Méthodes de distance pour l'inférence phylogénomique

Alexis Criscuolo

ACADÉMIE DE MONTPELLIER

UNIVERSITÉ MONTPELLIER II

- SCIENCES ET TECHNIQUES DU LANGUEDOC -

Thèse

présentée à l'Université des Sciences et Technologies du Languedoc
pour obtenir le diplôme de Doctorat

SPÉCIALITÉ : **BioInformatique**

ECOLE DOCTORALE : **Biologie des Systèmes Intégrés, Agronomie - Environnement**

Méthodes de distance pour l'inférence phylogénomique

ALEXIS CRISCUOLO

Soutenue le 5 décembre 2006 devant le jury composé par

M. Manolo GOUY, Directeur de Recherche CNRS/LBBE (Villeurbanne)	Rapporteur
M. Alain GUÉNOCHE, Directeur de Recherche CNRS/IML (Marseille)	Rapporteur
M. Vincent BERRY, Maitre de Conférence, Université Montpellier 2	Examinateur
M. François-Joseph LAPOINTE, Professeur, Université de Montréal	Examinateur
M. Christian MICHEL, Professeur, Université Strasbourg 1	Invité
M. Olivier GASCUEL, Directeur de Recherche CNRS/LIRMM	Co-Directeur de Thèse
M. Emmanuel J.P. DOUZERY, Professeur, Université Montpellier 2	Directeur de thèse

ACADÉMIE DE MONTPELLIER

UNIVERSITÉ MONTPELLIER II

- SCIENCES ET TECHNIQUES DU LANGUEDOC -

Thèse

présentée à l'Université des Sciences et Technologies du Languedoc
pour obtenir le diplôme de Doctorat

SPÉCIALITÉ : **BioInformatique**

ECOLE DOCTORALE : **Biologie des Systèmes Intégrés, Agronomie - Environnement**

Méthodes de distance pour l'inférence phylogénomique

ALEXIS CRISCUOLO

Soutenue le 5 décembre 2006 devant le jury composé par

M. Manolo GOUY, Directeur de Recherche CNRS/LBBE (Villeurbanne)	Rapporteur
M. Alain GUÉNOCHE, Directeur de Recherche CNRS/IML (Marseille)	Rapporteur
M. Vincent BERRY, Maitre de Conférence, Université Montpellier 2	Examinateur
M. François-Joseph LAPOINTE, Professeur, Université de Montréal	Examinateur
M. Christian MICHEL, Professeur, Université Strasbourg 1	Invité
M. Olivier GASCUEL, Directeur de Recherche CNRS/LIRMM	Co-Directeur de Thèse
M. Emmanuel J.P. DOUZERY, Professeur, Université Montpellier 2	Directeur de thèse

Table des matières

Remerciements **7**

Note **9**

Introduction **11**

1 D'une volonté de classifier le vivant à la reconstruction de son histoire évolutive **15**
1.1 Classifier le vivant en mettant en évidence les ressemblances 16
1.2 Classifier le vivant en induisant un ordre naturel et une organisation de la nature . 17
1.3 Classifier le vivant en rendant sensible les liens de parenté qui lient les espèces . 21

2 Définition des arbres phylogénétiques et de leurs techniques d'inférence **25**
2.1 L'arbre phylogénétique 26
2.1.1 Notions biologiques 26
2.1.2 Définitions et propriétés combinatoires 27
2.1.3 Les distances induites par les arbres phylogénétiques 29
2.2 Techniques algorithmiques d'inférence d'arbres 30
2.2.1 Le schéma agglomératif 30
2.2.2 Le schéma divisif 31
2.2.3 La procédure d'insertion 32
2.2.4 Les recherches locales 34

3 L'inférence phylogénétique **41**
3.1 Les différents types de données biologiques 42
3.2 Optimiser le critère du maximum de parcimonie 44
3.2.1 Calcul de la valeur de parcimonie d'un arbre phylogénétique 44
3.2.2 Recherche de l'(des) arbre(s) phylogénétique(s) le(s) plus parcimonieux . 46
3.3 Modèles probabilistes d'évolution des séquences génétiques 47
3.4 Optimiser le critère du maximum de vraisemblance 50
3.4.1 Calcul de la vraisemblance d'un arbre phylogénétique 50
3.4.2 Recherche de l'arbre phylogénétique le plus vraisemblable 52

3.5 Optimiser les critères basés sur les distances évolutives 53
3.5.1 Calculs de distances évolutives 53
3.5.2 Les critères basés sur les distances évolutives 57
3.5.3 Inférence d'arbres phylogénétiques à partir de distances 59
3.6 Estimation de la fiabilité d'un arbre phylogénétique inféré par l'optimisation d'un critère 67
3.6.1 Comparaison de deux arbres phylogénétiques 67
3.6.2 Evaluation de la fiabilité des branches d'un arbre phylogénétique 69

4 L'inférence phylogénomique 71
4.1 La combinaison basse 73
4.1.1 *Total evidence* 73
4.1.2 Analyse simultanée à partir d'une supermatrice de caractères 74
4.2 La combinaison haute 75
4.2.1 Les techniques de consensus d'arbres phylogénétiques 76
4.2.2 La généralisation des techniques de consensus au problème du superarbre 76
4.2.3 L'algorithme BUILD 77
4.2.4 Les adaptations de BUILD 79
4.2.5 Représentation matricielle binaire 80
4.2.6 Heuristiques de recherche locale par descente 83
4.2.7 Décomposition sous forme de quadruplets 83
4.2.8 Moyennes de distances additives d'arbre 85
4.3 La combinaison moyenne 86
4.3.1 Décomposition sous forme de quadruplets 87
4.3.2 Modèles d'évolution des génomes complets 88
4.3.3 Méthodes de distance basées sur les signatures génomiques 89
4.3.4 Méthodes de distance basées sur les scores de BLAST 90

5 Utilisation des distances évolutives en inférence phylogénomique 93
5.1 Préliminaires biologiques et mathématiques 94
5.1.1 Les informations biologiques propres à l'inférence phylogénétique 94
5.1.2 Quelques propriétés des distances additives d'arbre 95
5.2 La méthode SDM 97
5.2.1 Définition des paramètres optimaux des modèles PM et SSM à l'aide d'un critère WLS 97
5.2.2 Calcul des paramètres optimaux des modèles SSM et PM 100
5.2.3 Construction d'une supermatrice de distance avec les paramètres optimaux des modèles SSM et PM 102

5.2.4 Complexités algorithmiques du calcul des paramètres et de la supermatrice de distance . . . 102
5.3 Application et discussion . . . 104
5.3.1 Estimation des vitesses d'évolution relatives à chaque gène . . . 104
5.3.2 Inférence phylogénomique par combinaison moyenne . . . 107

6 Construction d'arbres à partir de (super)matrices de distance 113
6.1 Adaptation de la classe des algorithmes agglomératifs aux distances incomplètes 114
6.1.1 Critères d'agglomération . . . 114
6.1.2 Valuation des branches externes . . . 117
6.1.3 Réduction matricielle . . . 118
6.2 Adaptation des algorithmes NJ, UNJ, BIONJ et MVR aux distances incomplètes . 118
6.2.1 Calcul des paramètres w_i^* et λ_i^* . . . 119
6.2.2 Considération de la variance associée aux distances incomplètes . . . 121
6.2.3 Les critères d'agglomération : complexité algorithmique et performances . 123
6.3 Les algorithmes NJ*, UNJ*, BIONJ* et MVR* dans le cadre de l'inférence phylogénomique . . . 130
6.3.1 Description des algorithmes NJ*, UNJ*, BIONJ* et MVR* . . . 130
6.3.2 Utilisation des supermatrices de distance inférées par SDM . . . 131

7 Comparaison de différents scénarios d'inférence phylogénomique 135
7.1 Protocole de simulation . . . 136
7.2 Scénarios de combinaison basse . . . 137
7.2.1 Scénarios de combinaison basse utilisant des méthodes de distance . . . 137
7.2.2 Résultats et discussion . . . 139
7.3 Scénarios de combinaison moyenne . . . 140
7.3.1 Inférence d'arbre et complétion de distances manquantes à partir des supermatrices de distance inférées par SDM . . . 141
7.3.2 Résultats et discussion . . . 143
7.4 Scénarios de combinaison haute . . . 144
7.4.1 Les supermatrices de distance inférées à partir d'une collection d'arbres phylogénétiques . . . 145
7.4.2 Les techniques de représentation matricielle binaire MRP et MRH . . . 146
7.4.3 Résultats et discussion . . . 148
7.5 Les différents niveaux de combinaison en inférence phylogénomique . . . 151

8 Conclusions et perspectives 155

A Annexes du Chapitre 5 **159**

A.1 Preuve d'invariance topologique des algorithmes AddTree, NJ, UNJ, BioNJ et MVR à certaines déformations de la distance (Δ_{ij}) 159

A.1.1 Invariance topologique à la multiplication par un facteur 159

A.1.2 Invariance topologique à l'ajout d'une constante 161

A.2 Calcul des dérivées partielles de $f(v)$ 163

B Annexes du Chapitre 6 **165**

C Annexes du Chapitre 7 **167**

C.1 Calcul de la variance d'une distance estimée par complétion 167

C.1.1 Complétion additive 167

C.1.2 Complétion par quadruplets 169

Index **171**

Liste des figures **174**

Bibliographie **176**

Remerciements

Une thèse, c'est court. Encore une fois, Vincent Berry, Emmanuel J. P. Douzery et Olivier Gascuel avaient raison : les années passent plus vite qu'on ne le croit. Et ils ont eu raison sur bien d'autres points — plus scientifiques — abordés tout au long de cette thèse. Chacun avec sa manière, son domaine de spécialité et sa disponibilité, ils ont tout trois apporté leur contribution propre, dans un effort complémentaire, à l'affinement de la forme et du fond de chaque page de ce manuscrit de thèse. Améliorations de forme, il y en a eu dans mon écriture, guidée durant ces quelques années par leurs conseils éclairés. Je le reconnais ici-même : l'écriture scientifique a son style propre qui s'apprend au fil des si nombreuses corrections et remarques. Que les auteurs de celles-ci en soient remerciés. Améliorations de fond, il y en a eu aussi, nombreuses et parfois difficiles. La connaissance d'une branche scientifique ne s'acquiert pas en un jour. Le temps et la pratique améliore les premières intuitions grossières, et mes encadrants ont su détecter et affiner celles qui se sont révélées efficaces à la lumière de leur expérience. Qu'ils en soient remerciés également.

Une thèse, c'est long. On y rentre jeune étudiant, la tête pleine de quelques connaissances diverses et variées. On en sort plus agé et plus riche des expériences acquises dans les universités, les laboratoires, mais aussi dans la vie. Entre mon entrée et ma très future sortie, combien de souvenirs, de lieux, de peines, de joies ai-je accumulé, vécu ou subi ? Et parrallèlement à la présence constante de mes encadrants, combien d'autres personnes m'ont accompagné le temps d'un bout de chemin, quelquefois court, quelquefois long ? Cette thèse n'appartient à personne, et pourtant, derrière chaque chapitre se cachent les fantomes de ceux ou celles qui étaient à mes cotés, en amitié ou en amour, au moment où j'en apprenais ou en développais le thème. Je ne remercierai pas toutes celles qui m'ont donné la main durant cette thèse, par discrétion et par humilité, mais j'associe dans ma tête chaque chapitre à son ou ses visages, avec les sentiments qui les accompagneront toujours, qu'ils soient encore heureux ou devenus malheureux.

Note

Cette thèse décrit plusieurs nouvelles méthodes et nouveaux scénarios d'inférence phylogénomique. Ces nouvelles approches, cherchant à inférer des arbres à partir de plusieurs gènes, s'appuient sur de nombreuses méthodes d'inférence phylogénétique. Conséquemment, ce manuscrit de thèse présente et utilise un grand nombre de définitions, de notations, de critères et d'algorithmes. Chaque lecteur est invité à consulter l'index disponible à la fin de ce manuscrit, afin d'en avoir une lecture aussi confortable que précise.

Introduction

> Il a fallu plusieurs fois se procurer les machines, les construire, mettre la main à l'oeuvre, se rendre, pour ainsi dire, apprenti, et faire soi-même de mauvais ouvrages pour apprendre aux autres comment on en fait de bons.
>
> *Jean le Rond d' Alembert*

De tout temps, l'Homme a eu le besoin et la volonté de classifier ce que la nature lui offrait à l'observation. Conjointement, la nécessité d'expliquer l'apparition et la pluralité des formes vivantes amena nombre de réflexions sur d'éventuelles relations entre les différentes espèces. Fruit de nombreuses hypothèses et théories, la classification des espèces vivantes ou fossiles, ainsi que la schématisation de leur évolution sous la forme d'un arbre phylogénétique est actuellement devenue une norme en biologie de l'évolution.

La classification par arbre s'est beaucoup alimentée des progrès technologiques du siècle dernier. Ainsi, suite au développement du séquençage automatique de l'information génétique propre à chaque espèce, les données moléculaires ont été rapidement préférées aux observations morphologiques. Les séquences génétiques étant plus riches en information évolutive que les observations morphologiques, nombre d'hypothèses, de critères, de méthodes et d'algorithmes ont été développés afin de pouvoir traiter ces jeux de données particuliers et les interpréter avec autant d'exactitude que possible.

Depuis dix ans, la taille des banques de données *on-line* stockant les gènes et protéines séquencées croit exponentiellement. Ainsi, la banque européenne de données génomiques EMBL-BANK stockait moins de 1 millions de séquences génétiques en 1996, environ 5 millions en 1999 et 23 millions en 2003, et compte presque 80 millions de séquences génétiques en cette fin d'année 2006[1]. Cette énorme quantité de données porte à croire que l'on peut actuellement construire de grands arbres, définissant l'histoire évolutive de vastes ensembles d'espèces.

Néanmoins, un traitement informatique de ce nombre exponentiel de données est nécessaire pour pouvoir construire de tels arbres (outre les étapes préliminaires d'extraction d'information ou d'alignement de gènes). De plus, l'entreprise de séquençage n'étant pas équilibrée, certaines espèces sont sur-représentées, d'autres absentes, et il en va de même pour les gènes et les protéines, ce qui implique de vastes zones de données manquantes handicapant les méthodes

[1] pour plus de détail, cf l'URL `http://www3.ebi.ac.uk/Services/DBStats/`

d'inférence d'arbres. Enfin, plusieurs caractéristiques évolutives sont à prendre en compte afin de pouvoir construire des arbres phylogénétiques statistiquement et biologiquement fiables (*e.g.* vitesse d'évolution ou pression de sélection propre à chaque gène).

L'inférence phylogénétique, consistant à construire un arbre phylogénétique à partir d'un unique gène source (*e.g.* arbre de gènes ou arbre d'espèces), a conduit à la formalisation de plusieurs critères pour définir mathématiquement un arbre optimal, c'est-à-dire représentant au mieux l'histoire évolutive induite par un gène. Le critère du Maximum de Parcimonie définit le meilleur arbre comme étant celui qui implique le moins de mutations le long de ses branches. Les critères probabilistes optimisent plusieurs paramètres, soit afin de trouver l'arbre ayant la plus grande probabilité de représenter l'histoire évolutive induite par le gène, soit afin de maximiser la probabilité d'existence du gène étant donné un arbre. Les critères basés sur les distances évolutives définissent le meilleur arbre par comparaison avec une distance évolutive estimée entre chaque paire d'espèces.

L'inférence phylogénomique, consistant à construire un arbre phylogénétique à partir d'un grand nombre de gènes sources[2] (*e.g.* plus d'une centaine), a définit trois principales méthodologies de combinaison de données pour construire un arbre optimal. La *combinaison basse* s'appuie sur la concaténation des gènes sources. La *combinaison haute* considère l'ensemble des arbres obtenus par inférence phylogénétique à partir de chaque gène. La *combinaison moyenne* tente d'encoder le signal phylogénétique induit par chaque gène. Ces trois combinaisons amalgament ensuite toutes les informations phylogénétiques qu'elles ont extraites des gènes en un unique arbre.

Les travaux présentés dans cette thèse étudient ces trois critères phylogénétiques et ces trois familles de combinaison de données afin de proposer de nouveaux scénarios d'inférence phylogénomique dans le but de maximiser fiabilité et rapidité.

Le premier chapitre expose brièvement l'historique des hypothèses et théories scientifiques qui ont conduit à la modélisation de l'histoire évolutive sous la forme d'arbres phylogénétiques.

Le second chapitre introduit les notations, le vocabulaire ainsi que les principales propriétés propres aux arbres phylogénétiques. Il expose également les principales techniques algorithmiques d'inférence d'arbre sur lesquelles s'appuient la grande majorité des méthodes d'inférence phylogénétique et phylogénomique.

[2] La définition précise des inférences phylogénétique et phylogénomique adoptée dans ce manuscrit de thèse est issue d'un choix volontaire de l'auteur. Dans la littérature scientifique, la notion d'inférence phylogénétique a été très souvent employée pour désigner la construction d'un arbre de l'évolution inféré aussi bien à partir d'un que de plusieurs gènes. De même, la notion d'inférence phylogénomique y a été utilisée avec différents sens, tels que l'inférence d'arbre à partir de génomes complets ou d'une vaste collection de gènes. Cette thèse ne prétend pas imposer une terminologie, mais adopte une certaine sémantique qu'elle justifie en argumentant sur les difficultés propres, quoique complémentaires, à chacun de ces deux types d'inférence.

Le troisième chapitre est consacré à un état de l'art sur l'inférence phylogénétique. Des principaux critères d'optimalité aux différentes approches méthodologiques, son but est de fournir une description des grands principes qui la définissent.

Le quatrième chapitre décrit les principales techniques d'inférence phylogénomique et discute de la fiabilité et de l'utilité des trois grandes familles de combinaison de données génétiques.

Le cinquième chapitre décrit une nouvelle méthode de combinaison moyenne, nommée *Super Distance Matrix* (SDM ; Criscuolo et al., 2006), permettant, à la fois, d'estimer les vitesses d'évolution relatives de chaque gène source, et de calculer une distance évolutive entre les paires d'espèces définies par les gènes sources. L'application d'une méthode d'inférence phylogénétique sur cette dernière distance (parfois incomplète) permet d'inférer des arbres présentant de bonnes caractéristiques topologiques.

Le sixième chapitre décrit quatre nouveaux algorithmes, nommés NJ*, UNJ*, BIONJ* et MVR* (Criscuolo, 2006), pouvant constuire un arbre phylogénétique à partir d'une distance évolutive incomplète. Ces adaptations des algorithmes NJ (Saitou and Nei, 1987; Studier and Keppler, 1988), UNJ (Gascuel, 1997b), BIONJ (Gascuel, 1997a) et MVR (Gascuel, 2000) contiennent certaines améliorations qui les dédient aux distances de type SDM. Elles présentent des performances similaires au méthodes standards, telles que FITCH (Felsenstein, 1997) ou MW* (Makarenkov and Lapointe, 2004), avec des temps d'exécution extrèmement rapides.

Le septième chapitre décrit une vingtaine de nouveaux scénarios d'inférence phylogénomique issus de l'utilisation de la méthode SDM et des algorithmes NJ*, UNJ*, BIONJ* et MVR*. Chaque scénario est observé et discuté sur la base d'un protocole de simulation. Il montre entre autre que les techniques de combinaison basse et haute peuvent être très significativement améliorées, aussi bien en termes de qualité que de temps d'exécution, par l'utilisation de critères de distance.

Chapitre 1

D'une volonté de classifier le vivant à la reconstruction de son histoire évolutive

L'apparition de la conscience dans le règne animal est peut-être un aussi grand mystère que l'origine de la vie même. Cependant, il faut bien supposer, quoique cela pose un problème impénétrable, qu'il y a bien là un effet de l'évolution, un produit de la sélection naturelle.

Karl Popper

Les humains ne sont pas le résultat final d'un progrès évolutif prédictible mais plutôt une minuscule brindille sur l'énorme buisson arborescent de la vie qui ne repousserait sûrement pas si la graine de cet arbre était mise en terre une seconde fois.

Stephen Jay Gould

Sommaire

1.1 Classifier le vivant en mettant en évidence les ressemblances 16

1.2 Classifier le vivant en induisant un ordre naturel et une organisation de la nature 17

1.3 Classifier le vivant en rendant sensible les liens de parenté qui lient les espèces 21

Ce chapitre a pour objectif d'expliciter brièvement comment, au fil des siècles, l'idée naturelle de classification des espèces vivantes a progressé vers la notion d'évolution. Il cherche, à la fois, à apporter une connaissance moins "froidement" scientifique sur le thème de l'inférence d'arbres évolutifs, et à montrer que la représentation de l'évolution du vivant reste une branche active des sciences naturelles.

1.1 Classifier le vivant en mettant en évidence les ressemblances

Le besoin de classifier est une caractéristique humaine universelle. Ainsi, depuis l'aube de la mémoire humaine, nombre d'éléments ont été soumis au joug intellectuel du classement, des divers corps célestes aux différentes populations, en passant par l'ensemble des êtres vivants.

Si on classe chronologiquement les fragments écrits conservés jusqu'à nos jours, on peut citer, comme un des points de départ de cette volonté de classifier et d'expliquer la diversité du vivant, les différents philosophes regroupés sous l'étiquette "présocratiques". Ainsi Empédocle (490 à 435 av. J.-C.) considérait qu'un corps vivant n'est que la réunion aléatoire de différents organes isolés. Anaximandre de Milet (610 à 546 av. J.-C.) imaginait que la vie apparaissait suite à l'évaporation de l'eau au soleil par un processus de *génération spontanée*[1], tout comme Démocrite (460 à 370 av. J.-C.), qui pensait que la vie est issue de vers sortant de la boue, tout en admettant le rôle du hasard et d'une certaine forme de sélection dans ce processus.

C'est Aristote (384 à 322 av. J.-C.) qui fût le premier à effectuer une forme de synthèse des différentes théories de son époque. Il prétendit que la nature est le lieu de l'accidentel et que comme on ne peut discourir sur ce qui s'y produit nécessairement, on peut s'aventurer sur ce qui s'y produit le plus souvent. Il consacra une partie de sa vie à l'observation de la nature, en se basant sur l'idée moderne que la théorie doit rendre compte de ce qui est observé, et non l'inverse. La notion d'hérédité n'étant pas encore imaginée, Aristote affirma que les espèces vivantes sont d'essence naturelle et reprendra l'idée de génération spontanée. Ses oeuvres sur le sujet, principalement descriptives, représentent une première approche de classification du vivant. Parmi elles, *Des Parties des Animaux* expose une classification de nombre d'espèces animales suivant des critères morphologiques. Ces critères de classification du vivant, basés sur les similarités morphologiques observables, perdureront durant de nombreux siècles.

Une autre explication de la diversité du monde vivant fût initiée par les écrits bibliques. Dans le chapitre *Genèse*, la Bible explique que le monde fût créé tel qu'il est actuellement en six jours. Les végétaux fûrent créés le troisième jour, les animaux vivants dans l'eau et le ciel le cinquième jour, et les animaux vivants sur terre le sixième jour. La *Genèse* explique aussi que le Créateur du monde a créé *"l'homme à son image"* afin qu'il domine poissons, oiseaux, animaux domestiques ou non, ainsi que *"les reptiles qui rampent sur la terre"*. L'explication biblique du début de la vie sur terre induit deux grandes idées : le *créationnisme*[2], considérant que tout le vivant a été créé

[1] *Génération spontanée* ou *Abiogenèse* : Processus d'apparition d'un être vivant sans ascendant sous l'effet de facteurs physico-chimiques à partir de substances inorganiques.

[2] *Créationnisme* : Croyance selon laquelle l'Univers et les éléments le composant (entre autre les êtres vivants) ont été créés par une puissance divine. Cette doctrine peut se découper en deux courants de pensée : le *créationnisme fixiste* qui considère une unique, voire un petit nombre successif, de création(s) divine(s), et le *créationnisme évolutiste* qui considère que certaines formes peuvent évoluer à partir des formes initialement créées.

directement sous sa forme actuelle, et l'*anthropocentrisme*[3], plaçant l'être humain au sommet du règne animal. Cette conception chrétienne induit certains corrolaires sur les relations entre êtres vivants. Si la création divine est parfaite, alors rien de ce qui a été créé ne peut s'éteindre ; ainsi la notion d'extinction de groupes d'espèces est réfutée par la lecture des textes bibliques. Le principe de génération spontanée fût néanmoins réhabilité par Saint Augustin (354 - 430) qui, par son interprétation libérale de la Bible, opposa l'idée de *créationnisme évolutiste* à celle, originelle, de *créationnisme fixiste*.

Le Moyen-Age s'est nourri des conceptions chrétiennes et aristotéliciennes durant de nombreux siècles. Suivant consciencieusement ces doctrines, l'ensemble des sciences naturelles considéra que chaque être vivant a une place fixée suivant un plan divin et chercha simplement à cataloguer les différents liens entre les membres du vivant. La Renaissance, si riche en bouleversements idéologiques, semble ne pas avoir touché les conceptions que l'on se faisait sur le vivant. Même si Francesco Redi (1626 - 1698) ou, au siècle suivant, Lazzaro Spallanzani (1729 - 1799) tentèrent de montrer, par une approche expérimentale, que l'idée de génération spontanée était fausse, la principale tâche des naturalistes de l'époque fût de continuer à mettre en évidence les ressemblances morphologiques sans chercher à réfuter l'explication biblique. Cette tâche ne fût pas pour autant sans conséquence. Ainsi Joseph-Pitton de Tournefort (1656 - 1708) fût parmi les premiers à s'inspirer de ces nombreuses observations pour essayer de fournir des règles claires permettant de définir des genres naturels. Il permit de compléter les classifications d'Aristote en définissant les genres végétaux suivant la morphologie des fleurs et des fruits. Dans ses *Eléments de Botanique* (1694), il introduisit l'idée que les caractéristiques d'un genre donné doivent s'appliquer à tous ses membres.

S'inspirant des travaux de Tournefort et leur rendant hommage, Karl von Linné (1707 - 1778) apporta d'autres outils à la classifications du vivant en général, et des plantes en particulier. Initiée dans *Systema Naturae* (1735), c'est dans *Species Plantarum* (1753) qu'il applique systématiquement la nomenclature binominale permettant de dénommer précisément toutes les formes vivantes : le nom du genre suivi par le nom de l'espèce. Cette nomenclature est devenue une norme, toujours en vigueur actuellement. Il introduisit aussi la notion de groupes d'espèces définis par leurs organes sexuels. Convaincu par le principe du créationnisme fixiste, ces travaux lui valurent néanmoins une très forte notoriété qui freina, *a posteriori*, la propagation d'idées nouvelles sur l'évolution des êtres vivants.

1.2 Classifier le vivant en induisant un ordre naturel et une organisation de la nature

Dans son livre *Le Règne Animal Distribué selon son Organisation* (1817), Georges Cuvier (1769 - 1832) énonça le principe de corrélation des parties selon lequel chaque organe est lié

[3] *Anthropocentrisme* : Croyance considérant l'être humain au centre de l'Univers et supérieur à toute autre forme de vie.

dans son fonctionnement à tous les autres. Proche de Cuvier, Geoffroy Saint-Hilaire (1772 - 1844) eu l'intuition d'une unité fondamentale du vivant où tous les animaux dérivent d'un même plan d'organisation. Il observa de fortes similitudes morphologiques en comparant de nombreux squelettes. Même si des structures squelettiques sont globalement différentes, ces différences s'amenuisent localement : des os présents chez certains squelettes semblent simplement atrophiés chez d'autres, mais pas absents. Dans *Philosophie Anatomique* (1818) et *Histoire Naturelle des Mammifères* (1819), il énonça les lois des connexions (les organes conservent toujours les mêmes relations entre eux), de permanence (il ne se crée aucun organe nouveau) et du balancement (le développement d'un organe se fait au détriment d'un autre). Il introduisit ainsi la notion d'*homologie de connexion*[4] : un organe est homologue chez deux espèces si, sous quelque forme ou fonction que ce soit, il a les mêmes connexions avec d'autres organes. Cette notion induira l'idée d'*homologie d'ascendance*[5] par laquelle deux organes sont homologues s'ils dérivent d'un organe unique chez un hypothétique ancètre commun, et provoquera de violentes controverse avec Cuvier, *fixiste*[6] convaincu.

C'est dans ce contexte, où les idées créationnistes et fixistes affrontaient les premières intuitions évolutionnistes, que Jean-Baptiste Pierre Antoine de Monet, Chevalier de Lamarck (1744 - 1829) introduisit l'idée du *transformisme*[7] des formes vivantes. Instruit des travaux de Saint-Hilaire, Lamarck proposa, dans *Philosophie Zoologique* (1809), une théorie basée sur l'hérédité des caractères acquis. Cette théorie expose que l'environnement influence le développement des organes sous forme de mutations qui répondent aux modifications de cet environnement. Plus simplement, les organes évoluent suivant que l'environnement nécessite leur utilisation ou pas. Même si August Weissman (1834 - 1914) démontra plus tard l'impossibilité d'hérédité pour les caractères acquis, Lamarck demeure historiquement la première personne a avoir suggéré l'idée d'une évolution du vivant. Sa théorie du transformisme influencera beaucoup de biologistes qui verront ensuite l'évolution comme un processus *gradiste*[8], c'est à dire un parcours vers plus d'adaptation et plus de perfection, ceci par une inhérente tendance vers la complexité. Lamarck fût aussi le premier à tenter de modéliser sa vision de l'évolution gradiste du vivant sous la forme d'un arbre qui prend racine dans les espèces de vers et évolue vers les mammifères, en passant par les stades intermédiaires des insectes, des poissons, des reptiles

[4] *Homologie de position* ou *Homologie primaire* : Partage d'une même organisation fondamentale et des mêmes connexions avec les organes voisins.

[5] *Homologie d'ascendance* ou *Homologie secondaire* : Partage d'un même caractère par différentes espèces en raison d'une ascendance commune.

[6] *Fixisme* : Théorie privilégiant la stabilité ou la faible dynamique de nombreux processus physiques et biologiques.

[7] *Transformisme* : Théorie, opposée au *fixisme*, prônant la progression successive d'une partie ou de l'ensemble d'une forme vivante au cours des différentes générations.

[8] *Gradisme* : Théorie considérant que les êtres vivants évoluent, par *transformisme*, d'un état simple et primitif vers un état plus complexe et adapté.

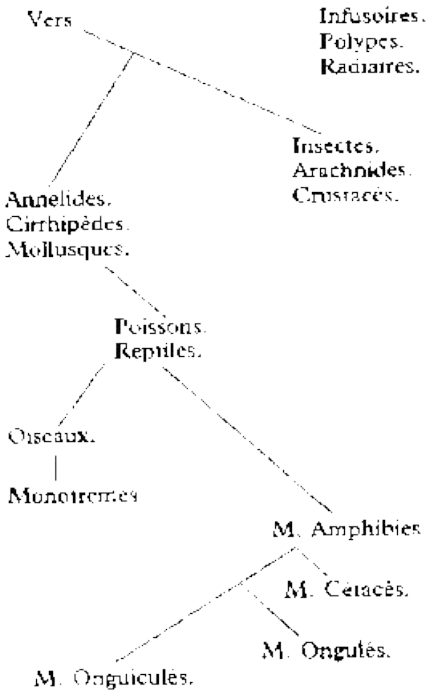

FIG. 1.1 – **Modélisation de l'évolution par Lamarck (1809)**

S'enracinant dans les vers, cet arbre explique l'évolution du vivant (du haut vers le bas), par un phénomène transformiste et gradiste, qui conduit l'ensemble des formes de vie simples vers plus de complexité et d'adaptation.

et des oiseaux (cf Figure 1.1).

Durant la première moitié du 19$^{\text{ième}}$ siècle, Charles Lyell (1797 - 1875) proposa, dans son livre *Principles of Geology* (1830), une théorie *uniformitariste*[9] sur la formation géologique de la surface de la terre. Cette théorie suggère que la Terre a été façonnée lentement sur une très longue période de temps par des forces toujours existantes. Cette vision géologique de la nature peut être vue comme un exemple illustratif de la prise de conscience des scientifiques de cette époque – principalement anglais – des conséquences macroscopiques d'une succession d'évènements microscopiques sur une grande échelle de temps. Cette prise de conscience scientifique s'illustra, pendant la même époque, dans les réflexions de différents naturalistes et biologistes sur la question de l'évolution du vivant. Parmi les plus connus, Charles Darwin (1809 - 1882) fut historiquement l'un des premiers à proposer une théorie de l'évolution réellement documentée. Dans l'édition de 1860 de son célèbre traité *The Origin of Species by means of Natural Selection*, Darwin relate consciencieusement l'historique des propositions scientifiques

[9] *Uniformitarisme* : Théorie géologique expliquant que les bouleversements passés constatés sont issus des causes et des mécanismes présents observés. C'est la succession lente et uniforme de ces causes et mécanismes qui aboutit aux bouleversements constatés.

quant à l'évolution des espèces vivantes. Il cite comme point de départ l'influence des théories transformistes et gradistes de Lamarck, ainsi que les observations morphologiques de Saint-Hilaire. Il effectue la description d'un mémoire, adressé en 1813 à la Société Royale de Londres par un certain docteur W.-C. Wells, où ce dernier admet distinctement le principe de la *sélection naturelle*[10], néanmoins uniquement appliqué à l'espèce humaine et à certains caractères uniquement. Ce mémoire expose que, premièrement, tous les animaux tendent à varier dans une certaine mesure et, deuxièmement, que les agriculteurs améliorent leurs animaux domestiques par la sélection. Partant de cette observation, Wells expose l'hypothèse que, parmi les populations humaines du centre de l'Afrique, certains membres ont pu accidentellement apparaitre en présentant une aptitude à résister à certaines maladies. Ainsi, ces derniers individus ont pu se multiplier. Les autres individus ont pu plus difficilement se reproduire par les conséquences conjointes de leur faible résistance aux maladies et de leur impossibilité de lutter contre les individus plus résistants. Darwin cite également, parmi de nombreuses références, les travaux et les idées de Patrick Matthew (1790 - 1894) qui, en 1831, prétendait que de nouvelles formes de vie peuvent apparaitre sous l'influence directe de la sélection naturelle en fonction des conditions d'existence dictées par l'environnement.

Darwin influença profondément la vision scientifique et philosophique du vivant grâce à l'argumentation détaillée de sa théorie. Proche de Lyell et sensible à ses théories géologiques, il exposa sa vision de l'évolution par le mécanisme de la sélection naturelle dans un premier essai en 1844 mais qu'il ne publiera pas, par souci des débats qu'il pourrait susciter dans l'Angleterre victorienne, choquée par tout ce qui est susceptible de remettre en cause un certain ordre moral bourgeois. Néanmoins, lorsque Alfred Russel Wallace (1823 - 1913), lui adressa un courier, en 1858, présentant des idées sur la variation et la sélection naturelle identiques à celles que Darwin avait formulées dans ses notes, il fût alors décidé que leur travaux seraient présentés conjointement devant la Société Linnéenne de Londres. Ces théories eûrent un impact immédiat sur le monde scientifique et Darwin se lança avec frénésie dans la rédaction de la version destinée au grand public, *On the origin of species*, paru en 1859. D'après Ernst Mayr (1904 - 2005), le *darwinisme*[11] peut se résumer en cinq points :

- l'*évolution* : le monde est suffisament ancien pour permettre l'évolution des espèces ;
- l'*ascendance commune* : toutes les espèces, tous les organismes vivants, ont un ancètre commun ; plus on remonte loin dans le passé, plus on découvre un lien de parenté avec de nombreuses espèces ;
- la *multiplication des espèces* : une espèce donnée peut donner naissance à des espèces filles qui évoluent différemment en raison de l'isolement géographique ;

[10] *Sélection naturelle* : Théorie expliquant que les êtres vivants les plus aptes à s'adapter à un environnement sont ceux qui y survivront le mieux, les espèces les moins aptes étant vouées à l'extinction.

[11] *Darwinisme* : Courant d'idées en adéquation avec la théorie de la *sélection naturelle* de Darwin.

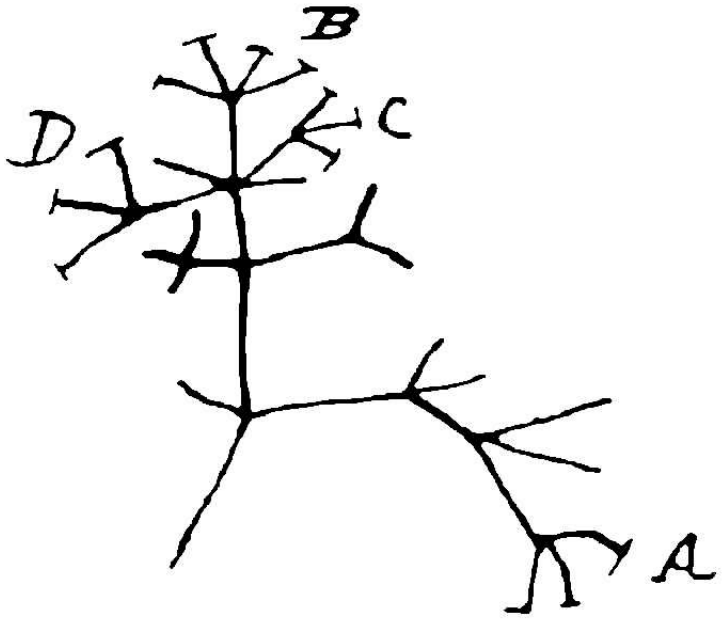

FIG. 1.2 – **Modélisation de l'évolution par Darwin (1837)**

L'extrémité de la longue branche pendante en bas à gauche représente l'espèce ancestrale. Les branches terminales marquées représentent les espèces contemporaines ; les lettres A, B et C modélisent les grands groupes d'espèces. Les autres branches terminales représentent les espèces éteintes.

- le *gradualisme* : l'évolution est un phénomène lent et progressif ;
- la *sélection naturelle* : dans une population animale ou végétale, les plus aptes survivent le mieux, se reproduisent avec plus de probabilité et leurs caractères sont transmis préférablement.

Darwin réutilisera, dès 1837, le modèle arboré de Lamarck pour représenter l'évolution du vivant mais enracinera cet arbre par l'espèce ancestrale et modélisera les branches externes pour représenter à la fois les espèces contemporaines et les espèces éteintes (cf Figure 1.2).

1.3 Classifier le vivant en rendant sensible les liens de parenté qui lient les espèces

Ce n'est que vers la fin du 19$^{\text{ième}}$ siècle, suite aux travaux de Grégor Mendel (1822 - 1884) et de Weissman, que l'on parla explicitement de la génétique lors des débats sur l'évolution du vivant. Un gène désigne une unité d'information transmise par un individu à sa descendance, par reproduction sexuée ou asexuée. Les notions d'information génétique et d'hérédité devinrent alors incontournables et nombre de théories fûrent avancées pour méler génétique et darwi-

nisme. Conséquemment, le domaine de la génétique des populations fût fondé indépendamment par Sewall Wright (1889 - 1988), John Burdon Sanderson Haldane (1892 - 1964) et Ronald Fisher (1890 - 1962). Ils appuyèrent leurs arguments sur le fait que lorsqu'une mutation directe survient dans un gène et si elle est favorable pour l'individu, son extension ultérieure dans la population dépend de plusieurs variables : la taille de la population, la longévité des générations, le degré de viabilité de la mutation et le taux avec lequel la même mutation réapparaît dans la descendance. Sachant qu'un allèle[12] donné est favorable sous certaines conditions environnementales, si celles-ci changent dans l'espace ou dans le temps, alors il se peut que cet allèle ne soit favorable que pour une partie seulement de la population. Si les conditions évoluent dans le temps, en général l'allèle devient inefficace. Etant donné que tous les individus contiennent habituellement un assortiment particulier d'allèles, le nombre total disponible pour la prochaine génération constitue un vaste réservoir potentiel de variables génétiques, le *pool génétique*. La reproduction garantit que les allèles seront réarrangés à chaque génération dans le processus de recombinaison. Dans une population stable, la fréquence avec laquelle un allèle réapparaît est proportionnelle au nombre total dans le pool génétique ; elle reste constante même si les allèles se recombinent différemment dans chaque individu. Par contre, si la fréquence génétique change d'une façon sensible, on assiste à une évolution. C'est ainsi que les mutations offrent une possibilité au pool génétique d'être réalimenté en nouveaux allèles. Ce processus complète la sélection naturelle qui, en modifiant la fréquence génétique donne aux allèles avantagés plus de chances de se reproduire. Cette théorie étant à l'origine supportée par des lois mathématiques, elle ne sera pas reconnue jusqu'à la fin des années 1930, époque à laquelle Theodosius Grigorovich Dobzhansky (1900 - 1975) la vérifia en laboratoire. Il démontra ainsi que l'adaptation génétique évolue dans les grandes populations de mouche des fruits et résulte d'un changement environnemental contrôlé. Dobzhansky prouva que les observations génétiques sont compatibles avec la sélection naturelle de Darwin, qui sont à la source des plus petites modifications de la fréquence génétique et donc des changements propres à l'évolution des caractères d'une population.

Cette nouvelle façon de comprendre la théorie de Darwin revitalisa pratiquement tous les champs de la biologie. Plusieurs écoles s'affrontèrent alors dans un débat cherchant à affirmer la meilleure manière de modéliser l'évolution du vivant à partir de données génétiques. Ainsi, alors que Robert Sokal et Peter Sneath jugèrent que l'évolution était encore basée sur des concepts flous et proposèrent la *taxonomie numérique*[13] s'appuyant sur des concepts de ressemblance pour la modéliser, Willi Hennig (1913 - 1976) postula, en 1950, que la systématique se doit d'exprimer le savoir évolutionniste. Hennig oeuvrant en faveur de la représentation de l'évolution du vivant sous la forme d'*arbres phylogénétiques*[14] , il rejoignit Darwin et sa conception arborée

[12] *Allèle* : Une des différentes formes sous laquelle un gène peut se présenter après mutation.

[13] *Taxonomie numérique* : Etude des relations entre espèces par l'estimation d'un indice de ressemblance entre chaque paire d'espèces.

[14] *Arbre phylogénétique* : Arbre modélisant l'histoire évolutive d'un groupe d'espèces.

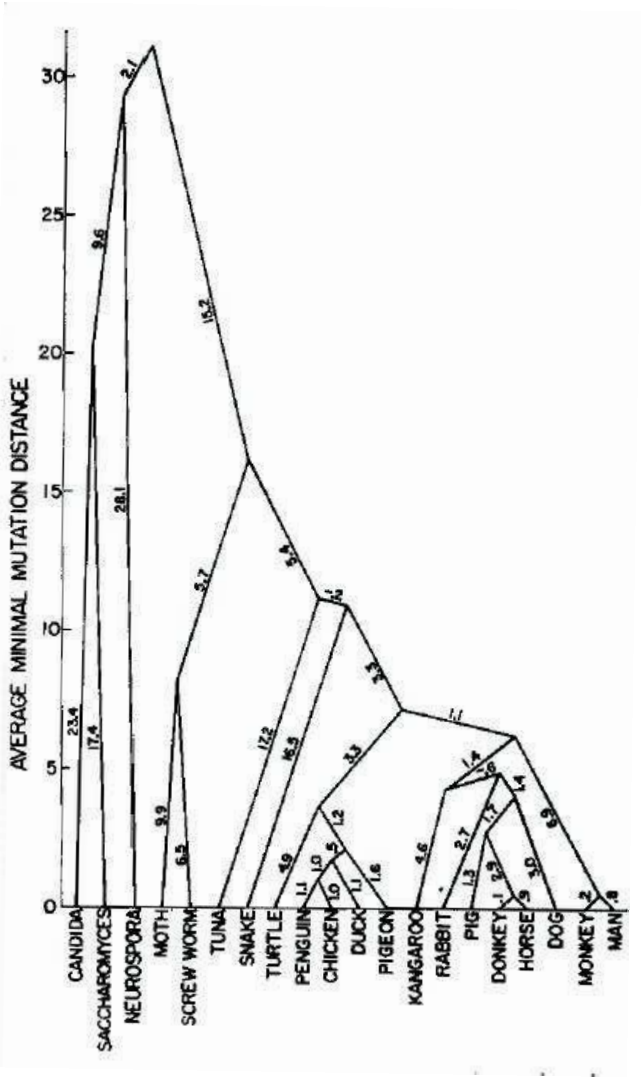

FIG. 1.3 – **Arbre phylogénétique inféré par Fitch et Margoliash (1967)**

(cf Figure 1.2). Hennig observa que si l'on compare les caractères de divers organismes, on constate diverses ressemblances et dissemblances. En terme évolutif, cela signifie que les caractères n'évoluent pas tous à la même vitesse. Ainsi, afin de modéliser l'évolution du vivant, Hennig postula qu'au lieu de chercher à mesurer la ressemblance globale, il faut analyser les caractères individuels de façon simultanée afin de faire émerger l'homologie d'ascendance les reliant. La reconstruction des liens de parenté entre espèces ne peut donc se faire qu'en précisant les états ancestraux et dérivés des caractères homologues.

La vision de Hennig évolua parrallèlement avec le développement de la technologie en général et de l'informatique en particulier. En 1953, James Watson et Francis Crick (1916 - 2004) découvrirent la structure en double hélice de l'ADN En 1977, Allan Maxam et Walter Gilbert, ainsi que Frédéric Sanger développèrent indépendamment les méthodes de séquençage d'ADN.

Dans le milieu des années 80, le développement de la PCR (*Polymerase Chain Reaction*) rendit possible le séquençage rapide des segments d'ADN et annonça le développement des bases de données génomiques. Conjointement, les travaux de nombreux algorithmiciens à partir des années 60, tels que Steve Farris ou Walter Fitch, permit de calculer, en utilisant l'outil informatique, des arbres optimaux (au sens des critères de parcimonie ou de distance) par l'analyse simultanée de différents caractères génétiques.

Historiquement, Emile Zuckerkandl et Linus Pauling, puis Walter Fitch et Emanuel Margoliash, apparaissent classiquement comme les pionniers des phylogénies moléculaires fondées sur des données de séquences génétiques. Ces derniers construisirent à partir des séquences protéiques du cytochrome c d'une vingtaine d'espèces (dont la plupart sont des Vertébrés), par une méthode de construction utilisant des distances évolutives entre chaque paire d'espèces, un arbre comportant seulement 3 différences (sur 18 groupements d'espèces) par rapport aux arbres phylogénétiques fondés sur les (jusqu'alors classiques) caractères morphologiques (cf Figure 1.3). Des travaux nombreux et célèbres sur les séquences issues de protéines furent menés dès le début des années 60 avec Zuckerkandl, Pauling et Margaret Dayhoff. Ils permirent de retrouver de fortes similarités entre phylogénies moléculaires et morphologiques de Vertébrés. Cette étape est importante car on comprit que, comme des copies d'un gène peuvent être soeurs par *spéciation*[15], elles peuvent être également soeurs par duplication[16] au sein d'un même génome, et que cette séparation physique suivie d'évolutions ultérieures autorise l'emploi des mêmes outils conceptuels de la reconstruction phylogénétique pour retracer l'histoire des génomes. Ces concepts sont toujours à l'oeuvre aujourd'hui, à l'époque des génomes complets et de la génomique. Carl Woese et Georges Fox montrèrent en 1977 par observations phylogénétiques que les organismes vivants se divisent en trois domaines : les Eubactéries, les Archées et les Eucaryotes (et non plus en "Procaryotes" et Eucaryotes) sur l'analyse des fragments de l'ARN ribosomique 16S. Ce concept novateur d'Archée est une des découvertes majeures du 20$^{\text{ième}}$ siècle.

Les chapitres 2, 3 et 4 de ce manuscrit de thèse exposent, avec plus de formalisme, les différentes notions qui caractérisent un arbre phylogénétique décrivant l'histoire évolutive d'un groupe d'espèces. Ils décrivent également les principales approches permettant, à partir des informations caractérisant chaque espèce, de modéliser les évènements évolutifs qui se sont produits entre elles.

[15] *Spéciation* : Désignation du processus évolutif par lequel de nouvelles espèces vivantes apparaissent à partir d'une espèce ancestrale. On désigne deux types principaux de spéciation : la *spéciation allopatrique*, désignant l'apparition d'une nouvelle espèce suite à l'isolement géographique d'une population, et la *spéciation sympatrique*, désignant l'apparition d'une nouvelle espèce à l'intérieur même de l'aire géographique de l'espèce ancestrale.

[16] *Duplication* : Désignation du processus évolutif par lequel de nouveaux brins d'ADN apparaissent par copie d'un brin initial

Chapitre 2

Définition des arbres phylogénétiques et de leurs techniques d'inférence

> Le concept de formalisation est habituellement employé dans les sciences de la Nature pour désigner le travail qui consiste à transposer une intuition théorique, dans une expression mathématique précise. [...] On sait, par exemple, qu'il y a très peu de formalisation en sociologie, aucune formalisation en histoire, mais par contre que la physique contient des formalisations tout a fait remarquables.
>
> *Serge Carfantan*

Sommaire

2.1 L'arbre phylogénétique 26

2.1.1 Notions biologiques 26

2.1.2 Définitions et propriétés combinatoires 27

2.1.3 Les distances induites par les arbres phylogénétiques 29

2.2 Techniques algorithmiques d'inférence d'arbres 30

2.2.1 Le schéma agglomératif 30

2.2.2 Le schéma divisif 31

2.2.3 La procédure d'insertion 32

2.2.4 Les recherches locales 34

Ce chapitre introduit l'objet autour duquel s'articulent la problématique et les résultats de ce manuscrit : l'arbre phylogénétique. La première partie définit et formalise biologiquement et mathématiquement ce qu'est un arbre phylogénétique en s'appuyant sur force exemples. La seconde partie explicite les principaux schémas algorithmiques permettant d'inférer un arbre phylogénétique.

