomedical Researh Foundation, Silver Spring, 1966.

[Efr79] B. Efron. Bootstrap methods: another look at the jackknife. *Ann. Statistics*, 7:1–26, 1979.

[EHH96] B. Efron, E. Halloran, and S. Holmes. Bootstrap confidence levels for phylogenetic trees. *Proc. Natl. Acad. Sci. USA*, 93:13429–13434, 1996.

[EJM76] G.F. Estabrook, C.S. Johnson, and F.R. McMorris. A mathematical foundation for the analysis of cladistic character compatibility. *Math. Biosci.*, 29:181–187, 1976.

[ESSW97a] P.L. Erdös, M.A. Steel, L.A. Székely, and T. Warnow. A few logs suffice to build (almost) all trees (i). soumis, 1997.

[ESSW97b] P.L. Erdös, M.A. Steel, L.A. Székely, and T.J. Warnow. Inferring big trees from short quartets. In *Proc. of ICALP'97*, 1997.

[Est72] G.F. Estabrook. Cladistic methodology: a discussion of the theoretical basis for the induction of evolutionary history. *Ann. Rev. Ecol. Syst.*, 3:427–456, 1972.

[EWOD88] M. Eigen, R. Winkler-Oswatitsch, and A. Dress. Statistical geometry in sequence space : A method of quantitative comparative sequence analysis. *Proc. Natl. Acad. Sci. USA*, 85:5913–5917, 1988.

[Far70] J.H. Farris. Methods for computing wagner trees. *Syst. Zool.*, 19:83–92, 1970.

[Far72] J.S. Farris. Estimating phylogenetic trees from distance matrices. *Am. Nat.*, 106:645–668, 1972.

[Far73] James S. Farris. A probability model for inferring evolutionary trees. *Syst. Zool.*, 22:250–256, 1973.

[Far77] J.S. Farris. Phylogenetic analysis under dollo's law. *Syst. Zool.*, 26:77–88, 1977.

[Far86] J.S. Farris. On the boundaries of phylogenetic systematics. *Cladistics*, 2:14–27, 1986.

[Fel78a] J. Felsenstein. Cases in which parsimony and compatibility methods will be positively misleading. *Syst. Zool.*, 27:401–410, 1978.

[Fel78b] J. Felsenstein. The number of evolutionary trees. *Syst. Zool.*, 27:27–33, 1978.

[Fel81] J. Felsenstein. Evolutionary trees from DNA sequences : a maximum likelihood approach. *J. Mol. Evol.*, 17:368–376, 1981.

[Fel85] J. Felsenstein85. Confidence Limites on Phylogenies : an approach using the bootstrap. *Evolution*, 39(4):783–791, 1985. (*Mesure de variabilite*).

[Fel88] J. Felsenstein. Phylogenies from molecular sequences: inference and reliability. *Annu. Rev. Genet.*, 22:521–565, 1988.

[Fel93] J. Felsenstein. Phylip (phylogeny inference package) version 3.5c. Department of genetics, University of Washington, Seattle., 1993.

[Fit71] Walter M. Fitch. Toward defining the course of evolution : minimum change for a specific tree topology. *Syst. Zool.*, 1971.

[Fit81] W. M. Fitch. A non-sequential method for constructing trees and hierarchical classifications. *J. Mol. Evol.*, 18:30–37, 1981.

[FK93] J. Felsenstein and H. Kishino. Is there something wrong with the bootstrap on phylogenies? a reply to Hillis and Bull. *Syst. Biol.*, 42:193–200, 1993.

[FK96] M. Farach and S. Kannan. Efficient algorithms for inverting evolution. In *Proceedings of the 28th Annual ACM Symposium on the Theory of Computing*, pages 230–235, 1996.

[FM67] Walter M. Fitch and Emanuel Margoliash. Construction of phylogenetic trees. *Science*, 155:279–284, 1967.

[Fou84] L.R. Foulds. Maximum savings in the steiner problem in phylogeny. *J. Theor. Biol.*, 107:471–474, 1984.

[FS86] J. Felsenstein and E. Sober. Parsimony and likelihood: an exchange. *Syst. Zool.*, 35:617–626, 1986.