2.1 L'arbre phylogénétique

Un arbre phylogénétique est à la fois un mode de représentation de l'histoire évolutive d'un groupe d'éléments et un objet mathématique simple à définir. Après une description de la vision biologique d'un arbre phylogénétique, les dernières parties introduisent à la fois les notations et les principales propriétés mathématiques et combinatoires propres à celui-ci.

2.1.1 Notions biologiques

L'hypothèse fondamentale en systématique phylogénétique est que l'histoire évolutive des espèces se déroule par spéciations successives. Suivant cette hypothèse, une espèce (ancestrale ou pas) peut, par spéciation, donner le jour à une nouvelle espèce. Plus schématiquement, dans l'arbre phylogénétique enraciné représenté dans la Figure 2.1, les espèces 1, 2 et 7 sont toutes différentes et les deux premières sont issues, par spéciation, d'une espèce ancestrale modélisée par le noeud 7.

L'arbre phylogénétique de la Figure 1.3 représente une histoire évolutive de vingt Vertébrés. Ces vingt *taxons* (espèces) sont représentés aux extrémités des vingt branches terminales. Chaque noeud interne de l'arbre peut être considéré comme une hypothétique espèce ancestrale et/ou un évènement de spéciation. Suivant un raisonnement similaire, la racine de l'arbre (*i.e.* le noeud interne situé en haut de la Figure 1.3) modélise à la fois une hypothétique espèce ancestrale à l'ensemble des vingt Vertébrés et un noeud interne sur lequel on peut connecter la racine d'un autre arbre phylogénétique (représentant l'histoire évolutive des Tuniciers, par exemple).

Ce mode de représentation de l'évolution du vivant induit directement une forme de *taxonomie*[1]. En effet, on distingue clairement que le groupe des Mammifères forme un *clade* (*i.e.* l'ensemble des descendants d'un noeud interne) composé par les espèces Kangourou, Lapin, Cochon, Ane, Cheval, Chien, Singe et Homme. Suivant cet arbre phylogénétique, les Mammifères forment ce qui est nommé un *groupe monophylétique* (*i.e.* qui correspond sans exception à l'ensemble des descendants d'une hypothétique espèce ancestrale). Le Kangourou est un Marsupial (*i.e.* développement partiel de l'embryon *in utero*) et se différencie des autres Mammifères de l'arbre phylogénétique de la Figure 1.3. Ces derniers sont appelés Mammifères placentaires (*i.e.* développement total de l'embryon *in utero*). Or on observe que les Mammifères placentaires (*i.e.* Lapin, Cochon, Ane, Cheval, Chien, Singe et Homme) ne forment pas un groupe monophylétique. Les Mammifères placentaires forment ainsi dans cet arbre un groupe dit *paraphylétique* (*i.e.* qui ne regroupe pas l'ensemble des descendants du plus récent ancêtre commun).

Les branches de l'arbre phylogénétique de la Figure 1.3 sont valuées. La valuation d'une branche correspond généralement à la quantité d'évolution estimée entre les deux espèces

[1] *Taxinomie* ou *Taxonomie* : Théorie et pratique de la classification des organismes vivants.

modélisées par les deux extrémités de cette branche. Autrement dit, plus la valuation est importante (*i.e.* plus la branche est longue), plus la divergence a été importante entre les deux espèces.

2.1.2 Définitions et propriétés combinatoires

Définitions

La notion d'arbre phylogénétique ainsi que ses différentes variantes (enraciné, binaire, ...) peuvent aisément se définir mathématiquement :

- un *graphe* $G = (V, E)$ est un objet combinatoire constitué d'un ensemble $V = \{v_i : 1 \leq i \leq |V|\} \neq \emptyset$ de $|V|$ sommets v_i et d'un ensemble d'arêtes $E \subseteq \{\{v_i, v_j\} : v_i, v_j \in V\}$,
- le degré d'un sommet v_i est le nombre d'arêtes reliées à v_i, *i.e.* $|\{v_j : \{v_i, v_j\} \in E\}|$,
- un chemin entre deux sommets v_s et v_p dans G est une liste d'arêtes $\{\{v_{i_x}, v_{i_{x+1}}\} : 0 \leq x \leq m\} \subseteq E$ telle que $i_0 = s$ et $i_m = p$.
- un graphe est dit *connexe* s'il existe au moins un chemin reliant chaque paire de sommets,
- un graphe connexe est dit *sans cycle* si, pour chaque paire de sommets, il n'existe qu'un unique chemin les reliant,
- un *arbre* est un graphe connexe sans cycle,
- un $X-$*arbre* est un arbre où il n'existe aucun sommet de degré 2, et où les sommets de degré 1 sont bijectivement associés aux éléments de l'ensemble X (Barthélemy and Guénoche, 1988),
- un $X-$arbre est de taille n s'il possède n sommets de degré 1, *i.e.* $|X| = n$,
- un $X-$arbre est dit *enraciné* en ρ s'il existe un unique sommet ρ de degré 2, appelé racine,
- un $X-$arbre est dit *binaire* si tous les noeuds internes autres que la racine (si elle existe) sont de degré 3,
- un arbre est dit phylogénétique si c'est un $X-$arbre enraciné[2].

Un exemple d'arbre phylogénétique enraciné et sa version non enracinée est représenté dans la Figure 2.1. Les $n = 6$ noeuds de degré 1 (*i.e.* 1, 2, 3, 4, 5 et 6) sont appelés *feuilles* et représentent chacun couramment une espèce (cf Figure 1.3) ou une séquence génétique connues. L'ensemble des feuilles d'un arbre T sera noté $\mathcal{L}_T$ dans la suite de ce manuscrit. Les noeuds de degré 3 (*i.e.* 7, 8, 9 et 10), dits *internes*, ne sont en général pas étiquetés. Les noeuds internes reliés à exactement deux feuilles (*i.e.* 7 et 10) induisent ce que l'on appelle une *cerise*. Une cerise peut aussi se définir comme une paire de feuilles reliées par un chemin composé d'un unique noeud interne (*i.e.* les deux paires $1\ 2$ et $5\ 6$).

[2] Dans un souci de simplicité, la suite de ce manuscrit désignera un arbre phylogénétique comme étant un $X-$arbre, qu'il soit enraciné ou non. La précision sera explicitement énoncée en cas d'ambiguité.

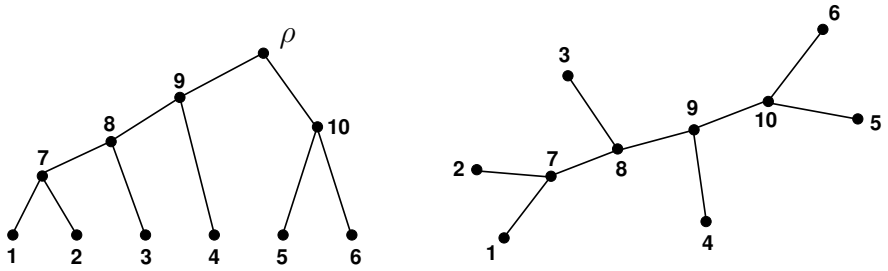

FIG. 2.1 – **Exemple d'arbres phylogénétiques**

A gauche : un arbre phylogénétique binaire enraciné en ρ ; à droite : le même arbre phylogénétique non enraciné.

Propriétés

Un arbre phylogénétique enraciné peut être représenté par l'ensemble des clades qu'il induit. Ainsi l'arbre phylogénétique enraciné de la Figure 2.1 (partie gauche) induit les clades $\{1, 2\}$, $\{1, 2, 3\}$, $\{1, 2, 3, 4\}$ et $\{5, 6\}$. Une feuille seule représente également un clade, ainsi que l'ensemble des n feuilles. Néanmoins, ces clades, dits *triviaux*, n'apportent aucune information sur la structure topologique de l'arbre. Un arbre phylogénétique enraciné binaire étant défini par $n - 2$ arêtes (ou branches[3]) internes (*i.e.* dont les deux extrémités sont des noeuds internes), il faut donc au plus $n - 2$ clades pour représenter cet arbre. D'une manière similaire, un arbre phylogénétique non enraciné peut être représenté par l'ensemble des *bipartitions* sur $\mathcal{L}_T$ induit par chacune des arêtes internes. En effet, le retrait d'une arête dans un arbre phylogénétique induit une bipartition de l'ensemble des feuilles. Ainsi les arbres phylogénétiques de la Figure 2.1 induisent les bipartitions $\{1, 2\}|\{3, 4, 5, 6\}$, $\{1, 2, 3\}|\{4, 5, 6\}$, $\{1, 2, 3, 4\}|\{5, 6\}$ ainsi que les bipartitions triviales propres à chacune des n feuilles. Un arbre phylogénétique non enraciné binaire étant défini par $n - 3$ branches internes, il faut donc au plus $n - 3$ bipartitions pour représenter un arbre phylogénétique quelconque (binaire ou pas).

Un arbre phylogénétique non enraciné binaire est composé par $n - 3$ branches internes et n branches externes (*i.e.* dont une des deux extrémités est une feuille). Il contient donc un total de $2n - 3$ branches. Soit A_n (resp. R_n) le nombre total d'arbres phylogénétiques binaires non enracinés (resp. enracinés) définis sur n feuilles. Comme chacun de ces A_n arbres contient $2n - 3$ branches, il existe $A_{n+1} = (2n - 3)A_n$ arbres non enracinés binaires définis sur $n + 1$ feuilles. Sachant que $A_3 = 1$, il existe donc $A_n = (2n - 5) \times (2n - 3) \times (2n - 1) \times ... \times 5 \times 3 = (2n - 5)!!$ arbres non enracinés binaires sur n feuilles (Felsenstein, 1978b). La Figure 2.5 représente les $A_5 = (2 \times 5 - 5)!! = 5 \times 3 \times 1 = 15$ différentes topologies d'arbre non enraciné binaire

[3]Les termes arête et branche seront des synonymes tout au long de ce manuscrit. Le premier terme (arête) appartient au vocabulaire mathématique et informatique, alors que le deuxième (branche) appartient plutôt au vocabulaire biologique.

définies sur $n = 5$ feuilles. Observant que, sur un arbre non enraciné binaire, il y a autant de racines possibles que de branches, il existe donc $R_n = A_{n+1} = (2n-3)!!$ arbres phylogénétiques enracinés binaires sur n feuilles. Par exemple, $R_{11} = A_{12}$ correspond à $654\ 729\ 075$. Ce nombre de topologies différentes augmentant de manière exponentielle, il représente une réelle difficulté lors de la recherche de l'arbre phylogénétique qui représente le mieux l'évolution d'un groupe de plus de 12 espèces.

2.1.3 Les distances induites par les arbres phylogénétiques

Soit T un arbre phylogénétique. Soit $\mathcal{T}_{ij}$ le nombre de branches composant l'unique chemin entre les feuilles i et j dans T (*e.g.* si T est l'arbre binaire non enraciné de la Figure 2.1 alors $\mathcal{T}_{12} = 2$ et $\mathcal{T}_{35} = 4$). Ainsi un arbre phylogénétique implique une mesure de distance entre chaque paire des n feuilles qui le composent. Ces distances sont le plus souvent stockées dans une matrice $(\mathcal{T}_{ij})$. Si T est un arbre *valué* (*i.e.* chaque branche a une longueur propre), le même raisonnement conduit à la matrice de distance (T_{ij}) où T_{ij} correspond à la somme de la longueur des branches composant l'unique chemin entre la paire de feuilles ij.

Ces deux types de *matrices de dissimilarité*, dites *distances additives* d'arbre, vérifient les proprités classiques des dissimilarités :

- symétrie : $\mathcal{T}_{ij} = \mathcal{T}_{ji}$ et $T_{ij} = T_{ji}$, pour tout $i, j \in \mathcal{L}_T$,
- réflexivité : $\mathcal{T}_{ij} = T_{ij} = 0 \iff i = j$, pour tout $i, j \in \mathcal{L}_T$,

mais se caractérisent en vérifiant l'*inégalité quadrangulaire* (Zaretskii, 1965; Buneman, 1971), pour tout $i, j, u, v \in \mathcal{L}_T$:

$$\mathcal{T}_{ij} + \mathcal{T}_{uv} \leq \max\left[\mathcal{T}_{iu} + \mathcal{T}_{jv}; \mathcal{T}_{iv} + \mathcal{T}_{ju}\right] \quad \text{et} \quad T_{ij} + T_{uv} \leq \max\Big[T_{iu} + T_{jv}; T_{iv} + T_{ju}\Big].$$

Cette inégalité implique directement l'*inégalité triangulaire*, pour tout $i, j, u \in \mathcal{L}_T$:

$$\mathcal{T}_{ij} \leq \mathcal{T}_{iu} + \mathcal{T}_{uj} \quad \text{et} \quad T_{ij} \leq T_{iu} + T_{uj}.$$

Une dissimilarité induite par un arbre phylogénétique est donc une distance, car elle vérifie les deux conditions de symétrie et de refléxivité, ainsi que l'inégalité triangulaire. De plus, conséquemment à l'inégalité quadrangulaire, les deux plus grandes des trois sommes $T_{ij} + T_{uv}$, $T_{iu} + T_{jv}$ et $T_{iv} + T_{uj}$ sont toujours égales. Enfin, il a été montré que si une matrice de distance (Δ_{ij}) symétrique et réflexive vérifie l'inégalité quadrangulaire, alors $(\Delta_{ij}) = (T_{ij})$ où T est un arbre phylogénétique valué unique (Zaretskii, 1965; Smolenskii, 1969; Simões-Pereira, 1969; Buneman, 1971).

Plusieurs cas particuliers de distance additive d'arbre existent. Si une distance additive (T_{ij}) vérifie l'*inégalité ultramétrique* (Hartigan, 1967), pour tout $i, j, u \in \mathcal{L}_T$:

$$T_{ij} \ \leq \ \max\Big[\, T_{iu} \,;\, T_{uj} \,\Big],$$

alors il existe un point ρ sur une des branches de T qui est équidistant à toutes les feuilles $i \in \mathcal{L}_T$ (Jardine et al., 1967; Sibson, 1972). Une *distance à centre* $T_{ij} = a_i + a_j$ est équivalente

à un arbre phylogénétique ne possédant qu'un unique noeud interne (*i.e.* un *arbre en étoile* ; cf Figure 2.2) et où chaque branche externe correspondant à la feuille i est de longueur a_i.

2.2 Techniques algorithmiques d'inférence d'arbres

Dans la majorité des cas, les méthodes de reconstruction d'arbres phylogénétiques reposent sur l'optimisation d'un critère permettant de comparer les différentes topologies d'arbre possibles définies sur un ensemble de n taxons. Autrement dit, étant donné un critère C permettant d'associer une valeur à un arbre phylogénétique, le meilleur arbre T, au sens du critère C, est celui qui correspond à la valeur $C(T)$ optimale. Le Chapitre 3 explicitera plusieurs de ces critères. Certaines méthodes d'inférence consistent à maximiser la vraisemblance d'un arbre. D'autres méthodes reviennent à minimiser l'écart quadratique entre une estimation des distances évolutives entre chaque paire d'espèces et la matrice additive induite par un arbre.

Une solution naïve consiste à considérer l'ensemble des topologies possibles afin de sélectionner celle qui optimise le critère C choisi. Néanmoins, pour un nombre n donné d'espèces, il existe un nombre exponentiel de topologies d'arbre phylogénétique (plus de 13 milliards de topologies d'arbre phylogénétique non enraciné binaire à 13 feuilles). Cette première solution devient donc rapidement impossible lorsque l'on dépasse une douzaine de feuilles. En effet, si on imagine que, correctement implémenté sur un ordinateur puissant, une seule milliseconde soit suffisante pour générer distinctement une des A_{13} topologies d'arbre phylogénétique non enraciné binaire à 13 feuilles et pour en estimer sa valeur de critère, il faudra compter plus de 22 semaines pour sélectionner sans ambiguité l'arbre de topologie optimale... La parallélisation d'un tel calcul sur x ordinateurs permet d'augmenter le seuil maximal de 13 taxons mais ne fait que translater le même problème sur un nombre fini x fois plus grand.

Pour la plupart des critères utilisés en inférence d'arbre phylogénétique, la recherche d'un arbre optimal est un problème NP-difficile. Dans ce cas, il est nécessaire d'utiliser des méthodes heuristiques (*i.e.* approchées), afin d'inférer, en un temps raisonnable, un arbre satisfaisant mais sans avoir la possibilité de savoir si celui ci est optimal. Ces heuristiques sont la plupart du temps basées sur des procédures algorithmiques simples qui sont décrites dans les trois sections suivantes. Les deux premiers types sont des approches gloutonnes. Le troisième type d'heuristiques consiste à effectuer une recherche locale en définissant le voisinage d'un arbre par modification de la topologie de celui-ci.

2.2.1 Le schéma agglomératif

Cette technique algorithmique prend comme point de départ un arbre en étoile (*i.e.* arbre phylogénétique non binaire constitué d'un unique noeud interne, le centre, auquel sont connectées les n feuilles). Le *schéma agglomératif* est schématisé dans la Figure 2.2 pour un arbre phylogénétique non enraciné. Lors de chaque étape, une paire de noeuds reliée au centre c est

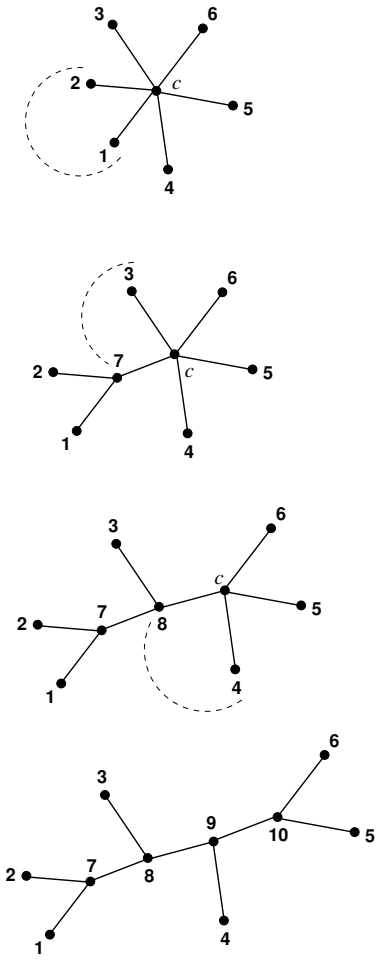

FIG. 2.2 – **Le schéma agglomératif pour construire un arbre phylogénétique non enraciné**
Partant d'un arbre en étoile de centre c (en haut), un arbre phylogénétique binaire a été obtenu (en bas) en agglomérant successivement les paires de noeuds $\{1, 2\}$, $\{3, 7\}$ puis $\{4, 8\}$.

sélectionnée et agglomérée à un nouveau noeud interne u. Le noeud interne u est relié au centre c, et la procédure est répétée itérativement jusqu'à ce que c soit de degré 3 ou, éventuellement, lorsque le critère à optimiser impose l'arrêt.

La première étape de la procédure schématisée dans la Figure 2.2 consiste à sélectionner la paire de feuilles $\{1, 2\}$ et à l'agglomérer au nouveau noeud interne $u = 7$, lui-même relié au noeud c. La procédure itérative est entièrement représentée jusqu'à l'obtention de l'arbre phylogénétique non enraciné binaire de la Figure 2.1.

2.2.2 Le schéma divisif

Le *schéma divisif* est schématisé dans la Figure 2.3 pour un arbre phylogénétique enraciné. Lors de chaque étape, un clade est sélectionné et divisé en deux nouveaux clades. La procédure

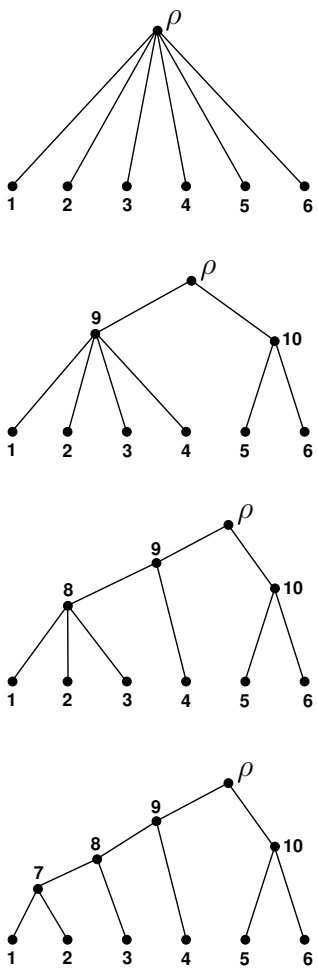

FIG. 2.3 – **Le schéma divisif pour construire un arbre phylogénétique enraciné**

Partant d'un arbre en étoile de centre r (en haut), un arbre phylogénétique binaire a été obtenu (en bas) en divisant successivement des clades en deux nouveaux clades : $\{1,2,3,4\}|\{5,6\}$, $\{1,2,3\}|\{4\}$ puis $\{1,2\}|\{3\}$.

est répétée itérativement à partir de chaque noeud interne u créé jusqu'à l'obtention d'un arbre binaire ou, éventuellement, lorsque le critère à optimiser impose l'arrêt.

La première étape de la procédure schématisée dans la Figure 2.3 consiste à créer la paire de clades $\{1,2,3,4\}|\{5,6\}$ et à les rattacher aux nouveaux noeuds internes $u = 7$ et 10 respectivement, eux-mêmes reliés à la racine ρ. La procédure itérative est entièrement représentée jusqu'à l'obtention de l'arbre phylogénétique enraciné binaire de la Figure 2.1.

2.2.3 La procédure d'insertion

Cette procédure algorithmique prend comme point de départ un arbre défini sur trois (resp. deux) feuilles dans le cadre de l'inférence d'un arbre phylogénétique non enraciné (resp. enraciné). En effet, il existe $(2 \times 3 - 5)!! = 1$ seule topologie d'arbre phylogénétique non enraciné

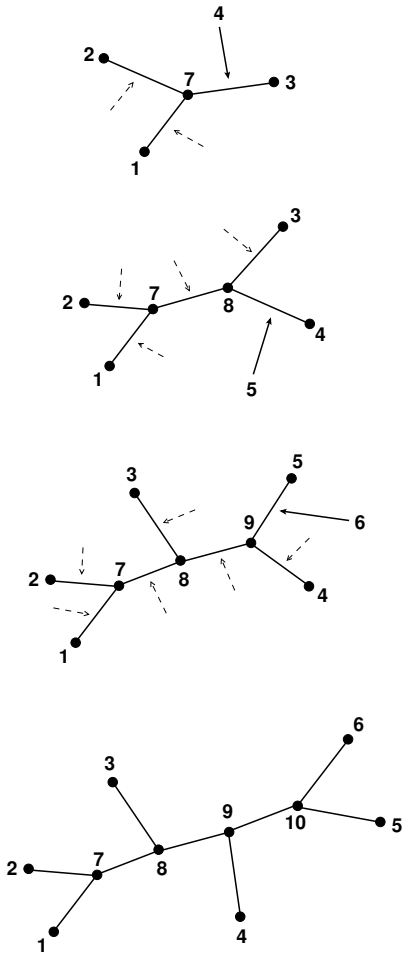

FIG. 2.4 – **La procédure d'insertion pour construire un arbre phylogénétique non enraciné**
Partant de l'unique topologie d'arbre définie sur les trois feuilles 1,2 et 3 (en haut), un arbre phylogénétique binaire a été obtenu (en bas) en insérant successivement les feuilles 4, 5 et 6.

sur trois feuilles, et également $(2\times 2-3)!! = 1$ seule topologie d'arbre phylogénétique enraciné sur deux feuilles.

A partir de cette unique topologie d'arbre, la procédure d'insertion consiste à insérer successivement une quatrième feuille sur chacune des trois branches et à calculer la valeur du critère pour chacune des trois nouvelles topologies ainsi obtenues. Le meilleur de ces arbres (au sens du critère) est alors sélectionné afin d'y insérer la cinquième feuille. La procédure est ainsi poursuivie itérativement jusqu'à la $n^{\text{ième}}$ feuille.

Cette procédure est schématisée dans la Figure 2.4. La première étape y représente l'arbre phylogénétique non enraciné défini sur les feuilles 1, 2 et 3 sur lequel on cherche à insérer la feuille 4. La topologie optimale est celle obtenue après insertion de la feuille 4 sur la branche $\{3, 7\}$. Cette topologie optimale est donc utilisée pour y insérer la feuille 5 suivant la même procédure. L'ensemble du processus d'insertion est représenté dans la Figure 2.4 jusqu'à l'ob-

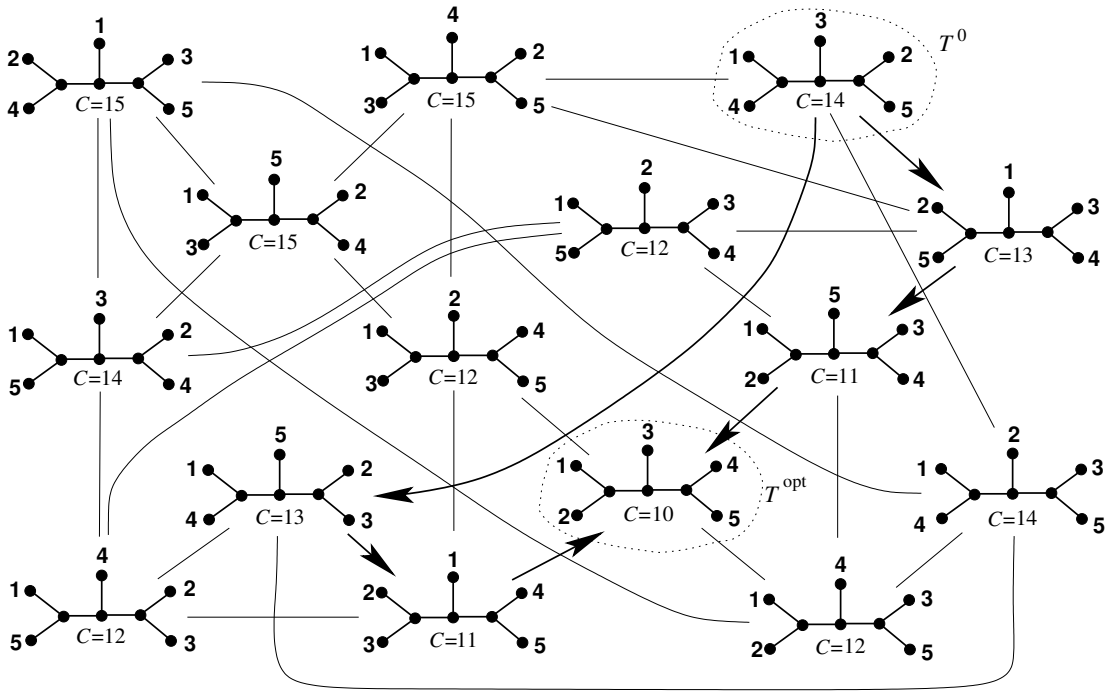

FIG. 2.5 – **Exemple d'une recherche locale**

Le voisinage est défini comme étant la permutation d'une paire de feuilles ij telle qie $\mathcal{T}_{ij} = 3$. Chaque arbre est relié aux quatre autres définissant son voisinage (*i.e.* T' est relié à T si $T' \in V(T)$). Le critère C à minimiser ici est volontairement artificiel, et est obtenu par la formule $C(T) = \sum_{i=1}^{4} \mathcal{T}_{i\ i+1}$. Une flèche relie un arbre T vers un arbre T' si $C(T') = \min_{T'' \in V(T)} [\ C(T'')\ ]$. Partant de l'arbre T^0, la recherche locale sélectionne dans son voisinage l'arbre induisant la plus petite valeur du critère C. On remarque que l'arbre T^{opt} peut être atteint par plusieurs chemin de recherche.

tention de l'arbre phylogénétique non enraciné binaire de la Figure 2.1.

Cette procédure et ses performances sont extrèmement dépendantes du choix des feuilles composant l'arbre de départ et de l'ordre d'insertion des $n-3$ autres feuilles. En effet, si, dans la Figure 2.4, on avait choisi d'insérer la feuille 6 sur les branches de l'arbre de départ défini sur les feuilles 1, 2 et 3, et si l'arbre optimal sélectionné sur les trois obtenus était de topologie $16|23$, alors il aurait été impossible de retrouver la topologie finale de la Figure 2.4.

2.2.4 Les recherches locales

Comme illustré dans la partie précédente, lorsque l'on construit un arbre phylogénétique suivant une procédure gloutonne, les choix effectués lors d'une étape ne sont pas remis en cause lors des étapes suivantes. Ainsi les techniques algorithmiques d'agglomération ou d'insertion, bien que souvent rapides, peuvent renvoyer des arbres relativement éloignés de la topologie optimale T^{opt} au sens du critère choisi. Il est néanmoins possible de se rapprocher de l'arbre optimal par un processus de recherche locale. Les recherches locales les plus courantes en

inférence phylogénétique sont les méthodes de descente.

Principe général

Soit C un critère que l'on cherche à minimiser (la description qui suit s'applique par symétrie avec un critère à maximiser). Une recherche locale par descente consiste à considérer un arbre de départ T^0 et un voisinage $V(T^0)$ de T^0. Si il existe au moins un arbre $T' \in V(T^0)$ tel que $C(T') < C(T^0)$, cela signifie que, trivialement, $T^0 \neq T^{\text{opt}}$, et qu'au moins l'un des arbres de $V(T^0)$ est plus proche de T^{opt} (au sens du critère). Le critère C permet de se déplacer dans l'espace de recherche constitué par l'ensemble des topologies d'arbre. Etant donné un arbre T, à chaque nouvel arbre $T' \in V(T)$ induisant une meilleure valeur de critère que $C(T)$, le processus de recherche locale peut être réitéré à partir de T'. Plusieurs possibilités sont envisageables :

- dès que l'on trouve, lors du parcours de $V(T)$, un arbre T' tel que $C(T') < C(T)$, on réitère une nouvelle recherche locale par descente à partir de T' ;
- on réitère une nouvelle recherche locale par descente à partir de chaque arbre $T' \in V(T)$ tel que $C(T') = \min_{T'' \in V(T)} \left[C(T'') \right]$.

Ces procédures sont itérativement reproduites tant qu'à partir d'au moins un arbre T, il existe au moins un arbre $T' \in V(T)$ tel que $C(T') < C(T)$. Soit C_{min} la valeur de critère minimale rencontrée lors d'une étape de recherche locale. Si, pour chacun des arbres T considérés à cette étape (*i.e.* $C(T) = C_{\text{min}}$), il n'existe aucun arbre $T' \in V(T)$ tel que $C(T') \leq C_{\text{min}}$, alors on peut supposer que $C_{\text{min}} = C(T^{\text{opt}})$ et la procédure de recherche locale par descente est arrêtée.

La Figure 2.5 schématise une recherche locale sur l'espace des arbres phylogénétiques non enracinés binaires définis sur $n = 5$ feuilles. Le critère C à minimiser est obtenu par la formule $C(T) = \sum_{i=1}^{4} \mathcal{T}_{i\ i+1}$. Ce critère consiste à dénombrer le nombre d'arêtes formant le chemin entre chaque paire de feuilles successives i et $i+1$. Pour chaque arbre T, le voisinage $V(T)$ est défini en permutant chaque paire de feuilles ij séparées par trois arêtes.

Améliorations

La technique de recherche locale par descente est connue pour être très dépendante du choix du voisinage V. Le choix d'un voisinage restreint laisse présager que la méthode de descente risque de s'arrêter dans un optimum local, *i.e.* $C_{\text{min}} \neq C(T^{\text{opt}})$. Le choix d'un voisinage plus grand offre une certaine garantie que la descente ne s'arrêtera pas dans un optimum local mais chaque étape sera d'autant plus longue que le voisinage sera de taille importante. Une solution intermédiaire consiste à lancer la méthode de descente à partir de plusieurs arbres de départ T^0. Plus le nombre d'arbres de départ est élevé, plus la probabilité de découvrir un trajet menant à T^{opt} est élevée. D'autres techniques peuvent néanmoins être employées pour éviter d'être bloqué au premier optimum local rencontré.

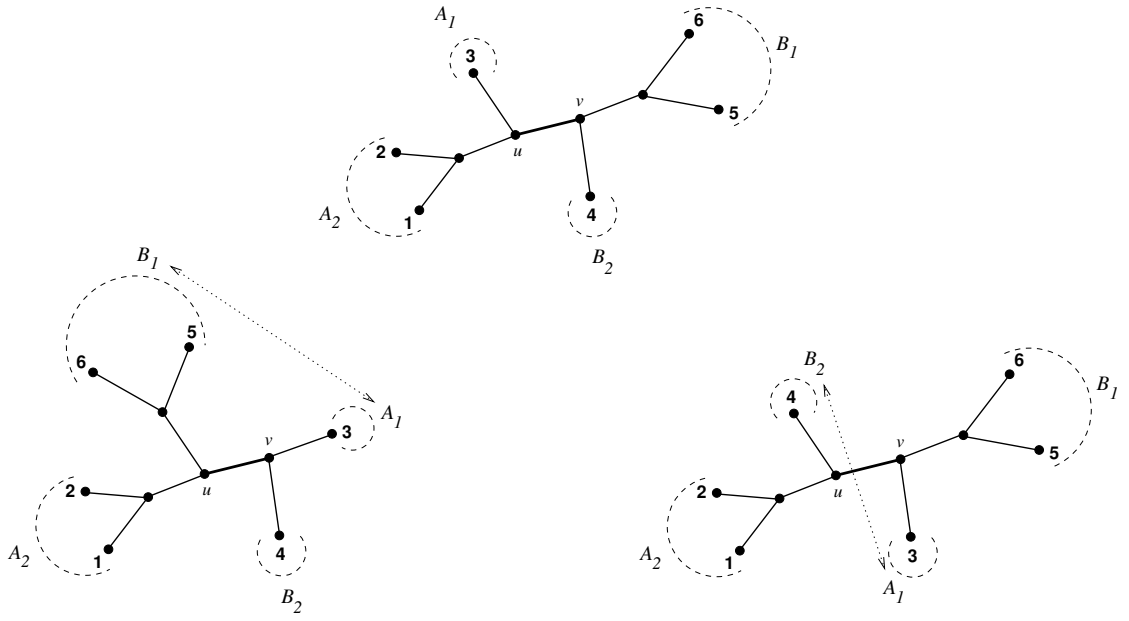

FIG. 2.6 – **Réarrangement de type NNI**

A partir d'une arête interne $\{u, v\}$, on peut définir une quadripartition $A_1|A_2|B_1|B_2$ des feuilles de l'arbre (en haut) ; deux nouveaux arbres phylogénétiques binaires ont été obtenus (en bas) en échangeant les sous-arbres $T_{|_{A_1}}$ et $T_{|_{B_1}}$ (à gauche), ainsi que les sous-arbres $T_{|_{A_1}}$ et $T_{|_{B_2}}$ (à droite).

La méthode du Recuit Simulé (Kirkpatrick et al., 1983) autorise, sous certaines probabilités, la considération d'arbres $T' \in V(T)$ tels que $C(T') > C(T)$.

La recherche Tabou (Glover, 1986; Hansen, 1986) choisit à chaque étape la meilleure solution $\min[C(T') : T' \in V(T)]$ même si $C(T') > C(T)$. Lorsqu'on atteint un optimum local par rapport au voisinage $V(T)$, la recherche Tabou se déplace vers la meilleure solution, même si elle est plus mauvaise. Néanmoins, pour interdire de revenir sur T, une mémorisation des dernières solutions visitées est nécessaire, ce qui implique une gestion de la mémoire parfois délicate voire coûteuse.

La recherche GRASP (*Greedy Randomized Adaptive Search Procedure* ; Feo and Resende, 1995) consiste à d'abord choisir une procédure heuristique (*e.g.* procédure d'insertion). Un arbre sur quatre feuilles est inféré par cette heuristique puis amélioré par recherche locale. L'arbre de quatre feuilles optimal est ensuite complété par une cinquième feuille (tirée aléatoirement ou pas) à l'aide de l'heuristique puis amélioré par recherche locale. Cette procédure est réitérée jusqu'à l'obtention d'un arbre défini sur les n feuilles.

La recherche locale à voisinages variables (RVV ; Mlanedovic and Hansen, 1997) consiste à essayer un voisinage différent lorsqu'on a atteint un optimum local par recherche locale avec un voisinage initial. Pour qu'une RVV soit efficace, il est recommandé d'utiliser des structures de voisinage qui soient complémentaires en ce sens qu'un optimum local pour un voisinage ne l'est

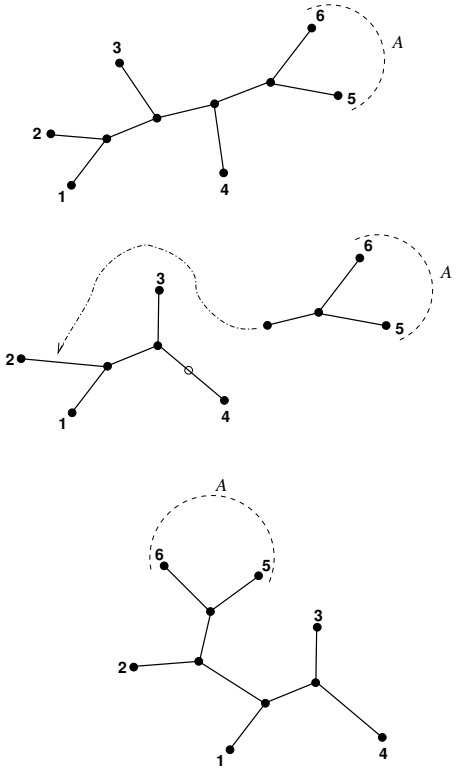

FIG. 2.7 – **Réarrangement de type SPR**

Etant donné un ensemble de feuilles $A \subset \mathcal{L}_T$, le sous-arbre enraciné correspondant $T_{|A}$ est détaché puis rebranché sur une autre arête du sous-arbre non enraciné restant $T_{|\mathcal{L}_T - A}$.

pas forcément pour un autre.

Une dernière technique pour améliorer les performances d'une recherche locale consiste à bruiter les données initiales lorsqu'on atteint un minimum local (Charon and Hudry, 1993). L'arbre optimal obtenu par une seconde recherche locale à partir des données bruitées est ensuite réutilisé dans une troisième recherche locale à partir des données initiales.

Voisinages d'arbres phylogénétiques binaires

La manière la plus courante pour définir le voisinage $V(T)$ d'un arbre phylogénétique T consiste à effectuer des mouvements topologiques sur T. Les trois types de mouvement topologique les plus utilisés en reconstruction phylogénétique par recherche locale sont les mouvements NNI (*Nearest-Neighbor Interchanges*), SPR (*Subtree Pruning and Regraphting*) et TBR (*Tree Bisection and Reconnection*) (Swofford et al., 1996, p. 484). La suite de cette partie décrit chacun de ces trois mouvements sur des arbres phylogénétiques non enracinés binaires (leur adaptation aux cas enracinés étant relativement simple).

Les mouvements topologiques NNI reposent sur le fait qu'une branche interne $\{u, v\}$ d'un

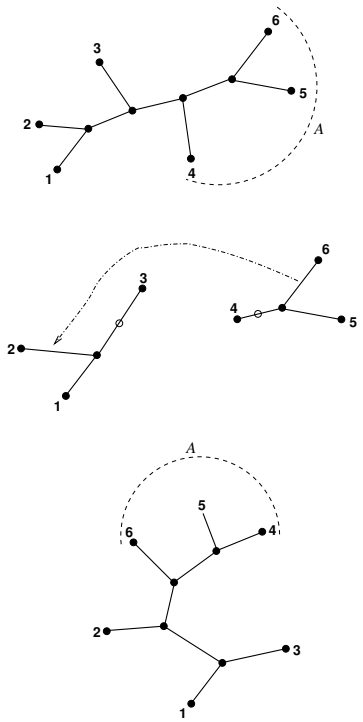

FIG. 2.8 – **Réarrangement de type TBR**

Etant donné un ensemble de feuilles $A \subset \mathcal{L}_T$, le sous-arbre non enraciné correspondant $T_{|A}$ est détaché. Une nouvelle racine y est définie et le sous-arbre $T_{|A}$ est rebranché par cette racine sur une autre arête du sous-arbre non enraciné restant $T_{|\mathcal{L}_T - A}$.

arbre phylogénétique non enraciné T induit une bipartition $A|B$ des feuilles de T, *i.e.* $A \subset \mathcal{L}_T$ et $B = \mathcal{L}_T - A$. Cette même branche interne induit également une quadripartition des feuilles de T. Ainsi $A = A_1|A_2$ telle que pour toutes les paires d'espèces $(a_1, a_2) \in A_1 \times A_2$, le chemin reliant a_1 à a_2 passe par u. Le même raisonnement s'applique pour définir $B = B_1|B_2$. Soient $T_{|A_1}$, $T_{|A_2}$, $T_{|B_1}$ et $T_{|B_2}$, les quatre sous-arbres de T induits par la quadripartition $A_1|A_2|B_1|B_2$. Pour chaque branche de T, le mouvement NNI définit deux nouveaux arbres en échangeant les sous-arbres $T_{|A_1}$ et $T_{|B_1}$ et en échangeant $T_{|A_1}$ et $T_{|B_2}$, comme schématisé dans la Figure 2.6. Ces deux seuls échanges sont suffisants car considérer les deux paires de sous-arbres $T_{|A_2}, T_{|B_1}$ et $T_{|A_2}, T_{|B_2}$ conduit aux deux mêmes nouvelles topologies. Comme un réarrangement NNI implique deux nouvelles topologies à partir de chaque branche interne d'un arbre, le voisinage NNI d'un arbre phylogénétique non enraciné binaire défini sur n feuilles est donc composé de $2 \times (n-3) = 2n-6$ nouvelles topologies (Robinson, 1971). En outre, une succession de mouvements NNI permet d'explorer l'espace constitué par les $(2n-5)!!$ arbres phylogénétiques (Robinson, 1971). La Figure 2.5 schématise l'ensemble de l'espace de recherche des arbres phylogénétiques à 5 feuilles, et représente, pour chacun d'entre eux, les $2 \times 5 - 6 = 4$ arbres définissant leur voisinnage NNI.

Etant donné un sous-ensemble de feuilles $A \subset \mathcal{L}_T$, un mouvement topologique SPR sur un arbre phylogénétique T consiste à détacher le sous-arbre enraciné $T_{|_A}$ de T et à l'insérer sur une des branches du sous-arbre non enraciné restant $T_{|_{\mathcal{L}_T - A}}$. La Figure 2.7 schématise ce mouvement. Lorsqu'un mouvement SPR est effectué, cela revient à sélectionner une des $2n-3$ arêtes de détachement (*i.e.* celle induisant la racine de $T_{|_A}$) puis à sélectionner une des $2n-4$ arêtes d'insertion. Néanmoins, l'ensemble des $(2n-3)(2n-4)$ arbres phylogénétiques ainsi obtenu n'est pas forcément composé d'arbres distincts. Il a été montré que le voisinage SPR d'un arbre phylogénétique non enraciné binaire est composé de $2(n-3)(2n-7)$ arbres distincts (Allen and Steel, 2001).

Un mouvement topologique TBR s'inspire du mouvement SPR en ce sens qu'il détache un sous-arbre $T_{|_A}$, mais ce dernier n'est pas considéré comme enraciné. Chaque arête de $T_{|_A}$ est considérée comme une racine potentielle et le sous-arbre $T_{|_A}$ est alors inséré dans $T_{|_{\mathcal{L}_T - A}}$ par une de ces racines potentielles. La Figure 2.8 schématise ce mouvement. Il a été montré que la taille du voisinage TBR d'un arbre phylogénétique non enraciné binaire T est de taille maximale $(2n-3)(n-3)^2$ et est dépendante de la topologie de T (Allen and Steel, 2001).

Un mouvement NNI est un mouvement SPR consistant à détacher un sous-arbre $T_{|_A}$ et à l'insérer sur une arête voisine de l'arête de détachement. Le voisinage NNI d'un arbre T est donc inclus dans le voisinage SPR de T. Suivant un raisonnement similaire, on observe que le voisinage SPR de T est inclus dans le voisinage TBR de T (Maddison, 1991).

Chapitre 3

L'inférence phylogénétique

Tous les étudiants de la nature doivent faire de ceci leur règle : quelque soit ce que l'esprit saisit avec une satisfaction toute particulière, ceci ne doit rester qu'à l'état de suspicion.

Francis Bacon

Il n'y a aucune certitude que le code des lois de la nature pourra un jour être dévoilé et lu, parce qu'il n'y a aucune certitude qu'une créature divine, même plus rationnelle que nous-mêmes, ait pu un jour formuler un tel code capable d'être lu.

Joseph Needham

Sommaire

3.1 Les différents types de données biologiques 42

3.2 Optimiser le critère du maximum de parcimonie 44

3.2.1 Calcul de la valeur de parcimonie d'un arbre phylogénétique 44

3.2.2 Recherche de l'(des) arbre(s) phylogénétique(s) le(s) plus parcimonieux 46

3.3 Modèles probabilistes d'évolution des séquences génétiques 47

3.4 Optimiser le critère du maximum de vraisemblance 50

3.4.1 Calcul de la vraisemblance d'un arbre phylogénétique 50

3.4.2 Recherche de l'arbre phylogénétique le plus vraisemblable 52

3.5 Optimiser les critères basés sur les distances évolutives 53

3.5.1 Calculs de distances évolutives 53

3.5.2 Les critères basés sur les distances évolutives 57

3.5.3 Inférence d'arbres phylogénétiques à partir de distances 59

3.6 Estimation de la fiabilité d'un arbre phylogénétique inféré par l'optimisation d'un critère 67

3.6.1 Comparaison de deux arbres phylogénétiques 67

3.6.2 Evaluation de la fiabilité des branches d'un arbre phylogénétique 69

Ce chapitre expose les principales techniques de l'inférence phylogénétique, c'est-à-dire les méthodologies permettant de construire un arbre phylogénétique décrivant l'évolution d'un ensemble d'espèces à partir d'un gène séquencé pour chacune de ces espèces. Après une brève description des types de données biologiques permettant d'inférer un arbre phylogénétique, les principaux critères d'inférence sont présentés : le maximum de parcimonie, le maximum de vraisemblance et les critères s'appuyant sur les distances évolutives entre espèces.

3.1 Les différents types de données biologiques

Caractères et distances

Les données biologiques à partir desquelles on peut inférer un arbre phylogénétique peuvent être classées en deux grands types : les caractères discrets et les mesures de distance. Toutes les techniques d'inférence basées sur les caractères cherchent à comparer les similitudes globales entre les séquences de caractères. Les autres méthodes, s'appuyant sur des mesures de distance quantifiant une quantité d'évolution entre chaque paire d'espèces, cherchent à retrouver la distance additive qui explique le mieux les mesures entre paires de taxons. Néanmoins, une relation existe entre ces deux types de technique. Les données formées par des caractères sont en effet souvent transformées en mesures de distance représentant une observation quantitative entre chaque paire de séquences de caractères. L'unique méthode permettant d'obtenir des mesures de distance directement à partir des molécules d'ADN est l'hybridation ADN-ADN (Johnson et al., 1970) consistant à estimer un taux de dissociation entre paires de séquences d'ADN soumises à de fortes températures. Inversement, des données continues telles que les mesures de distance peuvent être discrétisées en les classant, par exemple, suivant différents niveaux d'importance.

Les données composées de caractères discrets sont celles qui permettent d'associer un *état de caractère* x_{ip} à chaque taxon i pour chaque caractère p. Ces caractères peuvent être binaires (*e.g.* $x_{ip} = 1$ si le taxon i présente la caractéristique p, $x_{ip} = 0$ sinon), ou se composer de plus de valeurs discrètes. Ils peuvent s'appuyer sur des observations morphologiques (*e.g.* x_{ip} encode la couleur des yeux) ou moléculaires (*e.g.* le $p^{\text{ième}}$ nucléotide d'un gène séquencé pour l'espèce i est x_{ip}).

La plupart des méthodes d'inférence basées sur les caractères font l'hypothèse d'indépendance entre chaque caractère. Ces méthodes cherchent également à comparer des séquences de caractères homologues, c'est à dire hypothétiquement issus, parfois avec modifications, de l'état de ces caractères chez une espèce ancestrale.

$$
\begin{aligned}
S_1 &= \texttt{A G A A T A G C C A} \\
S_2 &= \texttt{A G G A T A G G A} \\
S_3 &= \texttt{A G T A T G G A}
\end{aligned}
$$

$$
\begin{array}{cc}
 & \texttt{0 1 2 3 4 5 6 7 8 9} \\
S_1 & \texttt{A G A A T A G C C A} \\
S_2 & \texttt{A G G A T A G G - A} \\
S_3 & \texttt{A G T A T - G G - A}
\end{array}
$$

FIG. 3.1 – **Exemple d'alignement de séquences d'ADN**

D'après Caraux et al. (1995) : les trois séquences S_1, S_2 et S_3 (en haut) ont été alignées (en bas) afin de respecter l'hypothèse de l'homologie de position.

Les données moléculaires

En pratique, l'utilisation de séquences de caractères moléculaires (*e.g.* ADN, ARN, protéines, ...) est la méthode la plus simple et la plus répandue. L'*ADN* est principalement formé par quatre molécules différentes, nommées *nucléotides* : l'adénine, la cytosine, la guanine et la thymine, respectivement encodées par les états de caractère A, C, G et T. Ces molécules sont reliées entre elles par des liaisons phosphodiester et constituent ainsi une longue chaîne. Le séquençage d'un gène (*i.e.* une partie de l'ADN codant une fonction dans l'organisme) revient à retranscrire la succession des nucléotides le définissant sous la forme d'un mot construit à partir de l'alphabet $\{\text{A , C , G , T}\}$. Si on dispose d'un ensemble de séquences génétiques homologues (*e.g.* gènes ayant la même séquence ancestrale), leur utilisation implique néanmoins une autre hypothèse : l'homologie de position[1]. Cette deuxième hypothèse implique que plusieurs séquences génétiques homologues, prélevées dans différentes espèces, ont conservé le même plan d'organisation induit par leur espèce ancestrale, et ce malgré les variations et mutations subies au cours de l'évolution reliant l'espèce ancestrale à celles utilisées pour le prélèvement. Ainsi, étant donné un ensemble de séquences génétiques correspondant à un unique gène prélevé et séquencé dans plusieurs espèces, la plupart des techniques d'inférence phylogénétique doivent passer par une étape intermédiaire d'alignement de ces séquences.

L'alignement de séquences génétiques homologues permet d'obtenir un jeu de données respectant l'hypothèse de l'homologie de position. La partie supérieure de la Figure 3.1 donne un

[1] D'après Darlu and Tassy (1993 , p. 38-40), on ne saurait que trop insister sur la différence entre séquences génétiques similaires et homologues, car l'une n'implique pas l'autre. De plus, l'homologie est un concept issu initialement de l'observation morphologique (cf page 18). Afin d'oter toute ambiguité, Fitch (1971) introduisit la notion d'*orthologie* pour qualifier l'homologie d'ascendance des gènes. Il introduisit également le terme *paralogie* qui désigne la similitude entre gènes issus d'un évènement de duplication le long d'un segment chromosomique. Ainsi, le concept d'homologie de position ne peut être considéré que dans le cadre de séquences orthologues.

exemple simple de trois séquences d'ADN S_1, S_2 et S_3. La partie inférieure de la Figure 3.1 représente un alignement possible de ces trois séquences. Cet alignement induit que lors de l'évolution de ces trois séquences :

- il n'y a eu aucun évènement mutationnel aux positions 0, 1, 3, 4, 6 et 9,
- il y a eu des évènements de substitution aux positions 2 et 7,
- il y a eu des évènements d'insertion / délétion aux positions 5 et 8.

Par exemple, les différents états de caractère au troisième site (*i.e.* position) $x_{12} = \texttt{A}$, $x_{22} = \texttt{G}$ et $x_{32} = \texttt{T}$ induisent des mutations de type substitution (*i.e.* mutation d'un nucléotide en un autre) dans la troisième position de ces séquences. Le signe -, nommé gap, induit que, lors de l'évolution de la séquence S_2 vers S_3, l'état de caractère A situé en sixième position a subit un évènement de délétion. Réciproquement, on peut aussi dire que, lors de l'évolution de la séquence S_3 vers S_2 (on ne connait pas *a priori* le sens de l'évolution), l'état de caractère A s'est inséré en sixième position.

3.2 Optimiser le critère du maximum de parcimonie

Les méthodes basées sur l'optimisation du maximum de parcimonie (MP) reposent sur l'idée que la représentation optimale des relations entre séquences de caractères sous forme d'arbre phylogénétique est celle qui implique le minimum d'évènements mutationels. Si, par exemple, les données initiales sont constituées d'un alignement de n séquences d'ADN, l'(les) arbre(s) optimisant le critère du maximum de parcimonie est celui (ceux), parmi les $(2n-5)!!$ arbres phylogénétiques non enracinés binaires T possibles et parmi toutes les séquences possibles aux noeuds internes de T, qui induit (induisent) le nombre minimal de mutations le long de ses (leurs) branches. Dans cette approche, les sites sont donc traités de manière indépendante. Ce dénombrement des mutations peut également varier suivant la pondération que l'on associe à chaque type de mutations et à chaque site. Le dénombrement des mutations induit par un arbre phylogénétique est également nommé la valeur de parcimonie ou longueur de cet arbre. Ainsi, le problème consistant à optimiser le critère MP peut s'énoncer plus simplement : étant donné un ensemble de séquences de caractères, quel(s) est (sont) l' (les) arbre(s) de longueur minimale qui explique(nt) ce jeu de données ? Ce problème ayant été montré NP-difficile (Foulds and Graham, 1982; Day and Sankoff, 1986), les parties suivantes explicitent comment il est solutionné de manière heuristique.

3.2.1 Calcul de la valeur de parcimonie d'un arbre phylogénétique

Etant donné un arbre phylogénétique T et un ensemble de séquences de caractères, le calcul de la valeur de parcimonie de T (souvent appelée parcimonie de Fitch) a été montré polynomial pour le critère MP non pondéré (Farris, 1970; Fitch, 1970; Fitch, 1971). Ce calcul a également été montré polynomial pour le critère MP généralisé (Sankoff and Rousseau, 1975; Sankoff and

Cedergreen, 1983). Ce dernier critère consiste à associer une pondération à chaque type de mutation et la valeur de parcimonie de l'arbre est alors souvent appelée parcimonie de Sankoff.

La parcimonie de Fitch

Un algorithme polynomial permet le calcul de la longueur d'un arbre phylogénétique enraciné binaire T pour la parcimonie de Fitch. Etant donné un ensemble de n séquences de caractères alignés sur ℓ sites, le calcul s'effectue suivant une procédure itérative à partir de chacun des ℓ sites. Cette procédure itérative consiste à calculer, pour chaque noeud interne u de T, un vecteur de parcimonie (X_u, p_u) où X_u est l'ensemble des états de caractère minimisant la longueur du sous-arbre T_u enraciné correspondant au noeud u, et p_u la valeur de parcimonie de T_u. A l'étape initiale de l'algorithme, seuls les n vecteurs de parcimonie des feuilles sont connus : si $u \in \mathcal{L}_T$, alors X_u ne contient que l'état de caractère associé à u dans l'alignement au site considéré, et $p_u = 0$. Si l'état de caractère est inconnu (*e.g.* un gap), alors X_u contient l'ensemble des états de caractère possibles. Comme l'arbre T est enraciné, chaque noeud interne u possède un fils gauche $l(u)$ et un fils droit $r(u)$ et les vecteurs de parcimonie (X_u, p_u) sont calculés à partir des deux vecteurs de parcimonie des noeuds $l(u)$ et $r(u)$ suivant la règle suivante (Fitch, 1971; Hartigan, 1973) :

$$\begin{array}{llll} \text{si} & X_{l(u)} \cap X_{r(u)} \neq \emptyset & \text{alors} & X_u = X_{l(u)} \cap X_{r(u)} \quad \text{et} \quad p_u = p_{l(u)} + p_{r(u)}, \\ \text{si} & X_{l(u)} \cap X_{r(u)} = \emptyset & \text{alors} & X_u = X_{l(u)} \cup X_{r(u)} \quad \text{et} \quad p_u = p_{l(u)} + p_{r(u)} + 1. \end{array}$$

Etant donné que T est enraciné à la racine ρ, la parcimonie de Fitch de l'arbre T pour un site donné est donc la valeur p_ρ. La parcimonie de Fitch de l'arbre T est obtenue en additionnant les valeurs p_ρ obtenues pour chacun des ℓ sites. Il a été montré que le choix de la racine ne modifie pas la valeur de la parcimonie de Fitch (Fitch, 1971; Hartigan, 1973). Ainsi on peut calculer la parcimonie de Fitch d'un arbre non enraciné en plaçant une racine sur l'une de ses branches. Le calcul du vecteur de parcimonie est effectué pour chaque noeud interne lors d'un parcours en profondeur de T. Lors de cette procédure algorithmique, chaque noeud n'est parcouru qu'une seule fois. Si les états de caractère sont définis sur un alphabet Σ, la complexité du calcul de la parcimonie de Fitch est donc de $O(n|\Sigma|)$ pour un site donné, et de $O(n\ell|\Sigma|)$ pour l'ensemble des ℓ sites.

La parcimonie de Sankoff

Le critère MP généralisé complète le critère MP non pondéré en considérant un coût $c_{x\to y}$ associé à chacune des $\Sigma(\Sigma-1)$ mutations de l'état de caractère x vers l'état $y \neq x$. Les Σ non-mutations peuvent également être associées à un coût $c_{x\to x}$. Le critère MP non pondéré est équivalent au critère MP généralisé avec $c_{x\to y} = 1$, pour tout état de caractère x, y ($x \neq y$). Le calcul de la parcimonie de Sankoff d'un arbre enraciné T s'effectue suivant une procédure algorithmique proche de celle employée pour le calcul de la parcimonie de Fitch. Chaque noeud

interne u de T est associé au vecteur de parcimonie $(p_{u,x})$ où $p_{u,x}$ représente la valeur de parcimonie de Sankoff du sous-arbre enraciné T_u si l'état de caractère au noeud u est x. A l'étape initiale de l'algorithme, les n vecteurs de parcimonie des feuilles sont calculés à partir du site considéré : si $u \in \mathcal{L}_T$, alors $p_{u,x} = 0$ si x est l'état de caractère associé à u et $p_{u,x} = +\infty$ si x est différent de l'état de caractère associé à u. Si l'état de caractère est inconnu, alors $p_{u,x} = 0$ pour tous les états de caractère x. Les vecteurs de parcimonie $(p_{u,x})$ des noeuds internes u sont calculés itérativement à partir des deux vecteurs de parcimonie des noeuds $l(u)$ et $r(u)$ suivant la règle suivante (Sankoff and Rousseau, 1975; Sankoff and Cedergreen, 1983) :

$$p_{u,x} = \min_{y \in \Sigma} \left[c_{y \to x} + p_{l(u),y} \right] + \min_{y \in \Sigma} \left[c_{y \to x} + p_{r(u),y} \right].$$

La parcimonie de Sankoff de l'arbre T est obtenue en additionnant les valeurs $\min_x [p_{\rho,x}]$ obtenues pour chacun des ℓ sites. Le choix de la racine peut modifier la valeur de la parcimonie de Sankoff si la fonction de coût n'est pas symétrique, *i.e.* $c_{x \to y} \neq c_{y \to x}$ Ainsi, pour obtenir la parcimonie de Sankoff d'un arbre non enraciné, il faut considérer chacune des $2n - 3$ racines potentielles. Si les états de caractères sont définis sur un alphabet Σ, alors l'application de la règle précédente s'effectue en $O(|\Sigma|^2)$. Ainsi la complexité du calcul de la parcimonie de Sankoff est de $O(n|\Sigma|^2)$ pour un site donné. Pour l'ensemble des ℓ sites, la complexité est de l'ordre de $O(n\ell|\Sigma|^2)$ si c est symétrique et de $O(n^2\ell|\Sigma|^2)$ si c n'est pas symétrique.

Cette approche présente l'avantage d'être très générique, car elle peut s'appliquer quelle que soit la définition du coût $c_{x \to y}$. Nénamoins, il est la plupart du temps très difficile d'avoir de bonnes estimations de ces différents coûts. Conséquemment, le critère MP généralisé n'est que très rarement utilisé en pratique.

3.2.2 Recherche de l'(des) arbre(s) phylogénétique(s) le(s) plus parcimonieux

La plupart des techniques d'inférence d'arbre minimisant les critères MP effectuent une recherche locale. Le logiciel DNAPARS du package PHYLIP (Felsenstein, 1993) cherche à optimiser la parcimonie de Fitch. Il offre une implémentation de la recherche GRASP, alternant, pour chaque espèce, un algorithme d'insertion et une recherche locale par descente sur un voisinage NNI. Le logiciel PAUP* (Swofford, 2002) propose l'implémentation d'une recherche locale par descente sur des voisinages NNI, SPR et TBR pour optimiser les parcimonies de Fitch et Sankoff. L'arbre de départ est construit par un algorithme d'insertion, où l'ordre d'insertion des espèces est aléatoire. Il est possible de lancer plusieurs recherches locales à partir de plusieurs arbres de départ. Le logiciel TNT (Goloboff et al., 2003) propose les mêmes types de recherche locale que PAUP* pour optimiser la parcimonie de Fitch mais offre des temps d'exécution bien plus rapides. Il permet en outre d'appliquer une technique de bruitage, appelée *parsimony ratchet* (Nixon, 1999), consistant, lorsque un minimum local est atteint, premièrement à pondérer aléatoirement certains sites et de continuer la recherche locale avec ce jeu de séquences bruitées, puis,

deuxièmement et lorsqu'un nouveau minimum local est atteint, à continuer la recherche locale avec le jeu de séquences initiales.

Il est courant d'observer que plusieurs arbres minimisent les critères MP. Dans ce cas de figure, une approche classique consiste à combiner l'ensemble des arbres phylogénétiques les plus parcimonieux en une unique topologie par une technique de consensus (pour plus de détails, cf Bryant, 2003). Le plus courant, nommé consensus strict, renvoie l'arbre phylogénétique enraciné (resp. non enraciné) contenant l'ensemble des clades (resp. des bipartitions de feuilles) induits par tous les arbres les plus parcimonieux.

Néanmoins, ces méthodes d'inférence sont basées sur des critères spécifiques. Les critères MP n'utilisent pas de modèles explicites caractérisant les changements entre séquences de caractères. De plus, il a été montré que l'optimisation des critères MP peut conduire à une estimation inconsistante des arbres optimaux (*i.e.* tous les arbres optimaux sont différents de l'arbre vrai) si on considère des caractères moléculaires de type ADN, ARN ou protéine (Felsenstein, 1978a). Ce biais, nommé artéfact d'attraction des longues branches (Felsenstein, 1978a; Hendy and Penny, 1989), s'explique par le fait que si beaucoup de mutations sont induites le long de deux branches d'un arbre phylogénétique, alors la minimisation des critères MP tendra à rapprocher ces deux branches dans les topologies des arbres phylogénétiques optimaux. Ces observations, qui peuvent être considérées comme une faiblesse méthodologique, ont donc mené à la considération de modèles d'évolution des séquences génétiques.

3.3 Modèles probabilistes d'évolution des séquences génétiques

Les méthodes d'inférence phylogénétique basées sur un modèle probabiliste de l'évolution s'appuient sur certaines hypothèses. Ces hypothèses générales sont pour la plupart simplificatrices mais néanmoins nécessaires pour permettre de décrire avec un cerain formalisme l'évolution des séquences génétiques. Ces différentes hypothèses sont décrites ci-dessous, ainsi que les implications pratiques dans la définition de différents modèles de l'évolution des séquences génétiques.

Les hypothèses axiomatiques sur l'évolution des séquences génétiques

Etant donnés un arbre T, une branche $\{u, v\}$ de T et une séquence génétique S_u associée à u, l'évolution le long de cette branche transforme S_u en une nouvelle séquence S_v associée au noeud v. Un modèle d'évolution décrit, pour chaque caractère composant S_u, les probabilités $p_{x \to y}$ de passage de l'état de caractère x initial, *i.e.* en u, à un état de caractère y final, *i.e.* en v. Si, par exemple, les séquences génétiques sont des brins d'ADN, les états de caractères x et y appartiennent à l'alphabet $\Sigma = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$. La plupart des modèles d'évolution décrivant

les transformations des séquences de caractères génétiques s'appuient sur les hypothèses suivantes.

- Les séquences évoluent majoritairement par le mécanisme de substitution nucléotidique. Les évènements mutationels de type insertion ou délétion ne sont pas traités.
- Les sites évoluent indépendamment les uns des autres. Les transformations affectant un site à un instant t ne sont affectées ni par les états des autres caractères de la séquence, ni par les transformations les affectant (hypothèse d'indépendance), ni par la place de ce site dans la séquence (hypothèse de distribution identique).
- Le processus d'évolution est markovien d'ordre 1, homogène et stationnaire. L'évolution d'une séquence ne dépend que de son état actuel à l'instant t et non de la suite d'évènements mutationels qui a conduit à cette séquence (processus markovien), et demeure le même pour toutes les branches de T (homogénéité). La probabilité π_x d'observer un état de caractère x ne dépend pas du moment de l'observation (stationnarité).
- Il ne peut se produire qu'au plus une mutation par unité de temps. Sur un intervalle minime de temps dt, il ne peut se produire qu'une unique mutation.

Quelques modèles classiques de l'évolution des séquences de type ADN

Conséquemment à ces différentes hypothèses, l'expression mathématique d'un modèle de substitution sur une séquence génétique est une matrice Q de taux de substitution par site et par unité de temps dt. Cette matrice, dite de taux instantanés, représente les taux de substitution pour tranformer un état de caractère x en un état de caractère y. Par la suite, le modèle d'évolution des séquences d'ADN sera choisi pour illustrer les modèles d'évolution. Pour les séquences d'ADN, le modèle d'évolution réversible le plus général (GTR ; *General Time Reversible* ; Lanave et al., 1984 ; Rodriguez et al., 1990) est représenté par la matrice de taux instantanés suivante :

$$Q_{\mathsf{GTR}} = \begin{pmatrix} \tilde{\lambda}_{\mathsf{A}} & \tilde{\mu}a\pi_{\mathsf{C}} & \tilde{\mu}b\pi_{\mathsf{G}} & \tilde{\mu}c\pi_{\mathsf{T}} \\ \tilde{\mu}a\pi_{\mathsf{A}} & \tilde{\lambda}_{\mathsf{C}} & \tilde{\mu}d\pi_{\mathsf{G}} & \tilde{\mu}e\pi_{\mathsf{T}} \\ \tilde{\mu}b\pi_{\mathsf{A}} & \tilde{\mu}d\pi_{\mathsf{C}} & \tilde{\lambda}_{\mathsf{G}} & \tilde{\mu}f\pi_{\mathsf{T}} \\ \tilde{\mu}c\pi_{\mathsf{A}} & \tilde{\mu}e\pi_{\mathsf{C}} & \tilde{\mu}f\pi_{\mathsf{G}} & \tilde{\lambda}_{\mathsf{T}} \end{pmatrix}, \tag{3.1}$$

où les $\tilde{\lambda}_x$ sont tels que la somme de chaque ligne est nulle, *e.g.* $\tilde{\lambda}_{\mathsf{T}} = -\tilde{\mu}(c\pi_{\mathsf{A}} + e\pi_{\mathsf{C}} + f\pi_{\mathsf{G}})$. Le facteur $\tilde{\mu}$ représente la vitesse à laquelle se produisent les différentes substitutions. Cette vitesse globale est modifiée par les taux relatifs a, b, c, d, e et f propres à chaque substitution tranformant l'état de caractère x en l'état de caractère y. Les paramètres π_{A}, π_{C}, π_{G} et π_{T} sont les probabilités d'observer les états A, C, G et T, respectivement. Ces derniers paramètres, assumés comme constants par l'hypothèse de stationnarité, sont le plus souvent estimés par les fréquences d'apparition dans les séquences.

Si on connait les probabilités d'apparition $A(t)$, $C(t)$, $G(t)$ et $T(t)$ des quatre nucléotides au temps t, on peut utiliser la matrice Q pour calculer les nouvelles probabilités au temps $t + dt$. En notant $L(t) = (A(t), C(t), G(t), T(t))$ le vecteur des probabilités des états de caractères à l'instant t, on obtient les probabilités au temps $t + dt$ par l'équation $L(t + dt) = L(t) + L(t)\, Q\, dt$, qui se réécrit $dL(t)/dt = L(t)\, Q$. La solution de cette dernière équation est $L(t) = L(0)e^{Qt}$, où $L(0)$ est le vecteur des probabilités ancestrales (Cox and Miller, 1965; Yang, 1994). Le calcul de $e^{Qt} = \sum_{p=1}^{+\infty}(Q^p t^p/p!)$ s'effectue en décomposant $Q = RDR^{-1}$ avec une matrice diagonale D et une matrice inversible R. On obtient alors, après calcul, le résultat $e^{Qt} = R\, e^{Dt}\, R^{-1}$. Or l'exponentielle d'une matrice diagonale D correspond à la matrice obtenue en prenant l'exponentielle de ses termes diagonaux. Ainsi, à partir de la matrice Q des taux de substitution instantanés, il est possible d'obtenir les probabilités $p_{x\to y}(t) = (e^{Qt})_{xy}$ de changement d'un état de caractère x en un état de caractère y durant un intervalle de temps t. Etant donné un arbre phylogénétique enraciné valué T et une séquence d'ADN correspondant à la racine de T, ces probabilités peuvent être utilisées pour simuler l'évolution de cette séquence le long des branches de T suivant un modèle prédéfini (Rambaut and Grassly, 1997).

Le modèle GTR est le plus général des modèles réversibles, *i.e.* considérant que $\pi_x p_{x\to y}(t) = \pi_y p_{y\to x}(t)$, pour tout couples d'états de caractère x, y. Il nécessite d'estimer les cinq des six paramètres a, b, c, d, e et f (il existe une dépendance linéaire entre eux), ainsi que les trois paramètres π_A, π_C, π_G (trois seulement sont nécessaires car $\pi_A+\pi_C+\pi_G+\pi_T = 1$). Le modèle GTR est donc un modèle à huit paramètres libres. Néanmoins il existe des modèles plus simples qui sont des cas particuliers du modèle GTR. Parmi ceux-là, le modèle JC (Jukes and Cantor, 1969) suppose que la probabilité d'observer un nucléotide est la même pour chacun des quatre (*i.e.* $\pi_A = \pi_C = \pi_G = \pi_T = 1/4$) et que tous les types de substitution ont le même taux d'apparition (*i.e.* $a = b = c = d = e = f = 1$). Si on pose le paramètre $\tilde{\alpha} = \tilde{\mu}/4$, alors $\tilde{\alpha}$ est l'unique paramètre du modèle JC caractérisé par la matrice de taux instantanés :

$$Q_{\text{JC}} = \begin{pmatrix} -3\tilde{\alpha} & \tilde{\alpha} & \tilde{\alpha} & \tilde{\alpha} \\ \tilde{\alpha} & -3\tilde{\alpha} & \tilde{\alpha} & \tilde{\alpha} \\ \tilde{\alpha} & \tilde{\alpha} & -3\tilde{\alpha} & \tilde{\alpha} \\ \tilde{\alpha} & \tilde{\alpha} & \tilde{\alpha} & -3\tilde{\alpha} \end{pmatrix}.$$

Il est alors aisé d'obtenir directement, à partir de Q_{JC}, les différentes probabilités de changement $p_{x\to y}(t) = (1 - e^{-4\tilde{\alpha}t})/4$ si $x \neq y$ et $p_{x\to x}(t) = (1 + 3e^{-4\tilde{\alpha}t})/4$ sinon. Un autre modèle, plus simple que le modèle GTR mais induisant plus de paramètres que le modèle JC, suppose également que les fréquences sont les mêmes (*i.e.* $\pi_A = \pi_C = \pi_G = \pi_T = 1/4$) mais distingue deux types d'évènement mutationnel. Les nucléotides formant les séquences d'ADN sont classés en deux catégories biochimiques, les *purines* (A et G) et les *pyrimidines* (C et T), en ce sens que les réactions chimiques permettant de transformer un nucléotide d'une famille en celui de l'autre famille sont moins probables. Une *transition* transforme un nucléotide d'une famille

en un nucléotide de la même famille (*i.e.* $A \leftrightarrow G$ et $C \leftrightarrow T$) et une *transversion* transforme un nucléotide d'une famille en un nucléotide de l'autre famille. Le modèle K2P (Kimura, 1980) suppose que les taux de tranversion sont égaux (*i.e.* $a = c = d = f = 1$) et que les taux de transitions sont plus élevés $b = e = \tilde{\kappa} \geq 1$. Si on pose les deux paramètres $\tilde{\alpha} = \tilde{\mu}\tilde{\kappa}/4$ et $\tilde{\beta} = \tilde{\mu}/4$, alors le modèle K2P est caractérisé par la matrice de taux instantanés :

$$Q_{\text{K2P}} = \begin{pmatrix} -\tilde{\alpha} - 2\tilde{\beta} & \tilde{\beta} & \tilde{\alpha} & \tilde{\beta} \\ \tilde{\beta} & -\tilde{\alpha} - 2\tilde{\beta} & \tilde{\beta} & \tilde{\alpha} \\ \tilde{\alpha} & \tilde{\beta} & -\tilde{\alpha} - 2\tilde{\beta} & \tilde{\beta} \\ \tilde{\beta} & \tilde{\alpha} & \tilde{\beta} & -\tilde{\alpha} - 2\tilde{\beta} \end{pmatrix},$$

où $\tilde{\kappa} = \tilde{\alpha}/\tilde{\beta}$ représente le taux de transition/transvertion. Lorsque $\tilde{\kappa} = 1$, on revient au modèle JC à un paramètre. On obtient alors directement les différentes probabilités de changement $p_{x \to y}(t) = (1 - e^{-4\tilde{\beta}t})/4$ si $x \neq y$ correspond à une transversion, $p_{x \to y}(t) = (1 - 2e^{-2(\tilde{\alpha}+\tilde{\beta})t} + e^{-4\tilde{\beta}t})/4$ si $x \neq y$ correspond à une transition et $p_{x \to y}(t) = (1 + 2e^{-2(\tilde{\alpha}+\tilde{\beta})t} + e^{-4\tilde{\beta}t})/4$ si $x = y$.

Si le paramètre $\tilde{\mu}$ reste constant, alors on fait l'hypothèse que tous les sites évoluent à la même vitesse. Si chaque site évolue suivant le même modèle mais à des vitesses différentes, alors le paramètre $\tilde{\mu}$ peut prendre des valeurs différentes pour chaque site. La loi gamma a été proposée pour modéliser la distribution des vitesses d'évolution pour chaque site (Yang, 1996a). L'aspect pratique de cette loi est qu'elle peut correspondre à une distribution proche d'une distribution gaussienne ou exponentielle suivant la valeur d'un unique paramètre α. Si $\alpha \leq 1$, alors la forme exponentielle indique que la plupart des sites évoluent à un taux faible mais que quelques sites peuvent évoluer à un taux très rapide. Si $\alpha > 1$, alors la forme gaussienne indique que nombre de sites évoluent à taux proche de la valeur moyenne.

3.4 Optimiser le critère du maximum de vraisemblance

Dans sa forme générale, la vraisemblance est la probabilité conditionnelle d'observer des données sous un modèle particulier. Etant donné un modèle d'évolution qui spécifie les probabilités d'observer des évènements de mutation sur un alignement de séquences génétiques, la vraisemblance L est la probabilité d'obtenir cet alignement de séquences étant donné un arbre phylogénétique T et les paramètres du modèle (Edwards and Cavalli-Sforza, 1967; Felsenstein, 1973). Cette partie décrit comment inférer un arbre phylogénétique T maximisant la vraisemblance L.

3.4.1 Calcul de la vraisemblance d'un arbre phylogénétique

Etant donné un alignement de séquences d'ADN $\mathbb{S}$, un arbre phylogénétique T enraciné valué et les paramètres d'un modèle d'évolution M, le principe général du calcul de la vraisemblance

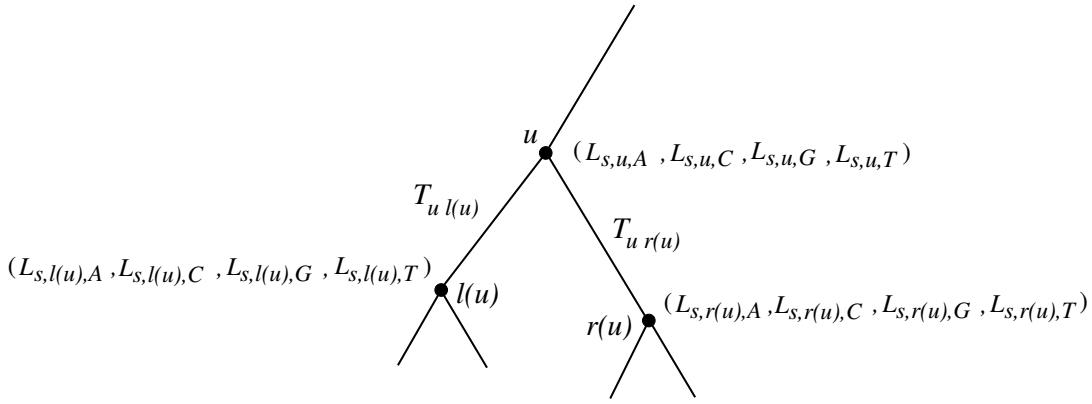

FIG. 3.2 – **Association d'un vecteur de vraisemblance à un noeud interne ainsi qu'à ses deux noeuds fils**

$L_{\mathbb{S}}$ consiste à estimer la vraisemblance L_s pour chaque site s de l'alignement $\mathbb{S}$ et, suivant l'hypothèse d'indépendance de chaque site, à poser $L_{\mathbb{S}} = \prod_s L_s$. Comme les vraisemblances sont des probabilités, le logarithme des vraisemblances est utilisé pour obtenir des valeurs plus facilement interprétables et calculables, *i.e.* $\ln L_{\mathbb{S}} = \sum_s \ln L_s$.

Pour calculer la valeur L_s, on associe un vecteur de vraisemblance $L_{s,u} = (L_{s,u,A}, L_{s,u,C}, L_{s,u,G}, L_{s,u,T})$ à chaque noeud u de T, où $L_{s,u,x}$ est la probabilité d'observer le nucléotide x au noeud u, comme schématisé dans la Figure 3.2. La plupart des modèles d'évolution des séquences d'ADN (*e.g.* JC, K2P, GTR) étant des modèles réversibles, la vraisemblance L_s ne dépend pas de la position de la racine, ce qui permet d'appliquer la procédure algorithmique sur un arbre T non enraciné. En pratique, on enracine l'arbre à partir de l'une de ses feuilles. A l'étape initiale de l'algorithme, seuls les n vecteurs de vraisemblance des feuilles sont connus : si $u \in \mathcal{L}_T$, alors $L_{s,u,x} = 1$ si x correspond à l'état de caractère x_{us} de la feuille u au site s, et $L_{s,u,x} = 0$ si $x \neq x_{us}$. Si l'état de caractère est inconnu, alors $L_{s,u,x} = 1$ pour tout état de caractère x. Comme l'arbre T est enraciné, chaque noeud interne u possède un fils gauche $l(u)$ et un fils droit $r(u)$ et les vecteurs de vraisemblance $L_{s,u}$ sont calculés à partir des deux vecteurs de vraisemblance des noeuds $l(u)$ et $r(u)$ suivant la formule :

$$L_{s,u,x} = \sum_{y \in \Sigma} \Big(p_{y \to x}(T_{u\,l(u)}) \;\; L_{s,l(u),y} \Big) \sum_{z \in \Sigma} \Big(p_{z \to x}(T_{u\,r(u)}) \;\; L_{s,r(u),z} \Big),$$

où Σ est l'ensemble des états de caractère possibles, et $T_{u\,l(u)}$ (resp. $T_{u\,r(u)}$) est la longueur de la branche $\{u, l(u)\}$ (resp. $\{u, r(u)\}$). Autrement dit, la probabilité $L_{s,u,x}$ d'observer l'état x au noeud u est le produit de la probabilité $\sum_{y \in \Sigma}(p_{y \to x}(T_{u\,l(u)}) L_{s,l(u),y})$ que x soit issu de y, l'état de caractère associé à $l(u)$, et de la probabilité $\sum_{z \in \Sigma}(p_{z \to x}(T_{u\,r(u)}) L_{s,r(u),z})$ que x soit issu de l'état z associé à $r(u)$ (Felsenstein, 1981). La vraisemblance L_s est donc obtenue par le produit $\prod_{x \in \Sigma} \pi_x L_{s,\rho,x}$ où ρ est la racine de T, et π_x la probabilité d'observation

de l'état de caractère x. Cette technique algorithmique permet ainsi de calculer le critère du maximum de vraisemblance à partir d'un arbre T binaire valué avec une complexité de l'ordre de $O(n\ell|\Sigma|^2)$, où $\Sigma = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$ (Felsenstein, 1981). Les différents paramètres du modèle M sont estimés par des procédures d'ajustement mathématique. Dans le cas où les branches de T ne sont pas valuées, leurs longueurs sont optimisées successivement ou simultanément par d'autres procédures de programmation mathématique.

3.4.2 Recherche de l'arbre phylogénétique le plus vraisemblable

Différentes approches ont été menées pour inférer un arbre optimisant le critère du maximum de vraisemblance (ML). Les logiciels CODONML, BASEML et AAML du package PAML (Yang, 1997) proposent au choix le schéma agglomératif ascendant, la procédure d'insertion ou une recherche locale par descente sur un voisinage NNI. Le logiciel FASTDNAML (Olsen et al., 1994) construit un arbre de départ par une procédure d'insertion, puis effectue une recherche locale par descente avec un voisinage SPR paramétrable (*i.e.* il est possible de définir le nombre maximal d'arêtes entre l'arête d'arrachage et l'arête de réinsertion). Le logiciel PHYML (Guindon and Gascuel, 2003) construit un arbre de départ avec une méthode de distance (cf partie suivante) puis effectue une recherche locale par descente avec un voisinage NNI augmenté (*i.e.* l'ensemble des mouvements NNI améliorant le critère sont préalablement stockés, puis un sous-ensemble topologiquement compatible de ces mouvements NNI est appliqué simultanément). Une variante, nommée PH-SPR (Hordijk and Gascuel, 2005), effectue une recherche de type RVV en définissant un voisinage SPR paramétrable lorsque la recherche locale initiale (par voisinage NNI augmenté) atteint un optimum local. Le logiciel RAXML-III (Stamatakis et al., 2005) construit un arbre de départ minimisant le critère du maximum de parcimonie par une procédure d'insertion, puis effectue une recherche de type RVV avec un voisinage SPR paramétrable (*i.e.* si aucun arbre optimal n'est trouvé dans le voisinage initial, les paramètres sont relachés afin d'aggrandir la taille des voisinages SPR). RAXML-III permet de relancer plusieurs fois cette procédure en utilisant autant d'arbres de départ qu'en offre la procédure d'insertion initiale.

Même si PHYML et RAXML-III demeurent actuellement les plus rapides dans la catégorie ML, ils n'en demeurent pas moins lents en pratique. Un total d'environ 40 minutes avec PHYML, et 8 heures avec RAXML-III ont été nécessaires pour retrouver l'arbre optimisant le critère ML à partir d'un alignement de 500 séquences sur 759 sites. Pour un alignement de 1000 séquences sur 5547 sites, environ 5 heures ont été nécessaires avec PHYML, et 6 jours avec RAXML-III (Stamatakis et al., 2005). Ainsi, même si l'optimisation du critère ML offre de très bonnes performances pour construire des arbres phylogénétiques fiables (Guindon and Gascuel, 2003), ces approches souffrent encore de temps d'exécution trop longs pour permettre de les employer sur des jeux de données de très grande taille (*e.g.* plus de 10000 séquences définies sur plus d'un million de sites). De plus, leur structure de données respectives nécessitent une grande quantité de mémoire lors de leur utilisation sur de tels jeux de données.

3.5 Optimiser les critères basés sur les distances évolutives

Les méthodes optimisant un critère de distance passent par une étape intermédiaire consistant à estimer la quantité d'évolution entre chaque paire de séquences de caractères. Cette distance peut être le nombre moyen de différences observées, mais aussi une valeur corrigée à l'aide d'un modèle d'évolution afin d'estimer le nombre moyen d'évènements mutationnels apparus entre les deux séquences. Après une brève description du lien de certaines distances évolutives avec les modèles d'évolution, les parties suivantes décrivent les critères et méthodes permettant d'inférer un arbre phylogénétique à partir d'un ensemble de distances.

3.5.1 Calculs de distances évolutives

Distances basées sur un modèle d'évolution

Etant donné un ensemble de séquences de caractères alignées, la distance Δ_{ij} entre deux séquences i et j est l'espérance du nombre de substitutions par site s'étant produit entre elles. Dans le cas où aucun modèle d'évolution n'est assumé ou dans le cas idéal où chaque différence observée correspond à un unique évènement de substitution, la distance de Hamming (*i.e.* le nombre total de différences normalisé par la taille des séquences) représente exactement Δ_{ij}. Cependant, lorsque les séquences de caractères sont issues de données moléculaires (*e.g.* ADN), il se peut qu'une différence observée soit la conséquence d'une succession de plusieurs évènements de substitution. Il devient alors nécessaire de tenir compte de ces substitutions masquées en utilisant des modèles d'évolution.

En pratique, lorsque l'on cherche à calculer des distances évolutives entre une paire ij de séquences d'ADN définies sur ℓ sites, on construit préalablement la matrice de dénombrement suivante (d'après Swofford et al., 1996) :

$$F_{ij} = \begin{pmatrix} \ell_{\mathsf{AA}}/\ell & \ell_{\mathsf{AC}}/\ell & \ell_{\mathsf{AG}}/\ell & \ell_{\mathsf{AT}}/\ell \\ \ell_{\mathsf{CA}}/\ell & \ell_{\mathsf{CC}}/\ell & \ell_{\mathsf{CG}}/\ell & \ell_{\mathsf{CT}}/\ell \\ \ell_{\mathsf{GA}}/\ell & \ell_{\mathsf{GC}}/\ell & \ell_{\mathsf{GG}}/\ell & \ell_{\mathsf{GT}}/\ell \\ \ell_{\mathsf{TA}}/\ell & \ell_{\mathsf{TC}}/\ell & \ell_{\mathsf{TG}}/\ell & \ell_{\mathsf{TT}}/\ell \end{pmatrix},$$

où ℓ_{xy} est le nombre de fois où l'état de caractère x est aligné dans la séquence i avec l'état de caractère y dans la séquence j. Quand l'alignement de séquences contient des gaps (ou des états de caractère manquants), le nombre ℓ de sites est remplacé par le nombre de couples xy tels que $x, y \in \Sigma = \{\mathsf{A},\mathsf{C},\mathsf{G},\mathsf{T}\}$. Cette dernière pratique est courante, bien que l'option MISSDIST=IGNORE la commandant dans PAUP* (Swofford, 2002), un des logiciels les plus utilisés, ne soit pas l'option par défaut. La distance de Hamming Hm_{ij} est alors trivialement obtenue par la formule :

$$\mathsf{Hm}_{ij} = 1 - \frac{\ell_{\mathsf{AA}} + \ell_{\mathsf{CC}} + \ell_{\mathsf{GG}} + \ell_{\mathsf{TT}}}{\ell},$$

et l'expression de sa variance V_{ij} associée par la formule :

$$V_{ij} = \frac{\mathsf{Hm}_{ij}(1 - \mathsf{Hm}_{ij})}{\ell}.$$

On peut également ne dénombrer que les différences moyennes de type transition :

$$\mathsf{Ts}_{ij} = \frac{\ell_{\mathsf{AG}} + \ell_{\mathsf{CT}} + \ell_{\mathsf{GA}} + \ell_{\mathsf{TC}}}{\ell}$$

et celles de type transversion :

$$\mathsf{Tv}_{ij} = 1 - \mathsf{Hm}_{ij} - \mathsf{Ts}_{ij}$$

dont les variances sont calculées de manière analogue à la variance de la distance de Hamming.

Une distance évolutive Δ_{ij} peut être vue comme la longueur de l'unique branche de l'arbre phylogénétique T défini sur i et j. Cette longueur T_{ij} peut être formulée d'une manière générale à l'aide des hypothèses d'un modèle évolutif :

$$T_{ij} = \left(\sum_{x \in \Sigma} \pi_x \sum_{y \in \Sigma - x} \pi_y R_{xy} \right) t, \tag{3.2}$$

où $R_{xy} \in \{a, b, c, d, e, f\}$ sont les taux de substitution instantanés de l'état de caractère x par y tels que définis dans la Formule (3.1) définissant Q_{GTR}. Comme le modèle GTR, ainsi que ses cas particuliers (*e.g.* JC, K2P) sont réversibles et homogènes, considérer deux séquences i et j ayant divergé à un instant t revient à considérer que i et j sont séparées par une durée $2t$. Si l'on considère le modèle JC (Jukes and Cantor, 1969), alors $R_{xy} = 1$ et $\pi_x = \pi_y = 1/4$, pour toute paire d'état de caractère $x, y \in \Sigma$. Ainsi, la Formule (3.2) se réécrit $T_{ij} = 3t/2$ avec le modèle JC. Conséquemment à la forme de la matrice Q_{JC}, il a été montré que la probabilité de substitution s'exprime avec la formule $p_{x \to y}(2t) = (1 - e^{-2\tilde{\mu}t})/4$ si $x \neq y$. Or la probabilité d'observer une différence entre i et j sur un site donné étant $P = \sum_{x \in \Sigma} \pi_x \sum_{y \in \Sigma - x} p_{x \to y}(2t)$, on a donc $P = 3(1 - e^{-2\tilde{\mu}t})/4$. Cette dernière égalité conduit, après calcul, à la formule $t = -\frac{1}{2} \ln \left(1 - \frac{4}{3}P\right)$, si l'on considère que $\tilde{\mu} = 1$ (*i.e.* la vitesse d'évolution est la même pour chacun des ℓ sites). Sachant que $T_{ij} = 3t/2$, on obtient $T_{ij} = -\frac{3}{4} \ln \left(1 - \frac{4}{3}P\right)$. En posant $P = \mathsf{Hm}_{ij}$, la formule analytique de la distance évolutive Δ_{ij} sous le modèle JC est donc :

$$\Delta_{ij} = -\frac{3}{4} \ln \left(1 - \frac{4}{3}\mathsf{Hm}_{ij}\right).$$

La variance V_{ij} de la distance évolutive Δ_{ij} sous le modèle JC est donnée par la formule suivante (Kimura and Ohta, 1972) :

$$V_{ij} = \frac{\mathsf{Hm}_{ij}(1 - \mathsf{Hm}_{ij})}{\ell(1 - 4\mathsf{Hm}_{ij}/3)^2}$$

Suivant un raisonnement similaire, on peut calculer la formule analytique de la distance Δ_{ij} et de la variance V_{ij} entre i et j sous le modèle K2P (Kimura, 1980) :

$$\Delta_{ij} = \frac{1}{2} \ln \upsilon_1 + \frac{1}{4} \ln \upsilon_2, \quad \text{et} \quad V_{ij} = \frac{1}{\ell} \Big(\upsilon_1^2 \mathsf{Ts}_{ij} + \upsilon_3^2 \mathsf{Tv}_{ij} - (\upsilon_1 \mathsf{Ts}_{ij} + \upsilon_3 \mathsf{Tv}_{ij})^2 \Big),$$

avec $\upsilon_1 = (1 - 2\mathsf{Ts}_{ij} - \mathsf{Tv}_{ij})^{-1}$, $\upsilon_2 = (1 - 2\mathsf{Tv}_{ij})^{-1}$ et $\upsilon_3 = (\upsilon_1 + \upsilon_2)/2$. Une écriture analytique existe également pour la distance du modèle le plus général, GTR (Rodrìguez et al., 1990) :

$$\Delta_{ij} = -\mathsf{Tr}\left[\Pi \ln\left(\Pi^{-1} F_{ij}\right)\right],$$

où $\mathsf{Tr}(M)$ est la trace (*i.e.* la somme des éléments diagonaux) de la matrice M et Π est la matrice diagonale formée par les fréquences π_A, π_C, π_G et π_T.

Le cas particulier des distances manquantes

Il peut arriver que certaines distances évolutives soient incalculables. Si, par exemple, $\mathsf{Hm}_{ij} \geq 0.75$, la distance JC ne peut être calculée. Ces grandes distances de Hamming provoquent une erreur lors de l'exécution du logiciel dédié DNADIST du package PHYLIP (Felsenstein, 1993). Dans ce cas de figure, Swofford et al. (1996, p. 458) suggèrent de supprimer la (les) séquence(s) impliquant ces distances manquantes, notées $\Delta_{ij} = \varnothing$. Si les séquences impliquant ces distances manquantes sont trop nombreuses, ils suggèrent alors de ne calculer que les distances de type transition Ts_{ij} (resp. transversion Tv_{ij}) si le problème est dû à une trop grande différence de type transversion (resp. transition). En dernier recours, Swofford et al. (1996, p. 458) proposent de remplacer les distances manquantes $\Delta_{ij} = \varnothing$ par une valeur arbitraire ou, plus rigoureusement, par $2\Delta_{\mathsf{max}}$ où $\Delta_{\mathsf{max}} = \max_{ij:\Delta_{ij} \neq \varnothing}[\,\Delta_{ij}\,]$. En effet, d'après l'inégalité quadrangulaire, on a :

$$\Delta_{ij} \;\leq\; \max\Big[\,\Delta_{iu} + \Delta_{jv}\;;\;\Delta_{iv} + \Delta_{ju}\,\Big] - \Delta_{uv} \;\leq\; 2\Delta_{\mathsf{max}}, \quad \forall i, j, u, v.$$

Le logiciel PAUP* (Swofford, 2002) effectue systématiquement cette dernière opération pour chaque $\Delta_{ij} = \varnothing$ rencontré.

Pour affiner cette dernière solution, plusieurs approches, cherchant à estimer les distances manquantes en utilisant celles existantes, ont été développées[2] La complétion ultramétrique (De Soete, 1984a; De Soete, 1984b; Lapointe and Kirsch, 1995) s'appuie sur l'inégalité ultramétrique. Si, d'après cette inégalité, $\Delta_{ij} \leq \max\left[\Delta_{iu}; \Delta_{uj}\right]$ pour tout triplet de taxons iju, alors cette propriété est aussi applicable au distances manquantes $\Delta_{ij} = \varnothing$. Ainsi, pour toute distance manquante $\Delta_{ij} = \varnothing$, la complétion ultramétrique est obtenue par la formule :

$$\Delta_{ij} = \min_{\substack{u \neq i,j \\ \Delta_{iu}, \Delta_{uj} \neq \varnothing}} \left[\max\left[\Delta_{iu}\;;\;\Delta_{uj}\right]\right]. \tag{3.3}$$

Cette estimation revient à rechercher la plus grande des valeurs possibles (avec la fonction $\max$) parmi celles autorisées par l'inégalité ultramétrique (avec la fonction $\min$). Suivant un raisonne-

[2] Méthodologiquement, c'est pour le cas de figure des distances non-estimables que les techniques de complétion de distances manquantes ont été développées. Néanmoins, historiquement, le développement d'algorithmes de complétion a connu un certain engouement dans les années 90 lorsque les praticiens de l'hybridation ADN-ADN souhaitèrent ne pas tenir compte de distances peu fiables, *e.g.* (Lapointe and Kirsch, 1995). La rapide ascension de l'inférence phylogénomique ces dernières années a également nécessité l'utilisation de méthodes de complétion (cf Chapitre 4).

ment similaire à partir de l'inégalité quadrangulaire, la complétion additive (Landry et al., 1996; Landry and Lapointe, 1997) est obtenue par la formule :

$$\Delta_{ij} = \min_{\substack{u,v \neq i,j \,:\, \Delta_{uv} \neq \varnothing \\ \Delta_{iu},\Delta_{uj},\Delta_{iv},\Delta_{vj} \neq \varnothing}} \Big[\max \big[\Delta_{iu} + \Delta_{jv} \;;\; \Delta_{iv} + \Delta_{ju} \big] - \Delta_{uv} \Big]. \tag{3.4}$$

Si $\tilde{m}$ est le nombre de distances manquantes, alors les procédures de complétion sont de complexité $O(\tilde{m}n)$ pour l'ultramétrique, et $O(\tilde{m}n^2)$ pour l'additive.

Guénoche et Grandcolas (1999) ont néanmoins remarqué que ces procédures pouvaient parfois surestimer la valeur de certaines distances manquantes, même lorsque (Δ_{ij}) est additive. En effet, un corrolaire de l'inégalité quadrangulaire explique que les deux plus grandes des trois sommes $S_1 = \Delta_{ij} + \Delta_{uv}$, $S_2 = \Delta_{iu} + \Delta_{jv}$ et $S_3 = \Delta_{iv} + \Delta_{ju}$ sont toujours égales. Ainsi, on a $S_1 = \max[S_2, S_3]$ si et seulement si $S_2 \neq S_3$. Conséquemment, si $\Delta_{ij} = \varnothing$, alors la complétion additive $\Delta_{ij} = \max[S_2, S_3] - \Delta_{uv}$ n'est exacte que si $S_2 \neq S_3$. La complétion est surestimée si $S_2 = S_3$, ce qui arrive lorsque i et j forment une cerise dans l'arbre équivalent à la distance additive.

Pour corriger ce biais, Guénoche et Grandcolas (1999) ont proposé deux critères pour controler d'éventuelles surestimations lors de l'application des techniques de complétion. Le premier critère, s'appuyant sur le fait qu'une distance respecte l'inégalité triangulaire, définit des bornes inférieure et supérieure pour chaque distance manquante $\Delta_{ij} = \varnothing$:

$$\max_{\substack{u \neq i,j \\ \Delta_{iu},\Delta_{uj} \neq \varnothing}} \Big| \Delta_{iu} - \Delta_{uj} \Big| \;\leq\; \Delta_{ij} \;\leq \min_{\substack{u \neq i,j \\ \Delta_{iu},\Delta_{uj} \neq \varnothing}} \Big[\Delta_{iu} + \Delta_{uj} \Big].$$

Le deuxième critère propose la calcul d'un facteur permettant d'estimer, dans le cas d'une distance proche mais non additive, l'écart entre les deux plus grandes des trois sommes S_1, S_2 et S_3. Pour chaque quadruplet de taxons $uvxy$ permettant de calculer ces trois sommes, on peut classer ces dernières de la plus petite à la plus grande : $S^{\text{min}}_{uvxy} \leq S^{\text{med}}_{uvxy} \leq S^{\text{max}}_{uvxy}$. On peut alors déterminer le facteur F_{uvxy} permettant de rendre les deux plus grandes sommes égales, *i.e.* $(1 + F_{uvxy}) S^{\text{med}}_{uvxy} = (1 - F_{uvxy}) S^{\text{max}}_{uvxy}$; d'où $F_{uvxy} = (S^{\text{max}}_{uvxy} - S^{\text{med}}_{uvxy}) / (S^{\text{max}}_{uvxy} + S^{\text{med}}_{uvxy})$. Le facteur global F, concernant l'ensemble de la matrice de distance (Δ_{ij}), est alors considéré comme étant le plus grand de tous les facteurs F_{uvxy}, *i.e.* $F = \max_{u,v,x,y}[F_{uvxy}]$. Ainsi, le deuxième critère considère que les deux sommes S_2 et S_3 sont significativement différentes si

$$\frac{\big| S_2 - S_3 \big|}{S_2 + S_3} \;\geq\; F. \tag{3.5}$$

Si les deux sommes vérifient ce critère, alors $\max[S_2, S_3] - \Delta_{uv}$ est un estimateur valide de $\Delta_{ij} = \varnothing$. Il est donc possible d'utiliser cet estimateur dans une procédure de complétion additive. Il est aussi possible d'effectuer la moyenne des estimateurs valides pour obtenir une autre complétion de $\Delta_{ij} = \varnothing$. Cette dernière technique utilisant les moyennes est appelée complétion par quadruplets (Guénoche and Grandcolas, 1999).