[FY91] W.M. Fitch and J. Ye. *Phylogenetic analysis of DNA sequences*, chapter Weighted parsimony: does it work?, pages 147–154. Oxford Univ. Press, 1991.

[Gal97] N. Galtier. *L'approche statistique en phylogénie moléculaire : influence des compositions en bases variables.* PhD thesis, Université Claude Bernard - Lyon I, 1997.

[Gal98] N. Galtier. Communication personnelle, 1998.

[Gas94] O. Gascuel. A note on sattah and tversky's, saitou and nei's, and studier and keppler's algorithms for inferring phylogenies from evolutionary distances. *Mol. Biol. Evol.*, 11:961–963, 1994.

[Gas97a] O. Gascuel. Communication personnelle, 1997.

[Gas97b] Olivier Gascuel. Concerning the NJ algorithm and its unweighted version UNJ. In B. Mirkin, F.R. McMorris, F.S. Roberts, and A. Rzhetsky, editors, *Mathematical Hierarchies and Biology*, pages 149–170. Am. Math. Society, Providence, USA, 1997.

[GGP+96] L.A. Goldberg, P.W. Goldberg, C.A. Phillips, E. Sweedyk, and T. Warnow. Minimizing phylogenetic number to find good evolutionary trees. *Disc. Appl. Math.*, 71:111–136, 1996.

[GHL91] D. Graur, W.A. Hide, and W.-H. Li. Is the guinea-pig a rodent? *Nature*, 351:649–652, 1991.

[GJ79] M. Garey and D. Johnson. *Computers and Intractability - A guide to the Theory of NP-Completeness.* Freeman, San Fransisco, 1979.

[GL96] O. Gascuel and D. Levy. A reduction for approximating a (non-metric) dissimilarity by a tree distance. *J. of Classification*, 1996.

[GP68] E.N. Gilbert and H.O. Pollack. Steiner minimal trees. *SIAM J. on Appl. Math.*, 16:1–29, 1968.

[Gué97a] A. Guénoche. Order distance associated with a hierarchy. *J. of Classification*, 14:101–115, 1997.

[Gué97b] A. Guénoche. Ordinal properties of tree distances. *Disc. Math.*, à paraître, 1997.

[Gus91a] D. Gusfield. Efficient algorithms for inferring evolutionnary trees. *Networks*, 21:19–28, 1991.

[Gus91b] D. Gusfield. The steiner tree problem, historical reconstruction and phylogeny. Technical report, Comp. Sci. Division, University of California, 1991.

[Har71] E.F. Harding. The probabilities of rooted tree shapes generated by random bifurcation. *Adv. Appl. Math.*, 3:44–77, 1971.

[Har72] J. Hartigan. Minimum mutation fits to a given tree. *Biometrics*, 29:53–65, 1972.

[HB93] D.M. Hillis and J.J. Bull. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst. Biol.*, 42:182–192, 1993.

[Hei89] J. Hein. A tree reconstruction method that is economical in the number of pairwise comparisons used. *Mol. Biol. Evol.*, 6(6):669–684, 1989.

[Hen50] W. Hennig. *Grundzünge einer Theorie der Phylogenetischen Systematik.* Deutscher Zentralverlag, Berlin, 1950.

[Hen66] W. Hennig. *Phylogenetic systematics.* University of Illinoy press, Urbana, 1966.

[Hen91] M.D. Hendy. A combinatorial description of the closest tree algorihtm for finding evolutionary trees. *Disc. Math.*, 96:51–58, 1991.

[HH92] D.M. Hillis and J.P. Huelsenbeck. Signal, noise, and reliability in molecular phylogenetic analyses. *Journal of Heredity*, 83:189–195, 1992. Examine la distribution des arbres les plus parcimonieux : gaussienne applatie.

[HH93] J.P. Huelsenbeck and D. M. Hillis. Success of phylogenetic methods in the four-taxon case. *Syst. Biol.*, 42(3):247–264, 1993.

[HHS94] D.M. Hillis, J.P. Huelsenbeck, and D.L. Swofford. Hobgoblin of phylogenetics? *Nature*, 369:363–364, 1994.