On remarque qu'un certain nombre de distances existantes sont nécessaires pour effectuer la complétion ultramétrique, additive ou par quadruplets d'une seule distance manquante. Ainsi il est possible d'obtenir une matrice de distance toujours incomplète après l'application d'une technique de complétion. Dans ce cas de figure, une possibilité consiste à réappliquer ce procédé itérativement jusqu'à l'obtention d'une matrice complète (Landry et al., 1996). Toutefois, il a été montré que si $(\Delta_{ij}) = (T_{ij})$ est une distance additive d'arbre, alors il ne nécessite qu'au moins $2n - 3$ distances (sous certaines conditions) pour retrouver l'arbre T (Guénoche and Leclerc, 2001).

3.5.2 Les critères basés sur les distances évolutives

Les critères de moindres-carrés LS

Minimisation d'un critère quadratique

Lorsque l'on cherche à inférer un arbre à partir d'une matrice de distance (Δ_{ij}), on considère que celle-ci est l'approximation d'un signal topologique (T_{ij}) correspondant à un arbre T valué. La distance additive (T_{ij}) n'est généralement pas égale à (Δ_{ij}) à cause d'un bruit dû à des biais méthodologiques, à des séquences trop courtes, à la présence d'états de caractères manquants dans l'alignement ou à une mauvaise estimation par le modèle d'évolution choisi. Un des grands principes des méthodes de distance consiste à ajuster deux grands types de paramètre (la topologie de T et ses longueurs de branche induisant les distances T_{ij}) afin de minimiser l'écart ε entre (Δ_{ij}) et (T_{ij}). Dans ce but, la plupart des méthodes de distance utilisent le critère des moindres carrés (LS ; *Least Squares*) suivant la formule :

$$\varepsilon = \sum_{\substack{i<j \\ i,j \in \mathcal{L}_T}} \mathcal{W}_{ij} \Big(\Delta_{ij} - T_{ij} \Big)^2, \tag{3.6}$$

où $\mathcal{W}_{ij}$ est un poids associé à chaque distance Δ_{ij}. Si $\mathcal{W}_{ij} = 1$ (Cavalli-Sforza and Edwards, 1967), alors la Formule (3.6) correspond au critère des moindres carrés ordinaires (OLS ; *Ordinary Least Squares*). Le critère OLS induit l'hypothèse que toutes les distances Δ_{ij} ont été estimées avec le même taux d'erreur. Si $\mathcal{W}_{ij}$ n'est pas le même pour chaque paire ij, alors la Formule (3.6) correspond au critère des moindres carrés pondérés (WLS ; *Weighted Least Squares*) qui cherche à exprimer l'incertitude des différentes estimations de Δ_{ij}. Un critère WLS avec $\mathcal{W}_{ij} = 1/V_{ij}$, où V_{ij} est la mesure de variance des estimateur de Δ_{ij}, se révèle optimal car il exploite une mesure théorique précise de l'incertitude de l'estimation de chaque distance Δ_{ij} (suivant certaines hypothèses statistiques, c'est un cas particulier du maximum de vraisemblance). Des expressions analytiques de V_{ij} existent pour la plupart des distances évolutives basées sur un modèle d'évolution (*e.g.* JC, K2P, GTR) ou non (*e.g.* distance de Hamming). Néanmoins , il est possible d'utiliser des approximations de ces variances. Ainsi, le critère WLS

avec $\mathcal{W}_{ij} = 1/\Delta_{ij}^2$ (Fitch and Margoliash, 1967) est couramment utilisé, et s'appuie sur l'observation que la plupart des variances V_{ij} sont proportionnelles à Δ_{ij}^2 (Kuhner and Felsenstein, 1994). Récemment, il a été expérimentalement montré (cf page 69) que les variances V_{ij} sont proportionnelles à $\Delta_{ij}^{1{,}823}$ (Sanjuán and Wróbel, 2005).

Solution matricielle

Soient :

- T un arbre phylogénétique non enraciné binaire,
- $A = (A_{\{i,j\}e})$ la matrice binaire décrivant la topologie de T, *i.e.* $A_{\{i,j\}e} = 1$ si la branche e appartient au chemin reliant i et j dans T, et $A_{\{i,j\}e} = 0$ sinon. Dans cette représentation, $\{i,j\}$ correspond à une ligne de A ; les paires de taxons ij sont donc considérées ici comme ordonnées les unes par rapport aux autres.
- $\Lambda = (l_e)$ le vecteur contenant la longueur $l(e)$ de chaque branche e,
- $\mathcal{D} = (\Delta_{\{i,j\}})$ le vecteur contenant les $n(n-1)/2$ distances Δ_{ij} (les paires $\{i,j\}$ sont dans le même ordre que dans $(A_{\{i,j\}e})$),
- $\mathcal{T} = (T_{\{i,j\}})$ le vecteur contenant les $n(n-1)/2$ distances T_{ij} (les paires $\{i,j\}$ sont dans le même ordre que dans $(A_{\{i,j\}e})$).

Le critère OLS se réécrit alors $\varepsilon_{\text{OLS}} = {}^{\text{t}}(\mathcal{T}-\mathcal{D})(\mathcal{T}-\mathcal{D})$, où ${}^{\text{t}}M$ est la transposée de la matrice M. En utilisant ces notations, l'estimation optimale des longueurs des branches de T, au sens du critère OLS, est alors donnée par la formule matricielle

$$\Lambda_{\text{OLS}} = ({}^{\text{t}}A\,A)^{-1}\,{}^{\text{t}}A\,\mathcal{D}. \tag{3.7}$$

Si $\mathcal{V}$ est la matrice diagonale contenant l'inverse des différentes pondérations $1/\mathcal{W}_{ij}$, alors le critère WLS se réécrit $\varepsilon_{\text{WLS}} = {}^{\text{t}}(\mathcal{T}-\mathcal{D})\mathcal{V}^{-1}(\mathcal{T}-\mathcal{D})$. On obtient alors les longueurs de branche optimales, au sens du critère WLS, par la formule

$$\Lambda_{\text{WLS}} = ({}^{\text{t}}A\,\mathcal{V}^{-1}\,A)^{-1}\,{}^{\text{t}}A\,\mathcal{V}^{-1}\,\mathcal{D}. \tag{3.8}$$

Cette dernière formule peut être appliquée en utilisant la matrice de variance-covariance $\mathcal{V}$ complète. Dans ce cas, on obtient le critère des moindres carrés généralisés (GLS ; *Generalized Least-Squares*) (Bulmer, 1991; Susko, 2003).

Il n'est pas rare que la résolution des équations (3.7) et (3.8) conduisent à certaines longueurs de branche négatives. Ces branches négatives ne correspondent à aucun processus biologique explicable. Une alternative revient à compléter le critère (3.6) par la contrainte $\Lambda_e \geq 0$, pour toute branche e (Lawson and Hanson, 1974).

Le traitement de matrices de distance incomplètes (*i.e.* contenant des distances manquantes) est possible en associant un poids nul $\mathcal{W}_{ij} = 0$ à chaque distance manquante $\Delta_{ij} = \varnothing$ (Swofford et al., 1996, p. 449). Cette procédure revient à considérer que l'absence d'estimateur pour une distance manquante induit une variance infinie $V_{ij} = +\infty$.

Les critères ME

Etant donnés T, un arbre phylogénétique non enraciné binaire valué, et Λ, l'ensemble des longueurs de branche $l(e)$ de T estimées à partir de $\mathcal{D}$, les méthodes dites de *Minimum Evolution* (ME) sélectionnent l'arbre T qui minimise le critère

$$l(T, \mathcal{D}) = \sum_{e} l(e). \tag{3.9}$$

Optimiser les critères ME revient donc à rechercher l'arbre phylogénétique de longueur (*i.e.* la somme des longueurs de branche) minimale. Conséquemment aux diverses manières de calculer les longueurs de branche, il existe autant de critère ME que d'estimations. Les expressions matricielles de ce critère sont $l_{\text{OLS}}(T, \mathcal{D}) = {}^{\text{t}}\mathbf{1}({}^{\text{t}}A\,A)^{-1}\,{}^{\text{t}}A\,\mathcal{D}$ pour une estimation OLS des longueurs de branche, et $l_{\text{WLS}}(T, \mathcal{D}) = {}^{\text{t}}\mathbf{1}({}^{\text{t}}A\,\mathcal{V}^{-1}\,A)^{-1}\,{}^{\text{t}}A\,\mathcal{V}^{-1}\,\mathcal{D}$ pour une estimation WLS des longueurs de branche, où $\mathbf{1}$ est un vecteur composé de 1.

Le nom *Minimum Evolution* ainsi que le critère (3.9) associé à une estimation OLS des longueurs de branches furent proposés par Rzhetsky et Nei (1992). Il a été démontré que si $(\Delta_{ij}) = (T_{ij})$, alors l'arbre optimal, au sens du critère l_{OLS}, est T (Rzhetsky and Nei, 1993; Denis and Gascuel, 2003). Autrement dit, il n'existe aucun autre arbre phylogénétique $T' \neq T$ tel que $l_{\text{OLS}}(T', \mathcal{D}) \leq l_{\text{OLS}}(T, \mathcal{D})$. Le critère ME est donc dit consistant dans sa version OLS. Il a été néanmoins démontré que le critère ME n'est pas consistant dans sa version WLS et GLS (Gascuel et al., 2001) : si $(\Delta_{ij}) = (T_{ij})$, alors il est parfois possible de trouver un arbre $T' \neq T$ tel que $l_{\text{WLS}}(T', \mathcal{D}) \leq l_{\text{WLS}}(T, \mathcal{D})$ ou $l_{\text{GLS}}(T', \mathcal{D}) \leq l_{\text{GLS}}(T, \mathcal{D})$, avec des matrices de variance-covariance $\mathcal{V}$ particulières.

Récemment, Pauplin (2000), en s'appuyant sur une nouvelle définition des distances entre sous-arbres, a proposé une formule simple pour estimer la longueur d'un arbre :

$$l_{\text{BME}}(T, \mathcal{D}) = \sum_{\substack{i<j \\ i,j \in \mathcal{L}_T}} 2^{1-\mathcal{T}_{ij}}\,\Delta_{ij}. \tag{3.10}$$

Ce critère, nommé BME (*Balanced Minimum Evolution*), a été montré consistant (Desper and Gascuel, 2004). Il a été également montré que le critère l_{BME} est un cas particulier du critère l_{WLS} avec $\mathcal{W}_{ij} \propto 2^{-\mathcal{T}_{ij}}$ dans la Formule (3.6) (Desper and Gascuel, 2004).

3.5.3 Inférence d'arbres phylogénétiques à partir de distances

Calcul de la longueur des branches d'un arbre phylogénétique minimisant un critère LS

La valuation des branches d'un arbre au sens du critère OLS est proposée, entre autre, par les logiciels FITCH (Felsenstein, 1997), PAUP* (Swofford, 2002) et FASTME (Desper and Gascuel, 2002). Ces implémentations s'appuient sur l'équation (3.7). Il a été montré que la valuation OLS des branches d'une topologie d'arbre phylogénétique binaire peut s'effectuer avec une complexité de l'ordre de $O(n^2)$ (Vach and Degens, 1991; Gascuel, 1997b). Soient (Δ_{ij}) une distance

évolutive, T un arbre phylogénétique non enraciné binaire non valué et $A, B \subset \mathcal{L}_T$ tels que $A \cap B = \emptyset$. Si on définit la distance moyenne entre A et B comme étant :

$$\Delta_{A|B} = \frac{1}{|A||B|} \sum_{(i,j)\in A\times B} \Delta_{ij}, \tag{3.11}$$

alors l'ensemble des distances moyennes entre toutes les paires de sous-arbres $T_{|_A}$ et $T_{|_B}$ s'effectue en $O(n^2)$ (Desper and Gascuel, 2002). Soit $e = \{u, v\}$ une branche interne de T induisant une quadripartition $A_1|A_2|B_1|B_2$ des feuilles de T (cf Figure 2.6). Si les précalculs de la Formule (3.11) sont effectués, alors il a été montré que l'estimation OLS de la longueur de la branche e s'obtient en $O(1)$ avec la formule suivante (Vach and Degens, 1991; Rzhetsky and Nei, 1993) :

$$l_{\text{OLS}}(e) = \frac{1}{2}\Big[\hat{\lambda}\big(\Delta_{A_1|B_1} + \Delta_{A_2|B_2}\big) + \big(1-\hat{\lambda}\big)\big(\Delta_{A_1|B_2} + \Delta_{A_2|B_1}\big) - \big(\Delta_{A_1|A_2} + \Delta_{B_1|B_2}\big)\Big],$$

où

$$\hat{\lambda} = \frac{|A_1||B_2| + |A_2||B_1|}{\big(|A_1| + |A_2|\big)\big(|B_1| + |B_2|\big)}.$$

Si $e = \{i, u\}$ est la branche externe correspondant à la feuille i, alors elle n'induit qu'une tripartition $\{i\}|A_1|A_2$ sur les feuilles de T et l'estimation OLS de la longueur $l(e)$ est calculée en $O(1)$ par :

$$l_{\text{OLS}}(e) = \frac{1}{2}\big(\Delta_{A_1|\{i\}} + \Delta_{A_2|\{i\}} - \Delta_{A_1|A_2}\big).$$

Cette technique de valuation de branche est implémentée dans le logiciel FASTME ainsi que la valuation BME. Les valuations OLS et BME de toutes les branches d'une topologie d'arbre phylogénétique à n feuilles s'effectuent avec une complexité de l'ordre de $O(n^2)$ et $O(n^2\phi_T)$, respectivement, où $\phi_T = \max_{i,j\in\mathcal{L}_T}[\ \mathcal{T}_{ij}\ ]$ représente le diamètre de T (Pauplin, 2000; Desper and Gascuel, 2002). La valuation WLS s'effectue avec une complexité de l'ordre de $O(n^3)$ (Bryant and Waddell, 1998; Makarenkov and Leclerc, 1999). La technique de Bryant et Waddell (1998) est implémentée dans le logiciel PAUP*. Une implémentation est également disponible dans le logiciel FITCH pour $\mathcal{W}_{ij} = 1/\Delta_{ij}$ et $\mathcal{W}_{ij} = (1/\Delta_{ij})^2$.

Recherche d'arbres minimisant un critère LS

Trouver l'arbre phylogénétique minimisant les critères de moindres-carrés OLS, WLS et GLS est un problème NP-difficile (Day, 1987; Day, 1996).

Néanmoins, de très nombreux travaux ont été effectués afin d'approcher la matrice additive (T_{ij}) qui minimise la version OLS du critère (3.6), *e.g.* (Vach and Degens, 1991; Hubert and Arabie, 1995). Parmi les différentes techniques, certaines sont basées sur la décomposition d'une distance additive en la somme d'une distance ultramétrique et d'une distance à centre, afin de simplifier la minimisation du critère (3.6) (Carroll and Pruzansky, 1980; Brossier, 1985). D'autres méthodes s'appuient sur des techniques de programmation mathématique afin de converger

vers T (*e.g.* Cunningham, 1978 ; De Soete, 1983), comme, par exemple, la modification itérative des distances ne respectant pas l'inégalité quadrangulaire (Roux, 1988), ainsi que l'inégalité triangulaire (Gascuel and Levy, 1996) jusqu'à l'obtention d'une distance additive d'arbre.

Le logiciel FITCH (Felsenstein, 1997) du package PHYLIP (Felsenstein, 1993) offre l'implémentation d'une recherche locale de type GRASP pour minimiser les versions OLS et WLS du critère (3.6). Pour les version WLS, il est possible de choisir entre $\mathcal{W}_{ij} = 1/\Delta_{ij}$ et $\mathcal{W}_{ij} = 1/\Delta_{ij}^2$. Pour chaque taxon successif i, FITCH construit un arbre phylogénétique T par une procédure d'insertion. Avant l'insertion d'un nouveau taxon dans T, ce dernier est utilisé comme point de départ d'une recherche locale par descente avec voisinage NNI afin de minimiser le critère (3.6). Il est également possible d'utiliser l'arbre renvoyé à la fin de la recherche GRASP comme point de départ pour une nouvelle recherche locale par descente avec voisinage SPR. Cette procédure complète peut être répétée un certain nombre de fois en utilisant des ordres aléatoires de taxons lors de la procédure d'insertion. La complexité de FITCH est citée par son auteur comme étant de $O(n^4)$ (Felsenstein, 1997).

L'algorithme MW (*Method of Weight* ; Makarenkov and Leclerc, 1999) du logiciel T-REX (Makarenkov, 2001) effectue une procédure d'insertion afin de minimiser les versions OLS et WLS du critère (3.6). L'implémentation de T-REX laisse le choix à l'utilisateur d'utiliser la pondération $\mathcal{W}_{ij}$ qu'il désire dans le cas WLS. L'algorithme s'appuie sur une procédure classique d'insertion de taxons. Comme la complexité de l'insertion d'un taxon y est de $O(n^2)$, l'algorithme MW a une complexité de $O(n^3)$. Néanmoins, le choix a été fait d'effectuer une procédure d'insertion complète de $n-2$ taxons à partir de chaque paire de taxons $i, j \in \mathcal{L}_T$. Ainsi la complexité totale de l'algorithme MW tel qu'implémenté dans T-REX est de l'ordre de $O(n^5)$.

Ces deux derniers logiciels sont les seuls permettant de traiter les matrices de distance incomplètes. Le logiciel FITCH permet ce traitement si on utilise l'option S (*Subreplicate* ; cf Felsenstein, 1993), et si on modifie le fichier contenant la matrice de distance incomplète en associant un poids $\mathcal{W}_{ij} = 0$ devant chaque distance manquantes $\Delta_{ij} = \varnothing$ et un poids $\mathcal{W}_{ij} = 1$ devant les autres distances $\Delta_{ij} \neq \varnothing$. Le logiciel T-REX offre une implémentation de ce même principe, nommée MWMODIF (Makarenkov, 2002), qui utilise l'algorithme MW. Il offre également une implémentation des complétions ultramétrique et additive (cf page 3.5.1), ainsi qu'une variante, nommée MW* (Makarenkov and Lapointe, 2004), consistant à appliquer une unique procédure de complétion ultramétrique ou additive (suivant le taux de distances manquantes), avant d'appliquer l'algorithme MW en associant un poids $\mathcal{W}_{ij} = 1$ aux distances originales, un poids $\mathcal{W}_{ij} = 1/2$ aux distances estimées par complétion et un poids $\mathcal{W}_{ij} = 0$ aux distances manquantes (s'il en reste).

Recherche d'arbres minimisant un critère ME

Même si le logiciel FITCH propose une recherche locale de type GRASP pour rechercher un arbre phylogénétique T minimisant les versions OLS et WLS du critère ME, les algorithmes les

plus couramment utilisés dans ce but demeurent les procédures agglomératives adaptées des célèbres algorithmes AddTree (Sattath and Tversky, 1977) et NJ (*Neighbor Joining* ; Saitou and Nei, 1987). Après une description précise de ces algorithmes, les parties suivantes seront dédiées à de récents algorithmes d'insertion et de recherche locale par descente exploitant la version OLS du critère ME ainsi que le critère BME, et présentant les complexités algorithmiques parmi les plus rapides à ce jour, tout critères confondus.

Les algorithmes agglomératifs

Le schéma générique des algorithmes agglomératifs (cf Chapitre 2) cherchant à optimiser les critères ME pour construire T est présenté dans la Figure 3.3. Pour plus de simplicité, l'ensemble des r taxons encore reliés au centre c à chaque itération du schéma agglomératif sera noté $\mathcal{L}_r$.

A chaque itération, ces algorithmes recherchent une paire de taxons à agglomérer dans $\mathcal{L}_r$ par l'optimisation d'un critère d'agglomération, puis estiment la longueur des deux branches externes ainsi créées avant de remplacer les deux taxons agglomérés par leur père dans la matrice de distance. La suite de cette section est divisée en deux parties : la description de plusieurs critères d'agglomération, puis la présentation de plusieurs manières d'effectuer les calculs des longueurs de branche et de réduction matricielle.

• *Les critères d'agglomération* — Basé sur le principe du schéma agglomératif, la première étape de ces algorithmes consiste à sélectionner la paire de feuilles à agglomérer. Une première approche a été proposée avec l'algorithme AddTree (Sattath and Tversky, 1977) et repose sur l'inégalité quadrangulaire. La distance évolutive (Δ_{ij}) n'étant généralement pas additive, il est rare d'observer que les deux plus grandes des trois sommes $S_1 = \Delta_{xy} + \Delta_{ij}$, $S_2 = \Delta_{ix} + \Delta_{jy}$ et $S_3 = \Delta_{iy} + \Delta_{jx}$ sont égales, pour tout quadruplet de taxons $ijxy$. Si on classe ces trois sommes sans ambiguité de la plus petite à la plus grande, *i.e.* $S^{\text{min}}_{ijxy} \leq S^{\text{med}}_{ijxy} \leq S^{\text{max}}_{ijxy}$, on peut raisonnablement espérer que les deux plus grandes restent relativement proches. Ainsi, sachant que $S^{\text{min}}_{ijxy} \leq \min\left[S^{\text{med}}_{ijxy}\ ;\ S^{\text{max}}_{ijxy}\right]$, si les taxons x et y sont voisins (et, conséquemment, i et j), alors $S_1 = S^{\text{min}}_{ijxy}$ et on obtient l'inégalité suivante :

$$\min\left[\Delta_{ix} + \Delta_{jy} - \Delta_{xy} - \Delta_{ij}\ ;\ \Delta_{iy} + \Delta_{jx} - \Delta_{xy} - \Delta_{ij}\right]\ \geq\ 0. \qquad (3.12)$$

Gascuel (1994) a proposé d'utiliser la fonction de Heaviside (Heaviside, 1893) pour modéliser cette dernière inéquation. Cette fonction, notée H, se définit par $H(t) = 1$ si $t \geq 0$, et $H(t) = 0$ si $t < 0$, où $t \in \mathbb{R}$. L'inégalité (3.12) est donc vérifiée si et seulement si

$$H\big(\Delta_{ix} + \Delta_{jy} - \Delta_{xy} - \Delta_{ij}\big)H\big(\Delta_{iy} + \Delta_{jx} - \Delta_{xy} - \Delta_{ij}\big) = 1.$$

Une paire de taxons xy formant une cerise dans un arbre T si et seulement si elle vérifie l'inégalité (3.12) pour tout ij, la paire de taxons à agglomérer est donc celle qui maximise le

critère

$$N_{xy} = \sum_{\substack{i,j \in \mathcal{L}_r \\ i \neq j}} H\big(\Delta_{ix} + \Delta_{jy} - \Delta_{xy} - \Delta_{ij}\big) H\big(\Delta_{iy} + \Delta_{jx} - \Delta_{xy} - \Delta_{ij}\big). \tag{3.13}$$

Le critère N_{xy} représente ainsi le nombre de fois où la paire xy peut être considérée comme une cerise potentielle dans T. Un arbre phylogénétique T induisant entre 2 et $\lfloor n/2 \rfloor$ cerises, le critère discret N_{xy} peut être maximisé par plusieurs paires de taxons. Le calcul de N_{xy} s'effectuant en $O(n^2)$ pour une paire xy donnée, la recherche des paires de taxons à agglomérer parmi les $O(n^2)$ paires potentielles induit une complexité de l'ordre de $O(n^4)$.

Une autre approche pour sélectionner la paire de taxons à agglomérer a été proposée par Saitou et Nei (1987) avec l'algorithme NJ. Considérant la topologie de l'arbre T obtenue après la première étape d'agglomération, *i.e.* $r = n$ (*e.g.* arbre obtenu après agglomération des feuilles $x = 1$ et $y = 2$ dans la Figure 2.2), l'estimation de la somme de la longueur des $r + 1$ branches suivant la version OLS du critère des moindres-carrés (3.6) est donnée par la formule suivante (Saitou and Nei, 1987) :

$$S_{xy} \;=\; \frac{1}{2}\Delta_{xy} \;+\; \frac{1}{2(r-2)} \sum_{i \in \mathcal{L}_r - \{x,y\}} \Big(\Delta_{xi} + \Delta_{yi}\Big) \;+\; \frac{2}{r-2} \sum_{\substack{i,j \in \mathcal{L}_r - \{x,y\} \\ i \neq j}} \Delta_{ij}.$$

Considérant la version OLS du critère ME, la paire xy minimisant le critère S_{xy} est optimale, car elle induit la topologie d'arbre T de longueur $l_{\mathsf{OLS}}(T)$ minimale. Comme le calcul de S_{xy} s'effectue en $O(n^2)$, la recherche de la paire optimale s'effectue avec une complexité de l'ordre de $O(n^4)$. Studier et Keppler (1988) ont proposé de remplacer le critère d'agglomération S_{xy} par le critère suivant :

$$Q_{xy} \;=\; R_x \;+\; R_y \;-\; (r-2)\Delta_{xy}, \tag{3.14}$$

où $R_z = \sum_{i \in \mathcal{L}_r} \Delta_{zi}$, pour tout $z \in \mathcal{L}_r$. Gascuel (1994) a démontré que maximiser Q_{xy} revient à minimiser S_{xy} en établissant la relation suivante :

$$S_{xy} \;=\; \frac{1}{2(r-2)} \left(\left(\sum_{\substack{i,j \in \mathcal{L}_r \\ i \neq j}} \Delta_{ij} \right) \;-\; Q_{xy} \right)$$

et en y observant que la somme de toutes les valeurs Δ_{ij} est une constante. Gascuel (1994) a également observé que la maximisation d'une version non-discrète du critère N_{xy} :

$$N'_{xy} \;=\; \sum_{\substack{i,j \in \mathcal{L}_r \\ i \neq j}} \big(\Delta_{ix} + \Delta_{jy} + \Delta_{iy} + \Delta_{jx} - 2\Delta_{xy} - 2\Delta_{ij}\big) \tag{3.15}$$

est équivalente à la maximisation de Q_{xy} en établissant la relation suivante :

$$N'_{xy} \;=\; (r-1)Q_{xy} \;-\; 4 \sum_{\substack{i,j \in \mathcal{L}_r \\ i \neq j}} \Delta_{ij}.$$

La maximisation de N'_{xy} s'effectue en $O(n^4)$ mais, si on calcule préalablement en $O(n)$ chacune des $O(n)$ variables R_z, la maximisation de Q_{xy} n'induit qu'une complexité de l'ordre de $O(n^2)$.

A première vue, le critère Q_{xy} représente donc une manière rapide d'agglomérer une paire de taxons xy en minimisant localement la version OLS du critère ME. Néanmoins, en s'appuyant sur une réécriture du critère Q_{xy} (Mirkin, 1996) :

$$Q_{xy} \;=\; 2\Delta_{xy} \;+\; \sum_{i \in \mathcal{L}_r - \{x,y\}} \Big(\Delta_{xi} + \Delta_{yi} - \Delta_{xy} \Big), \tag{3.16}$$

Gascuel (1997) a démontré, premièrement, que si $(\Delta_{ij}) = (T_{ij})$, alors maximiser Q_{xy} permet de retrouver une des cerises xy de T et, deuxièmement, qu'agglomérer la paire xy qui maximise Q_{xy} permet d'obtenir la topologie d'arbre dont la longueur des branches, ajustées par la version GLS du critère des moindres-carrés (3.6), est minimale. En effet, si $(\Delta_{ij}) \neq (T_{ij})$, la valeur $(\Delta_{xi} + \Delta_{yi} - \Delta_{xy})$ est l'estimation GLS de la valeur $(T_{xi} + T_{yi} - T_{xy})$ (Bulmer, 1991). Ces derniers résultats sont importants car ils démontrent que, lorsque, premièrement, on considère une distance additive, le critère de sélection Q_{xy} est consistant (voir aussi Bryant, 2005) et que, lorsque, deuxièmement, on considère une distance non-additive, le critère d'agglomération Q_{xy} permet de sélectionner la paire de taxons xy qui, agglomérée, produira à chaque étape l'arbre phylogénétique qui minimise les différentes versions du critère ME. De nombreuses méthodes de distance basées sur le schéma agglomératif inspiré par ADDTREE et NJ utilisent le critère Q_{xy}, telles que UNJ (Gascuel, 1997b), BIONJ (Gascuel, 1997a) ou MVR (Gascuel, 2000).

• *Les différents paramètres* — Après avoir agglomeré la paire de taxons xy à un nouveau noeud u, l'étape suivante consiste à estimer les longueurs T_{xu} et T_{yu}. Une fois ces longueurs estimées, les taxons x et y sont remplacés par u dans la matrice (Δ_{ij}) qui est réduite en calculant les nouvelles distances Δ_{ui}, pour tous les taxons i restant dans (Δ_{ij}). Les équations (3.17) et (3.18) dans la Figure 3.3 représentent, respectivement, la classe des formules d'estimation des longueurs de branche et la classe des formules de réduction. Si $(\Delta_{ij}) = (T_{ij})$ est une matrice additive, la valeur $(\Delta_{xi} + \Delta_{xy} - \Delta_{yi})/2$ est un estimateur de T_{xu}, pour tout $i \neq x, y$. La formule (3.17) correspond à la moyenne de ces différents estimateurs pondérée par les w_i. Si $(\Delta_{ij}) = (T_{ij})$ est une matrice additive et si u appartient au chemin reliant x (resp. y) à la feuille $i \neq x, y$ alors $\Delta_{xi} - T_{xu}$ (resp. $\Delta_{yi} - T_{yu}$) est un estimateur de Δ_{ui}. La formule (3.18) représente la moyenne de ces deux estimateurs pondérée par λ_i et $(1 - \lambda_i)$.

Dans la formule (3.17), l'algorithme NJ (Saitou and Nei, 1987; Studier and Keppler, 1988) utilise la valeur $w_i = w = 1/(2(r-2))$ comme pondération, ce qui correspond à une moyenne simple des estimateurs de T_{xu}. NJ utilise les valeurs $\lambda_i = \lambda = 1/2$ dans la formule de réduction (3.18), ce qui correspond également à une moyenne simple. Il a été récemment montré que l'algorithme NJ minimise, à chaque étape du schéma agglomératif, la longueur de l'arbre au sens du critère BME (Gascuel and Steel, 2006).

- $r = n$;
- `Tant que` $r > 2$

  `Faire` • Sélectionner la paire $x, y \in \mathcal{L}_r$ en optimisant un critère d'agglomération ;

  • Agglomérer x et y au nouveau noeud u ;

  • Estimer les longueurs T_{xu} et T_{yu} :

$$T_{xu} = \Delta_{xy} - T_{yu} = \frac{1}{2}\Delta_{xy} + \sum_{i \in \mathcal{L}_r - \{x,y\}} w_i(\Delta_{xi} - \Delta_{yi}) \tag{3.17}$$

  avec $\sum_{i \in \mathcal{L}_r - \{x,y\}} w_i = 1/2$;

  • Réduire la matrice de distance (Δ_{ij}) pour tout $i \neq x, y$:

$$\Delta_{ui} = \lambda_i \Delta_{xi} + (1 - \lambda_i)\Delta_{yi} - \lambda_i T_{xu} - (1 - \lambda_i)T_{yu} \tag{3.18}$$

  avec $\lambda_i \in [0, 1]$;

  • $r = r - 1$;
- Renvoyer T ;

FIG. 3.3 – **Schéma générique des algorithmes agglomératifs**

La donnée initiale est la matrice de distance (Δ_{ij}). Les équations (3.17) et (3.18) représentent, respectivement, la classe des formules d'estimation des longueurs de branche et la classe des formules de réduction. Les variables w_i et λ_i sont des pondérations liées aux Formules (3.17) et (3.18), respectivement. (d'après Gascuel, 2000)

L'algorithme UNJ (Gascuel, 1997b) utilise, dans la formule (3.17), les pondération w_i caractérisées par la formule suivante :

$$w_i = \frac{n_i}{2 \sum_{j \in \mathcal{L}_r - \{x,y\}} n_j} = \frac{n_i}{2(n - n_x - n_y)},$$

où, au départ (*i.e.* $r = n$), $n_j = 1$, pour tout $j = 1, ..., n$ et, à chaque étape, $n_u = n_x + n_y$ lors de l'agglomération de la paire xy au nouveau noeud u. Dans la formule de réduction (3.18), UNJ utilise les valeurs $\lambda_i = \lambda = n_x/(n_x + n_y)$. Ces deux formules définissant les paramètres w_i et λ_i permettent à UNJ de minimiser, à chaque étape du schéma agglomératif, la version OLS du critère ME.

L'algorithme BIONJ (Gascuel, 1997a) s'appuie sur une approximation linéaire de V_{ij} (*i.e.* $V_{ij} = \Delta_{ij}/\ell$) et utilise les paramètres $w_i = 1/(2(r-2))$ et

$$\lambda_i = \lambda = \frac{1}{2} + \frac{1}{2(r-2)V_{xy}} \sum_{j \in \mathcal{L}_r - \{x,y\}} (V_{yj} - V_{xj}). \tag{3.19}$$

L'estimation des longueurs de branche ne suit aucune version particulière du critère ME, à la

différence de NJ avec le critère BME et UNJ avec la version OLS du critère ME. Néanmoins, Gascuel (1994) a observé que si, à chaque étape r de réduction, on retranche une constante γ dans la formule (3.18), alors on obtient $\Delta'_{ui} = \Delta_{ui} - \gamma$, pour tout $i \neq x, y$, et $Q'_{xy} = Q_{xy} - 2\gamma$, pour tout $x, y \in \mathcal{L}_{r-1}$. Ainsi, la paire xy maximisant le critère d'agglomération Q_{xy} n'est pas modifié par la soustraction de la constante γ. Or, comme $\gamma = \lambda T_{xu} + (1-\lambda)T_{yu}$ est une constante dans la formule (3.18) calculée avec le paramètre λ de la formule (3.19), l'estimation des longueurs de branche n'a pas d'influence sur la topologie finale renvoyée par BIONJ. Par contre, la première partie de la formule (3.18) influençant la topologie à chaque étape de réduction/agglomération, la formule (3.19) est calculée de manière à minimiser la variance des estimateurs $\lambda\Delta_{xi} + (1-\lambda)\Delta_{yi}$, pour tout $i \in \mathcal{L}_r$. Ce résultat s'appuie sur un modèle simplifié des variances (V_{ij}) associées aux distances évolutives (Δ_{ij}) consistant à poser, à l'origine, que $V_{ij} = \Delta_{ij}$, et à réduire (V_{ij}) après chaque d'agglomération de xy au nouveau noeud u suivant la formule $V_{ui} = \lambda V_{xi} + (1-\lambda)V_{yi} - \lambda(1-\lambda)V_{xy}$.

Dans le cadre WLS et en utilisant une matrice de variance (V_{ij}) associée à la matrice de distance (Δ_{ij}), l'algorithme MVR (Gascuel, 2000) suit la même ligne que BIONJ en cherchant, lors de chaque itération, à minimiser la variance des estimateurs dans les équations (3.17) et (3.18) en utilisant les paramètres

$$w_i = \frac{\mu}{V_{xi} + V_{yi}} \quad \text{où} \quad \mu = \frac{1}{2}\left(\sum_{j \in \mathcal{L}_r - \{x,y\}} \frac{1}{V_{xj} + V_{yj}}\right)^{-1} \tag{3.20}$$

et

$$\lambda_i = \frac{V_{yi}}{V_{xi} + V_{yi}}, \tag{3.21}$$

respectivement. L'étape de réduction de la matrice de variance s'effectue suivant la formule $V_{ui} = \lambda_i V_{xi}$. Cet algorithme est la version simplifiée d'un autre plus complexe tenant compte des covariances associées à chaque distance. MVR utilise les Formules (3.20) et (3.21) en considérant que la matrice de variance-covariance $\mathcal{V}$ dans (3.8) est diagonale durant chacune de ses itération. Bien que fortement inspirées de NJ et UNJ, optimisant respectivement le critère BME et la version OLS du critère ME, l'algorithme MVR, tout comme BIONJ ne minimise pas explicitement un critère ME.

Les algorithmes NJ, UNJ, BIONJ et MVR utilisant le critère d'agglomération Q_{xy}, ils ont tous une complexité algorithmique de l'ordre de $O(n^3)$.

Les algorithmes d'insertion

Soient e et e' deux branches adjacentes (*i.e.* ayant un noeud interne en commun) d'un arbre phylogénétique T. Soient i une feuille à insérer dans T, et $T_{i,e}$ et $T_{i,e'}$ les deux arbres obtenus après insertion de i dans e et e', respectivement. Desper et Gascuel (2002) ont montré que la quantité $l_{\text{OLS}}(T_{i,e}) - l_{\text{OLS}}(T_{i,e'})$ peut se calculer en $O(1)$ grâce aux précalculs des distances

moyennes entre chaque paires de sous-arbres avec la formule (3.11). Un résultat similaire existe également pour le critère BME. Ils ont ainsi développé des algorithmes d'insertion cherchant à minimiser la version OLS de ME ainsi que le critère BME avec une complexité de l'ordre de $O(n^2)$ et $O(n^2\phi_T)$, respectivement, où ϕ_T représente le diamètre de T. Ces deux algorithmes sont implémentés dans le logiciel FASTME (Desper and Gascuel, 2004).

Les algorithmes de recherche locale par descente

Pour rechercher l'arbre minimisant les versions OLS et WLS du critère ME, le logiciel FITCH du package PHYLIP (Felsenstein, 1993) effectue la même recherche locale de type GRASP qu'il utilise pour minimiser les critère LS. Les complexités algorithmiques sont de l'ordre de $O(n^4)$ pour les versions OLS et WLS (Felsenstein, 1997).

D'une manière similaire aux algorithmes d'insertion décrits dans la partie précédente, Desper et Gascuel (2002) ont montré que les quantités $l_{\text{OLS}}(T) - l_{\text{OLS}}(T')$ et $l_{\text{BME}}(T) - l_{\text{BME}}(T')$, où T' est l'arbre T obtenu après un mouvement NNI, peuvent se calculer en $O(1)$. Ce résultat s'obtient grâce à des précalculs en $O(n^2)$. Rechercher le meilleur arbre d'un voisinage NNI s'effectue ainsi avec une complexité de l'ordre de $O(n)$, *i.e.* linéaire en la taille du voisinage.

3.6 Estimation de la fiabilité d'un arbre phylogénétique inféré par l'optimisation d'un critère

L'inférence phylogénétique est une des branches scientifiques où il est rarement possible de vérifier si le résultat est exact, *i.e.* si la topologie correspond réellement à la vraie histoire évolutive lorsque l'on considère des données réelles. Si une hypothèse est suggérée, il est souvent aisé de la vérifier en la testant dans diverses conditions expérimentales. Hélas, étant donnés un critère C et un arbre phylogénétique T, il est souvent impossible de vérifier si $C(T)$ représente la valeur optimale, et encore plus impossible en temps polynomial[3]. Plusieurs approches statistiques ont été toutefois développées pour tester l'optimalité d'un arbre par rapport à un autre ou pour évaluer la fiabilité d'un clade.

3.6.1 Comparaison de deux arbres phylogénétiques

Comparaison par rapport à un critère

Etant donné un critère C, nombre de tests ont été décrits pour observer si deux arbres phylogénétiques T^1 et T^2 sont significativement différents ou similaires. Ces différents tests ont été définis pour chacun des critères standards dédiés à l'inférence phylogénétique.

[3] On notera néanmoins l'approche expérimentale de Hillis *et al.* (1992), où l'évolution de bactériophages a pu être observée et représentée phylogénétiquement grâce au taux de mutation rapide du génome de ces organismes.

Pour les critères MP, un des tests les plus simples est le test WS (*Winning Sites* ; Prager and Wilson, 1988) qui consiste à effectuer un test des signes (MacStewart, 1941) sur le nombre de sites s informatifs (*i.e.* contenant au moins deux états de caractères distincts) induisant une plus petite valeur de parcimonie $\min\left[p_s(T^1); p_s(T^2)\right]$ pour les arbres T^1 et T^2. Par exemple, si les arbres T^1 et T^2 ont été construits à partir d'un alignement de séquences défini sur 100 sites, si $p_s(T^1) < p_s(T^2)$ pour 42 sites s et si $p_s(T^1) > p_s(T^2)$ pour 39 sites s, alors le test des signes s'effectue avec les deux valeurs 42 et 39 : la différence normalisée de ces deux valeurs correspond à une distribution normale, ce qui donne $p \sim 0.824$. On ne peut donc pas conclure que T^1 et T^2 sont significativement différents. Templeton (1983) a repris ce test en tenant compte, non plus du signe de la différence $p_s(T^1) - p_s(T^2)$, mais de la valeur de cette différence pour chaque site s.

Kishino et Hasegawa (1989) ont développé un test (*i.e. KH test*) applicable aussi bien avec les critères MP que ML. Soient Δp et $\Delta \ln L$ les différences de valeurs de parcimonie (à partir des sites informatifs) et de vraisemblance, respectivement, entre T^1 et T^2. Le test KH considère que T^1 et T^2 ne sont pas significativement différents si Δp ou $\Delta \ln L$ ne sont pas significativement différents de zéro (au sens d'une loi de probabilité de distribution normale). Plusieurs variantes ont ensuite été proposées afin de corriger certains biais statistiques du test KH (Shimodaira and Hasegawa, 1999; Shimodaira, 2002).

Guénoche et Garetta (2001) ont proposé plusieurs formules analytiques simples permettant d'estimer à quel point les distances additives (T^1_{ij}) et (T^2_{ij}) sont statistiquement proches d'une distance évolutive (Δ_{ij}). Ces formules permettent de tester le degré de similitude métrique entre (T^1_{ij}), (T^2_{ij}) et (Δ_{ij}) (*e.g.* variance résiduelle) ainsi que le degré de similitude topologique entre T^1, T^2 et (Δ_{ij}) (*e.g.* taux de quadruplets retrouvés).

Comparaison topologique

Une *distance topologique* $d(T^1, T^2)$ entre deux arbres phylogénétiques non enracinés définis sur le même ensemble de n taxons $\mathcal{L} = \mathcal{L}_{T^1} = \mathcal{L}_{T^2}$ est une mesure de dissemblance entre les topologies respectives de T^1 et T^2. Si $d(T^1, T^2) = 0$, alors les deux arbres sont identiques.

Sachant que chaque branche interne d'un arbre phylogénétique induit une bipartition de l'ensemble de ses feuilles, la distance de bipartition d_{RF} mesure le nombre de bipartitions de $\mathcal{L}$ induites par un arbre mais pas par l'autre (Bourque, 1978; Robinson and Foulds, 1979). Un arbre phylogénétique non enraciné dénombrant au plus $n - 3$ branches internes, la distance d_{RF} est couramment normalisée par $2n - 6$ pour la situer dans l'intervalle $[0, 1]$. Bien que ce soit la distance topologique la plus intuitive et la plus utilisée, sa variance, assez restreinte, en fait une mesure peu précise et généralement recommandée pour comparer des arbres relativement proches. Elle est de plus soumise à un biais ; si, en effet, T^2 est obtenu en effectuant un mouvement LPR (*Leaf Pruning and Regraphting*, *i.e.* un mouvement SPR où le sous-arbre arraché est défini sur une seule feuille) sur T^1, alors $d_{\mathsf{RF}}(T^1, T^2)$ sera d'autant plus élevée que l'arête de

réinsertion sera éloignée de l'arête de débranchage.

La distance d'agrément $d_{\mathsf{MAST}}(T^1, T^2)$ n'est pas du tout sensible à ce biais. Soit $T^1_{|_A}$ (resp. $T^2_{|_A}$) la restriction de l'arbre T^1 (resp. T^2) aux feuilles de $A \subset \mathcal{L}_{T^1} = \mathcal{L}_{T^2}$. La distance d'agrément se définit par la formule $d_{\mathsf{MAST}}(T^1, T^2) = \max\left[n - |A| : T^1_{|_A} = T^2_{|_A}\right]$. Elle dénombre les feuilles qui ne sont pas présentes dans la plus grande restriction commune à T^1 et T^2 (Gordon, 1980; Finden and Gordon, 1985; Goddard et al., 1994). Le coefficient normalisateur de cette distance est n. Néanmoins, cette distance topologique, tout comme d_{RF}, voit son espérance tendre vers 1 lorsque la taille n des arbres comparés T^1 et T^2 augmente. Plus précisément, il a été observé expérimentalement que $d_{\mathsf{MAST}}(T^1, T^2) \approx 1 - \sqrt{2/n}$ lorsque T^1 et T^2 sont tirés aléatoirement (Bryant et al., 2003)

La distance de quadruplets d_{quad} dénombre tous les sous-arbres de quatre feuilles (*i.e.* quadruplets) qui sont présents dans un arbre mais pas dans l'autre (Estabrook et al., 1985). Un arbre phylogénétique non enraciné induisant au plus C_n^4 quadruplets, on normalise couramment la distance d_{quad} par $2\mathsf{C}_n^4$. Cette distance topologique ne présente aucun des biais propres à d_{RF} et d_{MAST}, et sa variance élevée et peu sensible au type de distribution des arbres en fait une mesure de choix (Steel and Penny, 1993).

La distance par chemin est calculée par la formule suivante (Steel and Penny, 1993) :

$$d_{\mathsf{path}} = \sqrt{\sum_{i,j \in \mathcal{L}} \left(\mathcal{T}^1_{ij} - \mathcal{T}^2_{ij} \right)^2}.$$

Autrement dit, cette distance calcule la norme L^2 entre les matrices de distance $(\mathcal{T}^1_{ij})$ et $(\mathcal{T}^2_{ij})$, mais la norme L^1 peut également être utilisée (Williams and Clifford, 1971). La normalisation de d_{path} est difficile car cette distance ne correspond pas à un dénombrement. Néanmoins, les normalisateurs $(\mathsf{C}_n^2)^{\frac{1}{2}}$ ou C_n^2 sont des choix possibles (Steel and Penny, 1993). Son espérance est très dépendante de la topologie des arbres et présente les mesures les plus élevées lorsque les arbres sont des chenilles (*caterpillars*, *i.e.* des arbres ne contenant que deux cerises).

3.6.2 Evaluation de la fiabilité des branches d'un arbre phylogénétique

Une des mesures de fiabilité des branches d'un arbre phylogénétique les plus utilisées est celle obtenue par bootstrap. Originellement, la technique du bootstrap est un outil statistique qui repose sur la création de pseudo-données à partir d'un nombre limité d'observations, et qui permet d'obtenir des résultats statistiques d'une manière simple et efficace (Efron, 1979; Efron, 1981; Efron, 1982; Diaconis and Efron, 1983; Efron and Tibshirani, 1993).

Un échantillon bootstrap d'une séquence S définie sur ℓ sites est obtenu en effectuant ℓ tirages aléatoires d'états de caractères dans S avec remise. Ainsi, par exemple, la séquence $S'_1 = \mathtt{A\ G\ A\ A\ T\ T\ G\ C\ C\ A}$ est un échantillon bootstrap de la séquence $S_1 = \mathtt{A\ G\ A\ A\ T\ A\ G\ C\ C\ A}$ de la Figure 3.1 obtenu en tirant deux fois le site 4, en ne tirant pas le site 5 et en tirant une seule fois les autres sites. Un échantillon bootstrap se caractérise

donc par autant d'états de caractère que la séquence initiale et les mêmes états de caractères que la séquence originale, mais avec des fréquences d'apparition potentiellement différentes.

Si, par exemple, on souhaite estimer la variance V_{ij} d'une distance Δ_{ij} entre deux séquences de caractère S_i et S_j, alors la technique du bootstrap consiste à

- générer N échantillons bootstrap $S_i^1 S_j^1$, $S_i^2 S_j^2$, ... , $S_i^N S_j^N$ de la paire de séquences,
- calculer les N distances Δ_{ij}^1, Δ_{ij}^2, ... , Δ_{ij}^N à partir de chacun des échantillons bootstrap,
- calculer la variance V_{ij} des N estimateurs Δ_{ij}^1, Δ_{ij}^2, ... , Δ_{ij}^N de la distance originale Δ_{ij}.

Cette technique, complétée par une régression linéaire, a permis d'estimer la variance d'une distance évolutive comme étant de la forme $V_{ij} \propto \Delta_{ij}^{1.823}$ (Sanjuán and Wróbel, 2005).

Suivant un principe similaire à l'échantillonage d'une seule séquence, un échantillon bootstrap d'un alignement de séquences $\mathbb{S}$ est obtenu en effectuant ℓ tirages aléatoires de sites dans $\mathbb{S}$ avec remise.

Felsenstein (1985) fût le premier à suggérer d'utiliser la technique du bootstrap pour associer une probabilité aux branches internes d'un arbre phylogénétique T inféré à partir d'un alignement S de séquences. Il préconisa de générer N échantillons bootstrap à partir de S et d'inférer les N arbres T^1, T^2, .. , T^N à partir de chaque échantillon S^1, S^2, ... , S^N. Une branche interne de T induisant une bipartition de $\mathcal{L}_T$, chaque branche interne de T est alors étiquetée par le pourcentage de fois où cette même bipartition est retrouvée par les arbres T^1, T^2, .. , T^N.

Même si les valeurs de bootstrap sont couramment utilisées comme estimateurs de la fiabilité des branches internes, un grand débat reste ouvert quant au bien-fondé et à la solidité de ces estimateurs (Hillis and Bull, 1993; Felsenstein and Kishino, 1993; Berry and Gascuel, 1996; Berry et al., 2000).

Chapitre 4

L'inférence phylogénomique

On fait la science avec des faits, comme on fait une maison avec des pierres ; mais une accumulation de faits n'est pas plus une science qu'un tas de pierres n'est une maison.

Henri Poincaré

Au début d'une science, les scientifiques peuvent être fiers d'avoir découvert des centaines de lois. Mais au fur et à mesure que les lois se font plus nombreuses, ils deviennent de plus en plus mécontents de cet état de choses ; ils commencent alors à rechercher des principes sous-jacents.

Rudolph Carnap

Sommaire

4.1 La combinaison basse 73

4.1.1 *Total evidence* 73

4.1.2 Analyse simultanée à partir d'une supermatrice de caractères 74

4.2 La combinaison haute 75

4.2.1 Les techniques de consensus d'arbres phylogénétiques 76

4.2.2 La généralisation des techniques de consensus au problème du superarbre 76

4.2.3 L'algorithme Build 77

4.2.4 Les adaptations de Build 79

4.2.5 Représentation matricielle binaire 80

4.2.6 Heuristiques de recherche locale par descente 83

4.2.7 Décomposition sous forme de quadruplets 83

4.2.8 Moyennes de distances additives d'arbre 85

4.3 La combinaison moyenne 86

4.3.1 Décomposition sous forme de quadruplets 87

4.3.2 Modèles d'évolution des génomes complets 88

4.3.3 Méthodes de distance basées sur les signatures génomiques 89

4.3.4 Méthodes de distance basées sur les scores de BLAST 90

Un des objectifs de la taxonomie phylogénétique est de construire des arbres définis sur un très grand nombre d'espèces, voire, à plus long terme, de reconstruire l'Arbre de la Vie, modélisant l'histoire évolutive de l'ensemble des êtres vivants (contemporains ou fossiles). Néanmoins, ces grands arbres se doivent aussi de renvoyer une représentation relativement fidèle du processus d'évolution des séquences depuis lesquelles ils sont construits. Il n'est en effet pas rare d'observer des différences topologiques entre deux arbres phylogénétiques sur un même ensemble d'espèces mais inférés à partir de deux gènes différents. Ce phénomène s'explique par les différentes pressions de sélection subies par chaque gène, ou d'éventuels transferts horizontaux entre eux au cours de l'histoire évolutive (Doolittle, 1999). Il s'explique également par les erreurs stochastiques impliquées par le fait que certains gènes sont définis sur un petit nombre de sites (*e.g.* moins de 200). A la lumière de cette dernière hypothèse, il a été observé (Huelsenbeck and Hillis, 1993; Hillis et al., 1994; Philippe and Douzery, 1994; Huelsenbeck, 1995; Hillis, 1996; Wiens, 1998a; Wiens, 1998b; Wiens, 1998c; Wiens and Servedio, 1998) que l'augmentation de la quantité de gènes à analyser permettait d'obtenir une représentation arborée plus représentative de l'évolution du vivant contemporain (*i.e.* moins dépendante de la pression de sélection subie par chaque gène). De plus, d'autres études tendent à prouver que l'augmentation du nombre d'espèces à représenter lors d'une reconstruction phylogénétique améliore aussi la qualité de l'arbre inféré (Lecointre et al., 1993; Graybeal, 1998; Hillis, 1998; Rannala et al., 1998; Wiens, 1998c). Malheureusement, l'augmentation du nombre d'espèces et de la quantité de séquences conduit à la création de données incomplètes. Cette incomplétude admet deux principales causes :

- certains gènes disparaissant au cours de l'histoire évolutive, on ne peut disposer du prélèvement de certains gènes pour chaque espèces vivantes ;
- le séquençage du génome n'est pas mené uniformément, ce qui conduit à la surreprésentation de certains gènes ou de certaines espèces par rapport à d'autres.

De très nombreuses techniques ont été développées afin de traiter ces données génétiques incomplètes pour pouvoir s'en servir de support à l'inférence d'arbres. Toutes ces techniques sont regroupées sous l'étiquette nommée *inférence phylogénomique* afin de les distinguer des méthodes d'*inférence phylogénétique*, ces dernières ne travaillant qu'à partir d'un unique gène. Ce chapitre décrit les principales approches d'inférence phylogénomique en adoptant la formalisation de Schmidt (2003 ; ch. 7) :

- la *combinaison basse* regroupe les méthodes cherchant à inférer un arbre phylogénétique directement à partir de la concaténation des séquences de caractères incomplètes ;
- la *combinaison haute* regroupe les méthodes de superarbre, consistant à inférer un arbre phylogénétique à partir de chaque données distinctes (*e.g.* moléculaire, morphologique), puis à combiner ces arbres en une unique topologie : le superarbre ;
- la *combinaison moyenne* désigne les méthodes passant par une étape d'interprétation de

l'information phylogénétique induite par les jeux de données initiaux qui ne soit pas les arbres eux-mêmes, puis utilise cette information pour inférer un arbre.

4.1 La combinaison basse

La combinaison basse s'appuie directement sur les données initiales. La technique la plus courante consiste à prendre plusieurs matrices de caractères et à les concaténer. On obtient ainsi un jeu de données plus vaste tant sur le nombre d'espèces que sur la quantité de données génétiques. On applique ensuite une méthode de reconstruction phylogénétique sur cette "supermatrice" de caractères (Miyamoto, 1985). Une brève discussion préliminaire est présentée sur le principe méthodologique de la combinaison basse, nommé *total evidence*, souvent considéré à tort comme le nom donné aux techniques de concaténation et d'inférence.

4.1.1 *Total evidence*

La Figure 4.1 schématise la concaténation de trois alignements de gènes. Ce type de combinaison basse est souvent cité comme relevant du principe dit de *total evidence* (TE). Ce terme a été introduit dans le domaine de l'inférence phylogénomique par Kluge (1989). Ce dernier a emprunté le terme TE au philosophe et logicien Rudolf Carnap (1950). Pour Carnap (1950), TE se réfère à un principe de logique inductive consistant à accepter ou rejeter certaines hypothèses durant un acte de prise de décision. Un cas particulier de ce principe explicite le degré de croyance que l'on voue à certaines hypothèses et les actions que l'on peut mener à la lumière de ce degré de croyance (*e.g.* construire un avion conséquemment à la croyance aux lois de l'aérodynamique ; Rieppel, 2005). Plus explicitement, une des interprétations du principe TE peut être explicitée comme "l'information observable maximale" (*total observational knowledge* ; Carnap, 1997) au moment de la prise de décision. Conséquemment, Kluge (1989) considéra – à tort (Lecointre and Deleporte, 2005) – que TE est une procédure consistant à considérer l'ensemble maximal des données moléculaires dans le but d'inférer un arbre phylogénétique s'appuyant sur un ensemble efficace (*i.e.* maximal) de connaissances (*i.e.* l'ensemble des données moléculaires), et prit le parti de concaténer un ensemble d'alignements de séquences pour y appliquer une méthode d'inférence phylogénétique minimisant un critère MP. Or, TE, dans sa définition originale, désigne l'ensemble des informations phylogénétiques disponibles et ce que l'on peut en faire dans un contexte d'inférence phylogénomique (Lecointre and Deleporte, 2005). TE, suivant la définition originale de Carnap (1950), peut donc plutôt être considéré comme un synonyme de phylogénomique, alors que le principe méthodologique de Kluge (1989) se doit plutôt d'être qualifié d'analyse simultanée (Nixon and Carpenter, 1996; Lecointre and Deleporte, 2005).

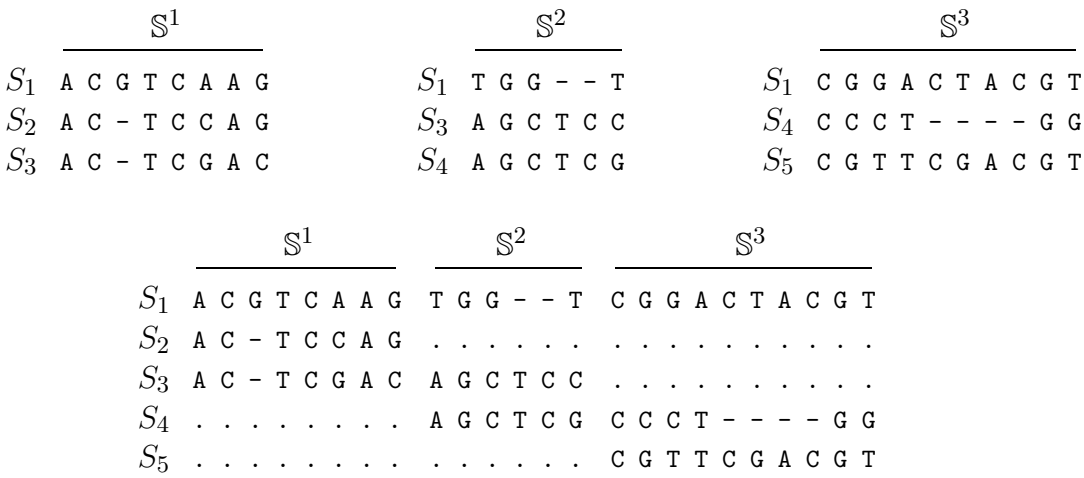

FIG. 4.1 – **Exemple de supermatrice de caractères**

A partir de trois alignements de séquences d'ADN $\mathbb{S}^1$, $\mathbb{S}^2$ et $\mathbb{S}^3$ (en haut), une supermatrice de caractère a été construite en concaténant les trois alignements de séquences et en figurant les états de caractère manquants par un point (en bas).

4.1.2 Analyse simultanée à partir d'une supermatrice de caractères

Il n'est pas rare d'observer des différences topologiques entre deux arbres phylogénétiques définis sur le même ensemble d'espèces mais inférés à partir de deux gènes différents. Ce fait, nommé incongruence entre gènes, est souvent provoqué par le phénomène d'homoplasie (Sober, 1988). Par homoplasie, on entend l'apparition indépendante d'états de caractère similaires chez des taxons éloignés, impliquant souvent un phénomène d'attraction des longues branches, en particulier si on utilise les critères MP ou de distance. L'homoplasie est subdivisée en convergence (apparition indépendante d'un même état de caractère) et réversion (apparition d'un état de caractère ayant l'apparence d'un état de caractère ancestral). Une forte hétérogénéité des taux d'évolution entre sites est une des causes de l'homoplasie lorsque l'on considère des séquences de caractères moléculaires. Les transferts horizontaux de gènes sont également responsables de l'incongruence entre gènes. Ainsi, lorsque l'on souhaite inférer un arbre phylogénétique à partir d'une collection de gènes, il est souvent recommandé d'effectuer des tests d'incongruence entre gènes, *e.g.* test ILD (Farris et al., 1995). Une première approche consiste à éliminer du jeu de données initial le(s) gène(s) présentant une forte incongruence par rapport aux autres. Une autre approche revient à détecter et éliminer seulement les états de caractères responsables de l'incongruence. Ces deux approches ne sont pas incompatibles avec le principe TE (Lecointre and Deleporte, 2005).

Les techniques de combinaison basse peuvent souvent être handicapées par l'apparition de nombreuses données manquantes (*e.g.* 46% d'états de caractère manquants dans l'exemple de la Figure 4.1). Les études de phylogénomiques récentes basées sur la combinaison basse offrent souvent plus de 50% d'états de

caractères manquants, *e.g.* (Gatesy et al., 2002; Driskell et al., 2004).

Les deux critères les plus utilisés pour inférer un arbre à partir d'une supermatrice de caractères sont les critères MP et ML. Les critères MP présentent l'avantage de pouvoir traiter tous les types d'état de caractère et de pouvoir construire des supermatrices de caractères à partir d'états de caractère nucléaire, protéique ou morphologique (*e.g.* Gatesy *et al.*, 2002). Le critère ML permet d'utiliser des modèles probabilistes de l'évolution et certains travaux récents tentent d'améliorer les modèles actuels en tenant compte de l'hétérogénéité des vitesses d'évolution de chaque gène concaténé (Yang, 1996a; Pupko et al., 2002; Bevan et al., 2006), *i.e.* associer et estimer un paramètre $\tilde{\mu}_p$ pour chaque gène G^p dans la matrice Q de taux instantanés présentée dans la Formule (3.1) (cf page 48).

Peu de logiciels d'inférence phylogénétique ont été conçus dans le but d'effectuer une analyse simultanée par combinaison basse. Un des seuls existant, *MrBayes* (Huelsenbeck and Ronquist, 2001), et plus particulièrement la version 3 (Ronquist and Huelsenbeck, 2003), permet d'effectuer une inférence phylogénomique par combinaison basse en associant un modèle évolutif à différentes partitions de la supermatrice de caractères (*e.g.* à chaque gène). Malheureusement, à cause de sa demande élevée en temps et en mémoire, l'approche probabiliste implémentée dans le logiciel *MrBayes*, bien que produisant des arbres phylogénétiques de bonne qualité, reste difficilement utilisable (comme nombre d'approches probabilistes) lorsque les supermatrices de caractères sont de tailles relativement importantes, *i.e.* plusieurs centaines de taxons et de gènes. (Williams and Moret, 2003). Ainsi, l'analyse simultanée de supermatrices de caractères induisant d'autant plus de paramètres à estimer qu'on y définit de partitions, elle demeure une approche fiable mais difficilement utilisable en pratique sur de très grands jeux de données phylogénomiques.

4.2 La combinaison haute

La combinaison haute cherche à amalgamer les arbres phylogénétiques inférés séparément à partir de chaque jeu de données (*e.g.* matrices de distance, alignements de séquences, caractères morphologiques). Initialement issue des techniques de consensus (*i.e.* combinaison de k arbres sources définis sur le même ensemble de n taxons), la combinaison haute a impliqué le développement des méthodes de *superarbre* (*i.e.* combinaison de k arbres sources définis sur des ensembles recouvrants mais non semblables de taxons) qui consistent à représenter au mieux dans un seul arbre (l'arbre consensus ou le superarbre) l'information topologique contenue dans la collection $\mathcal{C}_T = \{T^1, T^2, ...T^p, ..., T^k\}$ des arbres sources.

Le problème du superarbre a connu un grand intérêt ces dernières années. Cette partie décrit brièvement les différentes techniques algorithmiques développées pour résoudre ce problème.

4.2.1 Les techniques de consensus d'arbres phylogénétiques

Les premières tentatives pour amalgamer les topologies d'une collection $\mathcal{C}$ d'arbres sources ont été les méthodes de consensus. Ces dernières constituent un cas particulier des méthodes de superarbres car elles ne considèrent que des arbres sources définis sur le même ensemble de taxons, *i.e.* $\mathcal{L}_{T^p} = \mathcal{L}$, pour tout arbre source T^p. Historiquement, le problème du consensus ainsi que la première méthode est dûe à Adams (1972). Nombre de techniques ont été développées depuis (Bryant, 2003), parmi les plus connues :

- l'arbre de consensus strict qui ne contient que les clades (resp. bipartitions) communs à tous les arbres sources,
- l'arbre de consensus majoritaire qui contient tous les clades (resp. bipartitions) apparaissant dans plus de la moitié des arbres sources,

Il n'y a pas de bonnes ou de mauvaises méthodes de consensus, il n'y a que des consensus propres à ce que l'on en attend. Le consensus strict demeure très utilisé car il ne sélectionne que les clades (resp. bipartitions) universellement accepté(e)s par les arbres sources. En association avec les techniques de bootstrap, le consensus majoritaire est également très utilisé pour représenter les proportions d'apparitions aux branches d'un arbre.

4.2.2 La généralisation des techniques de consensus au problème du superarbre

Plusieurs adaptations algorithmiques des techniques de consensus d'arbres ont été proposées.

Superarbre de consensus strict

Il existe un algorithme appliquant le principe du consensus strict sur une collection $\mathcal{C}_T$ d'arbres sources enracinés compatible (Day and Sankoff, 1986; Gordon, 1986). Cet algorithme considère comme point de départ le consensus strict des arbres sources restreints aux taxons apparaissant dans tous les arbres de $\mathcal{C}_T$. Cet arbre est considéré comme un squelette dans lequel il insère les sous-arbres restants. Si plusieurs solutions d'insertion sont possibles sur une branche du squelette, un unique noeud interne y est créé sur lequel sont attachés les différents sous-arbres.

Une collection $\mathcal{C}_T$ d'arbres phylogénétiques sources est dite compatible si il existe au moins un arbre T tel que chaque arbre source est un sous-arbre de T. Si $\mathcal{C}_T$ est compatible et que les arbres sources sont enracinés, cet algorithme renvoie un superarbre avec une complexité polynomiale, nommé superarbre de consensus strict (Gordon, 1986) ; sinon, l'algorithme ne renvoie rien. Le superarbre de consensus strict possède la propriété d'être le consensus strict de tous les superarbres possibles de $\mathcal{C}_T$.

Superarbre semi-strict

Etant donné une collection $\mathcal{C}_T$ d'arbres phylogénétiques enracinés, Lanyon (1993) a décomposé les clades de chaque arbre source T^p en deux ensembles : les clades observés et les clades possibles. Les clades observés sont les clades non-triviaux composant l'arbre T^p. Les clades possibles sont tout ceux obtenus en décomposant les éventuelles multifurcations de T^p, ou en insérant les taxons dans $\mathcal{L}_{\mathcal{C}_T} - \mathcal{L}_{T^p}$, où $\mathcal{L}_{\mathcal{C}_T} = \cup_p \mathcal{L}_{T^p}$. Si on considère les deux arbres sources de la Figure 4.2, les clades observés sont les clades $\{1,2\}$, $\{1,2,3\}$, $\{4,5\}$ et $\{1,3\}$, $\{4,5,6\}$, $\{5,6\}$. Les clades possibles sont obtenus par toutes les insertions possibles du taxon 6 dans le premier arbre (*i.e.* $\{1,6\}$, $\{2,6\}$, $\{3,6\}$, $\{4,6\}$, $\{5,6\}$, $\{1,2,6\}$, $\{4,5,6\}$, $\{1,2,3,6\}$) et toutes les insertions possibles du taxon 2 dans le deuxième arbre (*i.e.* $\{1,2\}$, $\{2,3\}$, $\{2,4\}$, $\{2,5\}$, $\{2,6\}$, $\{1,2,3\}$, $\{2,5,6\}$, $\{2,4,5,6\}$).

Le superarbre de consensus semi-strict (Lanyon, 1993) est composé par tous les clades observés et possibles universellement acceptés par les arbres sources (*i.e.* appartenant à l'intersection de tous les ensembles de clades induits par chaque arbre source) tels que :

- un clade n'est contredit par aucun autre, ou
- un clade observé n'est contredit que par certains clades possibles (qui sont éliminés en conséquence).

Ainsi, dans l'exemple de la Figure 4.2, l'intersection de tous les ensembles de clades est composée des clades observés $\{1,2\}$, $\{5,6\}$, $\{1,2,3\}$, $\{4,5,6\}$ et du clade possible $\{2,6\}$. Ce dernier contredisant tous les autres, il est éliminé de l'analyse.

Goloboff et Pol (2002) ont démontré que l'algorithme de Lanyon (1993) renvoie un superarbre pouvant contenir des clades en contradiction avec certaines combinaisons des arbres sources. Ils ont proposé une heuristique permettant d'obtenir un superarbre semi-strict ayant de meilleures propriétés, ainsi qu'une démonstration que la généralisation du consensus majoritaire est impossible dans le cadre des problèmes de superarbre.

4.2.3 L'algorithme BUILD

L'algorithme BUILD (Aho et al., 1981) teste si une collection $\mathcal{C}_T$ d'arbres sources enracinés est compatible, et, dans l'affirmative, renvoie un superarbre T compatible avec l'ensemble des arbres de $\mathcal{C}_T$. Il construit un graphe G composé de n sommets correspondant aux taxons dans $\mathcal{L}_{\mathcal{C}_T} = \cup_p \mathcal{L}_{T^p}$, où chaque paire de sommets est reliée si la paire de feuilles correspondante est contenue dans au moins un clade non trivial dans au moins un arbre source. Si G est non connexe, chaque clade correspondant à chaque composante connexe est créé suivant le schéma divisif, et le processus est réitéré pour chaque composante connexe sur la collection des arbres sources restreints aux feuilles de la composante. Si G est composé d'une unique composante connexe de plus de deux sommets, alors $\mathcal{C}$ est incompatible et l'algorithme BUILD s'arrête. Sa complexité algorithmique est de l'ordre de $O(kn^2)$ dans le cas général, et de $O(k\sqrt{n})$ si les

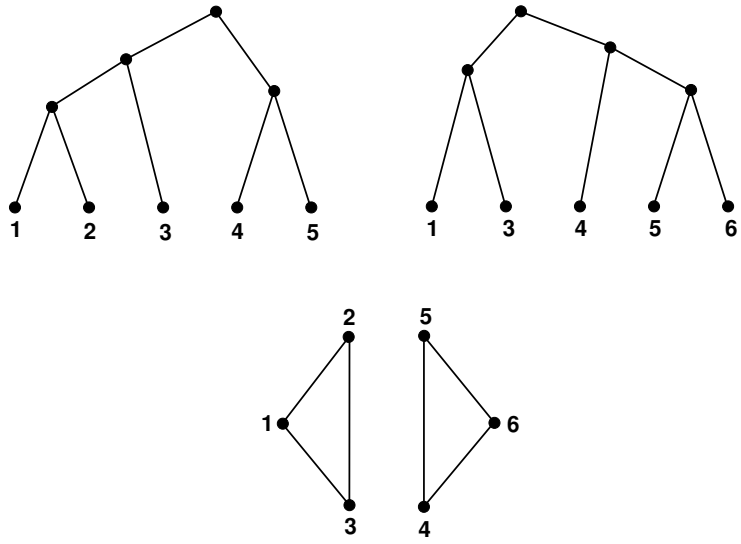

FIG. 4.2 – **Construction du graphe représentant l'information topologique d'une collection de deux arbres**

Le graphe non connexe (en bas) représente les paires de feuilles contenues dans au moins un clade non trivial de la collection constituée de deux arbres sources enracinés (en haut).

arbres sources sont tous binaires (Henzinger et al., 1999).

La Figure 4.2 schématise la construction de G à partir de deux arbres phylogénétiques enracinés T^1 et T^2. Le graphe G étant formé de deux composantes connexes, le schéma divisif utilisé par BUILD produira les deux clades $\{1,2,3\}$ et $\{4,5,6\}$. La procédure sera ensuite réitérée à partir de chacune des deux composantes connexes. Ainsi un graphe sera construit pour représenter l'information topologique de la collection $T^1_{|\{1,2,3\}}$ et $T^2_{|\{1,2,3\}}$ (resp. $T^1_{|\{4,5,6\}}$ et $T^2_{|\{4,5,6\}}$) , et — si le nouveau graphe est formé de plusieurs composantes connexes — le schéma divisif sera appliqué sur le superarbre pour diviser le clade $\{1,2,3\}$ (resp. $\{4,5,6\}$).

L'algorithme BUILD, pour sa simplicité et sa rapidité, a été employé dans plusieurs cas de figures théoriques :

- la construction de tous les superarbres enracinés compatibles avec un ensemble de triplets ou d'arbres enracinés compatibles (Constantinescu and Sankoff, 1995; Ng et al., 2000; Semple, 2002),
- la polynomialité du problème consistant à trouver le superarbre d'une collection d'arbres enracinés compatibles ayant tous une même bipartition de feuilles en commun (Bryant et al., 2004),
- l'écriture d'un algorithme polynomial pour la construction d'un superarbre d'une collection de taille bornée de quadruplets compatibles (Bryant et al., 2004).

4.2.4 Les adaptations de BUILD

BUILD est un algorithme performant par sa simplicité et sa rapidité mais demeure difficilement utilisable car il n'opère que sur des collections d'arbres sources enracinés compatibles. Des améliorations ont ainsi été proposées.

L'algorithme MINCUTSUPERTREE et ses variantes

L'algorithme BUILD décompose l'information topologique contenue dans les arbres enracinés sources et l'exprime dans un graphe G. Chaque composante connexe de G représente un clade du superarbre à construire. La compatibilité des arbres sources garantit l'existence de différentes composantes connexes dans les graphes G construits à chaque itération. L'algorithme MINCUTSUPERTREE (MC) ajoute une étape intermédiaire qui supprime des arêtes dans G afin de créer des composantes connexes dans le cas d'une collection d'arbres sources non-compatible (Semple and Steel, 2000). Cette dernière étape est appelée une *coupe* de graphe et est définie par un ensemble d'arêtes. Si les arêtes d'un graphe sont valuées (*i.e.* chacune associée à une valeur), une *coupe minimale* est une coupe telle que la somme des arêtes la composant est minimale. La recherche des coupes minimales d'un graphe s'effectue avec une complexité polynomiale (Gomory and Hu, 1961).

Plus précisément, l'algorithme MC procède de la manière suivante :

- chaque arête $\{u, v\}$ de G est valuée par le nombre de fois où la paire de feuilles uv est contenue dans un clade non trivial dans la collection d'arbres sources.
- Les arêtes $\{u, v\}$ de valuation k sont contractées, *i.e.* $u = v$ dans le nouveau graphe G' ainsi obtenu. Le but de l'algorithme MC étant de créer des composantes connexes dans G s'il n'en contient qu'une, cette étape garantit qu'une paire de feuilles uv ne se retrouvera pas séparée (*i.e.* dans le superarbre ainsi créé, on doit interdire que la racine appartienne aux noeuds internes composant l'unique chemin reliant u et v) si elle est présente dans un clade non-trivial dans chacun des k arbres sources de $\mathcal{C}_T$.
- Si G' est non connexe, l'algorithme MC recherche toutes les coupes minimales de G', puis supprime toutes les arêtes appartenant à une coupe minimale.

L'algorithme MC permet de construire un superarbre jouissant de nombreuses propriétés combinatoires avec une complexité de l'ordre de $O(kn^5)$ (Semple and Steel, 2000).

Néanmoins, le choix des arêtes à supprimer si G' est non connexe peut être modifié suivant différents critères. Page (2002) a ainsi proposé un algorithme similaire, appelé MODIFIEDMINCUTSUPERTREE (MMC), qui diffère par ce critère de suppression. Des résultats de simulation montrent que MMC améliore significativement les résultats obtenus par MC (Eulenstein et al., 2004).

Utilisation d'informations supplémentaires aux noeuds des arbres sources

Si les noeuds internes des arbres sources sont associés à un niveau taxonomique supérieur (*e.g.* le noeud interne correspondant à tous les mammifères dans l'arbre de la Figure 1.3 est étiqueté *Mammals*), alors ces informations supplémentaires peuvent être utilisées pour inférer des superarbres induisant une meilleure résolution (*i.e.* induisant plus de noeuds internes de degré 3). L'inférence phylogénétique se situant souvent à différents niveaux taxonomiques (*e.g.* arbre phylogénétique général des grands groupes de Mammifères, arbre phylogénétique particulier des espèces formant le groupe des Primates), les méthodes de superarbres se doivent de pouvoir combiner des arbres sources modélisant l'évolution du vivant à ces différents niveaux (Page, 2004).

Plusieurs adaptations de BUILD ont été proposées dans ce but, telles que les algorithmes RANKEDTREE (Bryant et al., 2004; Semple et al., 2004) et ANCESTRALBUILD (Daniel and Semple, 2004; Berry and Semple, 2006). Suivant le même principe que BUILD, ces algorithmes ne renvoient rien si la collection d'arbres sources enracinés n'est pas compatible, et calculent un superarbre en temps polynomial si les arbres sources sont compatibles.

4.2.5 Représentation matricielle binaire

La représentation matricielle (MR ; *Matrix Representation*) binaire d'une collection d'arbres phylogénétiques consiste à encoder dans plusieurs vecteurs de caractères binaires l'ensemble des clades (resp. bipartitions) induits par chaque arbre source enraciné (resp. non enraciné). La représentation MR est obtenue par la concaténation de l'ensemble des vecteurs binaires. Il existe différentes manières de représenter un graphe en général (Poincaré, 1901; Ponstein, 1966), et, conséquemment, un arbre en particulier (Farris et al., 1970). Brooks (1981), Baum (1992), Doyle (1992), Mishler (non publié ; cité dans Ragan, 1992) et Ragan (1992) ont proposé d'utiliser l'encodage binaire suivant, pour chaque noeud interne u de chaque arbre source enraciné T^p de $\mathcal{C}_T$:

- chaque taxon descendant de u (*i.e.* appartenant au clade induit par u) est codé par '1',
- les autres taxons de $\mathcal{L}_{T^p}$ (*i.e.* non descendant de u) sont codés par '0',
- les taxons restant (*i.e.* non présents dans $\mathcal{L}_{T^p}$) sont codés par un état de caractère manquant.

Cet encodage est schématisé dans la Figure 4.3 et peut facilement s'adapter au cas d'arbres phylogénétiques T non enracinés, pour chaque branche e induisant une bipartition $A|B$ de $\mathcal{L}_{T^p}$:

- chaque taxon appartenant à A est codé par '1',
- chaque taxon appartenant à B est codé par '0',
- les taxons restant (*i.e.* non présents dans $\mathcal{L}_{T^p}$) sont codés par un état de caractère manquant.

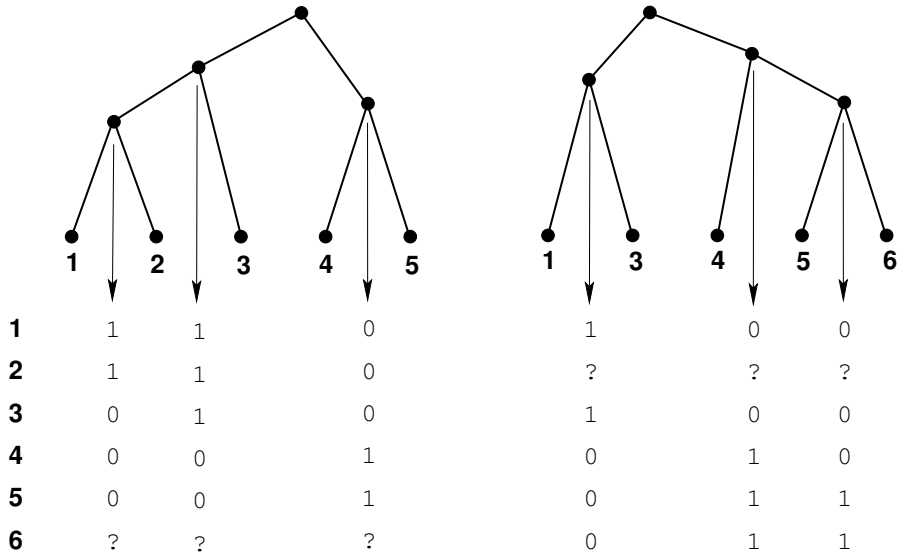

FIG. 4.3 – **Construction de la représentation matricielle binaire de l'information topologique d'une collection de deux arbres**

L'encodage binaire (en bas) proposé par Baum (1992) et Ragan (1992) représente l'ensemble des clades correspondant à chaque noeud interne de la collection constituée de deux arbres sources enracinés (en haut). Les feuilles et les racines n'ont pas été représentées pour plus de clarté.

Ainsi, la représentation MR d'une collection de k arbres non enracinés contient k sites de moins que la représentation MR des k mêmes arbres enracinés. De plus, l'encodage binaire d'arbres enracinés induit une connaissance *a posteriori* de l'orientation des arbres ; en effet, cet encodage implique que l'état ancestral de la racine de chaque arbre correspond à un vecteur entièrement composé de '0'.

Purvis (1995) a remarqué que certaines topologies peuvent avoir une influence excessive par rapport à d'autre dans une MR construite suivant l'encodage de Baum (1992) et Ragan (1992), et a attribué ce fait à la forte dépendance entre les vecteurs binaires issus d'un seul arbre. La redondance du signal topologique entre les vecteurs binaires d'un arbre étant la cause d'éventuels biais, il a proposé un nouvel encodage binaire, pour chaque noeud interne u de chaque arbre enraciné T^p :

- chaque taxon descendant de u est codé par '1',
- chaque taxon descendant du père de u et non-descendant de u est codé par '0',
- les taxons restant sont codés par un état de caractère manquant.

Ronquist (1996) a néanmoins démontré que ce nouvel encodage était imparfait et que le biais décrit par Purvis (1995) n'était pas dû à une redondance de l'information mais à la taille respective des arbres sources. Afin de corriger le biais observé par Purvis (1995), Ronquist (1996) a suggéré également une autre variante qui consiste à pondérer chaque vecteur binaire par une valeur de fiabilité associée au noeud u ou à la branche e interne correspondant (*e.g.* va-

leur de bootstrap). Cette technique de pondération est actuellement une méthode standard d'inférence MR, bien qu'elle soit critiquable et critiquée (*e.g.* Wilkinson *et al.*, 2004).

On remarquera qu'il existe plusieurs autres manières de construire la représentation MR d'une collection d'arbres, telles que la représentation matricielle binaire par triplets enracinés qui encode l'ensemble des triplets enracinés induits par une collection d'arbres enracinés (Nelson and Ladiges, 1992; Nelson and Ladiges, 1994; Wilkinson et al., 2001).

Représentation Matricielle avec Parcimonie (MRP)

Baum (1992), Mishler (non publié ; cité dans Ragan, 1992) et Ragan (1992) ont tous trois suggéré d'inférer un superarbre à partir de la représentation MR en minimisant le critère MP non pondéré. En effet, si T est un arbre phylogénétique binaire enraciné défini sur n feuilles, alors sa représentation MR contient $2n-1$ sites et la valeur de parcimonie de Fitch de T non enraciné est de $2n-1$, car chaque branche correspond à une unique mutation $0 \leftrightarrow 1$. Toutefois, seuls les $n-2$ vecteurs binaires correspondant aux noeuds internes sont utilisés en pratique, car les $n+1$ vecteurs binaires induits par les feuilles et la racine ne sont pas informatifs pour le critère MP. D'un autre point de vue, la parcimonie de Fitch permet d'inférer les arbres phylogénétiques qui induisent une histoire évolutive minimisant le nombre d'homoplasies (Sober, 1988), ce qui justifie son utilisation car il permet ainsi de minimiser le nombre de conflits entre clades (Eulenstein et al., 2004). Cette méthode de combinaison haute est appelée MRP (*Matrix Representation with Parsimony* ; Ragan, 1992) et demeure actuellement la plus utilisée.

Une technique courante pour enraciner le superarbre MRP consiste à rajouter dans la représentation MR un taxon *root* uniquement constitué de '0' modélisant l'état de caractère ancestral de toutes les racines des arbres phylogénétiques sources enracinés. La feuille *root* du superarbre indique ainsi comment l'enraciner.

Comme il est très courant qu'une collection d'arbres sources induise une MR impliquant plusieurs superarbres distincts minimisant le critère MP non pondéré, la technique standard consiste à considérer le superarbre MRP comme étant le consensus strict des arbres optimaux au sens de la parcimonie de Fitch. Grâce à ce consensus, le superarbre MRP possède la propriété d'être identique au superabre de consensus strict lorsqu'il est calculé à partir d'une collection d'arbres sources compatible (Thorley, 2000).

Une adaptation naturelle (car utilisant la connaissance *a posteriori* de l'orientation des arbres sources enracinés) consiste à minimiser la parcimonie de Sankoff en posant $c_{0\to1} = 1$ et $c_{1\to0} = +\infty$ (Bininda-Emonds and Bryant, 1998), d'autant plus que, connaissant la racine du superarbre, la complexité de cette approche est du même ordre que celle utilisant la parcimonie de Fitch, *i.e.* $O(kn)$. Néanmoins, cette approche n'est curieusement jamais utilisée en pratique (cf toutefois les quelques expérimentations de Bininda-Emonds et Bryant, 1998).

Représentation Matricielle avec "Chiquenaude" (MRF)

Si MRP infère le superarbre minimisant le nombre d'homoplasies induits par la MR, une autre technique consiste à modifier un nombre minimal d'états de caractère binaire (*i.e.* transformer des `0` en `1`, ou inversement) afin de supprimer toute homoplasie, *i.e.* afin que la représentation MR obtenue contienne un ensemble de caractères compatibles.

Le changement de plusieurs états de caractère binaire dans une représentation matricielle MR^1 donne une nouvelle représentation matricielle MR^2. La distance $d(\text{MR}^1, \text{MR}^2)$ représente le nombre de différence entre MR^2 et MR^1 (Chen et al., 2003). Si $\text{MR}_{\mathcal{C}_T}$ est la représentation matricielle d'une collection $\mathcal{C}_T$ d'arbres sources, alors la méthode MRF (*Matrix Representation with Flipping* ; Eulenstein et al., 2004) consiste à rechercher le(s) superarbre(s) T minimisant la distance $d(\text{MR}_{\mathcal{C}_T}, \text{MR}_T)$, où MR_T est la représentation matricielle de T (Chen et al., 2003). La recherche de(s) superarbre(s) MRF est un problème NP-difficile (Chen et al., 2006) et nécessite l'utilisation d'heuristiques (Chen et al., 2003; Eulenstein et al., 2004; Chen et al., 2006). Néanmoins, cette heuristique offre des performances en simulation globalement similaires à celles de MRP (Eulenstein et al., 2004).

4.2.6 Heuristiques de recherche locale par descente

Il existe une très grande variété de mesures pour comparer la similitude topologique entre deux arbres (cf partie 3.6.1, page 68). Ainsi, si d est une distance topologique normalisée et T un superarbre de la collection $\mathcal{C}_T = \{T^1, T^2, ...T^p, ..., T^k\}$, alors T représente d'autant mieux l'information topologique de $\mathcal{C}_T$ que la valeur

$$\Delta(T, \mathcal{C}_T) \; = \; \sum_{1 \leq p \leq k} d\big(\, T_{|_{\mathcal{L}_{T^p}}} \, , \, T^p \, \big)$$

est minimale. Cette valeur $\Delta(T, \mathcal{C}_T)$ peut être considérée comme un critère à minimiser.

La méthode MSS (*Most Similar Supertree* ; Creevey and McInerney, 2005) utilise la distance par chemin d_{path} (définie avec la norme L^1) dans le critère $\Delta(T, \mathcal{C}_T)$, et effectue une recherche locale par descente avec voisinage NNI ou SPR pour rechercher l'(les) superarbre(s) minimisant ce critère. Cette méthode a été implémentée dans le logiciel CLANN (Creevey and McInerney, 2005) et complétée par deux autre méthodes similaires : MSF (*Maximum Splits Fit*) utilisant la distance de partition d_{RF}, et MQF (*Maximum Quartet Fit*) utilisant la distance de quadruplets d_{quad}.

4.2.7 Décomposition sous forme de quadruplets

Nombre de méthodes existent pour inférer un arbre à partir d'un ensemble de quadruplets. La décomposition par quadruplets consiste à considérer l'ensemble des sous-arbres binaires non enracinés de quatre feuilles de chacun des arbres sources de $\mathcal{C}_T$, puis de les combiner pour

inférer un superarbre. Malheureusement, ce problème a été montré NP-complet (Steel, 1992) et son application efficace impliqua dans un premier temps l'emploi d'algorithmes exponentiels (Ben-Dor et al., 1998; Robinson-Rechavi and Graur, 2001).

Plus précisément, étant donné l'ensemble $Q(\mathcal{C}_T)$ des quadruplets induits par les arbres sources dans $\mathcal{C}_T$, le superarbre T est celui qui minimise le critère $\sum_{q \in Q(\mathcal{C})} \big[-\ln \hat{f}(q) \big]$, où $\hat{f}(q)$ est la fréquence d'appartition du quadruplet $q \in Q(\mathcal{C}_T)$ dans les arbres de $\mathcal{C}_T$. Une heuristique, nommée QUARTET-PUZZLING (Strimmer and von Haeseler, 1996), cherche à minimiser ce critère dans le cadre de l'inférence phylogénétique (*i.e.* $Q(\mathcal{C}_T)$ n'est pas obtenu à partir d'une collection d'arbres mais à partir d'un alignement de séquences). Son emploi dans le cadre de la combinaison haute a été suggéré (Pisani and Wilkinson, 2002; Wilkinson et al., 2004b) mais abandonné avec les performances offertes en simulation par un nouvel algorithme proposé par Willson (1999).

L'algorithme de Willson (1999)

Etant donné un quadruplet de taxons distincts $ijxy$ issu de $\mathcal{L}_{\mathcal{C}_T}$, il existe exactement trois topologies binaires définies sur $ijxy$, notées q^1_{ijxy}, q^2_{ijxy} et q^3_{ijxy} ; leur numérotation est telle que $w(q^1_{ijxy}) \leq w(q^2_{ijxy}) \leq w(q^3_{ijxy})$ où $w(q) = -\ln \hat{f}(q)$, *i.e.* q^1_{ijxy} est la topologie de quadruplet la plus représentée dans $\mathcal{C}_T$. Lors de la première étape, l'algorithme de Willson (1999) sélectionne l'arbre $T = q^1_{ijxy}$ qui maximise la fonction $w(q^2_{ijxy}) - w(q^1_{ijxy})$ sur l'ensemble des quadruplets $ijxy$.

La suite de l'algorithme utilise le schéma algorithmique d'insertion mais recherche, à chaque itération, le meilleur taxon x à insérer dans T. Etant donné un quadruplet $ijxy$ et un arbre T' tel que $\{i, j, x, y\} \subseteq \mathcal{L}_{T'}$, la fonction ex est définie par

$$ex(T', ijxy) = \begin{cases} w(T'_{|_{\{i,j,x,y\}}}) - w(q^2_{ijxy}) = w(q^1_{ijxy}) - w(q^2_{ijxy}) & \text{si } T'_{|_{\{i,j,x,y\}}} = q^1_{ijxy}, \\ w(T'_{|_{\{i,j,x,y\}}}) - w(q^1_{ijxy}) & \text{sinon.} \end{cases}$$

La maximisation de ex sert à sélectionner la meilleure topologie de T complétée avec l'insertion d'un taxon x. Ainsi, un taxon x inséré sur une branche de T produit l'arbre T' et permet de définir la fonction d'inconsistence locale (Willson, 1999) :

$$LI(T', x) = \max_{\{i,j,x,y\} \subseteq \mathcal{L}_{T'}} \big[\, ex(T', ijxy) \, \big].$$

A chaque itération de l'algorithme d'insertion, l'arbre T' sélectionné est celui qui minimise le critère $LI(T', x) - LI(\tilde{T}', x)$, où $\tilde{T}'$ est un arbre obtenu par l'insertion de x sur une autre branche de T. Essayant, à chaque itération, d'insérer tous les taxons $x \notin \mathcal{L}_T$ sur toutes les branches de T, la complexité algorithmique d'une itération est de l'ordre de $O(n^5)$, sachant qu'il existe $O(n^3)$ quadruplets $ijxy$ avec x fixé. Ainsi, l'algorithme de Willson (1999) est de l'ordre de $O(n^6)$, mais différents procédés permettent de réduire cette complexité en $O(n^4)$ (Ranwez and Gascuel, 2001).

Malgré plusieurs modifications de la fonction w de pondération des quadruplets, plusieurs simulations ont montré que la combinaison haute par décomposition sous forme de quadruplets ne permet pas d'observer de meilleures performances que la méthode MRP (Piaggio-Talice et al., 2004). On notera que des observations similaires ont été effectuées dans le cadre de l'inférence phylogénétique (Ranwez and Gascuel, 2001).

4.2.8 Moyennes de distances additives d'arbre

Lapointe et Cucumel (1997) ont été les premiers à essayer de construire un superarbre à partir de l'information induite par les longueurs de branche de chaque arbre source de $\mathcal{C}_T$. La collection d'arbres sources peut alors être considérée comme la collection $\mathcal{C}_{(T)} = \left\{(T^1_{ij}), (T^2_{ij}), ..., (T^p_{ij}), ..., (T^k_{ij})\right\}$ des matrices de distance additive équivalentes aux arbres phylogénétiques sources.

Dans le cadre du consensus d'arbres (*i.e.* $\mathcal{L}_{T^p} = \mathcal{L}_{\mathcal{C}_{(T)}}$, pour tout p), il a été montré que la distance (Δ^c_{ij}) représentant le mieux, au sens OLS ou WLS, la collection $\mathcal{C}_{(T)}$ de distances additives, *i.e.* minimisant la fonction

$$\sum_{1\leq p\leq k} \sum_{i,j\in\mathcal{L}_{\mathcal{C}_{(T)}}} \mathcal{W}_{ij}\Big(\ \Delta^c_{ij}\ -\ T^p_{ij}\ \Big)^2,$$

est également celle qui minimise la fonction suivante (Lapointe and Cucumel, 1997) :

$$\sum_{i,j\in\mathcal{L}_{\mathcal{C}}} \mathcal{W}_{ij}\Big(\ \Delta^c_{ij}\ -\ \overline{T}_{ij}\ \Big)^2 \quad \text{avec} \quad \overline{T}_{ij} = \frac{1}{k}\sum_{1\leq p\leq k} T^p_{ij}.$$

Le calcul de $(\overline{T}_{ij})$ s'effectuant avec une complexité de l'ordre de $O(kn^2)$, les auteurs ont proposé cette méthode pour inférer l'arbre consensus T^c (*Average Consensus Tree* ; Lapointe and Cucumel, 1997) de la collection $\mathcal{C}_T$ en appliquant une méthode minimisant un critère LS. Cette technique peut également être utilisée dans le cadre plus général du problème des superarbres en utilisant un calcul de moyennes adapté aux ensembles de feuilles $\mathcal{L}_{T^p}$ partiellement recouvrants des arbres de $\mathcal{C}$ (Lapointe and Cucumel, 1997) :

$$\overline{T}_{ij} = \frac{1}{k_{ij}} \sum_{\substack{1\leq p\leq k \\ i,j\in\mathcal{L}_{T^p}}} T^p_{ij}, \tag{4.1}$$

où k_{ij} est le nombre de fois où la paire ij est incluse dans chaque arbre source $\mathcal{L}_{T^p}$, *i.e.* $k_{ij} = \Big|\ \{T^p : i,j \in \mathcal{L}_{T^p}, 1\leq p\leq k\}\ \Big|$. Un superarbre peut alors être inféré à partir de la supermatrice de distance $(\overline{T}_{ij})$ à l'aide d'une méthode d'inférence phylogénétique pouvant accepter les distances incomplètes. En effet, si $x \in \mathcal{L}_{T^1}$ et $x \notin \mathcal{L}_{T^2}$, et si $y \notin \mathcal{L}_{T^1}$ et $y \in \mathcal{L}_{T^2}$, alors la distance $\overline{T}_{xy}$ sera inconnue dans la supermatrice de distance $(\overline{T}_{ij})$ si $\mathcal{C}_{(T)} = \left\{(T^1_{ij}), (T^2_{ij})\right\}$.

Un problème propre à cette méthode, appelée ACS (*Average Consensus Supertree*), est posé par l'hétérogénéité des vitesses d'évolution des gènes correspondant à chaque arbre source T^p. Un gène induisant une vitesse d'évolution rapide induira des valeurs de distance additive élevées dans (T^p_{ij}), alors qu'un gène lent induira des distances dans (T^p_{ij}) proches de zéro. Ce problème peut être corrigé en inférence phylogénomique ML dans le cadre de la combinaison basse en associant un paramètre modélisant la vitesse d'évolution relative $\tilde{\mu}$ à chaque gène (cf partie 4.1.2, page 74). Dans le cadre des distances additives, ce biais peut être corrigé en remplaçant la distance T^p_{ij} dans la Formule (4.1) par la distance normalisée $\alpha_p T^p_{ij}$, où α_p est une constante associée à chaque matrice de distance additive (T^p_{ij}). La valeur de α_p peut être l'inverse de la plus grande distance dans (T^p_{ij}), *i.e.* $\alpha_p = \left(\max_{i,j\in\mathcal{L}_{T^p}}[T^p_{ij}]\right)^{-1}$ (Lapointe and Cucumel, 1997) ou l'inverse de la plus grande des distances définies par une paire ij commune à toutes les matrices additives, *i.e.* $\alpha_p = \left(\max_{i,j\in\cap_m\mathcal{L}_{T^m}}[T^p_{ij}]\right)^{-1}$, lorsque celle-ci existe (Lapointe and Levasseur, 2004). Néanmoins, ces normalisations peuvent être également biaisées par la présence de deux longues branches dans un arbre source T^p ; en effet, si x et y correspondent à deux très longues branches externes et que les autres branches sont de longueur relativement petite, alors $T^p_{xy} >> T^p_{ij}$, pour toute paire $i, j \in \mathcal{L}_{T^p} - \{x, y\}$ et la normalisation par α_p fera tendre toutes les distances T^p_{ij} vers zéro, même si le gène correspondant n'induit pas une vitesse d'évolution rapide globale dans l'ensemble de l'arbre T^p.