[HP82] M.D. Hendy and D. Penny. Branch and bound algorithms to determine minimal evolutionary trees. In *Mathematical Biosciences*, volume 59, pages 277–290. Elsevier Science Publishing, 1982.

[HP89] M.D. Hendy and D. Penny. A framework for the quantitative study of evolutionary trees. *Syst Zool*, 38(4):297–309, 1989.

[HPS94] M.D. Hendy, D. Penny, and M.A. Steel. A discrete fourier analysis for evolutionary trees. *Proc. Natl. Acad. Sci. USA*, 91:3339–3343, 1994.

[HPT97] D. Huson, H. Paschke, and U. Toenges. Splitstree 2.0.d5. Package, 1997.

[HRW92] F.K. Hwang, D. Richards, and P. Winter. The steiner tree problem. *Ann. Disc. Math.*, 53, 1992.

[HY64] S.L. Hakimi and S.S. Yau. Distance matrix of a graph and its realizability. *Quaterly of Appl. Math.*, 22:305–317, 1964.

[JC69] Thomas H. Jukes and Charles R. Cantor. Evolution of protein molecules. In H. N. Munro, editor, *Mammalian Protein Metabolism*, volume III, chapter 24, pages 21–132. Academic Press, New York, 1969.

[JL92] T. Jiang and Ming Li. Some optimizations problems related to molecular biology. Technical report, Mc Master Univ. and University of Waterloo, 1992.

[JN90] L. Jin and M. Nei. Limitations of the evolutionary parsimony method of phylogenetic analysis. *Mol. Biol. Evol.*, 7:82–102, 1990.

[Kea97] P.E. Kearney. A six-point condition for ordinal matrices. *J. of Computational Biology*, to appear, 1997.

[KF69] A.G. Kluge and J.S. Farris. Quantitative phyletics and the evolution of anurans. *Syst. Zool.*, 18(1):1–32, 1969.

[KF94] M.K. Kuhner and J. Felsenstein. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.*, 11:459–468, 1994.

[KH90] Hirohisa Kishino and Masami Hasegawa. Converting distance to time : Application to human evolution. In Russsell F. Doolittle, editor, *Molecular evolution : computer analysis of protein and nucleic acid sequences*, volume 183 of *Methods In Enzymology*, chapter 34, pages 550–570. Academic Press, 1990.

[KHM97] P.E. Kearney, R.B. Hayward, and H. Meijer. Inferring evolutionary trees from ordinal data. In *Proc. of the 8th Ann. ACM-SIAM Symposium on Discrete Algorithms*, pages 418–426, 1997.

[Kim80] Motoo Kimura. A simple method for estimating evolutionary rates base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, 16:111–120, 1980.

[Kim93] J. Kim. Improving the accuracy of phylogenetic estimation by combining different methods. *Syst. Biol*, 42:331–340, 1993.

[Kri86] M. Krivanek. Np-hard problems in hierarchical-tree clustering. *Acta Informatica*, 23:311–323, 1986.

[KS73] M. G. Kendall and A. Stuart. *The Advanced Theory of Statistics*, volume II : Inference and relationship. Griffin, London, 1973.

[Kum96] S. Kumar. A stepwise algorithm for finding minimum evolution trees. *Mol. Biol. Evol.*, 13:584–593, 1996.

[KW95] S. Kannan and T. Warnow. Tree reconstruction from partial orders. *SIAM J. Comp.*, 24(3):511–519, 1995.

[Lak86] J.A. Lake. In defense of bacterial phylogeny. *Nature*, 321:657–658, 1986.

[Lak87] James A. Lake. A rate-independent technique for analysis of nucleic acid sequences : Evolutionary parsimony. *Mol. Biol. Evol.*, 4(2):167–191, 1987.

[Lap97] F.-J. Lapointe. How to validate phylogenetic trees? a stepwise procedure. In *Actes des 5emes rencontres de la SFC*, 1997.

[LeQ69] W.J. LeQuesne. A method of selection of characters in numerical taxonomy. *Syst. Zool.*, 18(2):201–205, 1969.