4.3 La combinaison moyenne

Face à l'augmentation exponentielle du nombre de taxons étudiés et de la quantité de gènes séquencés, la combinaison basse tend, à moyen terme, à devenir inutilisable pour l'étude phylogénomique de très grandes collections de gènes (voire de génomes complets). Une des plus grandes supermatrices de caractères construites à ce jour est définie par la concaténation des alignements de 1131 protéines, induisant 469497 sites pour 70 taxons (Driskell et al., 2004). Cette supermatrice de caractères étant formée par des protéines (*i.e.* codées sur un alphabet de 20 états de caractères), une inférence phylogénomique à l'aide des critères MP ou ML se révèle très longue et excessivement solliciteuse en mémoire. De plus, cette supermatrice étant formée de 92% d'états de caractère manquants, les arbres MP inférés peuvent se révéler extrèmement peu fiables, et une analyse ML par bootstrap trop longue.

Néanmoins, la combinaison basse offre de forts enrichissements dans la recherche de l'histoire évolutive des grands groupes d'espèces. Si deux arbres T^1 et T^2 inférés à partir de deux gènes distincts présentent des incongruences significatives (*i.e.* des dissimilitudes entre les topologies des sous-arbres $T^1_{|_{\mathcal{L}_{T^1}\cap\mathcal{L}_{T^2}}}$ et $T^2_{|_{\mathcal{L}_{T^1}\cap\mathcal{L}_{T^2}}}$, il a été observé, sur un exemple théorique précis, que l'analyse simultanée (par un critère MP) de la supermatrice de caractères obtenue par concaténation des deux gènes permet d'obtenir un nouvel arbre induisant une histoire évolutive différente de celles déduites de T^1 et T^2 (Barrett et al., 1991; Pisani and Wilkinson, 2002)

Des résultats similaires ont été observés de nombreuses fois avec, le plus souvent, des arbres inférés par combinaison basse plus fiables (*e.g.* valeurs de bootstrap élevée) que ceux obtenus par l'analyse séparée de chaque gène, *e.g.* (Eernisse and Kluge, 1993; Chippindale and Wiens, 1994; de Queiroz et al., 1995; Lafay et al., 1995; Smith et al., 1995; Hasegawa et al., 1997; Bond and Hedin, 2006). Autrement dit, l'accumulation de données implique un meilleur recouvrement du signal topologique (Nixon, 1999; Huelsenbeck et al., 1996; Wiens, 1998c; Wiens, 1998b). Allant dans le sens du principe *total evidence*, un fort taux de bruit évolutif dans différents gènes (*e.g.* homoplasie, forte hétérogénéité des taux d'évolution entre sites, séquences courtes, ...) implique donc des inférences phylogénétiques de mauvaise qualité et une inférence phylogénomique de meilleure qualité. Conséquemment, comme un arbre issu de l'inférence phylogénétique peut être le miroir de nombreux bruits évolutifs, un superarbre inféré par la combinaison haute de ces différentes inférences phylogénétiques n'est que le résultat d'une accumulation d'incongruences. Ce fait est la principale source de critiques des techniques de combinaison haute (Springer and de Jong, 2001; Gatesy et al., 2002; Gatesy and Springer, 2004; Lecointre and Deleporte, 2005).

Afin de ne pas trop globalement interpréter chaque donnée moléculaire considérée séparément (combinaison haute) et de pouvoir traiter de très grands jeux de données en exploitant l'information offerte par l'ensemble des gènes et protéines disponibles (combinaison basse), les techniques de combinaison moyenne (se situant méthodologiquement entre les deux) ont été développées afin de pouvoir conduire une inférence phylogénomique de bonne qualité avec des temps d'exécution efficaces en pratique. La combinaison moyenne consiste à passer par une étape intermédiaire d'interprétation du signal évolutif induit par l'ensemble des données initiales. Cette interprétation, moins éloignée des données que celle faite par la combinaison haute en considérant l'ensemble des arbres, s'efforce de tenir compte d'informations propres aux données phylogénomiques (*e.g.* vitesse d'évolution propre à chaque gène, ordonnancement des gènes le long d'un génome, taux de similarité entre segments chromosomiques). Elle s'efforce également de traduire les différentes informations phylogénétiques sous la forme de structures de données de taille moins importante que celle des importantes collections de gènes, dans le but de pouvoir les traiter informatiquement avec des complexités algorithmiques (en temps et en espace mémoire) les rendant utilisables en pratique.

4.3.1 Décomposition sous forme de quadruplets

Le terme de combinaison moyenne (*medium level combination*) a été introduit par Schmidt (2003) pour classifier l'adaptation de l'algorithme QUARTET-PUZZLING (Strimmer and von Haeseler, 1996) au cas d'une collection de gènes. Initialement développé pour l'inférence phylogénétique, Schmidt et al. (2002) ont proposé l'algorithme TREE-PUZZLE qui, étant donné une collection de gènes, décompose l'information phylogénétique induite par chaque gène en un ensemble de quadruplets, puis amalgame ces quadruplets suivant le principe algorithmique

de QUARTET-PUZZLING (Strimmer and von Haeseler, 1996; Strimmer and von Haeseler, 1997; Strimmer et al., 1997; Schmidt et al., 2002).

La décomposition sous forme de quadruplets dans le cadre de la combinaison moyenne permet de traiter plus rapidement les collections de gènes que la même approche en combinaison haute, car elle ne passe pas par une étape d'inférence phylognénétique à partir de chaque gène. Cette technique permet aussi d'éliminer le handicap causé par la présence de données manquantes lors d'une approche par combinaison basse. Toutefois, elle est sujette aux biais propres aux méthodes de quadruplets (Ranwez and Gascuel, 2001).

4.3.2 Modèles d'évolution des génomes complets

Certaines espèces ayant leur génome complet séquencé, il a été proposé plusieurs modèles d'évolution de génomes afin d'y appliquer des techniques d'inférence phylogénomique. Un des modèles les plus courants est la représentation des différents gènes le long d'un chromosome par un tableau de nombres ordonnés, où chaque nombre est associé à un gène (Nadeau and Taylor, 1984). Deux gènes homologues sont numérotés par un même nombre et l'inversion d'un gène le long d'un chromosome est codé par la valeur négative de son nombre associé. Ainsi, par exemple, les deux segments chromosomiques $[\,1|\,2|\,3|\,4|-8|-7|-6|-5|\,9|\,10]$ et $[\,1|\,2|-6|-5|-4|-3|\,7|\,8|\,9|\,10]$ sont tout deux issus du segment chromosomique ancestral $[\,1|\,2|\,3|\,4|\,5|\,6|\,7|\,8|\,9|\,10]$ par la permutation signée (*i.e.* l'inversion et le changement de signe) des segments de gènes $[\,5|\,6|\,7|\,8]$ et $[\,3|\,4|\,5|\,6]$, respectivement.

Plusieurs techniques ont été développées pour inférer un arbre phylogénétique à partir de cet encodage des segments chromosomiques. Etant donné l'ensemble des paires de gènes possibles (*e.g.* (1,2), (1,3), (1,4), (1,5), ...), la technique MPBE (*Maximum Parsimony on Binary Encoding* ; Cosner et al, 2000) consiste à associer l'état de caractère binaire '1' aux paires de gènes adjacents dans le segment chromosomique, et l'état de caractère '0' aux paires de gènes non adjacents. Une méthode minimisant un critère MP permet d'inférer un arbre à partir de cet encodage.

Les méthodes d'inférence d'arbre à partir des distances entre espèces sont souvent utilisées dans ces études. Le calcul de la distance de Hamming à partir de l'encodage MPBE permet d'obtenir une matrice de distance. Le nombre minimal de permutations signées séparant deux segments chromosomiques (Sankoff et al., 1992; Uno and Yagiura, 2000), ainsi que le nombre de paires de gènes adjacents dans un segment chromosomique qui ne l'est pas dans l'autre (Sankoff et al., 2000) représente d'autres alternatives de distance.

Les critères ML peuvent aussi être aisément adaptés au cas de l'inférence d'arbres phylogénétiques à partir de génomes complets. Néanmoins cette approche nécessite la description de modèles d'évolution propres à ce type de données et implique des complexités algorithmiques élevées conséquemment au grand nombre de possibilités d'états de caractère chromosomique à chaque noeud interne de l'arbre à inférer. Un modèle simplifié, consistant à limiter le nombre

CCC	GCC	CGC	GGC	CCG	GCG	CGG	GGG
0	0	0	0	0	0	1	0
ACC	TCC	AGC	TGC	ACG	TCG	AGG	TGG
0	0	0	0	2	0	0	1
CAC	GAC	CTC	GTC	CAG	GAG	CTG	GTG
0	1	0	1	0	0	0	0
AAC	TAC	ATC	TTC	AAG	TAG	ATG	TTG
0	1	0	0	1	0	0	0
CCA	GCA	CGA	GGA	CCT	GCT	CGT	GGT
0	0	0	1	0	0	2	1
ACA	TCA	AGA	TGA	ACT	TCT	AGT	TGT
0	1	0	0	1	0	0	0
CAA	GAA	CTA	GTA	CAT	GAT	CTT	GTT
1	0	1	0	0	0	0	0
AAA	TAA	ATA	TTA	AAT	TAT	ATT	TTT
0	0	0	0	0	0	0	0

FIG. 4.4 – **Exemple de représentation CGR d'une signature génomique**

A partir de la séquence d'ADN S_1 de la Figure 4.1, une signature génomique a été construite contenant le nombre d'occurences de chaque mot de taille $m = 3$ construit à partir de l'alphabet $\Sigma = \{\text{A,C,G,T}\}$. La séquence S_1 de la Figure 4.1 (cf page 74) ayant été obtenue par la concaténation de trois sous-séquences, les quatre mots AGT, GTG, GTC et TCG induits par cette concaténation ne sont pas dénombrés dans la signature génomique.

d'états de caractère à chaque noeud interne, a été récemment proposé et permet ainsi des inférences phylogénomiques suivant les techniques probabilistes propres aux critères ML (Dicks, 2000; Savva et al., 2003).

4.3.3 Méthodes de distance basées sur les signatures génomiques

Une signature génomique est une représentation numérique et graphique de l'ordonnancement des états de caractères composant un gène ou un ensemble de gènes. Etant donné une séquence de caractères nucléotidiques S de taille M (sans aucun état de caractère manquant), une signature génomique est un tableau de taille 4^m contenant le nombre d'occurences de chaque mot de taille $m \in [1, M]$ fixée construit à partir de l'alphabet $\Sigma = \{\text{A,C,G,T}\}$. La Figure 4.4 représente la signature génomique de la séquence S_1 de la Figure 4.1 obtenue par concaténation (cf page 74). La structure multi-échelle en deux dimensions de la signature génomique schématisée dans la Figure 4.4 est nommée CGR (*Chaos Game Representation* ; Jeffrey, 1990 ; Deschavanne et al., 1999).

Si on normalise les valeurs contenues dans la signature génomique d'une séquence S de taille M par le nombre de mots de taille m qu'elle peut contenir (*i.e.* $M - m + 1$), il a été observé que chaque espèce est globalement caractérisée par une unique signature génomique normalisée et que la signature génomique normalisée d'un génome entier est très similaire à la

signature obtenue à partir d'une partie de ce génome, *i.e.* d'un, de plusieurs ou de l'ensemble de ses gènes (Deschavanne et al., 1999). Le calcul de la représentation CGR d'une signature génomique s'effectue linéairement en $O(M+m)$ (Jeffrey, 1990) et présente l'avantage de ne pas nécessiter d'étape préliminaire d'alignement.

Chaque (concaténation de) séquence(s) pouvant être caractérisée par une signature génomique, des distances entre chaque paire de signatures génomiques peuvent aisément être calculées (*e.g.* distance euclidienne, distance χ^2). Il a été montré, en utilisant les outils de Guénoche et Garreta (2001), que ces distances induisent un fort signal phylogénétique, parfois meilleur que celui induit par les distances évolutives estimées directement sur les alignements de séquence (Chapus et al., 2005).

4.3.4 Méthodes de distance basées sur les scores de BLAST

Etant données deux séquences de caractères, un score BLAST (*Basic Local Alignment Search Tool*) est un indice exprimant les différences entre ces deux séquences (pour plus de détails, cf Altschul *et al.*, 1990). Etant donnés deux génomes X et Y, une HSP (*High-scoring Segment Pair*) est une paire de sous-mots de X et Y, respectivement, présentant un score BLAST indiquant une forte similarité. Un grand nombre de HSPs peut être détecté entre X et Y, en particulier si on tolère les intersections entre deux HSPs.

Etant donné un ensemble de génomes, une technique de combinaison moyenne, nommée GDBP (*Genome BLAST Distance Phylogeny*; Henz *et al.*, 2005), consiste à rechercher, pour chaque paire de génomes X et Y, l'ensemble $\mathcal{H}(X,Y)$ des HSPs, puis à estimer une distance entre X et Y suivant la formule :

$$\Delta_{XY} \;=\; -\ln\left(\frac{\sum_{(x,y)\in\mathcal{H}(X,Y)}\left(\ell_x+\ell_y\right)}{2\min(\ell_X,\ell_Y)}\right),$$

où (x,y) représente une HSP et ℓ_S représente la longueur de la séquence S. La fonction $\min$ a été utilisée afin de compenser une éventuelle forte hétérogénéité entre les tailles ℓ_X et ℓ_Y de deux génomes X et Y. Différentes variantes sont possibles, comme ne pas utiliser la fonction logarithme, ou éliminer certaines HSPs afin d'empêcher toute intersection entre les éléments de $\mathcal{H}(X,Y)$. On peut également ordonnancer les différentes HSPs suivant leur score BLAST ou leur longueur respective, avant d'éliminer les HSPs partiellement recouvrantes. D'autres distances ont également été proposées en utilisant d'autres mesures de ressemblance à l'intérieur de chaque HSP (Auch et al., 2006).

Ce chapitre se clôture volontairement par la description de plusieurs méthodes de combinaison moyenne. Un des buts de l'inférence phylogénomique par combinaison moyenne étant d'interpréter des informations phylogénétiques sous la forme d'une structure de données permettant

son exploitation avec une complexité algorithmique peu élevée en temps et en espace mémoire, on remarque de beaucoup d'approches passent par l'estimation d'une distance évolutive entre chaque taxon.

Le chapitre suivant introduit une nouvelle méthode de combinaison moyenne, nommée SDM (*Super Distance Matrix*), permettant de combiner une collection de matrices de distance en une unique *supermatrice de distance*.

Chapitre 5

Utilisation des distances évolutives en inférence phylogénomique

L'objectivité, *i.e.* le réel physique, est un consensus construit à partir d'expériences de sujets.

Jean Largeault

Sommaire

5.1 Préliminaires biologiques et mathématiques 94

5.1.1 Les informations biologiques propres à l'inférence phylogénétique 94

5.1.2 Quelques propriétés des distances additives d'arbre 95

5.2 La méthode SDM 97

5.2.1 Définition des paramètres optimaux des modèles PM et SSM à l'aide d'un critère WLS 97

5.2.2 Calcul des paramètres optimaux des modèles SSM et PM 100

5.2.3 Construction d'une supermatrice de distance avec les paramètres optimaux des modèles SSM et PM 102

5.2.4 Complexités algorithmiques du calcul des paramètres et de la supermatrice de distance 102

5.3 Application et discussion 104

5.3.1 Estimation des vitesses d'évolution relatives à chaque gène 104

5.3.2 Inférence phylogénomique par combinaison moyenne 107

Étant donnée une collection de gènes, une méthode d'inférence phylogénomique efficace se caractérise par une extraction complète des informations phylogénétiques induites par chaque gène (*e.g.* signal topologique, vitesse d'évolution de chaque gène) et un assemblage précis de cet ensemble d'informations, tout en impliquant une complexité algorithmique peu élevée, ce qui permet de l'appliquer sur de grandes collections de gènes.

Or, en inférence phylogénétique, les méthodes exploitant les distances évolutives entre espèces à partir d'un gène présentent en général un rapport fiabilité/rapidité relativement élevé. En inférence phylogénomique, les techniques de combinaison moyenne ont pour but de maximiser ce même rapport qualitatif.

A la lumière de ces observations, ce chapitre introduit une nouvelle méthode d'inférence phylogénomique, nommée SDM (*Super Distance Matrix* ; Criscuolo et al., 2006), permettant d'effectuer la combinaison moyenne de l'ensemble des matrices de distance évolutive directement inférées à partir de chaque gène initial. SDM calcule, avec une complexité polynomiale, une supermatrice de distance contenant une grande partie de l'information phylogénétique induite par la collection de matrices de distance, en s'appuyant, en particulier, sur l'estimation de la vitesse d'évolution relative de chaque gène. L'application d'une méthode d'inférence d'arbres sur cette supermatrice de distance permet ainsi d'obtenir l'histoire évolutive du groupe d'espèces défini par la collection de gènes.

5.1 Préliminaires biologiques et mathématiques

Le principe *Total Evidence* (consistant à tenir compte du plus grand nombre d'informations ; cf page 74) appliqué au problème de l'inférence phylogénomique implique la considération de l'ensemble de l'information phylogénétique induite par une collection de gènes. Après une brève description des différentes types d'information phylogénétique, cette partie définit deux modèles phylogénomiques et leur adaptation aux collections de matrices de distance en s'appuyant sur certaines propriétés des distances additives d'arbre.

5.1.1 Les informations biologiques propres à l'inférence phylogénétique

Lors de l'étude phylogénétique d'un gène suivant un modèle d'évolution, plusieurs données se doivent d'être définies et exploitées :

- le modèle théorique d'évolution des séquences moléculaires,
- les probabilités d'évènements mutationels entre chaque état de caractère,
- l'hétérogénéité des vitesses d'évolution entre sites,
- la topologie de l'arbre phylogénétique T exprimant l'histoire évolutive des séquences considérées,
- les longueurs $l(e)$ des branches e de T exprimant la vitesse d'évolution entre chaque espèce (contemporaine aux feuilles de l'arbre, ou hypothétique aux noeuds internes).

Lorsque l'on considère une collection de gènes, il a été montré, pour une topologie fixée, que ces différents paramètres peuvent varier dans chaque gène (Pupko et al., 2002), ce qui implique une forme d'incongruence. Les techniques de combinaison basse basées sur les critères ML cherchent, la plupart du temps, la topologie T et les longueurs de branche $l(e)$ à partir de la supermatrice de caractères obtenue par concaténation des différents gènes G^1, G^2, ..., G^k.

Un modèle phylogénomique, nommé SM (*Separate Model* ; Pupko et al., 2002), considère que chaque gène G^p induit des longueurs de branche $l_m(e)$ différentes dans la topologie T (Yang, 1996b). Un autre modèle, nommé PM (*Proportional Model* ; Pupko et al., 2002), considère que chaque gène G^p induit une vitesse d'évolution α_p, qui s'applique comme un facteur multiplicatif sur les longueurs de branche $l(e)$ de la topologie optimale T. Il a été montré que des méthodes ML d'inférence phylogénomique basées sur les modèles SM et PM permettent d'améliorer significativement la vraisemblance et la fiabilité des arbres phylogénétiques obtenus (Pupko et al., 2002).

Ainsi, les méthodes d'inférence phylogénomique doivent s'alimenter des informations phylogénétiques propre à chaque gène afin de construire des arbres de meilleure qualité. Or, les critères MP ne permettent pas une telle exploitation des données (mis à part une pondération souvent difficile à estimer pour le calcul de la parcimonie de Sankoff), alors que les critères ML impliquent des temps de calcul souvent longs par la considération d'un plus grand nombre de paramètres. L'utilisation des critères de distance devrait permettre de se situer dans un niveau intermédiaire entre les rapports qualitatifs des critères MP et ML.

5.1.2 Quelques propriétés des distances additives d'arbre

Si deux arbres phylogénétiques valués T^1 et T^2 sont compatibles (*i.e.* $\mathcal{L}_{T^1} = \mathcal{L}_{T^2}$ et, éventuellement, l'arbre T^2 peut être obtenu en contractant plusieurs branches de T^1), alors la distance (Δ_{ij}) obtenue par l'opération $\Delta_{ij} = T^1_{ij} + T^2_{ij}$ est une distance additive d'arbre équivalente à un arbre T compatible avec T^1 et T^2 (cf Barthélemy and Guénoche, 1988, pour une présentation plus complète). Cette propriété permet d'associer aux distances additives d'arbre des paramètres modifiant certains signaux phylogénétiques tout en laissant invariant le signal topologique. Cette partie détaille deux types de paramètres : les facteurs scalaires et les distances à centre.

Invariance du signal topologique à la multiplication par un facteur scalaire

Dans le cadre des distances évolutives, une première approche pour tenir compte de l'hétérogénéité des vitesses d'évolution entre gènes a été proposée par Lapointe et Cucumel (1997) avec la méthode ACS. Disposant d'une collection $\mathcal{C}_{(T)} = \left\{(T^1_{ij}), (T^2_{ij}), ..., (T^p_{ij}), ..., (T^k_{ij})\right\}$ de matrices de distance additive, ces auteurs proposèrent de normaliser chaque distance (T^p_{ij}) à l'aide du facteur $\alpha_p = (\max_{i,j \in \mathcal{L}_{T^p}}[T^p_{ij}])^{-1}$. Ainsi la collection $\left\{(\alpha_1 T^1_{ij}), (\alpha_2 T^2_{ij}), ..., (\alpha_p T^p_{ij}), ..., (\alpha_k T^k_{ij})\right\}$ est entièrement composée de distances comprises dans l'intervalle $[0, 1]$.

Cette opération de déformation s'appuie sur la propriété que la multiplication d'une distance additive (T_{ij}) par un facteur $\alpha \neq 0$ ne modifie pas la topologie de l'arbre équivalent T. En effet, si cette propriété est fausse, alors il existe un quadruplet de feuilles $\{i, j, x, y\} \subset \mathcal{L}_T$ tel que,

suivant l'inégalité quadrangulaire :

$$T_{ij}+T_{xy} \;<\; T_{ix}+T_{jy} \;=\; T_{iy}+T_{jx} \quad \text{et} \quad \alpha T_{ix}+\alpha T_{jy} \;<\; \alpha T_{ij}+\alpha T_{xy} \;=\; \alpha T_{iy}+\alpha T_{jx}.$$

On obtient alors les deux équation et inéquation suivantes :

$$T_{ix}+T_{jy} \;=\; T_{iy}+T_{jx} \quad \text{et} \quad \alpha(T_{ix}+T_{jy}) \;<\; \alpha(T_{iy}+T_{jx}),$$

qui impliquent une contradiction pour toute valeur de α.

Un raisonnement similaire s'applique sur les distances évolutives (Δ^p_{ij}) directement estimées à partir d'un gène G^p. L'Annexe A montre que les algorithmes agglomératifs les plus courants, *i.e.* NJ et BIONJ, ainsi que leurs variantes, UNJ et MVR, sont invariables topologiquement à la multiplication par un facteur α_p, *i.e.* ils renvoient la même topologie d'arbre lorsqu'ils sont appliqués sur (Δ^p_{ij}) ou $(\alpha_p\Delta^p_{ij})$. Toutes les autres méthodes de distance (*e.g.* FITCH, MW) sont invariables topologiquement dans la plupart des cas de figure. Il est donc possible de considérer un paramètre α_p permettant d'estimer la vitesse d'évolution du gène G^p sans modifier l'information topologique contenue dans la matrice de distance (Δ^p_{ij}). Le modèle PM est ainsi envisageable en utilisant une collection de matrices de distance.

Invariance du signal topologique à l'ajout d'une distance à centre

Une distance à centre $(a_i + a_j)$ est équivalente à un arbre en étoile où la branche externe correspondant à la feuille i est de longueur a_i. L'ajout d'une distance à centre $(a_i + a_j)$ à une distance additive (T_{ij}) modifie, dans l'arbre équivalent T, la longueur $l(e_i)$ de la branche externe e_i correspondant à chaque feuille i. La longueur de chaque branche externe est alors de $l(e_i) + a_i$. Cette opération ne modifie donc pas la topologie de l'arbre équivalent T. En effet, dans le cas contraire, alors il existe un quadruplet de feuilles $\{i, j, x, y\} \subset \mathcal{L}_T$ tel que, suivant l'inégalité quadrangulaire :

$$\begin{cases} T_{ij}+T_{xy} \;<\; T_{ix}+T_{jy} \;=\; T_{iy}+T_{jx}, \\ T_{ix}+a_i+a_x+T_{jy}+a_j+a_y < T_{ij}+a_i+a_j+T_{xy}+a_x+a_y, \\ T_{ij}+a_i+a_j+T_{xy}+a_x+a_y = T_{iy}+a_i+a_y+T_{jx}+a_j+a_x. \end{cases}$$

On obtient alors les deux équation et inéquation suivantes :

$$T_{ix}+T_{jy} \;=\; T_{iy}+T_{jx} \quad \text{et} \quad T_{ix}+T_{jy} \;<\; T_{iy}+T_{jx},$$

qui impliquent une contradiction.

Un raisonnement similaire s'applique sur les distances évolutives (Δ^p_{ij}) directement estimées à partir d'un gène G^p. Les algorithmes NJ, UNJ, BIONJ et MVR sont invariables topologiquement à l'ajout d'une distance à centre $(a_i + a_j)$ (cf Gascuel, 1994, ainsi que l'Annexe A). Toutes les autres méthodes de distance (*e.g.* FITCH, MW) sont invariables topologiquement dans la plupart des cas de figure. Il est donc possible de considérer une matrice de distance à centre $(a_{ip} + a_{jp})$ permettant de modifier les longueurs de chaque branche externe sans modifier l'information topologique contenue dans la matrice de distance (Δ^p_{ij}).

Un modèle d'inférence phylogénomique dédié aux matrices de distance

La considération de k facteurs multiplicatifs α_p modifiant les matrices de distance d'une collection $\mathcal{C}_{(\Delta)} = \left\{(\Delta_{ij}^1), (\Delta_{ij}^2), ..., (\Delta_{ij}^p), ..., (\Delta_{ij}^k)\right\}$ permet de déduire un modèle similaire au modèle PM. Ce modèle considère chaque matrice de distance déformée $(\alpha_p \Delta_{ij}^p)$.

En complément, la modification locale (*i.e.* propre à chaque gène G^p) de l'estimation des longueurs de branches externes des arbres phylogénétiques issus de chacune des k matrices de distance (Δ_{ij}^p) à l'aide de k distances à centre $(a_{ip} + a_{jp})$ permet d'améliorer le modèle PM. Ce nouveau modèle considère chaque matrice de distance déformées $(\alpha_p \Delta_{ij}^p + a_{ip} + a_{jp})$.

Le modèle SM ne peut néanmoins pas être appliqué. En effet, si $n_p = |\mathcal{L}_{\Delta^p}|$ représente le nombre de taxons définis par chaque gène G^p, alors les $k(n_p - 3)$ branches internes ne peuvent être localement modifiées par la considération des seules distances à centre $(a_{ip} + a_{jp})$. Toutefois, les branches externes des arbres issus de l'inférence phylogénétique étant généralement plus longues que les branches internes, l'essentiel de la variance V_{ij}^p associée à la distance évolutive Δ_{ij}^p est supportée par les deux branches externes correspondant à i et j. De plus, les $k(n_p - 3)$ branches internes sont globalement modifiées par les k facteurs multiplicatifs α_p. Ce modèle, dédié aux matrices de distance, est donc appelé semi-séparé (SSM ; *Semi-Separate Model*).

5.2 La méthode SDM

Etant donné une collection de k matrices de distance évolutive $\mathcal{C}_{(\Delta)} = \left\{(\Delta_{ij}^1), (\Delta_{ij}^2), ..., (\Delta_{ij}^p), ..., (\Delta_{ij}^k)\right\}$, cette partie explique comment estimer les paramètres α_p et a_{ip} du modèle SSM à l'aide d'un critère de moindres-carrés pondérés, et comment les utiliser dans le cadre de la combinaison moyenne.

5.2.1 Définition des paramètres optimaux des modèles PM et SSM à l'aide d'un critère WLS

Une première approche

Etant donnés les différents paramètres α_p et a_{ip} permettant de déformer chacune des k matrices de distance (Δ_{ij}^p) en une nouvelle matrice de distance $(\alpha_p \Delta_{ij}^p + a_{ip} + a_{jp})$ contenant le même signal topologique, l'estimation de la valeur optimale du vecteur $v = (\alpha_1, ..., \alpha_p, ..., \alpha_k, ..., a_{ip}, ...)$ contenant ces $O(k + \sum_p n_p)$ paramètres est obtenue par la minimisation du critère de moindres-carrés pondérés suivant :

$$f(v) = \sum_{\substack{i,j \in \mathcal{L}_{\mathcal{C}_{(\Delta)}} \\ i \neq j \\ k_{ij} \geq 2}} \sum_{\substack{1 \leq p \leq k \\ i,j \in \tilde{\mathcal{L}}_{\Delta^p}}} w_p \Big(\alpha_p \Delta_{ij}^p + a_{ip} + a_{jp} - \overline{\Delta}_{ij} \Big)^2, \tag{5.1}$$

où

- w_p est un poids associé à chaque matrice de distance (Δ_{ij}^p),
- $k_{ij} = \big|\{i,j \in \mathcal{L}_{\Delta^p} : 1 \leq p \leq k\}\big|$ représente le nombre d'apparition de la paire ij dans la collection $\mathcal{C}_{(\Delta)}$,
- $\tilde{\mathcal{L}}_{\Delta^p} = \big\{i \in \mathcal{L}_{\Delta^p} : \exists j \in \mathcal{L}_{\Delta^p} - \{i\}\ ,\ k_{ij} \geq 2\big\}$ est défini par l'ensemble des paires de taxons ij qui apparaissent dans au moins deux matrices, *i.e.* $k_{ij} \geq 2$.
- $\overline{\Delta}_{ij}$ est la moyenne des distances déformées $(\alpha_p \Delta_{ij}^p + a_{ip} + a_{jp})$ pondérée par chaque w_p, *i.e.*

$$\overline{\Delta}_{ij} = \frac{1}{W_{ij}} \sum_{\substack{1 \leq p \leq k \\ i,j \in \tilde{\mathcal{L}}_{\Delta^p}}} w_p(\alpha_p \Delta_{ij}^p + a_{ip} + a_{jp}) \qquad \text{avec} \qquad W_{ij} = \sum_{\substack{1 \leq p \leq k \\ i,j \in \tilde{\mathcal{L}}_{\Delta^p}}} w_p.$$

Le but de la méthode SDM consiste à rechercher la valeur des paramètres α_p et a_{ip} de manière à minimiser la variance pondérée de chaque estimateur déformé de la distance évolutive entre chaque paire ij. Autrement dit, le critère (5.1) est la sommation, pour toutes les paires ij, du critère suivant :

$$\mathcal{V}_{ij} = \sum_{\substack{1 \leq p \leq k \\ i,j \in \tilde{\mathcal{L}}_{\Delta^p}}} w_p \Big(\alpha_p \Delta_{ij}^p + a_{ip} + a_{jp} - \overline{\Delta}_{ij}\Big)^2,$$

qui représente, pour une paire de taxons ij fixée, la variance pondérée des différentes distances déformées $(\alpha_p \Delta_{ij}^p + a_{ip} + a_{jp})$.

Si $k_{ij} = 0$, alors $\mathcal{V}_{ij}$ ne peut être calculé. Si $k_{ij} = 1$, alors $\overline{\Delta}_{ij} = \alpha_p \Delta_{ij}^p + a_{ip} + a_{jp}$, avec p fixé, et $\mathcal{V}_{ij} = 0$. Ainsi, la variance pondérée $\mathcal{V}_{ij}$ n'est informative qu'avec les paires de taxons $i,j \in \tilde{\mathcal{L}}_{\Delta^p}$, *i.e.* telles que $k_{ij} \geq 2$, et le critère (5.1) :

$$f(v) = \sum_{\substack{i,j \in \mathcal{L}_{\mathcal{C}_{(\Delta)}} \\ i \neq j \\ k_{ij} \geq 2}} \mathcal{V}_{ij}$$

n'est informatif qu'avec ces paires de taxons particulières. Certaines distances Δ_{ij}^p n'étant pas informatives dans le calcul du critère (5.1), les paramètres a_{ip} et a_{jp} associés n'ont pas besoin d'être estimés. Ainsi, pour chaque (Δ_{ij}^p), seuls les $\tilde{n}_p = |\tilde{\mathcal{L}}_{\Delta^p}| \leq n_p$ paramètres déformateurs a_{ip} sont significatifs dans le critère (5.1).

Si $w_p = 1$, alors la minimisation de la version OLS du critère (5.1) permet d'obtenir les $k + \sum_p \tilde{n}_p$ paramètres α_p et a_{ip} minimisant chaque variance non-pondérée $\mathcal{V}_{ij}$. Les pondérations w_p permettent d'associer un indice de fiabilité à l'ensemble des distances dans la matrice (Δ_{ij}^p) dans la version WLS du critère (5.1). Sachant que la variance V_{ij}^p de la distance Δ_{ij}^p estimée à partir d'un alignement de gènes G^p est toujours inversement proportionnel au nombre ℓ_p de sites, il est cohérent de poser $w_p = \ell_p$. D'un autre coté, une matrice de distance contenant peu de taxons a une moindre influence qu'une matrice contenant un

nombre élevé de taxons. Pour compenser l'éventuelle hétérogénéité de taille des différentes matrices de distance, on peut utiliser la pondération $w_p = 1/(\tilde{n}_p(\tilde{n}_p - 1))$. Une pondération intermédiaire est représentée par $w_p = 1/\tilde{n}_p$. D'autre possibilités de pondération sont possibles, telles que $w_p = 1/V_{ij}^p$, ou une combinaison des premières pondérations, par exemple $w_p = \ell_p/(\tilde{n}_p(\tilde{n}_p - 1))$ ou $w_p = \ell_p/\tilde{n}_p$.

Affinage du critère WLS pour le modèle SSM

Afin d'éviter un trop grand allongement (ou un trop grand raccourcissement) des branches externes correspondant au taxon i dans les différents arbres phylogénétiques induits par chaque matrices de distance (Δ_{ij}^p) telles que $i \in \tilde{\mathcal{L}}_{\Delta^p}$, une contrainte linéaire est nécessaire :

$$\sum_{\substack{1\leq p\leq k \\ i\in\tilde{\mathcal{L}}_{\Delta^p}}} a_{ip} = 0, \quad \forall i \in \tilde{\mathcal{L}}_{\mathcal{C}_{(\Delta)}} = \bigcup_{1\leq p\leq k} \tilde{\mathcal{L}}_{\Delta^p}. \tag{5.2}$$

Pour une raison similaire à celle nécessitant la contrainte (5.2), mais généralisée à l'ensemble des branches externes de l'arbre phylogénétique induit par chaque matrice de distance (Δ_{ij}^p), une seconde contrainte linéaire est nécessaire :

$$\sum_{i\in\tilde{\mathcal{L}}_{\Delta^p}} a_{ip} = 0, \ \ \forall p = 1, 2, ..., k-1. \tag{5.3}$$

Cette deuxième contrainte évite que l'ensemble des branches externes d'un arbre soient trop longues, et ainsi d'étouffer les longueurs de branche interne. La contrainte linéaire (5.3) correspondant à $p = k$, *i.e.*

$$\sum_{i\in\tilde{\mathcal{L}}_{\Delta^k}} a_{ik} = 0 \tag{5.4}$$

est inutile car induite par l'ensemble des contraintes linéaires (5.2) et (5.3). En effet, d'après les contraintes (5.2), on a

$$a_{ik} = -\sum_{\substack{1\leq p\leq k-1 \\ i\in\tilde{\mathcal{L}}_{\Delta^p}}} a_{ip}, \ \ \forall i \in \tilde{\mathcal{L}}_{\mathcal{C}_{(\Delta)}},$$

ce qui implique

$$\sum_{i\in\tilde{\mathcal{L}}_{\mathcal{C}_{(\Delta)}}} a_{ik} = -\sum_{i\in\tilde{\mathcal{L}}_{\mathcal{C}_{(\Delta)}}} \left(\sum_{\substack{1\leq p\leq k-1 \\ i\in\tilde{\mathcal{L}}_{\Delta^p}}} a_{ip} \right) = -\sum_{\substack{1\leq p\leq k-1 \\ i\in\tilde{\mathcal{L}}_{\Delta^p}}} \left(\sum_{i\in\tilde{\mathcal{L}}_{\mathcal{C}_{(\Delta)}}} a_{ip} \right).$$

Comme on a

$$\sum_{i\in\tilde{\mathcal{L}}_{\mathcal{C}_{(\Delta)}}} a_{ik} = \sum_{i\in\tilde{\mathcal{L}}_{\Delta^k}} a_{ik} \qquad \text{et} \qquad \sum_{i\in\tilde{\mathcal{L}}_{\mathcal{C}_{(\Delta)}}} a_{ip} = \sum_{i\in\tilde{\mathcal{L}}_p} a_{ip} = 0,$$

on retrouve donc bien la contrainte linéaire (5.4) à partir des contraintes (5.2) et (5.3).

Le cas particulier du modèle PM

Si on pose $a_{ip} = 0$ pour toute matrice de distance (Δ_{ij}^p) et tout taxon i, alors la minimisation du critère (5.1) permet d'obtenir les différentes valeurs optimales α_p du modèle PM au sens des moindres-carrés pondérés. En effet, dans ce cas, le critère (5.1) devient :

$$f(\alpha_1, \alpha_2, ..., \alpha_p, ..., \alpha_k) = \sum_{\substack{i,j \in \mathcal{L}_{\mathcal{C}_{(\Delta)}} \\ i \neq j \\ k_{ij} \geq 2}} \sum_{\substack{1 \leq p \leq k \\ i,j \in \tilde{\mathcal{L}}_{\Delta^p}}} w_p \Big(\alpha_p \Delta_{ij}^p - \overline{\Delta}_{ij} \Big)^2,$$

avec

$$\overline{\Delta}_{ij} = \frac{1}{W_{ij}} \sum_{\substack{1 \leq p \leq k \\ i,j \in \tilde{\mathcal{L}}_{\Delta^p}}} w_p (\alpha_p \Delta_{ij}^p).$$

Or $f(\mathbf{0}) = 0$, où $\mathbf{0}$ est un vecteur uniquement composé de 0 (cette observation est également valable avec le modèle SSM). Pour éviter cette solution triviale induite par le modèle PM, Pupko et al. (2002) ont proposé d'utiliser la contraint linéaire suivante :

$$\sum_{1 \leq p \leq k} \alpha_p = k, \tag{5.5}$$

afin que tous les paramètres α_p soient égaux à 1 en moyenne. Cette contrainte permet d'éviter la solution triviale $(\alpha_1, \alpha_2, ..., \alpha_p, ..., \alpha_k) = \mathbf{0}$ et offre la possibilité d'interpréter les vitesses d'évolution $1/\alpha_p$ relatives de chaque gène G^p.

5.2.2 Calcul des paramètres optimaux des modèles SSM et PM

Le calcul des paramètres optimaux α_p et a_{ip} par la minimisation du critère (5.1) sous les contraintes linéaires (5.2), (5.3) et (5.5) est un problème polynomial, aussi bien pour le modèle SSM que PM. Cette partie décrit les différents calculs permettant d'aboutir à cette conclusion.

Le modèle SSM

Le vecteur v^* contenant les paramètres optimaux du modèle SSM est celui qui résoud le système suivant :

$$\begin{cases} \min_v f(v) & & \\ h^{(1)}(v) = 0 & & \text{avec} \quad h^{(1)}(v) = \Big(\sum_{1 \leq p \leq k} \alpha_p \Big) - k \\ h_i^{(2)}(v) = 0 & \forall i \in \mathcal{L}_{\mathcal{C}_{(\Delta)}} & \text{avec} \quad h_i^{(2)}(v) = \sum_{\substack{1 \leq p \leq k \\ i \in \tilde{\mathcal{L}}_{\Delta^p}}} a_{ip} \\ h_p^{(3)}(v) = 0 & \forall p = 1, 2, ..., k-1 & \text{avec} \quad h_p^{(3)}(v) = \sum_{i \in \tilde{\mathcal{L}}_{\Delta^p}} a_{ip}. \end{cases}$$

Ce problème de minimisation du critère quadratique (5.1) sous les contraintes linéaires $h^{(1)}(v) = 0$, $h_i^{(2)}(v) = 0$ et $h_p^{(3)}(v) = 0$ revient à résoudre le système linéaire suivant :

$$\begin{cases} \frac{\partial}{\partial \alpha_m} f(v) \ + \ \lambda \frac{\partial}{\partial \alpha_m} h^{(1)}(v) \ = \ 0, & \forall m = 1, ..., k, \\ \frac{\partial}{\partial a_{im}} f(v) \ + \ \mu_i \frac{\partial}{\partial a_{im}} h_i^{(2)}(v) \ + \ \eta_m \frac{\partial}{\partial a_{im}} h_m^{(3)}(v) \ = \ 0, & \forall m = 1, ..., k-1, \ \forall i \in \tilde{\mathcal{L}}_{\Delta^m}, \\ \frac{\partial}{\partial a_{ik}} f(v) \ + \ \mu_i \frac{\partial}{\partial a_{ik}} h_i^{(2)}(v) \ = \ 0, & \forall i \in \tilde{\mathcal{L}}_{\Delta^k}, \\ h^{(1)}(v) \ = \ 0, & \\ h_i^{(2)}(v) \ = \ 0, & \forall i \in \mathcal{L}_{\mathcal{C}_{(\Delta)}}, \\ h_p^{(3)}(v) \ = \ 0, & \forall p = 1, 2, ..., k-1, \end{cases} \tag{5.6}$$

où λ, μ_i et η_m sont les multiplicateurs de Lagrange associés aux contraintes linéaires $h^{(1)}(v) = 0$, $h_i^{(2)}(v) = 0$ et $h_p^{(3)}(v) = 0$, respectivement. Comme on a (cf Annexe A) :

$$\frac{\partial}{\partial \alpha_m} f(v) = 2 w_m \sum_{\substack{i,j \in \tilde{\mathcal{L}}_{\Delta^m} \\ i \neq j}} \Delta_{ij}^m \Big(\alpha_m \Delta_{ij}^m + a_{im} + a_{jm} \ - \ \overline{\Delta}_{ij} \Big),$$

$$\frac{\partial}{\partial a_{im}} f(v) = 4 w_m \sum_{j \in \tilde{\mathcal{L}}_{\Delta^m} - \{i\}} \Big(\alpha_m \Delta_{ij}^m + a_{im} + a_{jm} \ - \ \overline{\Delta}_{ij} \Big),$$

alors le système linéaire (5.6) peut être réécrit de manière analytique :

$$\begin{cases} \sum\limits_{\substack{i,j \in \tilde{\mathcal{L}}_{\Delta^p} \\ i \neq j}} \Delta_{ij}^p \left(\alpha_p \Delta_{ij}^p + a_{ip} + a_{jp} \ - \ \overline{\Delta}_{ij} \right) + \lambda = 0, & \forall p = 1, ..., k, \\ w_p \sum\limits_{j \in \tilde{\mathcal{L}}_{\Delta^p} - \{i\}} \Big(\alpha_p \Delta_{ij}^p + a_{ip} + a_{jp} \ - \ \overline{\Delta}_{ij} \Big) + \mu_i + \eta_p = 0, & \forall p = 1, ..., k-1, \ \forall i \in \tilde{\mathcal{L}}_{\Delta^p}, \\ \sum\limits_{j \in \tilde{\mathcal{L}}_{\Delta^k} - \{i\}} \Big(\alpha_k \Delta_{ij}^k + a_{ik} + a_{jk} \ - \ \overline{\Delta}_{ij} \Big) + \mu_i = 0, & \forall i \in \tilde{\mathcal{L}}_{\Delta^k}, \\ \sum\limits_{1 \leq p \leq k} \alpha_p = k, & \\ \sum\limits_{\substack{1 \leq p \leq k \\ i \in \tilde{\mathcal{L}}_{\Delta^p}}} a_{ip} = 0, & \forall i \in \mathcal{L}_{\mathcal{C}_{(\Delta)}}, \\ \sum\limits_{i \in \tilde{\mathcal{L}}_{\Delta^p}} a_{ip} = 0, & \forall p = 1, ..., k-1. \end{cases} \tag{5.7}$$

Le modèle PM

Le vecteur $v^* = (\alpha_1^*, \alpha_2^*, ..., \alpha_k^*)$ contenant les paramètres optimaux du modèle PM est celui qui résoud le système suivant :

$$\begin{cases} \min_v \left[\sum\limits_{\substack{i,j \in \mathcal{L}_{\mathcal{C}_{(\Delta)}} \\ i \neq j \\ k_{ij} \geq 2}} \sum\limits_{\substack{1 \leq p \leq k \\ i,j \in \tilde{\mathcal{L}}_{\Delta^p}}} w_p \Big(\alpha_p \Delta_{ij}^p - \overline{\Delta}_{ij} \Big)^2 \right] & \text{avec } \overline{\Delta}_{ij} = \frac{1}{W_{ij}} \sum\limits_{\substack{1 \leq p \leq k \\ i,j \in \tilde{\mathcal{L}}_{\Delta^p}}} w_p (\alpha_p \Delta_{ij}^p) \\ h^{(1)}(v) = 0 & \text{avec } h^{(1)}(v) = \Big(\sum\limits_{1 \leq p \leq k} \alpha_p \Big) - k \end{cases}$$

Suivant un raisonnement similaire à celui conduisant aux paramètres optimaux pour le modèle SSM, les k paramètres α_p^* sont obtenus par la résolution du système linéaire suivant :

$$\begin{cases} \sum\limits_{\substack{i,j \in \tilde{\mathcal{L}}_{\Delta^p} \\ i \neq j}} \Delta_{ij}^p \Big(\alpha_p \Delta_{ij}^p \; - \; \overline{\Delta}_{ij} \Big) \; + \; \lambda = 0, \quad \forall p = 1, 2, ..., k, \\ \\ \sum\limits_{1 \leq p \leq k} \alpha_p = k. \end{cases} \tag{5.8}$$

5.2.3 Construction d'une supermatrice de distance avec les paramètres optimaux des modèles SSM et PM

Soient α_p^* (et a_{ip}^*) les paramètres optimaux des modèles PM et SSM. La méthode SDM consiste, après avoir calculé ces paramètres, à inférer une supermatrice de distance $(\Delta_{ij}^{\text{SDM}})$ à l'aide de la formule :

$$\Delta_{ij}^{\text{SDM}} \; = \; \frac{1}{W_{ij}} \sum_{\substack{1 \leq p \leq k \\ i,j \in \mathcal{L}_{\Delta^p}}} w_p \big(\alpha_p^* \Delta_{ij}^p + a_{ip}^* + a_{jp}^* \big). \tag{5.9}$$

On pose $a_{ip}^* = 0$ si l'on considère le modèle PM. Chaque élément Δ_{ij}^{SDM} est la moyenne pondérée des distances déformées $\alpha_p^* \Delta_{ij}^p + a_{ip}^* + a_{jp}^*$. L'application d'une méthode d'inférence d'arbre phylogénétique (pouvant traiter d'éventuelles distances manquantes) sur $(\Delta_{ij}^{\text{SDM}})$ permet ainsi de définir un scénario d'inférence phylogénomique par la combinaison moyenne d'une collection de gènes.

5.2.4 Complexités algorithmiques du calcul des paramètres et de la supermatrice de distance

Le système linéaire (5.7), défini avec $\tilde{n} + 2k + \sum_{1 \leq p \leq k} \tilde{n}_p \in O(k\tilde{n})$ équations et variables, où $\tilde{n} = |\mathcal{L}_{\mathcal{C}_{(\Delta)}}|$, se construit avec une complexité algorithmique de l'ordre de $O(k^2 \tilde{n}^2)$

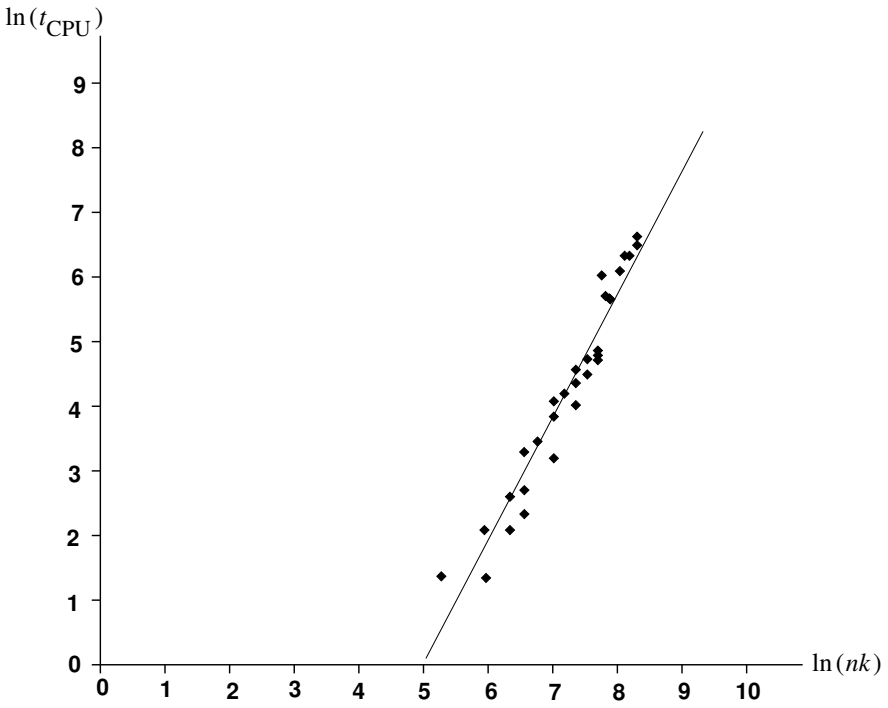

FIG. 5.1 – **Estimation de la complexité en pratique de la résolution du système linéaire de la méthode SDM**

Chaque point représente le logarithme du temps d'exécution de la résolution du système linéaire (5.1) en fonction des valeurs nk. La droite représente la régression linéaire de ce nuage de points.

et se résoud avec une complexité algorithmique de l'ordre de $O(k^3\tilde{n}^3)$, à l'aide de l'algorithme du pivot (Gauss, 1876). Néanmoins, ce système linéaire étant très creux (*e.g.* entre 50% et 85% de coefficients nuls observés en pratique), sa résolution induit une complexité en pratique de l'ordre de $O(k^\zeta\tilde{n}^\zeta)$, avec $\zeta < 3$. Afin d'estimer la valeur de ζ en pratique, trente collections de matrices de distance ont été générées avec $n = 50, 100, 150, 200, 250, 300$ et $k = 4, 8, 12, 16, 20,$ puis traitées avec une implémentation de la construction et de la résolution du système linéaire en JAVA 1.4 utilisant la librairie de résolution de systèmes linéaires MTJ[1]. Pour chaque paire (n, k), le temps t_{CPU} nécessaire à l'exécution de ce programme a été mesuré. Considérant que t_{CPU} est de la forme $\beta(nk)^\zeta$, la droite de régression linéaire $\ln(t_{\text{CPU}}) = \zeta \ln(nk) + \ln\beta$ a été estimée. La Figure 5.1 représente le nuage de points ainsi obtenu ainsi que la droite de régression linéaire. Les valeurs $\zeta = 1.855$ et $\ln\beta = -9.4281$ permettent d'affirmer que la complexité de la résolution du système linéaire (5.6) est au plus quadratique en pratique (sur les cas considérés).