[Leq72] W.J. Lequesne. Further studies based on the uniquely derived character concept. *Syst. Zool.*, 21(3):281–288, 1972.

[LLW85] W.-H. Li, C.C. Luo, and C.-I. Wu. *Evolution of DNA sequences*, volume Molecular Evolutionary Genetics, chapter ., pages 1–94. Plenum Press, New-York, 1985.

[LPLG93] G. Lecointre, H. Philippe, H.L.V. Lê, and H. Le Guyader. Species sampling has a major impact on phylogenetic inference. *Mol. Phylogenet. Evol.*, 2:205–224, 1993.

[LSHP94] P.J. Lockhart, M.A. Steel, M.D. Hendy, and D. Penny. Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol. Biol. Evol.*, 11(4):605–612, 1994.

[MC91] Michael M. Miyamoto and Joel Cracraft, editors. *Phylogenic Analysis of DNA Sequences.* Oxford University Press, New York - Oxford, 1991.

[Mea81] C. Meacham. A manual method for character compatibility. *Taxon*, 30:591–600, 1981.

[MF87] K.B. Mullis and F.A. Faloona. Specific synthesis of dna in vitro via polymerase catalyzed chained reaction. *Methods. in Enzym.*, 155:335–350, 1987.

[MS96] V. Moulton and M.A. Steel. Retractions of finite distance functions onto tree metrics. Technical Report 148, Univ. of Canterbury, 1996.

[Nei87] M. Nei. *Molecular Evolutionary Genetics.* Columbia University Press, New-York, 1987.

[Nei91] M. Nei. Relative efficiencies of different tree-making methods for molecular data. In M.M. Miyamoto and J. Cracraft, editors, *Phylogenetic analysis of DNA sequences.* Oxford Univ. Press, 1991.

[Ney71] J. Neyman. *Statistical decision theory and related topics*, chapter Molecular studies of evolution: a source of novel statistical problems, pages 1–27. Academic, New York: Academic, 1971.

[Nut04] G.H.F. Nutall. *Blood immunity and blood relationship.* Cambridge Univ. Press, Cambridge, 1904.

[PHH87] David Penny, Michael D. Hendy, and I. M. Henderson. Reliability of evolutionary trees. *Cold Spring Harbor Symposia on Quantitative Biology*, 52:857–862, 1987.

[PHS91] D. Penny, M.D. Hendy, and M.A. Steel. Testing the theory of descent. In M. M. Miyamoto and J. Cracraft, editors, *Phylogenetic analysis of DNA sequences*, pages 155–183. 1991.

[PHS92] David Penny, Michael D. Hendy, and Michael A. Steel. Progress with methods for constructing evolutionary trees. *Tree*, 7:73–78, 1992.

[QMB83] L.H. Qu, B. Michot, and J.P. Bachellerie. Improved methods for structure probing in large rnas: a rapid heterologous sequencing approach is coupled to the direct mapping of nuclease accessible sites. application to the 5' terminal domain of eukaryotic. *Nucl. Acid. Res.*, 11:5903–5920, 1983.

[RF81] D. F. Robinson and L. R. Foulds. Comparison of phylogenetic trees. *Math. Biosci.*, 53:131–147, 1981.

[RN92] Andrey Rzhetsky and Masatoshi Nei. Statistical properties of the ordinary least-squares, generalized least-squares, and minimum-evolution methods of phylogenetic inference. *J. Mol. Evol.*, 35:367–375, 1992.

[RN93] A. Rzhetsky and M. Nei. Theoretical foundation of the minimum-evolution method or phylogenetic inference. *Mol. Biol. Evol.*, 10:1073–1095, 1993.

[Roo95] D.S. Roos. A biologist's perspective on interesting questions and practical difficulties in phylogeny reconstruction. In *Actes du Phylogeny Workshop DIMACS*, DIMACS tech. rep. 95-58. Princeton University, 1995.

[Rou88] M. Roux. *Recent Developments in clustering and data analysis*, chapter Techniques of approximation for building two tree structures. Academic Press, 1988.

[RSWY97] K. Rice, M.A. Steel, T. Warnow, and S. Yooseph. Getting better topology estimates of difficult evolutionary trees. Submitted, 1997.

[RW97] K. Rice and T. Warnow. Parsimony is hard to beat! In *Computing and Combinatorics, Proc. of COCOON*, number 1276 in LNCS, pages 124–133. Springer, 1997.

[Sai88] N. Saitou. Property and efficiency of the maximum likelihood method for molecular phylogeny. *J. Mol. Evol.*, 27:261–273, 1988.

[Sai90] N. Saitou. Maximum likelihood methods. In Russsell F. Doolittle, editor, *Molecular evolution : computer analysis of protein and nucleic acid sequences*, volume 183 of *Methods In Enzymology*, chapter 36, pages 584–598. Academic Press, 1990.

[San89] M.J. Sanderson. Confidence limits on phylogenies: the bootstrap revisited. *Cladistics*, 5:113–129, 1989.

[SBPH84] J.T. Staley, M.P. Bryant, N. Pfennig, and J.G. Holt. *Bergey's Manual of Systematic Bateriology.* Williams and Wilkins, Baltimore, 1984.

[SI89] N. Saitou and T. Imanishi. Relative efficiencies of the Fitch-Margoliash, Maximum-Parsimony, Maximum-Likelihood, Minimum-Evolution, and Neighbor-Joining methods of phylogenetic tree construction in obtaining the correct tree. *Mol. Biol. Evol.*, 6(5):514–525, 1989.

[Sid94] A. Sidow. Answer to stewart. *Nature*, 367:26, 1994.

[SK88] J.A. Studier and K.J. Keppler. A note on the neighbor-joining method of saitou and nei. *Mol. Biol. Evol.*, 5(6):729–731, 1988.

[Smo69] Y.A. Smolenskii. A method for linear recoding of graphs. *U.S.S.R. Comput. Math. Phys.*, 2:396–397, 1969.

[SN87] N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstruction phylogenetic trees. *Mol. Biol. Evol.*, 4(4):406–425, 1987.

[SNC77] F. Sanger, S. Nicklen, and A.R. Coulson. Dna sequencing with chain-terminating inhibitors. *Proc. Natl. Sci. USA*, 74:5463–5467, 1977.

[SO90] David L. Swofford and Gary J. Olsen. Phylogeny reconstruction. In David M. Hillis and Craig Moritz, editors, *Molecular Systematics*, chapter 11, pages 411–501. Sinauer Associates, Inc., Sunderland, Massachusetts, USA, 1990.

[Sob83] E. Sober. Parsimony in systematics: philosophical issues. *Ann. Rev. Ecol. Syst.*, 14:335–357, 1983.

[Sob85] E. Sober. A likelihood justification of parsimony. *Cladistics*, 1:209–233, 1985.

[SOWH96] D.L. Swofford, G.J. Olsen, P.J. Wadell, and D.M. Hillis. Phylogenetic inference. In D.M. Hillis, D.M. Moritz, and B.K. Mable, editors, *Molecular systematics (2nd edition)*, pages 407–514. Sunderland, USA, 1996.

[SR75] D. Sankoff and P. Rousseau. Locating the vertices of steiner tree in an arbitrary metric space. *Math. Progr.*, 9:240–246, 1975.

[SS63] R.R. Sokal and P.H. Sneath. *Principles of Numerical Taxonomy.* Freeman, San Francisco, USA, 1963.

[SS73] P.H. Sneath and R.R. Sokal. *Numerical Taxonomy.* Freeman, San Francisco, 1973.

[SSH94] M.A. Steel, M.A. Székely, and M.D. Hendy. Reconstructing trees from sequences whosesites evolve at variable rates. *J. Comp. Biol.*, 1(2):153–163, 1994.

[ST77] S. Sattah and A. Tversky. Additive similarity trees. *Psychometrika*, 42:319–345, 1977.

[Ste88a] M.A. Steel. Distribution on the symmetric difference metric on phylogenetic trees. *SIAM J. Disc. Math.*, 1:541–551, 1988.

[Ste88b] M.A. Steel. *Distributions on bicoulored evolutionary trees.* PhD thesis, Massey Univ., NZ, 1988.

[Ste92] M. Steel. The complexity of reconstructing trees from qualitative characters and subtrees. *J. of Classification*, 9:91–116, 1992.