Concernant le modèle PM, le système linéaire (5.8) se construit avec une complexité de l'ordre de $O(k^2\tilde{n}^2)$ et se résoud avec une complexité de l'ordre de $O(k^3)$. Lors de l'utilisation de

[1] disponible à l'URL `https://mtj.dev.java.net/`

ce modèle, les temps d'exécution ayant été très rapides dans tous les cas de figure (*e.g.* toujours moins de 5 secondes par collection de matrices de distance), aucune étude de complexité en pratique n'a été effectuée.

La supermatrice de distance $(\Delta_{ij}^{\text{SDM}})$ se construit avec une complexité de l'ordre de $O(kn^2)$ avec l'ensemble des distances disponibles, *i.e.* induites par toutes les paires $i, j \in \mathcal{L}_{\Delta^p}$, pour tout $p = 1, 2, ..., k$.

La méthode SDM permet ainsi d'inférer une supermatrice de distance avec une complexité théorique de $O(k^3n^3)$ si l'on utilise le modèle SSM, et de $O(k^3n^2)$ si l'on utilise le modèle PM (la complexité en pratique étant bien moindre).

5.3 Application et discussion

Dans le but d'illustrer les performance de la méthode SDM[2] avec les modèles SSM et PM, cette partie décrit l'analyse d'un jeu de données déjà utilisé dans un contexte de combinaison basse par Gatesy et al. (2002). Ce jeu de données était composé à l'origine de cinquante-sept sources de caractères comprenant trois ensembles d'états de caractère morphologique et les séquences de cinq protéines, un transposon, trente-trois gènes nucléaires et quinze gènes d'ADN mitochondrial. Le jeu de données ré-analysé dans cette partie n'est plus composé que des quarante-huit séquences d'ADN (les données morphologiques et le transposon n'induisent pas de vitesses d'évolution, et les séquences de protéine ne sont pas comparables avec celles d'ADN lorsque l'on pondère la Formule (5.1) avec la taille ℓ_p des séquences). Ces quarante-huit gènes définissent un groupe de soixante-quinze mammifères placentaires et leur concaténation produit une supermatrice de caractères comprenant 36639 sites et 68% d'états de caractère manquants. Le Tableau 5.1 résume les quarante-huit gènes formant le jeu de données étudié dans cette partie (colonnes G^p).

5.3.1 Estimation des vitesses d'évolution relatives à chaque gène

La méthode SDM permet de calculer une estimation des vitesses d'évolution $1/\alpha_p$ de chaque gène suivant deux modèles. Le modèle PM associe un paramètre α_p à chaque gène G^p et estime les valeurs de chaque α_p en résolvant le système linéaire (5.8). Cette approche est très similaire à celle proposée par Bevan *et al.*, 2006. Le modèle SSM complète le modèle PM en ajoutant un paramètre a_{ip} associé à chaque taxon i dans chaque gène G^p.

Pour chacun des quarante-huit gènes, une matrice de distance a été estimée suivant un modèle d'évolution GTR. La méthode SDM a été appliquée suivant les modèle PM et SSM, avec $w_p = \ell_p/(\tilde{n}_p(\tilde{n}_p - 1))$ utilisé comme pondération dans le critère (5.1) pour compenser la forte hétérogénéité dans les tailles n_p des différentes matrices de distance (cf Tableau 5.1, colonnes n_p). Le temps d'exécution de SDM a été d'environ 4 secondes avec le modèle PM et

[2] implémentée en JAVA 1.4 et disponible à l'URL `http ://www.lirmm.fr/ criscuol/soft/sdm`

		$1/\alpha_p$				$1/\alpha_p$	
G^p	n_p	SSM	PM	G^p	n_p	SSM	PM
ZFX	23	0.3565	0.3585	BRCA1	23	1.1918	1.2466
BDNF	24	0.4638	0.4515	PLCB4	24	1.3775	1.4297
APP	24	0.4704	0.4666	CO1 mtDNA	21	1.4218	1.3933
CNR1	22	0.5040	0.4821	TNF-A	23	1.4277	1.4529
ADBR2	24	0.5268	0.5197	protamine P1	26	1.4696	1.5293
EDG1	22	0.5273	0.5175	Thyroglobuline	24	1.4864	1.4774
RAG2	22	0.5297	0.5261	SPTBN1	24	1.4971	1.4823
BMI1	20	0.6150	0.5871	CO3 mtDNA	21	1.5230	1.5045
ATP7A	24	0.6312	0.6375	β-casein intron 7	16	1.5672	1.5784
RAG1	19	0.6368	0.6067	γ-fibriogène	31	1.5905	1.7316
TYR	18	0.7296	0.7089	CO2 mtDNA	33	1.6017	1.6420
16S rDNA	46	0.7716	0.7899	ND1 mtDNA	21	1.6514	1.6414
A2AB	19	0.8164	0.7976	cytochrome b mtDNA	75	1.7352	1.7632
Thyrotropine	24	0.8708	0.8360	NADH4L mtDNA	21	1.9097	1.9004
ADORA3	23	0.8869	0.9321	NADH3 mtDNA	21	1.9356	1.9547
CREM	23	0.8989	0.9581	ATP6 mtDNA	21	1.9384	1.9457
12S rDNA	73	0.9934	1.0417	α-lactalbumine	10	2.0280	1.9078
PNOC	23	0.9939	1.0061	κ-casein exon 4	45	2.0295	2.2988
MGF	24	1.0234	0.9846	NADH4	21	2.0349	2.0381
PRKC1	24	1.0520	1.0393	β-casein exon 7	51	2.1054	2.3423
vWF	28	1.1445	1.1616	NADH5	21	2.1063	2.0920
κ-casein intron 3	26	1.1500	1.1348	NADH2	21	2.3438	2.3717
STAT5	24	1.1743	1.2165	NADH6	21	2.4154	2.4659
IRBP	31	1.1849	1.2333	ATP8 mtDNA	21	3.7558	4.0724

TAB. 5.1 – **Description du jeu de données de Gatesy et al. (2002)**

Les colonnes G^p contiennent le nom des gènes, les colonnes n_p contiennent le nombre de taxons par gènes et les colonnes $1/\alpha_p$ contiennent les vitesses d'évolution relatives estimées par la méthode SDM avec le modèle SSM et PM.

d'environ 30 secondes avec le modèle SSM sur un PC Pentium IV 1.8GHz (1Go RAM).

Les estimations par SDM des différentes vitesses d'évolution relatives de chacun des quarante-huit gènes sont représentées dans le Tableau 5.1 (colonnes $1/\alpha_p$). La méthode SDM permet ainsi d'observer une forte hétérogénéité des vitesses d'évolution entre les quarante-huit gènes. Le gène nucléaire ZFX est le plus lent (PM : 0.3585 et SSM : 0.3565) et le gène mitochondrial ATP8 est le plus rapide (PM : 4.0724 et SSM : 3.7558). Le rapport entre le gène le plus rapide et le plus lent est donc d'environ 10.5

Comparaison des modèles PM et SSM

L'estimation des vitesses d'évolution $1/\alpha_p$ effectuée par SDM avec le modèle SSM est relativement similaire à celle du modèle PM. Cette partie décrit un protocole de simulation ayant pour

but d'observer les similitudes entre les estimations de $1/\alpha_p$ avec ces deux modèles.

Protocole de simulation

Une analyse phylogénétique à partir de chacun des 48 gènes G^p a été effectuée par le logiciel PHYML suivant le modèle GTR. Cette analyse a permis de collecter les paramètres M^p du modèle GTR pour chacun des 48 gènes G^p. L'inférence phylogénomique par combinaison basse a été également effectuée avec le logiciel PHYML suivant le modèle GTR+Γ (avec estimation du paramètre α de la loi gamma modélisant l'hétérogénéité des vitesses d'évolution entre sites) afin d'obtenir un arbre modèle $T_{\mathcal{M}}$.

Les paramètres M^p d'un gène sur trois dans le Tableau 5.1 (*i.e.* ZFX, CNR1, RAG2, RAG1, A2AB, CREM, MGF, κ-casein intron 3, BRCA1, TNF-A, SPTBN1, γ-fibriogène, cytochrome b mtDNA, ATP6 mtDNA, NADH4, NADH2) ont été utilisés conjointement avec les vitesses d'évolution relatives originales (colonnes $1/\alpha_p$ SSM du Tableau 5.1) pour simuler l'évolution le long des branches de $T_{\mathcal{M}}$ avec le logiciel SEQ-GEN (Rambaut and Grassly, 1997), afin d'obtenir une collection de 16 alignements de séquences d'états de caractère nucléaire définis sur 75 taxons.

Certains taxons ont été aléatoirement supprimés dans chacun des 16 alignements de gènes avec une probabilité de suppression de 25% et 75% afin d'obtenir une collection de 16 alignements de séquences G'^m. Néanmoins un ensemble d'au moins quatre taxons a été laissé en commun entre chaque paire de gènes. Une collection de 16 matrices de distance GTR a été estimée à partir des 16 gènes partiellement supprimés G'^m, et a été analysée par la méthode SDM suivant les modèles PM et SSM.

Ce protocole de simulation a été répété 500 fois pour chaque taux de suppression de taxons (25%, 75%) et chacun des deux modèles PM et SSM.

Pour chacun des deux modèles (PM et SSM), chacun des deux taux de suppression (25% et 75%) et chacune des 500 simulations effectuées, les 16 vitesses d'évolution relatives inférées par SDM (*estimated rates*) ont été représentées graphiquement en fonction des vitesses d'évolution relatives originales (*real rates*) dans les quatre graphiques de la Figure 5.2. Dans chacun des quatre graphiques, une courbe relie les 16 moyennes des 500 estimations de vitesse d'évolution relative. L'écart-type des 16 moyennes est également représenté.

Résultats

On observe que quel que soit le taux de suppression et quel que soit le modèle choisi, la méthode SDM retrouve globalement bien en moyenne les vitesses d'évolution relatives de chacun des 16 gènes. On remarque que l'écart-type d'une estimation est d'autant plus élevée que la vitesse relative est importante. Malgré une bonne estimation en moyenne, les vitesses d'évolution relatives sont plus souvent sur-estimées que sous-estimées. Ce biais est dû au fait que les paramètres α_p calculés par SDM correspondent à l'inverse de la vitesse d'évolution relative.

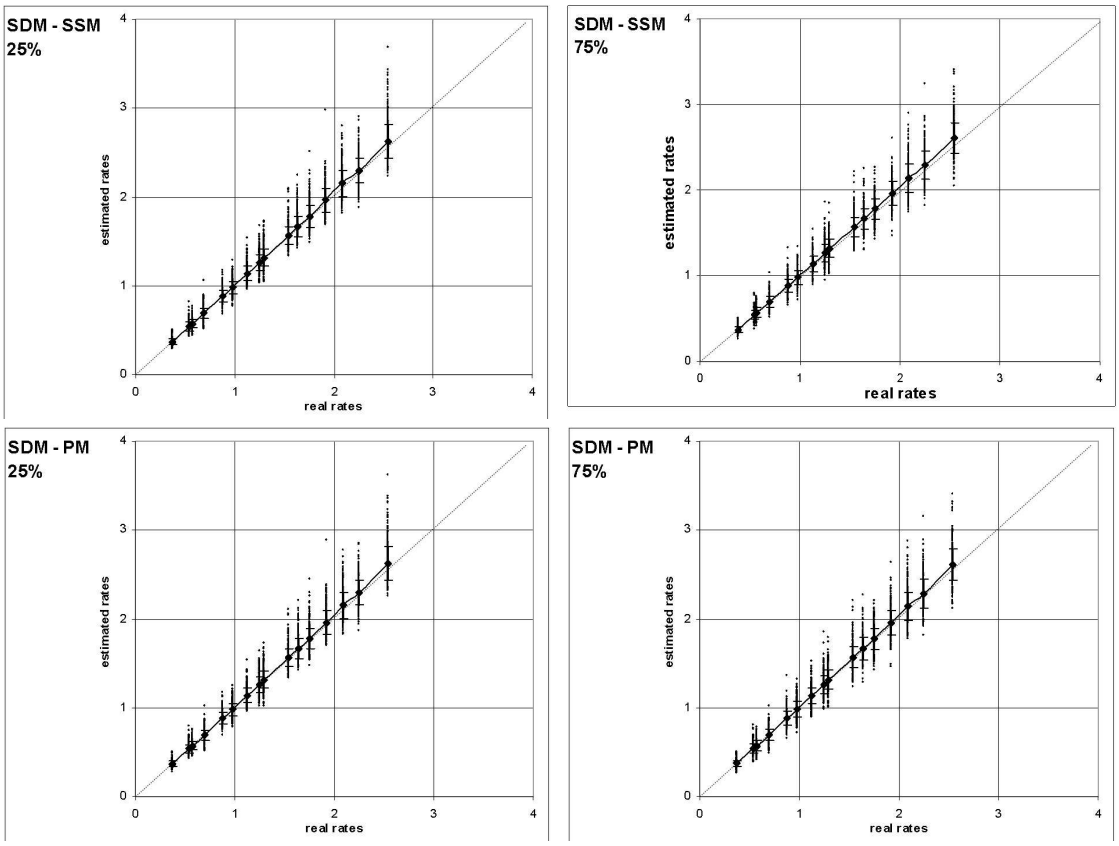

FIG. 5.2 – **Observation de la stabilité des estimations des vitesses d'évolution relatives par SDM**

Ces simulations montrent que le modèle PM permet d'obtenir une estimation des vitesses d'évolution relatives similaires à celle du modèle SSM, mais avec une complexité algorithmique théorique et pratique inférieure d'un facteur n. Elles montrent aussi que la taille des gènes (en nombre de taxons) n'a pas d'impact particulier sur la qualité d'estimation des vitesses $1/\alpha_p$ de la méthode SDM.

5.3.2 Inférence phylogénomique par combinaison moyenne

La méthode SDM ayant été appliquée sur les quarante-huit gènes du jeu de données de Gatesy et al. (2002), deux supermatrices de distance $(\Delta_{ij}^{\text{SDM}})$ ont été construites suivant les modèles PM et SSM. Le gène cytochrome b étant séquencé pour chacun des 75 taxons (cf Tableau 5.1), les deux supermatrices de distance ne contiennent aucune distance manquante. Ainsi deux arbres phylogénétiques ont pu être contruits avec le logiciel FASTME. FASTME a été appliqué de manière à rechercher les arbres minimisant le critère BME. Les vraisemblances $\ln L$ des deux arbres ont été estimées avec le logiciel PHYML suivant le modèle GTR+Γ.

L'arbre inféré à partir de la supermatrice de distance $(\Delta_{ij}^{\text{SDM}})$ correspondant au modèle PM

a une vraisemblance de l'ordre de $\ln L = -332224$, alors que l'arbre inféré à partir de $(\Delta_{ij}^{\text{SDM}})$ correspondant au modèle SSM a une vraisemblance $\ln L = -331532$. L'utilisation du modèle SSM, plus complexe que PM, a donc permis de construire un arbre plus fiable au sens du critère ML. Même si l'arbre $T_{\mathcal{M}}$, inféré par l'analyse simultanée de la supermatrice de caractère obtenue par concaténation des 48 gènes, a une vraisemblance $\ln L = -330354$, le temps d'exécution pour inférer $T_{\mathcal{M}}$ a été de presque 5 heures sur un PC Pentium IV 1.8 GHz (1Go RAM). Les scénarios d'inférence phylogénomique utilisant la méthode SDM ont tout deux duré moins de 30 secondes, grâce au fait que les supermatrices de distance $(\Delta_{ij}^{\text{SDM}})$ ne sont pas incomplètes et que FASTME nécessite moins d'une seconde pour construire un arbre phylogénétique de 75 taxons.

Comparaison des modèles PM et SSM

Comme les scénarios de combinaison moyenne utilisant SDM suivi d'une méthode d'inférence d'arbre basé sur un critère de distance présentent des performances différentes sur le jeu de données de Gatesy *et al.* (2002), des simulations ont été effectuées afin d'observer si les modèles PM et SSM permettent d'inférer des arbres de qualité similaire ou différente. Cette partie décrit le protocole de simulation, suivi des résultats.

Protocole de simulation

• *Génération des arbres modèles* — Un arbre modèle défini sur $n = 48$ taxons a été généré à l'aide du logiciel R8S (Sanderson, 2003) suivant le processus de Yule-Harding (Yule, 1925; Harding, 1971). L'arbre modèle ultramétrique enraciné UT ainsi obtenu respecte l'hypothèse de l'horloge moléculaire en garantissant une distance de 1 entre la racine et chacune des feuilles. Une déviation est créée par rapport à cette hypothèse en multipliant chaque branche par $(1+X)$ où X suit une loi exponentielle de paramètre $0.2/(0.001 + U)$, U suivant une loi uniforme (Guindon and Gascuel, 2003). Soit tbl la longueur totale de cet arbre. On obtient l'arbre non-ultramétrique T de longueur totale 1 en divisant chaque longueur de branche par tbl.
On génère, à partir de T, k arbres T^p en multipliant chaque branche par $0.4 + 8.6V_p$ où V_p suit une loi uniforme. La valeur de V_p est propre à chaque arbre T^p. Les k arbres sources T^p possèdent ainsi la même topologie que l'arbre T et une vitesse d'évolution propre variant uniformément entre 0.4 et 9.0.
• *Génération des données* — Une collection de k alignements de $n = 48$ séquences sur ℓ_p sites a été générée à l'aide du logiciel SEQ-GEN (Rambaut and Grassly, 1997) à partir de chaque arbre T^p suivant le modèle K2P. La valeur de ℓ_p est tirée uniformément entre 200 et 1000 sites et est propre à chaque arbre T^p.
Pour chaque alignement, certains taxons ont été aléatoirement supprimés suivant une certaine probabilité de délétion (Eulenstein et al., 2004). Deux valeurs de probabilité ont été utilisées :

25% pour une suppression faible et 75% pour une suppression forte. Un recouvrement d'au moins quatre taxons entre chaque paire de gènes a été néanmoins préservé afin de conserver une histoire évolutive et une information topologique communes significatives entre les k gènes. Ce processus de génération a été répété 500 fois pour chaque valeur de $k = 2, 4, 6, ..., 20$ et chacune des deux probabilités de suppression, aboutissant ainsi à $500 \times 10 \times 2$ jeux de données portant sur 48 taxons.

• *Inférence phylogénomique par combinaison moyenne* — A partir des k alignements partiellement effacés, une collection $\mathcal{C}_{(\Delta)}$ de k matrices de distance a été estimée suivant le modèle K2P. Une supermatrice de distance SDM a été calculée à partir de la collection $\mathcal{C}_{(\Delta)}$ pour chacun des deux modèles PM et SSM. Chaque matrice (Δ_{ij}^{p}) a été pondérée dans le critère (5.1) par la taille ℓ_p des séquences correspondantes. Si la supermatrice de distance $(\Delta_{ij}^{\text{SDM}})$ ne vérifiait pas l'inégalité triangulaire, elle a été tranformée en métrique par l'ajout de la constante positive $c = \max_{x,y,z \in \mathcal{L}_{\mathcal{C}_{(\Delta)}}} \left[\Delta_{xz}^{\text{SDM}} - \Delta_{xy}^{\text{SDM}} - \Delta_{yz}^{\text{SDM}} \quad \text{tq} \quad \Delta_{xz}^{\text{SDM}}, \Delta_{xy}^{\text{SDM}}, \Delta_{yz}^{\text{SDM}} \neq \varnothing \right]$ à chacune de ses valeurs non-diagonales. Cette opération revient à ajouter une distance à centre $(c_i + c_j)$ avec $c_i = c/2$, et ne déforme donc pas le signal topologique induit par la supermatrice de distance.

Une phylogénie a ensuite été inférée avec le programme FITCH, à même de traiter d'éventuelles valeurs manquantes dans la supermatrice de distance $(\Delta_{ij}^{\text{SDM}})$.

• *Mesure de la précision topologique et représentation des résultats* — Les performances des différentes combinaisons moyennes ont été comparées en utilisant la distance topologique d_{quad} entre la topologie inférée $\hat{T}$ et l'arbre modèle T.

Pour chaque $k = 2, 4, 6, ..., 20$ et chacun des deux modèles PM et SSM, la moyenne des 500 valeurs d_{quad} obtenues a été calculée et reproduite sous forme de graphiques dans la Figure 5.3. Ces graphiques représentent les valeurs d_{quad} moyennes observées en fonction de k pour 25% et 75% de suppression de taxons.

Résultats

Comme attendu, toutes les courbes des graphiques de la Figure 5.3 sont décroissantes. Les arbres modèles T sont d'autant mieux retrouvés (*i.e.* leur distance d_{quad} avec $\hat{T}$ diminue) que le nombre k de matrices sources augmente. Comme attendu aussi, les arbres phylogénétiques $\hat{T}$ inférés sont d'autant plus proches des arbres modèles T que le taux de suppression des taxons est faible, vu que, dans ce cas, l'inter-recouvrement des taxons de chaque matrice est important.

L'utilisation du modèle SSM permet dans tous les cas d'obtenir un arbre $\hat{T}$ plus proches de l'arbre modèle T. Ainsi, l'utilisation du modèle SSM, plus riche en paramètres que PM, présente un réel intérêt lorsque l'on cherche à effectuer une inférence phylogénomique par combinaison moyenne de bonne qualité. Néanmoins, ces meilleures performances s'obtiennent par des temps de calcul un peu plus élevés, bien que raisonnables. Ainsi, même si une ou deux secondes ont

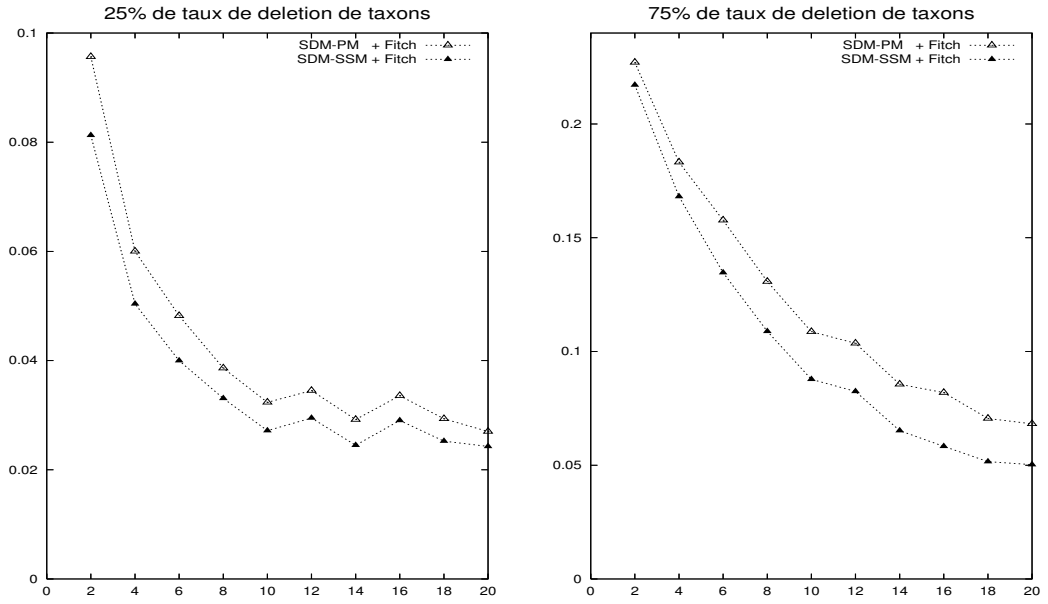

FIG. 5.3 – **Observation du recouvrement topologique des scénarios de combinaison moyenne utilisant SDM**

L'axe des abscisses représente les différentes valeurs de k. L'axe des ordonnées représente les distances topologiques d_{quad} moyennes observées pour les deux modèles PM et SSM.

suffit à SDM pour construire la supermatrice de distance $(\Delta_{ij}^{\text{SDM}})$ pour $k \leq 10$ pour chacun des deux modèles PM et SSM, il a nécessité un temps de presque 10 secondes pour $k = 20$ avec le modèle SSM, contre moins de deux secondes pour $k = 20$ avec le modèle PM.

En termes de temps d'exécution, le point faible du scénario d'inférence phylogénomique nommé SDM+FITCH est la méthode de distance FITCH. Pour $k \geq 10$, une durée de 23 secondes a été nécessaire à FITCH pour reconstruire un arbre phylogénétique à partir de $(\Delta_{ij}^{\text{SDM}})$, ce qui correspond à un temps d'exécution jusqu'à dix fois plus important que celui de SDM. La complexité de FITCH, de l'ordre de $O(n^4)$, représente donc un certain handicap pour l'étude phylogénomique par combinaison moyenne de collections de gènes définissant des groupes de plus d'une centaine d'espèces. Ce problème n'est pas posé lorsque la supermatrice de distance ne contient pas de distance manquante car on peut alors utiliser des algorithmes d'inférence d'arbres beaucoup plus rapides (*e.g.* $O(n^2)$ avec FASTME). Toutefois, lorsque $(\Delta_{ij}^{\text{SDM}})$ contient au moins une distance manquante, l'utilisation d'algorithmes avec des complexités plus élevées devient nécessaire (*e.g.* $O(n^4)$ avec FITCH, ou $O(n^5)$ avec MW*).

Pour remédier à ce dernier problème, le chapitre suivant décrit une nouvelle approche pour constuire rapidement un arbre à partir de la supermatrice de distance $(\Delta_{ij}^{\text{SDM}})$, qu'elle soit complète ou non. Cette approche s'appuie sur une adaptation des algorithmes agglomératifs

présentés dans la Figure 3.3 (cf page 65) afin de garantir une complexité de l'ordre de $O(n^3)$.

Le dernier protocole de simulation présenté dans ce présent chapitre (cf page 108) sera ensuite réutilisé pour observer les performances relatives de ces nouveaux algorithmes d'inférence d'arbre par rapport à celles des algorithmes existant (*e.g.* FITCH, MW*). Les supermatrices de distance inférés par SDM ne seront alors utilisées qu'avec le modèle SSM, et le signal topologique induit par les supermatrices de distance $(\Delta_{ij}^{\text{SDM}})$ sera également comparé avec celui induit par les supermatrices de distance construites par la méthode ACS.

Chapitre 6

Construction d'arbres à partir de (super)matrices de distance

> Il n'y a qu'une méthode pour inventer, qui est d'imiter. Il n'y a qu'une méthode pour bien penser, qui est de continuer quelque pensée ancienne et éprouvée.
>
> *Alain (Emile-Auguste Chartier, dit)*

Sommaire

6.1 Adaptation de la classe des algorithmes agglomératifs aux distances incomplètes 114

6.1.1 Critères d'agglomération 114

6.1.2 Valuation des branches externes 117

6.1.3 Réduction matricielle 118

6.2 Adaptation des algorithmes NJ, UNJ, BIONJ et MVR aux distances incomplètes 118

6.2.1 Calcul des paramètres w_i^* et λ_i^* 119

6.2.2 Considération de la variance associée aux distances incomplètes 121

6.2.3 Les critères d'agglomération : complexité algorithmique et performances 123

6.3 Les algorithmes NJ*, UNJ*, BIONJ* et MVR* dans le cadre de l'inférence phylogénomique 130

6.3.1 Description des algorithmes NJ*, UNJ*, BIONJ* et MVR* 130

6.3.2 Utilisation des supermatrices de distance inférées par SDM 131

Nombre d'algorithmes existent pour inférer un arbre phylogénétique à partir d'une matrice de distance. Toutefois, si au moins une seule valeur est absente, la grande majorité de ces algorithmes demeurent inutilisables.

L'apparition récente des supermatrices de distance dans le cadre de l'inférence phylogénomique a rendu indispensable l'utilisation de méthodes de distance pouvant traiter des

distances incomplètes (Lapointe and Cucumel, 1997; Criscuolo et al., 2006). Malheureusement, les rares algorithmes efficaces implémentés actuellement présentent des complexités élevées (*e.g.* de l'ordre de $O(n^4)$ ou plus) ou n'utilisent pas l'ensemble de l'information phylogénétique disponible (*e.g.* variances associées aux distances éolutives).

Ce chapitre introduit l'adaptation du schéma générique des algorithmes agglomératifs de la Figure 3.3 (cf page 65) au cas des distances incomplètes, et explicite la réécriture des équations caractérisant les différentes étapes du schéma agglomératif pour les algorithmes NJ, UNJ, BIONJ et MVR. Ces nouveaux algorithmes, nommés respectivement NJ*, UNJ*, BIONJ* et MVR* (Criscuolo, 2006), permettent d'inférer un arbre à partir d'une distance complète ou incomplète avec une complexité de l'ordre de $O(n^3)$. Plusieurs simulations démontrent que ces adaptations permettent, en association avec la méthode SDM, de développer des scénarios d'inférence phylogénomique de combinaison moyenne produisant des arbres topologiquement fiables, avec des temps d'exécution très rapides.

6.1 Adaptation de la classe des algorithmes agglomératifs aux distances incomplètes

Le schéma générique de la classe des algorithmes agglomératifs de la Figure 3.3 (cf page 65) décrit la construction itérative d'un arbre phylogénétique T à partir d'une matrice de distance complète (Δ_{ij}). Chaque itération se décompose en trois grandes étapes :

- recherche de la paire de taxons $x, y \in \mathcal{L}_r = \{1, 2, ..., r\}$ à agglomérer au nouveau noeud interne u,
- estimation de la longueur des deux branches externes T_{xu} et T_{yu},
- remplacement de x et y par u dans (Δ_{ij}), et calcul des nouvelles distances Δ_{ui}, pour tout $i \neq x, y$.

Cette partie décrit l'adaptation de ces trois étapes si la matrice de distance (Δ_{ij}) est incomplète.

6.1.1 Critères d'agglomération

Plusieurs critères d'agglomération ont été définis si (Δ_{ij}) est complète, parmi lesquels N_{xy}, N'_{xy} et Q_{xy} (cf pages 63-63). Cette partie explique comment adapter ces trois critères en les considérant comme des quantités moyennes de proximité topologique dans l'arbre T à inférer.

Les critères N^*_{xy} et $\tilde{N}^*_{xy}$

Le critère d'agglomération N_{xy}, défini par la Formule (3.13) :

$$N_{xy} = \sum_{\substack{i,j \in \mathcal{L}_r \\ i \neq j}} H\big(\Delta_{ix} + \Delta_{jy} - \Delta_{xy} - \Delta_{ij}\big) H\big(\Delta_{iy} + \Delta_{jx} - \Delta_{xy} - \Delta_{ij}\big),$$

où H est la fonction de Heaviside (cf page 63), représente le nombre de fois où la paire xy peut être considérée comme une cerise potentielle dans l'arbre T. Si on pose $C_{xy} = \big\{(i,j) \in \mathcal{L}_r \times \mathcal{L}_r : i \neq x,y,j\ ,\ j \neq x,y,i\big\}$, alors le critère $N_{xy}/|C_{xy}|$ représente une quantité moyenne de proximité topologique dans T. Si $N_{xy}/|C_{xy}| = 1$, alors la paire xy est toujours une cerise, pour toute topologie d'arbre définie sur les quadruplets de taxons $xyij$. De plus, maximiser $N_{xy}/|C_{xy}|$ tend à maximiser N_{xy}.

L'avantage du critère d'agglomération $N_{xy}/|C_{xy}|$ est que, exprimant une quantité moyenne, il ne nécessite pas l'ensemble des distances contenues dans (Δ_{ij}) pour être calculé. Conséquemment, si on pose

$$C^*_{xy} = \big\{(i,j) \in \mathcal{L}_r \times \mathcal{L}_r : i \neq x,y,j\ ,\ j \neq x,y,i\ ,\ \Delta_{ix}, \Delta_{jy}, \Delta_{iy}, \Delta_{jx}, \Delta_{xy}, \Delta_{ij} \neq \varnothing\big\}$$

et

$$N^*_{xy} = \sum_{i,j \in C^*_{xy}} H\big(\Delta_{ix} + \Delta_{jy} - \Delta_{xy} - \Delta_{ij}\big) H\big(\Delta_{iy} + \Delta_{jx} - \Delta_{xy} - \Delta_{ij}\big),$$

alors $N^*_{xy}/|C^*_{xy}| = N_{xy}/|C_{xy}|$ lorsque (Δ_{ij}) est une matrice de distance complète, mais seul le critère $N^*_{xy}/|C^*_{xy}|$ exprime une quantité moyenne de proximité topologique lorsque (Δ_{ij}) est incomplète.

On observe que le critère $N^*_{xy}/|C^*_{xy}|$ nécessite une importante quantité d'information pour être exprimé, *i.e.* les six distances $\Delta_{ix}, \Delta_{jy}, \Delta_{iy}, \Delta_{jx}, \Delta_{xy}, \Delta_{ij}$ doivent exister pour chaque quadruplet de taxons distincts $ijxy$. Ainsi, par exemple et pour une paire de taxons $\hat{i}\hat{j}$ donnée, si seule la distance $\Delta_{\hat{i}x}$ est inconnue, alors la paire $\hat{i}\hat{j}$ n'est pas utilisée pour le calcul de $N^*_{xy}/|C^*_{xy}|$. Or, dans ce cas, seule la valeur $H\big(\Delta_{\hat{i}x} + \Delta_{\hat{j}y} - \Delta_{xy} - \Delta_{\hat{i}\hat{j}}\big)$ ne peut être calculée, et on perd l'information topologique exprimée par la valeur $H\big(\Delta_{\hat{i}y} + \Delta_{\hat{j}x} - \Delta_{xy} - \Delta_{\hat{i}\hat{j}}\big)$. Afin de compenser cette perte d'information, on définit la fonction suivante :

$$\tilde{N}^*_{xy} = \left[\sum_{\substack{i,j \in \tilde{C}^*_{xy} \\ i<j}} H\big(\Delta_{ix} + \Delta_{jy} - \Delta_{xy} - \Delta_{ij}\big)\right] + \left[\sum_{\substack{i,j \in \tilde{C}^*_{yx} \\ i<j}} H\big(\Delta_{iy} + \Delta_{jx} - \Delta_{xy} - \Delta_{ij}\big)\right],$$

avec

$$\tilde{C}^*_{xy} = \big\{(i,j) \in \mathcal{L}_r \times \mathcal{L}_r : i \neq x,y,j\ ,\ j \neq x,y,i\ ,\ \Delta_{ix}, \Delta_{jy}, \Delta_{xy}, \Delta_{ij} \neq \varnothing\big\},$$

qui additionne les deux valeurs $H\big(\Delta_{ix} + \Delta_{jy} - \Delta_{xy} - \Delta_{ij}\big)$ et $H\big(\Delta_{iy} + \Delta_{jx} - \Delta_{xy} - \Delta_{ij}\big)$ au lieu de les multiplier. Cette fonction se réécrit plus simplement :

$$\tilde{N}^*_{xy} = \sum_{i,j \in \tilde{C}^*_{xy}} H\big(\Delta_{ix} + \Delta_{jy} - \Delta_{xy} - \Delta_{ij}\big)$$

et permet de définir le critère d'agglomération $\tilde{N}^*_{xy}/|\tilde{C}^*_{xy}|$. Si (Δ_{ij}) est une distance additive, alors $\tilde{N}^*_{xy} = N^*_{xy}$ si (Δ_{ij}) est complète, et $\tilde{N}^*_{xy} \geq N^*_{xy}$ si (Δ_{ij}) est incomplète. Ainsi

le critère d'agglomération $\tilde{N}^*_{xy}/|\tilde{C}^*_{xy}|$ exprime plus d'information topologique que $N^*_{xy}/|C^*_{xy}|$ lorsque (Δ_{ij}) est incomplète.

Les critères N'^*_{xy} et $\tilde{N}'^*_{xy}$

Le critère d'agglomération N'_{xy}, défini par la Formule (3.15) (cf page 63) :

$$N'_{xy} \;=\; \sum_{\substack{i,j\in\mathcal{L}_r\\ i\neq j}} \big(\Delta_{ix}+\Delta_{jy}+\Delta_{iy}+\Delta_{jx}-2\Delta_{xy}-2\Delta_{ij}\big)$$

est une version non-discrète du critère N_{xy}. Suivant le même raisonnement ayant conduit aux critères $N^*_{xy}/|C^*_{xy}|$ et $\tilde{N}^*_{xy}/|\tilde{C}^*_{xy}|$, on définit les fonctions

$$N'^*_{xy} \;=\; \sum_{i,j\in C^*_{xy}} \big(\Delta_{ix}+\Delta_{jy}+\Delta_{iy}+\Delta_{jx}-2\Delta_{xy}-2\Delta_{ij}\big)$$

et

$$\tilde{N}'^*_{xy} \;=\; \sum_{i,j\in \tilde{C}^*_{xy}} \big(\Delta_{ix}+\Delta_{jy}-\Delta_{xy}-\Delta_{ij}\big),$$

qui permettent d'obtenir les critères d'agglomération $N'^*_{xy}/|C^*_{xy}|$ et $\tilde{N}'^*_{xy}/|\tilde{C}^*_{xy}|$.

Les critères Q^*_{xy} et $\tilde{Q}^*_{xy}$

Le critère d'agglomération Q_{xy}, défini par la Formule (3.16) (cf page 64) et normalisé par $r-2$:

$$\frac{Q_{xy}}{r-2} \;=\; \frac{2}{r-2}\Delta_{xy} \;+\; \frac{1}{r-2}\sum_{i\in\mathcal{L}_r-\{x,y\}} \Big(\Delta_{xi}+\Delta_{yi}-\Delta_{xy}\Big),$$

exprime une quantité moyenne dans la somme normalisée par $r-2$. Si la paire de taxons xy est agglomérée au noeud u, alors cette moyenne $\overline{\Delta_{xi}+\Delta_{yi}-\Delta_{xy}} = 2\overline{T_{ui}}$ représente l'acentralité moyenne du chemin entre x et y par rapport aux autres taxons $i\neq x,y$. Comme ce calcul de moyenne renvoie une valeur que (Δ_{ij}) soit complète ou pas, si on pose

$$\hat{S}^*_{xy} = \Big\{ i\in\mathcal{L}_r-\{x,y\} : \Delta_{xi},\Delta_{yi}\neq\varnothing \Big\},$$

alors deux critères d'agglomération pour distances incomplètes sont définissables :

$$Q^*_{xy} \;=\; \frac{2}{|\hat{S}^*_{xy}|}\Delta_{xy} \;+\; \frac{1}{|\hat{S}^*_{xy}|}\sum_{i\in\hat{S}^*_{xy}} \Big(\Delta_{xi}+\Delta_{yi}-\Delta_{xy}\Big)$$

$$\tilde{Q}^*_{xy} \;=\; \frac{2}{r-2}\Delta_{xy} \;+\; \frac{1}{|\hat{S}^*_{xy}|}\sum_{i\in\hat{S}^*_{xy}} \Big(\Delta_{xi}+\Delta_{yi}-\Delta_{xy}\Big).$$

Le critère Q^*_{xy} considère que le facteur normalisateur $|S^*_{xy}|$ s'applique à l'ensemble de la formule, alors que $\tilde{Q}^*_{xy}$ considère qu'il ne s'applique qu'à l'estimateur d'acentralité de la paire xy. Si on définit les variables

$$S^*_{xy} = \big\{ i \in \mathcal{L}_r : \Delta_{xi}, \Delta_{yi} \neq \varnothing \big\} \qquad \text{et} \qquad R^*_{xy} = \sum_{i \in S^*_{xy}} \Big(\Delta_{xi} + \Delta_{yi} \Big),$$

alors on obtient les deux écritures suivantes après calculs (cf Annexe B) :

$$Q^*_{xy} \;=\; \frac{R^*_{xy}}{|S^*_{xy}| - 2} \;-\; \Delta_{xy}$$

$$\tilde{Q}^*_{xy} \;=\; \frac{R^*_{xy}}{|S^*_{xy}| - 2} \;-\; \left(\frac{r-4}{r-2} + \frac{2}{|S^*_{xy}| - 2} \right) \Delta_{xy},$$

qui présentent l'avantage de s'appuyer sur l'ensemble des taxons $i \in S^*_{xy}$ induisant toutes les paires de distance disponibles $\Delta_{xi}, \Delta_{yi} \neq \varnothing$.

On observe que $Q^*_{xy} = \tilde{Q}^*_{xy} = Q_{xy}/(r-2)$ lorsque (Δ_{ij}) est une matrice de distance complète. Or comme $r-2$ est une constante, maximiser $Q_{xy}/(r-2)$ revient à maximiser Q_{xy}.

6.1.2 Valuation des branches externes

Après agglomération de la paire de taxons xy au nouveau noeud interne u, la longueur T_{xu} de la branche externe créée est estimée par la classe de fonctions définie par la Formule (3.17) (cf page 65) :

$$T_{xu} \;=\; \frac{1}{2}\Delta_{xy} \quad + \sum_{i \in \mathcal{L}_r - \{x,y\}} w_i (\Delta_{xi} - \Delta_{yi}) \qquad \text{avec} \qquad \sum_{i \in \mathcal{L}_r - \{x,y\}} w_i = \frac{1}{2}.$$

La longueur de la branche externe correspondant au taxon y est naturellement obtenue par l'équation $T_{yu} = \Delta_{xy} - T_{xu}$.

Si $(\Delta_{ij}) = (T_{ij})$ est une distance additive et si la paire xy est une cerise de T, alors pour tout taxon $i \neq x, y$, la valeur $(\Delta_{xy} + \Delta_{xi} - \Delta_{yi})/2$ est égale à la longueur T_{xu}. La classe de fonctions définie par la Formule (3.17) pouvant se réécrire :

$$T_{xu} = \sum_{i \in \mathcal{L}_r - \{x,y\}} w_i (\Delta_{xy} + \Delta_{xi} - \Delta_{yi}),$$

elle représente une moyenne, pondérée par les w_i, des estimations de T_{xu} à partir de tous les taxons $i \neq x, y$. Cette moyenne pouvant être calculée même si certaines distances sont inconnues, on définit une nouvelle classe de fonctions :

$$T^*_{xu} \;=\; \Delta_{xy} - T^*_{yu} \;=\; \frac{1}{2}\Delta_{xy} + \sum_{i \in \hat{S}^*_{xy}} w^*_i (\Delta_{xi} - \Delta_{yi}) \quad \text{avec} \quad \sum_{i \in \hat{S}^*_{xy}} w^*_i = \frac{1}{2} \qquad (6.1)$$

permettant de calculer les longueurs T_{xu} et T_{yu} à partir d'une matrice de distance (Δ_{ij}) incomplète. Seules les pondérations w^*_i sont à redéfinir.

6.1.3 Réduction matricielle

Après l'agglomération de la paire de taxons xy au nouveau noeud interne u et l'estimation de la longueur des deux branches externes, les taxons x et y sont remplacés par u dans la matrice de distance (Δ_{ij}) et les nouvelles distances Δ_{ui} sont calculées à partir de la classe de fonctions définie par la Formule (3.18) (cf page 65) :

$$\Delta_{ui} = \lambda_i \Delta_{xi} + (1-\lambda_i)\Delta_{yi} - \lambda_i T_{xu} - (1-\lambda_i)T_{yu} \quad \text{avec} \quad \lambda_i \in [0,1].$$

Sachant que u appartient au chemin entre les taxons x et $i \neq x, y$, la valeur $\Delta_{xi} - T_{xu}$ est une estimation de la distance Δ_{ui}. Suivant le même raisonnement, la valeur $\Delta_{yi} - T_{yu}$ est une autre estimation de la distance Δ_{ui}. Ainsi la Formule (3.18) peut se réécrire :

$$\Delta_{ui} = \lambda_i\big(\Delta_{xi} - T_{xu}\big) + (1-\lambda_i)\big(\Delta_{yi} - T_{yu}\big).$$

Elle représente alors la moyenne pondérée par λ_i et $(1-\lambda_i)$ de ces deux estimations. Cette moyenne pouvant être calculée même si certaines distances sont inconnues, on définit une nouvelle classe de fonctions :

$$\Delta^*_{ui} = \begin{cases} \lambda^*_i\big(\Delta_{xi} - T^*_{xu}\big) + (1-\lambda^*_i)\big(\Delta_{yi} - T^*_{yu}\big) & \text{si} \quad \Delta_{xi} \neq \varnothing, \Delta_{yi} \neq \varnothing, \\ \Delta_{xi} - T^*_{xu} & \text{si} \quad \Delta_{xi} \neq \varnothing, \Delta_{yi} = \varnothing, \\ \Delta_{yi} - T^*_{yu} & \text{si} \quad \Delta_{xi} = \varnothing, \Delta_{yi} \neq \varnothing, \\ \varnothing & \text{si} \quad \Delta_{xi} = \Delta_{yi} = \varnothing, \end{cases} \tag{6.2}$$

permettant de calculer les distances Δ^*_{ui} en estimant le paramètre λ^*_i avec les valeurs disponibles dans (Δ_{ij}). Le paramètre λ^*_i ne nécessite pas d'être calculé si une seule des deux distances Δ_{xi} et Δ_{yi} est inconnue, car la distance Δ^*_{ui} est alors obtenue en effectuant la moyenne d'une unique valeur, c'est à dire cette valeur elle-même.

6.2 Adaptation des algorithmes NJ, UNJ, BIONJ et MVR aux distances incomplètes

Les algorithmes NJ, UNJ, BIONJ et MVR appartiennent à la classe des algorithmes agglomératifs applicables aux distances complètes et modélisée dans la Figure 3.3 (cf page 65). Leur fonctionnement itératif est similaire et ils ne diffèrent que par les informations qu'ils utilisent (*e.g.* distances, variances), ainsi que par le calcul des différents paramètres w_i et λ_i, à chaque itération.

La Figure 6.1 résume l'adaptation de la classe des algorithmes agglomératifs aux distances incomplètes. Cette partie explicite l'adaptation des quatre algorithmes NJ, UNJ, BIONJ et MVR. Elle décrit le calcul des nouveaux paramètres w^*_i et λ^*_i, puis discute du choix d'un critère d'agglomération parmi les critères N^*_{xy}, $\tilde{N}^*_{xy}$, N'^*_{xy}, $\tilde{N}'^*_{xy}$, Q'^*_{xy} et $\tilde{Q}'^*_{xy}$.

- $r = n$;
- `Tant que` $r > 2$

  `Faire` • Sélectionner la paire $x, y \in \mathcal{L}_r$ ($\Delta_{xy} \neq \varnothing$) en optimisant un critère d'agglomération ;

  - Agglomérer x et y au nouveau noeud u ;
  - Estimer les longueurs T_{xu} et T_{yu} :

$$T^*_{xu} \;=\; \Delta_{xy} - T^*_{yu} \;=\; \frac{1}{2}\Delta_{xy} + \sum_{i \in \hat{S}^*_{xy}} w^*_i (\Delta_{xi} - \Delta_{yi}) \tag{6.1}$$

  avec $\sum_{i \in \hat{S}^*_{xy}} w^*_i = 1/2$;

  - Réduire la matrice de distance (Δ_{ij}) pour tout $i \neq x, y$:

$$\Delta^*_{ui} = \begin{cases} \lambda^*_i \big(\Delta_{xi} - T^*_{xu}\big) + (1 - \lambda^*_i)\big(\Delta_{yi} - T^*_{yu}\big) & \text{si} \quad \Delta_{xi} \neq \varnothing, \Delta_{yi} \neq \varnothing, \\ \Delta_{xi} - T^*_{xu} & \text{si} \quad \Delta_{xi} \neq \varnothing, \Delta_{yi} = \varnothing, \\ \Delta_{yi} - T^*_{yu} & \text{si} \quad \Delta_{xi} = \varnothing, \Delta_{yi} \neq \varnothing, \\ \varnothing & \text{si} \quad \Delta_{xi} = \Delta_{yi} = \varnothing, \end{cases} \tag{6.2}$$

  avec $\lambda^*_i \in [0, 1]$;

  - Réduire la matrice de variance (V_{ij}) associée à (Δ_{ij}) si celle-ci est utilisée ;
  - $r = r - 1$;

- Renvoyer T ;

FIG. 6.1 – **Schéma générique des algorithmes agglomératifs adaptés aux distances incomplètes**

Les équations (6.1) et (6.2) représentent, respectivement, la classe des formules d'estimation des longueurs de branche et la classe des formules de réduction.