[Ste93] Caro-Beth Stewart. The powers and pitfalls of parsimony. In *Nature*, volume 361, pages 603–607. february 1993.

[Ste94a] M.A. Steel. Recovering a tree from the leaf colourations it generates under a markov model. *Appl. Math. Letter*, 7:19–24, 1994.

[Ste94b] K.B. Stewart. Answer to sidow. *Nature*, 367:26–27, 1994.

[Ste97] M.A. Steel. communication personnelle, 1997.

[SvH96] K. Strimmer and A. von Haeseler. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.*, 13(7):964–969, 1996.

[Swo90] D.L. Swofford. Paup: phylogenetic analysis using parsimony. version 3.0. *Illinois Natl. Hist. Surv.*, 1990. Champaign, IL,.

[Tho73] E. A. Thompson. The method of minimum evolution. *Ann. Hum. Genet.*, 36:333–340, 1973.

[TN84] Fumio Tajima and Masatoshi Nei. Estimation of evolution distance between nucleotide sequences. *Mol. Biol. Evol.*, 1(3):269–285, 1984.

[TN94a] N. Takezaki and M. Nei. Inconsistency of the maximum parsimony method when the rate of nucleotide, substitution is constant. *J. Mol. Evol.*, 39:210–218, 1994.

[TN94b] Y. Tateno and M. Nei. Relative efficiency of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. *Mol. Biol. Evol.*, 11:261–277, 1994.

[TNT82] Yoshio Tateno, Masatoshi Nei, and Fumio Tajima. Accuracy of estimed phylogenetic trees from molecular data : I. distantly related species. *J. Mol. Evol.*, 18:387–404, 1982.

[Tor46] E. Torricelli. *Opere di Evangelista Torricelli*, chapter De maximis et minimis, pages 79–97. Geometrica, 1646.

[War93] T. Wareham. On the computational complexity of inferring evolutionnary trees. Technical Report 9301, Memorial University of Newfoundland, 1993.

[War94] T. Warnow. Tree compatibility and inferring evolutionary history. *J. of Algorithms*, 16:388–407, 1994.

[WF90] Patrick L. Williams and Walter M. Fitch. Phylogeny determination using dynamically weighted parsimony method. In Russsell F. Doolittle, editor, *Molecular evolution : computer analysis of protein and nucleic acidsequences*, volume 183 of *Methods In Enzymology*, chapter 38, pages 615–626. Academic Press, 1990.

[WH88] W.C. Wheeler and R.L. Honeycutt. Paired sequences difference in ribosomal rnas evolution and phylognetic implications. *Mol. Biol. Evol.*, 1988.

[WLS87] K.H. Wolfe, W.-H. Li, and P.M. Sharp. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear dnas. *Proc. Natl. Acad. Sci. USA*, 84:9054–9058, 1987.

[Yan93] Z. Yang. Maximum likelihood estimation of phylogeny from dna sequences when substitution rates differ over sites. *Mol. Biol. Evol.*, 10(6):1396–1401, 1993.

[Zar65] K. Zarestkii. Construction d'un arbre sur la base d'un ensemble de distances entre ses feuilles. *Upekki Math. Nauk.*, 1965. (en Russe).

[ZL92a] Andrey Zharkikh and Wen-Hsiung Li. Statistical properties of bootstrap estimation of phylogenies variability from nucleotide sequences. I. four taxa with a molecular clock. *Mol. Biol. Evol.*, 9(6):1119–1147, 1992.

[ZL92b] Andrey Zharkikh and Wen-Hsiung Li. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences: II. four taxa without a molecular clock. *J. Mol. Evol.*, 35:356–366, 1992.

[ZL95] A. Zharkikh and W.-H. Li. Estimation of confidence in phylogeny: the complete-and-partial bootstrap technique. *Mol. Phylogenet. Evol.*, 4(1):44–63, 1995.

[ZP62] E. Zuckerlandl and L. Pauling. *Horizons in Biochemistry*, chapter Molecular disease, evolution, and genic heterogeneity, pages 189–225. Elsevier, Amsterdam, 1962.