6.2.1 Calcul des paramètres w^*_i et λ^*_i

Les paramètres de NJ

L'algorithme NJ est défini par les paramètres

$$w_i = w = \frac{1}{2(r-2)} \tag{6.3}$$

et

$$\lambda_i = 1/2. \tag{6.4}$$

Il correspond donc à l'algorithme effectuant des moyennes non-pondérées dans les équations (3.17) et (3.18). Son adaptation aux distances incomplètes, nommée NJ*, se définit donc par les

paramètres :

$$w_i^* = w^* = \frac{1}{2\big(|S_{xy}^*| - 2\big)} = \frac{1}{2|\hat{S}_{xy}^*|} \quad \text{et} \quad \lambda_i^* = \lambda^* = \frac{1}{2},$$

car ils impliquent des moyennes non pondérées dans les Formules (6.1) et (6.2). On vérifie aisément que

$$\sum_{i \in \hat{S}_{xy}^*} w_i^* \;=\; \sum_{i \in \hat{S}_{xy}^*} \frac{1}{2|\hat{S}_{xy}^*|} \;=\; \frac{|\hat{S}_{xy}^*|}{2|\hat{S}_{xy}^*|} \;=\; \frac{1}{2}.$$

Les paramètres de UNJ

L'algorithme UNJ est défini par les paramètres

$$w_i = \frac{n_i}{2 \sum\limits_{j \in \mathcal{L}_r - \{x,y\}} n_j} = \frac{n_i}{2(n - n_x - n_y)} \quad \text{et} \quad \lambda_i = \lambda = \frac{n_x}{n_x + n_y},$$

où, au départ (*i.e.* $r = n$), $n_j = 1$, pour tout $j = 1, ..., n$ et, à chaque étape, $n_u = n_x + n_y$ lors de l'agglomération de la paire xy au nouveau noeud u.

La moyenne dans la Formule (6.1) ne s'effectuant qu'avec les taxons $i \in \hat{S}_{xy}^*$, les nouveaux paramètres w_i^* et λ_i^* se calculent en n'utilisant également que ces mêmes taxons :

$$w_i^* = \frac{n_i}{2 \sum\limits_{j \in \hat{S}_{xy}^*} n_j} \tag{6.5}$$

et

$$\lambda_i^* = \lambda^* = \frac{n_x}{n_x + n_y}. \tag{6.6}$$

On vérifie aisément que

$$\sum_{i \in \hat{S}_{xy}^*} w_i^* \;=\; \sum_{i \in \hat{S}_{xy}^*} \frac{n_i}{2 \sum\limits_{j \in \hat{S}_{xy}^*} n_j} \;=\; \frac{\sum\limits_{i \in \hat{S}_{xy}^*} n_i}{2 \sum\limits_{j \in \hat{S}_{xy}^*} n_j} \;=\; \frac{1}{2}.$$

Sachant que $n_x, n_y \geq 1$, on vérifie également que $\lambda^* \in [0, 1]$.

Les paramètres de BIONJ

L'algorithme BIONJ est défini par les paramètres $w_i = w = 1/(2(r-2))$ et

$$\lambda_i = \lambda = \frac{1}{2} + \frac{1}{2(r-2)V_{xy}} \sum_{j \in \mathcal{L}_r - \{x,y\}} (V_{yj} - V_{xj}).$$

Le paramètre w^* est donc similaire à celui de NJ*. Pour le calcul de λ^*, on observe que la somme dans la Formule (3.19) correspond à la moyenne des différences entre les variances V_{yj} et V_{xj},

pour tout $j \in \mathcal{L}_r - \{x, y\}$, *i.e.*

$$\frac{1}{r-2} \sum_{j \in \mathcal{L}_r - \{x,y\}} (V_{yj} - V_{xj}).$$

Certaines des distances associées Δ_{yj} et Δ_{xj} étant inconnues, le calcul de λ^* s'appuie sur les taxons $j \in \hat{S}^*_{xy}$, *i.e.*

$$\lambda^*_i = \lambda^* = \frac{1}{2} + \frac{1}{2|\hat{S}^*_{xy}|V_{xy}} \sum_{j \in \hat{S}^*_{xy}} (V_{yj} - V_{xj}). \tag{6.7}$$

Les paramètres de MVR

Etant donné une matrice de variance (V_{ij}) associée à la distance (Δ_{ij}), l'algorithme MVR est défini par les paramètres

$$w_i = \frac{\mu}{V_{xi} + V_{yi}} \quad \text{où} \quad \mu = \frac{1}{2} \left(\sum_{j \in \mathcal{L}_r - \{x,y\}} \frac{1}{V_{xj} + V_{yj}} \right)^{-1}$$

et

$$\lambda_i \; = \; \frac{V_{yi}}{V_{xi} \; + \; V_{yi}}. \tag{6.8}$$

Comme les algorithmes agglomératifs 6.1 ne considèrent, à chaque itération, que les taxons $i \in \hat{S}^*_{xy}$ pour estimer les longueurs de branche, on définit :

$$w^*_i = \frac{\mu^*}{V_{xi} + V_{yi}} \quad \text{où} \quad \mu^* = \frac{1}{2} \left(\sum_{j \in \hat{S}^*_{xy}} \frac{1}{V_{xj} + V_{yj}} \right)^{-1}. \tag{6.9}$$

Les paramètres λ^*_i sont les mêmes que λ_i. Le paramètre w^*_i n'étant employé dans la Formule (6.1) que lorsque $\Delta_{xi}, \Delta_{yi} \neq \varnothing$, il est donc calculable car les variances associées, V_{xi} et V_{yi}, existent dans ce cas de figure. Conséquemment, les paramètres λ^*_i sont également calculables. On vérifie aisément que $\lambda^* \in [0, 1]$ et que

$$\sum_{i \in \hat{S}^*_{xy}} w^*_i \; = \; \mu^* \sum_{i \in \hat{S}^*_{xy}} \big(V_{xi} + V_{yi}\big)^{-1} \; = \; \frac{\sum\limits_{i \in \hat{S}^*_{xy}} \big(V_{xi} + V_{yi}\big)^{-1}}{2 \sum\limits_{j \in \hat{S}^*_{xy}} \big(V_{xj} + V_{yj}\big)^{-1}} \; = \; \frac{1}{2}.$$

6.2.2 Considération de la variance associée aux distances incomplètes

Les algorithmes BIONJ et MVR utilisent une matrice de variance (V_{ij}) associée à (Δ_{ij}), et applique une opération de réduction à chaque itération. Néanmoins, comme BIONJ et MVR ne considèrent pas les mêmes modèles, les équations de réduction de variance sont très différentes.

Le modèle de variance-covariance de BIONJ

L'algorithme BIONJ s'appuie sur l'approximation linaire $V_{ij} = \Delta_{ij}/\ell$. Il initialise ainsi la matrice de variance suivant cette approximation avant d'exécuter le schéma agglomératif. Si (Δ_{ij}) est une matrice de distance incomplète, alors on pose initialement

$$\begin{cases} V^*_{ij} = \Delta_{ij}/\ell & \text{si} \quad \Delta_{ij} \neq \varnothing, \\ V^*_{ij} = +\infty & \text{si} \quad \Delta_{ij} = \varnothing. \end{cases}$$

En effet, la variance associée à une distance Δ_{ij} étant d'autant plus grande que Δ_{ij} est imprécise, une distance inconnue (*i.e.* qui ne peut être estimée) correspond à une variance infinie.

A chaque itération de BIONJ, la matrice de variance (V_{ij}) est réduite suivant l'équation (cf page 65) :

$$V_{ui} = \lambda V_{xi} + (1-\lambda)V_{yi} - \lambda(1-\lambda)V_{xy}.$$

Dans le cadre des matrices de distance incomplètes, comme la Formule (6.2) implique que

$$\begin{cases} \lambda^*_i = 1 & \text{si} \quad \Delta_{xi} \neq \varnothing \quad \text{et} \quad \Delta_{yi} = \varnothing, \\ \lambda^*_i = 0 & \text{si} \quad \Delta_{xi} = \varnothing \quad \text{et} \quad \Delta_{yi} \neq \varnothing, \end{cases}$$

alors l'équation décrivant la réduction de la matrice de variance (V_{ij}) se réécrit :

$$V^*_{ui} = \begin{cases} \lambda^* V_{xi} + (1-\lambda^*)V_{yi} - \lambda^*(1-\lambda^*)V_{xy} & \text{si} \quad \Delta_{xi} \neq \varnothing \quad \text{et} \quad \Delta_{yi} \neq \varnothing, \\ V_{xi} & \text{si} \quad \Delta_{xi} \neq \varnothing \quad \text{et} \quad \Delta_{yi} = \varnothing, \\ V_{yi} & \text{si} \quad \Delta_{xi} = \varnothing \quad \text{et} \quad \Delta_{yi} \neq \varnothing, \\ +\infty & \text{si} \quad \Delta_{xi} = \Delta_{yi} = \varnothing. \end{cases}$$

Si, durant une itération, une agglomération est incorrecte (*i.e.* ne correspond pas à l'arbre vrai), on n'observe alors plus l'inégalité triangulaire de la matrice de variance (V_{ij}). Conséquemment, on n'observe pas systématiquement la condition

$$-\sum_{j \in \hat{S}^*_{xy}} V_{xy} \;\leq\; \sum_{j \in \hat{S}^*_{xy}} (V_{yj} - V_{xj}) \;\leq\; \sum_{j \in \hat{S}^*_{xy}} V_{xy}$$

garantissant que $\lambda^* \in [0,1]$. Dans ce cas de figure, on pose arbitrairement :

$$\begin{cases} \lambda^* = 0 & \text{si} \quad \lambda^* < 0 \\ \lambda^* = 1 & \text{si} \quad \lambda^* > 1. \end{cases}$$

Le modèle de variance de MVR

L'algorithme MVR ne définit pas de relation linéaire entre V_{ij} et Δ_{ij}, mais considère que les covariances sont toujours nulles à chaque itération. Cette approche permet ainsi d'associer toutes les valeurs possibles à chaque V_{ij} lors de l'initialisation de la matrice de variance. Par contre, si (Δ_{ij}) est une matrice de distance incomplète, alors $V^*_{ij} = +\infty$ si $\Delta_{ij} = \varnothing$.

L'algorithme MVR effectue la réduction matricielle de la variance (V_{ij}) en utilisant l'équation suivante (cf page 66) :

$$V_{ui} = \frac{V_{xi}V_{yi}}{V_{xi} + V_{yi}} = \left(V_{xi}^{-1} + V_{yi}^{-1}\right)^{-1}.$$

Ainsi, la réduction matricielle de la variance pour l'algorithme MVR* se calcule avec la formule :

$$V_{ui}^{*} = \begin{cases} \left(V_{xi}^{-1} + V_{yi}^{-1}\right)^{-1} & \text{si} \quad \Delta_{xi} \neq \varnothing \quad \text{et} \quad \Delta_{yi} \neq \varnothing, \\ V_{xi} & \text{si} \quad \Delta_{xi} \neq \varnothing \quad \text{et} \quad \Delta_{yi} = \varnothing, \\ V_{yi} & \text{si} \quad \Delta_{xi} = \varnothing \quad \text{et} \quad \Delta_{yi} \neq \varnothing, \\ +\infty & \text{si} \quad \Delta_{xi} = \Delta_{yi} = \varnothing. \end{cases}$$

6.2.3 Les critères d'agglomération : complexité algorithmique et performances

Six critères d'agglomération ($N_{xy}^{*}/|C_{xy}^{*}|$, $\tilde{N}_{xy}^{*}/|\tilde{C}_{xy}^{*}|$, $N_{xy}'^{*}/|C_{xy}^{*}|$, $\tilde{N}_{xy}'^{*}/|\tilde{C}_{xy}^{*}|$, Q_{xy}^{*} et $\tilde{Q}_{xy}^{*}$) ont été définis afin d'utiliser la classe des algorithmes agglomératifs (cf Figure 6.1) pour les matrices de distance incomplètes.

Cette partie explicite les techniques permettant d'optimiser la complexité algorithmique de leur maximisation. Ces techniques sont très largement inspirées de la réduction de la complexité d'un facteur n de l'algorithme ADDTREE proposée par Elemento et Gascuel (2002).

Une discussion, basée sur des simulations, est ensuite menée pour définir le(s) critère(s) induisant la meilleure quantité moyenne de proximité topologique.

Complexité algorithmique

Les critères $N_{xy}^{*}/|C_{xy}^{*}|$, $N_{xy}'^{*}/|C_{xy}^{*}|$, $\tilde{N}_{xy}^{*}/|\tilde{C}_{xy}^{*}|$ *et* $\tilde{N}_{xy}'^{*}/|\tilde{C}_{xy}^{*}|$

Pour chaque paire de taxons xy, les deux critères $N_{xy}^{*}/|C_{xy}^{*}|$ et $N_{xy}'^{*}/|C_{xy}^{*}|$ se calculent en utilisant toutes les autres paires de taxons $i, j \in C_{xy}^{*}$. Ils impliquent donc une complexité de l'ordre de $O(r^2)$ pour une paire xy fixée. Le même raisonnement s'applique pour les deux critères $\tilde{N}_{xy}^{*}/|\tilde{C}_{xy}^{*}|$ et $\tilde{N}_{xy}'^{*}/|\tilde{C}_{xy}^{*}|$. Ainsi, la maximisation de ces quatre critères s'effectue en $O(r^4)$. Comme les algorithmes agglomératifs s'effectuent en $O(n)$ itérations et comme on a $1 \leq r \leq n$, l'utilisation de ces quatre critères implique une complexité algorithmique de l'ordre de $O(n^5)$.

Si on considère le critère $N_{xy}'^{*}/|C_{xy}^{*}|$ avec

$$\tilde{C}_{xy}^{*} = \left\{(i,j) \in \mathcal{L}_r \times \mathcal{L}_r : i \neq x,y,j \ , \ j \neq x,y,i \ , \ \Delta_{ix}, \Delta_{jy}, \Delta_{xy}, \Delta_{ij} \neq \varnothing\right\}$$

et

$$\tilde{N}_{xy}^{*} = \sum_{i,j \in \tilde{C}_{xy}^{*}} H\left(\Delta_{ix} + \Delta_{jy} - \Delta_{xy} - \Delta_{ij}\right),$$

il est possible de réduire cette complexité. Si, avant l'application de l'algorithme (*i.e.* $r = n$), on précalcule les ensembles $\tilde{C}^*_{ij}$, ainsi que les valeurs $\tilde{N}^*_{ij}$, pour tout $i, j \in \mathcal{L}_n$, alors une mise à jour des $O(n^2)$ informations est possible à chaque itération de l'algorithme. Lorsque les taxons x et y sont agglomérés au noeud u, il suffit de :

1. retrancher de chaque $\tilde{N}^*_{ij}$ $(i, j \neq x, y)$ les valeurs $H\big(\Delta_{ix} + \Delta_{jy} - \Delta_{xy} - \Delta_{ij}\big)$ et $H\big(\Delta_{iy} + \Delta_{jx} - \Delta_{xy} - \Delta_{ij}\big)$ (lorsqu'elles y existent), et de les supprimer dans l'ensemble $\tilde{C}^*_{ij}$,
2. calculer $\tilde{N}^*_{ui}$ et $\tilde{C}^*_{ui}$, pour tout $i \neq x, y$,
3. compléter $\tilde{N}^*_{ij}$ et $\tilde{C}^*_{ij}$ avec les valeurs induites par les nouvelles distances Δ^*_{ui} et Δ^*_{uj}, pour tout $i, j \neq x, y$.

L'étape 1 s'effectue en $O(r^2)$. L'étape 2 s'effectue en $O(r^2)$ pour chaque $i \neq x, y$, et donc en $O(r^3)$ pour l'ensemble des taxons $i \neq x, y$. L'étape 3 s'effectue en $O(r^2)$. Les précalculs s'effectuant en $O(n^4)$, l'application des trois étapes de mise à jour des ensembles $\tilde{C}^*_{ij}$ et des valeurs $\tilde{N}^*_{ij}$ à chacune des $O(n)$ itérations impliquent une complexité algorithmique globale de l'ordre de $O(n^4)$.

Un raisonnement similaire s'applique aux trois autres critères $N^*_{xy}/|C^*_{xy}|$, $N'^*_{xy}/|\tilde{C}^*_{xy}|$ et $\tilde{N}'^*_{xy}/|\tilde{C}^*_{xy}|$.

Les critères Q^*_{xy} et $\tilde{Q}^*_{xy}$

Si on considère les deux critères

$$Q^*_{xy} = \frac{R^*_{xy}}{|S^*_{xy}| - 2} - \Delta_{xy} \qquad \text{et} \qquad \tilde{Q}^*_{xy} = \frac{R^*_{xy}}{|S^*_{xy}| - 2} - \left(\frac{r-4}{r-2} + \frac{2}{|S^*_{xy}| - 2}\right)\Delta_{xy},$$

avec

$$S^*_{xy} = \big\{i \in \mathcal{L}_r : \Delta_{xi}, \Delta_{yi} \neq \varnothing\big\} \qquad \text{et} \qquad R^*_{xy} = \sum_{i \in S^*_{xy}} \Big(\Delta_{xi} + \Delta_{yi}\Big),$$

alors leur maximisation s'effectue en $O(r^3)$ à chaque itération de l'algorithme, ce qui implique une complexité totale de l'ordre de $O(n^4)$.

Néanmoins, une mise à jour des ensembles S^*_{ij} et des valeurs R^*_{ij} à chaque itération de l'algorithme est également possible. Lorsque x et y ont été agglomérés au noeud u, il suffit de :

1. retrancher de chaque R^*_{ij} $(i, j \neq x, y)$ les valeurs $\Delta_{xi} + \Delta_{yi}$ (lorsqu'elles y existent), et de supprimer les taxons x et y dans l'ensemble S^*_{ij},
2. calculer R^*_{ui} et S^*_{ui}, pour tout $i \neq x, y$,
3. compléter R^*_{ij} et S^*_{ij} avec les valeurs induites par les nouvelles distances Δ^*_{ui}, pour tout $i, j \neq x, y$.

Les précalculs de S^*_{ij} et R^*_{ij}, pour tout $i, j \in \mathcal{L}_n$ s'effectuant en $O(n^3)$ et les trois étapes s'effectuant en $O(r^2)$ à chacune des $O(n)$ itérations, la complexité totale de l'algorithme est donc de l'ordre de $O(n^3)$.

On remarque que de telles opérations ne sont plus nécessaires lorsque (Δ_{xi}) ne contient pas (ou plus) de distance manquante. Dans ce cas, sachant que la maximisation de Q^*_{xy} et $\tilde{Q}^*_{xy}$ est équivalente à la maximisation de $Q_{xy} = R_x + R_y - (r-2)\Delta_{xy}$ telle que définie par la Formule (3.14), il suffit, à chaque itération de l'algorithme, de précalculer chaque valeur R_z en $O(r)$, pour tout $z \in \mathcal{L}_r$, pour obtenir le même résultat. Les précalculs des valeurs R_z et la maximisation de Q_{xy} impliquent également une complexité algorithmique totale de l'ordre de $O(n^3)$, mais le nombre d'opérations étant moins élevé que celui impliqué par une mise à jour des valeurs S^*_{ij} et R^*_{ij} à chaque itération, la complexité en moyenne est moins importante.

Quantité de proximité topologique

Comme les fonctions $\tilde{N}^*_{xy}$ et $\tilde{N}'^*_{xy}$ ont été montrées moins contraintes en termes de distances manquantes que les fonctions N^*_{xy} et N'^*_{xy}, ce sont les critères $\tilde{N}^*_{xy}/|\tilde{C}^*_{xy}|$ et $\tilde{N}'^*_{xy}/|\tilde{C}^*_{xy}|$ qui doivent être choisis parmi ces quatre. En effet, si, dans l'exemple de la Figure 6.2 (représentant un arbre phylogénétique T et sa matrice de distance additive (T_{ij}) rendue incomplète par la suppression des trois distances T_{bd}, T_{ce} et T_{df}), on considère que les distances T_{ac} et T_{af} sont également manquantes, alors les fonctions N^*_{xy} et N'^*_{xy} sont incalculables pour les paires de taxons ab, cd et ef. L'utilisation des critères $N^*_{xy}/|C^*_{xy}|$ et $N'^*_{xy}/|C^*_{xy}|$ conduit donc dans cet exemple à une agglomération de feuilles incorrecte, ce qui n'est pas le cas des deux autres critères $\tilde{N}^*_{xy}/|\tilde{C}^*_{xy}|$ et $\tilde{N}'^*_{xy}/|\tilde{C}^*_{xy}|$.

Les parties suivantes illustrent les performances des critères d'agglomération Q^*_{xy}, $\tilde{Q}^*_{xy}$, $\tilde{N}^*_{xy}/|\tilde{C}^*_{xy}|$ et $\tilde{N}'^*_{xy}/|\tilde{C}^*_{xy}|$ en termes de distribution de leurs valeurs, et du taux de recouvrement des cerises de la topologie d'un arbre modèle. Différents protocoles de simulation permettent de montrer qu'une combinaison de plusieurs de ces critères offrent de très bonnes performances tout en conservant une complexité de l'ordre de $O(n^3)$.

Distribution des valeurs de critère

La Figure 6.2 représente une distribution des valeurs des quatre critères d'agglomération Q^*_{xy}, $\tilde{Q}^*_{xy}$, $\tilde{N}^*_{xy}/|\tilde{C}^*_{xy}|$ et $\tilde{N}'^*_{xy}/|\tilde{C}^*_{xy}|$. A partir d'une matrice additive contenant trois distances manquantes, ces quatre critères ont été calculés pour chaque paire de taxons xy telle que $T_{xy} \neq \varnothing$, puis représentés graphiquement dans l'ordre croissant.

On remarque que les critères Q^*_{xy} et $\tilde{Q}^*_{xy}$ sont incapables de retrouver une des trois cerises de l'arbre modèle T (*i.e.* ab, cd et ef), alors que les trois plus grandes valeurs des critères $\tilde{N}^*_{xy}/|\tilde{C}^*_{xy}|$ et $\tilde{N}'^*_{xy}/|\tilde{C}^*_{xy}|$ correspondent aux trois cerises de T. Les meilleures performances de $\tilde{N}^*_{xy}/|\tilde{C}^*_{xy}|$ et de $\tilde{N}'^*_{xy}/|\tilde{C}^*_{xy}|$ s'expliquent par le fait que ces critères expriment une quantité

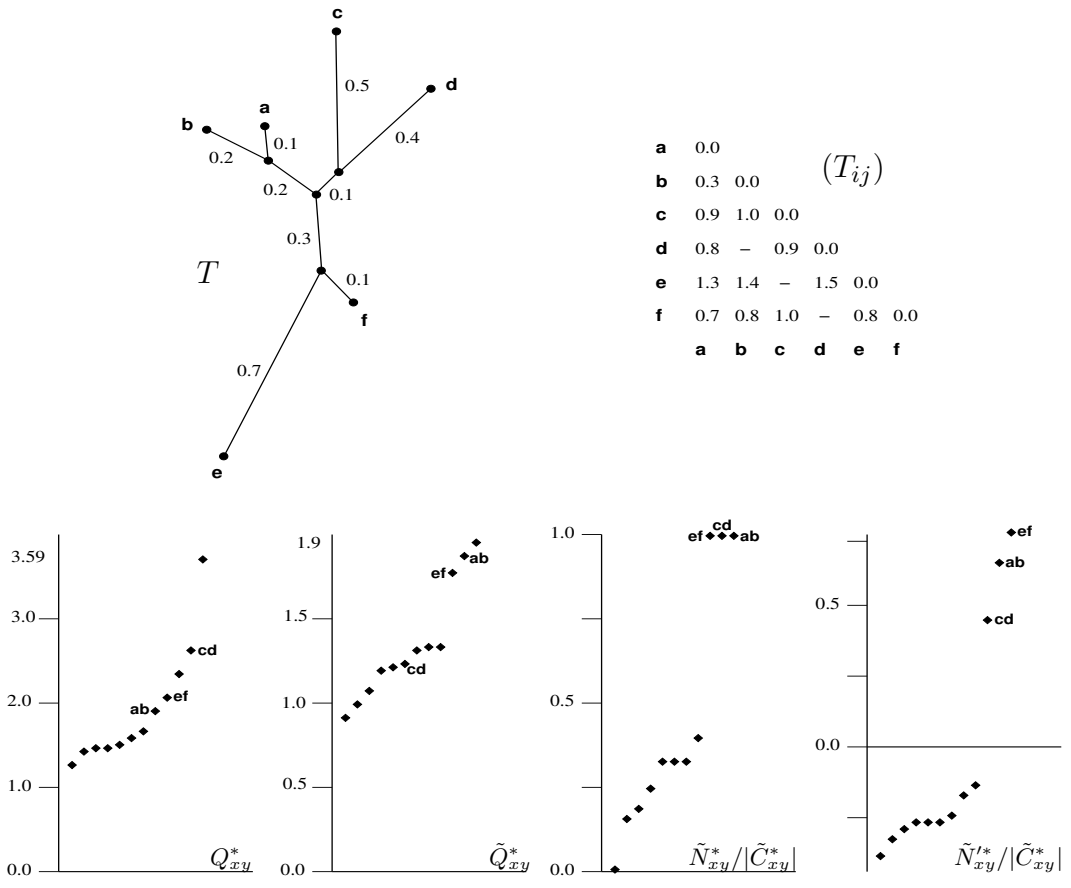

FIG. 6.2 – **Distribution des valeurs des quatre critères d'agglomération Q^*_{xy}, $\tilde{Q}^*_{xy}$, $\tilde{N}^*_{xy}/|\tilde{C}^*_{xy}|$ et $\tilde{N}'^*_{xy}/|\tilde{C}^*_{xy}|$**

Les trois distances T_{bd}, T_{ce} et T_{df} ont été supprimées dans la matrice additive (T_{ij}) (en haut à droite) calculée à partir de l'arbre phylogénétique T (en haut à gauche). Les quatre graphiques (en bas) représentent la distribution des valeurs des critères Q^*_{xy}, $\tilde{Q}^*_{xy}$, $\tilde{N}^*_{xy}/|\tilde{C}^*_{xy}|$ et $\tilde{N}'^*_{xy}/|\tilde{C}^*_{xy}|$ appliqués sur toutes les paires de taxons xy. Pour chaque distribution, les valeurs correspondant aux trois cerises ab, cd et ef de T ont été signalées.

moyenne de proximité topologique basée sur des quadruplets de taxons, alors que Q^*_{xy} et $\tilde{Q}^*_{xy}$ se basent sur des triplets. Les quadruplets de taxons exprimant plus d'informations topologiques que les triplets, les critères d'agglomération s'appuyant sur ces derniers expriment une quantité de proximité topologique moins informative.

Les deux critères Q^*_{xy} et $\tilde{Q}^*_{xy}$ s'appuyant sur les triplets de taxons, ils sont donc tous deux incapables d'effectuer systématiquement la sélection d'une des paires correctes de taxons à agglomérer. Néanmois, le fait qu'ils ne s'appuient que sur les triplets de taxons implique que ces deux critères sont plus rapides que $\tilde{N}^*_{xy}/|\tilde{C}^*_{xy}|$ et $\tilde{N}'^*_{xy}/|\tilde{C}^*_{xy}|$. De plus, l'observation des distributions de valeurs montre que les deux critères Q^*_{xy} et $\tilde{Q}^*_{xy}$ détectent les cerises de T (*e.g.* le premier quart des valeurs maximales contient toujours au moins une valeur associée à

k	Q^*_{xy}	$\tilde{Q}^*_{xy}$
2	**0.2395**	0.2867
4	**0.2009**	0.2775
6	**0.1766**	0.2487
8	**0.1524**	0.2193
10	**0.1314**	0.1859
12	**0.1228**	0.1680
14	**0.1024**	0.1448
16	**0.0941**	0.1240
18	**0.0798**	0.1044
20	**0.0760**	0.0929

TAB. 6.1 – **Performances de NJ* avec les deux critères d'agglomération Q^*_{xy} et $\tilde{Q}^*_{xy}$**
La colonne k indique le nombre de matrices de distance (obtenues à partir des gènes délétés avec 75% de délétion) utilisées pour construire chacune des 500 supermatrices de distance avec la méthode SDM. Les colonnes Q^*_{xy} et $\tilde{Q}^*_{xy}$ représentent la moyenne des 500 valeurs d_{quad} observées pour chaque valeur de k. Les valeurs écrites en caractères gras sont les plus petites pour chaque ligne.

une cerise de T), mais admettent une certaine marge d'erreur. Une comparaison sur une plus vaste échelle est donc menée pour mesurer l'écart entre les performances respectives de ces deux critères.

Comparaison des critères Q^*_{xy} et $\tilde{Q}^*_{xy}$

Afin d'observer les performances des critères Q^*_{xy} et $\tilde{Q}^*_{xy}$ en terme de recouvrement topologique d'un arbre modèle, les supermatrices de distance générées par la méthode SDM lors des simulations décrites dans la partie 5.3.2 (cf page 108) ont été ré-analysées avec l'algorithme agglomératif de la Figure 6.1 en utilisant les paramètres NJ*. Seules les supermatrices de distance $(\Delta^{\text{SDM}}_{ij})$ générées dans le cadre des 75% de délétion de taxons ont été ré-analysées, car elles contiennent un nombre important de distances manquantes (*i.e.* entre 11% et 42%), contrairement aux supermatrices de distance générées dans le cadre des 25% de délétion de taxons (*i.e.* entre 0% et 8%). Chaque arbre inféré $\hat{T}$ a été comparé à l'arbre modèle T en utilisant la distance topologique d_{quad} mesurant le nombre moyen de quadruplets présents dans un arbre mais pas dans l'autre. Pour chaque valeur de $k = 2, 4, ..., 20$, la distance d_{quad} moyenne est représentée dans le Tableau 6.1.

On observe que pour toutes les valeurs de k, l'utilisation du critère Q^*_{xy} permet d'inférer des arbres phylogénétiques plus proches de l'arbre modèle que lorsqu'on utilise le critère $\tilde{Q}^*_{xy}$. Si on définit l'ensemble

$$\varnothing_i = \big\{ j \neq i : \Delta_{ij} = \varnothing \big\}, \tag{6.10}$$

il a été remarqué, durant ces simulations, que le critère $\tilde{Q}^*_{xy}$ a une tendance à privilégier les paires de taxons xy induisant peu de valeurs manquantes, *i.e.* correspondant à une faible valeur de $|\varnothing_x \cup \varnothing_y|$. Inversement, la maximisation du critère Q^*_{xy} a une tendance à minimiser la valeur $|S^*_{xy}| = r - |\varnothing_x \cup \varnothing_y|$, Conséquemment, le critère Q^*_{xy} tendant à sélectionner les paires xy induisant beaucoup de distances manquantes (*i.e.* maximisant $|\varnothing_x \cup \varnothing_y|$), son utilisation permet de faire disparaitre un plus grand nombre de valeurs manquantes à chaque itération. La matrice (Δ_{ij}) contenant donc moins de valeurs manquantes à chaque itération, un plus grand nombre de paires de taxons sont testées par le critère d'agglomération, ce qui augmente la probabilité de sélectionner celles correspondant à une cerise.

Combinaison de différents critères d'agglomération

Comme observé précédemment, le critère Q^*_{xy}, induisant une complexité de l'ordre de $O(n^3)$, exprime une quantité de proximité topologique moins importante que les critères $\tilde{N}^*_{xy}/|\tilde{C}^*_{xy}|$ et $\tilde{N}'^*_{xy}/|\tilde{C}^*_{xy}|$. Malheureusement, ces deux derniers critères impliquent une complexité plus élevée de l'ordre de $O(n^4)$.

Afin d'obtenir un compromis efficace entre rapidité et efficacité, le critère Q^*_{xy} peut être utilisé pour obtenir une première classification de l'ensemble des paires de taxons xy. Ces différentes paires de taxons étant classées par ordre décroissant de leur valeur de critère Q^*_{xy}, les critères $\tilde{N}^*_{xy}/|\tilde{C}^*_{xy}|$ et $\tilde{N}'^*_{xy}/|\tilde{C}^*_{xy}|$ sont ensuite appliqués sur les s premières paires afin de sélectionner les deux taxons x et y à agglomérer.

Le critère $\tilde{N}^*_{xy}/|\tilde{C}^*_{xy}|$ est privilégié dans cette seconde étape car il est basé sur un dénombrement, ce qui implique un faible biais numérique dans son calcul. De plus, comme son maximum peut être atteint par plusieurs paires de taxons différentes, et comme la minimisation, à chaque itération de l'algorithme agglomératif, du nombre de distances manquantes permet d'augmenter la probabilité de sélectionner la paire correcte à l'itération suivante, une troisième étape consiste à sélectionner, parmi les différentes paires xy maximisant le critère $\tilde{N}^*_{xy}/|\tilde{C}^*_{xy}|$, celle qui maximise la fonction $|\varnothing_x| + |\varnothing_y|$, où $\varnothing_i$ est définie par la Formule (6.10). Ainsi, la paire xy sélectionnée est celle qui correspond au plus grand nombre de distances manquantes. Si, éventuellement, plusieurs paires de taxons restent encore disponibles à l'agglomération, une quatrième étape consiste à sélectionner celle qui maximise le critère $\tilde{N}'^*_{xy}/|\tilde{C}^*_{xy}|$.

Si on considère $s = 5$ dans l'exemple de la Figure 6.2, alors les cinq paires de taxons correspondant aux cinq plus grandes valeurs de Q^*_{xy} sont sélectionnées lors de la première étape. On observe que les trois cerises ab, cd et ef font parties des paires de taxons sélectionnées. Durant la deuxième étape, le critère $\tilde{N}^*_{xy}/|\tilde{C}^*_{xy}|$ ne retient que les trois cerises ; en effet, le troisième graphique indique que $\tilde{N}^*_{ab}/|\tilde{C}^*_{ab}| = \tilde{N}^*_{cd}/|\tilde{C}^*_{cd}| = \tilde{N}^*_{ef}/|\tilde{C}^*_{ef}| = 1$. Observant que $|\varnothing_a| + |\varnothing_b| = 1$, $|\varnothing_c| + |\varnothing_d| = 3$ et que $|\varnothing_e| + |\varnothing_f| = 2$, la troisième étape sélectionne la paire cd pour l'agglomération. Ainsi durant l'itération suivante, il n'y a plus aucune distance manquante dans la matrice (T_{ij}). Même si le critère $\tilde{N}'^*_{xy}/|\tilde{C}^*_{xy}|$ n'a pas été utilisé dans cet exemple,

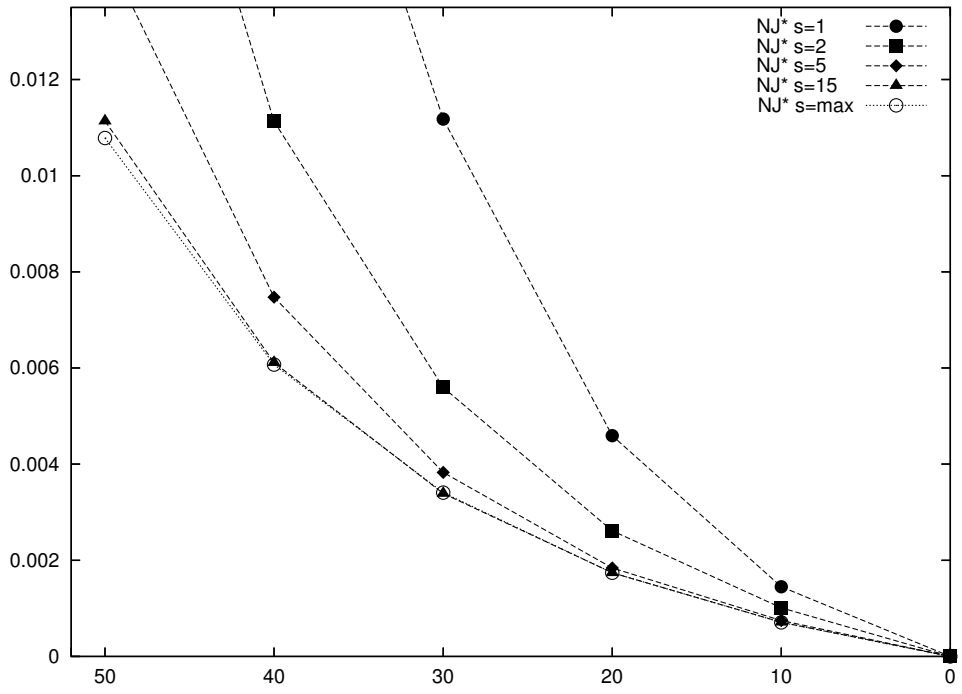

FIG. 6.3 – **Performances de la combinaison de plusieurs critères d'agglomération**

L'axe des abcisses indique les différentes valeur de $P_{miss} = 50\%, 40\%, 30\%, 20\%, 10\%, 0\%$. L'axe des ordonnées représente la moyenne des 500 distances d_{quad} entre l'arbre inféré $\hat{T}$ et l'arbre modèle $T_{\mathcal{M}}$.

de nombreux cas ont été observés en pratique où la troisième étape n'arrivait pas à choisir une unique paire xy à agglomérer. La quatrième étape reste donc nécessaire pour empêcher une agglomération aléatoire entre les paires de taxons sélectionnées durant la troisième étape.

A chaque itération, la maximisation de Q^*_{xy} s'effectue en $O(r^2)$. Pour une paire de taxons xy fixée, le calcul de $\tilde{N}^*_{xy}/|\tilde{C}^*_{xy}|$, de $\tilde{N}'^*_{xy}/|\tilde{C}^*_{xy}|$ et de $|\varnothing_x| + |\varnothing_y|$ s'effectue en $O(r^2)$. Ainsi, la combinaison des quatre critères Q^*_{xy}, $\tilde{N}^*_{xy}/|\tilde{C}^*_{xy}|$, $|\varnothing_x| + |\varnothing_y|$ et $\tilde{N}'^*_{xy}/|\tilde{C}^*_{xy}|$ permet de sélectionner la paire de taxons à agglomérer avec une complexité de l'ordre de $O(sr^2)$ à chaque itération, ce qui implique une complexité totale de $O(sn^3) = O(n^3)$ lorsque s est une constante.

Afin de définir une valeur numérique pour le paramètre s, une matrice de distance additive a été calculée à partir de l'arbre $T_{\mathcal{M}}$ inféré par combinaison basse à l'aide d'un critère ML dans le protocole de simulation décrit dans la partie 5.3.1 (cf page 106). Cet arbre modèle défini sur $n = 75$ taxons permet ainsi de considérer une distance additive $(T_{\mathcal{M}\,ij})$ avec une distribution des valeurs correspondant à une certaine réalité biologique. Pour chaque valeur $P_{miss} = 50\%, 40\%, 30\%, 20\%, 10\%, 0\%$, un ensemble de 500 matrices de distance a été généré aléatoirement à partir de $(T_{\mathcal{M}\,ij})$, chacune contenant une proportion P_{miss} de distances manquantes. L'algorithme agglomératif défini avec la combinaison des quatre critères d'agglomération et avec les paramètres NJ* a été appliqué sur chacune des 6×500 matrices de

distance additives incomplètes pour obtenir un arbre $\hat{T}$. La distance d_{quad} a été calculée entre chaque arbre $\hat{T}$ inféré et l'arbre modèle $T_{\mathcal{M}}$. Pour chaque valeur de P_{miss}, la moyenne des 500 distances topologiques d_{quad} a été représentée graphiquement en fonction de plusieurs valeurs de s dans la Figure 6.3.

La courbe du graphique de la Figure 6.3 correspondant à $s = 1$ montre les performances de l'algorithme agglomératif en $O(n^3)$ utilisé avec le seul critère Q^*_{xy}. Pour $s =$ max, l'algorithme agglomératif a été utilisé avec les trois critères $\tilde{N}^*_{xy}/|\tilde{C}^*_{xy}|$, $|\varnothing_x| + |\varnothing_y|$ et $\tilde{N}'^*_{xy}/|\tilde{C}^*_{xy}|$ appliqués sur toutes les paires de taxons, induisant une complexité de l'ordre de $O(n^4)$. Les courbes correspondant aux valeurs intermédiaires de s démontrent que la combinaison des quatre critères permet d'améliorer très significativement la qualité des arbres phylogénétiques inférés. On observe également qu'une constante s relativement faible (*i.e.* $s = 15$) permet d'inférer des arbres d'aussi bonne qualité qu'avec la version en $O(n^4)$ de l'algorithme agglomératif.

La même expérience a été effectuée en remplaçant le critère Q^*_{xy} par $\tilde{Q}^*_{xy}$. D'une manière très similaire aux résultats du Tableau 6.1, l'utilisation de $\tilde{Q}^*_{xy}$ a permis d'observer de moins bonnes valeurs d_{quad} moyennes avec $s = 1$ (*e.g.* pour $P_{miss} = 20\%$, $d_{\text{quad}} = 0.0046$ avec Q^*_{xy} et $d_{\text{quad}} = 0.0072$ avec $\tilde{Q}^*_{xy}$). De moins bonnes performances ont également été observées avec $s = 2$ (*e.g.* pour $P_{miss} = 20\%$, $d_{\text{quad}} = 0.0026$ avec Q^*_{xy} et $d_{\text{quad}} = 0.0038$ avec $\tilde{Q}^*_{xy}$). Toutefois, l'utilisation du critère $\tilde{N}^*_{xy}/|\tilde{C}^*_{xy}|$ sur un plus grand nombre s de paires de taxons permet de compenser les mauvaises performances de $\tilde{Q}^*_{xy}$. Ainsi, à partir de $s = 5$, les valeurs d_{quad} moyennes sont globalement similaires que l'on utilise Q^*_{xy} ou $\tilde{Q}^*_{xy}$ (quoique toujours très légèrement moins bonnes avec $\tilde{Q}^*_{xy}$).

6.3 Les algorithmes NJ*, UNJ*, BIONJ* et MVR* dans le cadre de l'inférence phylogénomique

Suite aux résultats théoriques et pratiques précédents, cette partie définit les algorithmes NJ*, UNJ*, BIONJ* et MVR* permettant d'inférer un arbre phylogénétique à partir d'une distance incomplète. Après une description formelle de ces algorithmes, une discussion explique comment utiliser les supermatrices de distance inférées par la méthode SDM ainsi que leur variance associée afin de définir de nouveaux scénarios d'inférence phylogénomique par combinaison moyenne.

6.3.1 Description des algorithmes NJ*, UNJ*, BIONJ* et MVR*

Les quatres algorithmes NJ*, UNJ*, BIONJ* et MVR* sont issus de la classe des algorithmes agglomératifs définie dans la Figure 6.1 pour les distances incomplètes. A chaque itération, la recherche de la paire de taxons à agglomérer s'effectue par la maximisation successive des critères

Q^*_{xy}, $\tilde{N}^*_{xy}/|\tilde{C}^*_{xy}|$, $|\varnothing_x| + |\varnothing_y|$ et $\tilde{N}'^*_{xy}/|\tilde{C}^*_{xy}|$ telle que décrite dans la partie précédente.

L'algorithme NJ*

Le calcul des paramètres w^* et λ^* s'effectue en $O(1)$ avec les Formules (6.3) et (6.4) pour l'algorithme NJ*. Ce dernier implique donc une complexité de l'ordre de $O(n^3)$.

L'algorithme UNJ*

Les paramètres w^*_i et λ^* de l'algorithme UNJ* se calculent en $O(r)$ et $O(1)$, respectivement, avec les Formules (6.5) et (6.6). La complexité algorithmique de UNJ* est donc de $O(n^3)$.

L'algorithme BIONJ*

Pour l'algorithme BIONJ*, les paramètres w^* et λ^* se calculent en $O(1)$ et $O(r)$, respectivement, avec les Formules (6.3) et (6.7). La réduction de la matrice de variance (V_{ij}) s'effectuant en $O(n)$, l'algorithme BIONJ* implique donc une complexité de l'ordre de $O(n^3)$.

L'algorithme MVR*

Pour l'algorithme MVR*, les paramètres w^*_i et λ^*_i se calculent en $O(r)$ et $O(1)$, respectivement, avec les Formules (6.9) et (6.8). La réduction de la matrice de variance (V_{ij}) s'effectuant en $O(n)$, l'algorithme MVR* implique donc une complexité de l'ordre de $O(n^3)$.

6.3.2 Utilisation des supermatrices de distance inférées par SDM

Comme montré dans le Chapitre 5, la méthode SDM permet de développer plusieurs scénarios d'inférence phylogénomique (suivant le modèle ou la méthode d'inférence d'arbre choisi). Etant donné une collection $\mathcal{C}_{(\Delta)} = \left\{(\Delta^1_{ij}), (\Delta^2_{ij}), ..., (\Delta^p_{ij}), ..., (\Delta^k_{ij})\right\}$ de k matrices de distance directement estimées à partir de k gènes différents, la méthode SDM calcule un ensemble de paramètres α_p et a_{ip} servant à déformer les différentes matrices de distance sans modifier l'information topologique induite par chacune d'entre elles. Ces différents paramètres sont calculés de manière à minimiser l'éloignement réciproque entre chaque paire de matrices de distance, au sens OLS ou WLS Les paramètres sont ensuite utilisés pour calculer une supermatrice de distance $(\Delta^{\text{SDM}}_{ij})$, avec la Formule (5.9) (cf page 102) :

$$\Delta^{\text{SDM}}_{ij} = \frac{1}{W_{ij}} \sum_{\substack{1\leq p\leq k \\ i,j\in\tilde{\mathcal{L}}_{\Delta^p}}} w_p\big(\alpha_p\Delta^p_{ij} + a_{ip} + a_{jp}\big) \qquad \text{où} \qquad W_{ij} = \sum_{\substack{1\leq p\leq k \\ i,j\in\tilde{\mathcal{L}}_{\Delta^p}}} w_p,$$

qui contient l'ensemble du signal topologique induit par la collection $\mathcal{C}_{(\Delta)}$.

Le Chapitre 5 décrit l'application d'une méthode d'inférence d'arbre sur $(\Delta^{\text{SDM}}_{ij})$ et définit donc un scénario d'inférence phylogénomique par combinaison moyenne, nommé SDM+FITCH.

Ce scénario utilisant le logiciel FITCH (Felsenstein, 1997) du package PHYLIP (Felsenstein, 1993) a montré de bonnes performances en simulation. Néanmoins, FITCH avec une complexité de l'ordre de $O(n^4)$, se révèle être la partie la plus longue en temps calcul du scénario SDM+FITCH. Une alternative efficace en terme de rapidité consiste donc à remplacer FITCH par les algorithmes NJ*, UNJ*, BIONJ* ou MVR*.

Les parties suivantes développent cette idée en montrant que la variance associée à $(\Delta_{ij}^{\text{SDM}})$ peut être facilement calculée, et en observant les performances des cinq scénarios SDM+FITCH, SDM+NJ*, SDM+UNJ*, SDM+BIONJ* et SDM+MVR* sur des données simulées.

Utilisation de la variance associée à la supermatrice de distance calculée par SDM

Les algorithmes NJ* et UNJ* n'utilisent que la supermatrice de distance $(\Delta_{ij}^{\text{SDM}})$. L'algorithme BIONJ* utilise une matrice de variance (V_{ij}) associée mais appuie le calcul des paramètres la caractérisant sur une estimation simplifiée de V_{ij}, *i.e.* $V_{ij} \propto \Delta_{ij}^{\text{SDM}}$.

L'algorithme MVR* utilise également la variance associée à $(\Delta_{ij}^{\text{SDM}})$ mais ne s'appuie *a priori* sur aucun modèle d'estimation particulière de cette variance. Ainsi, l'utilisation de MVR* avec $(\Delta_{ij}^{\text{SDM}})$ nécessite l'estimation mathématique de $V_{ij}^{\text{SDM}} = \text{Var}(\Delta_{ij}^{\text{SDM}})$. Après calcul, on obtient une première expression de la variance de Δ_{ij}^{SDM} :

$$V_{ij}^{\text{SDM}} = \frac{1}{W_{ij}^2}\left(\sum_{\substack{1\leq p\leq k\\ i,j\in\mathcal{L}_{\Delta^p}}} w_p^2\alpha_p^2\text{Var}(\Delta_{ij}^p) \quad + \quad 2\sum_{\substack{1\leq p<q\leq k\\ i,j\in\mathcal{L}_{\Delta^p}\cap\mathcal{L}_{\Delta^q}}} w_p w_q\alpha_p\alpha_q\text{Cov}(\Delta_{ij}^p,\Delta_{ij}^q)\right),$$

où $\text{Cov}(X,Y)$ correspond à la covariance de X et Y. Sachant que $\text{Cov}(X,Y)=0$ si X et Y sont indépendantes, cette expression se simplifie :

$$V_{ij}^{\text{SDM}} = \frac{1}{W_{ij}^2}\sum_{\substack{1\leq p\leq k\\ i,j\in\mathcal{L}_{\Delta^p}}} w_p^2\alpha_p^2\text{Var}(\Delta_{ij}^p). \tag{6.11}$$

Il est très courant de considérer l'estimation $\text{Var}(\Delta_{ij}^p) = (\Delta_{ij}^p)^\zeta/\ell_p$, où ℓ_p est la longueur des séquences génétiques S_i^p et S_j^p à partir desquelles la distance Δ_{ij}^p a été estimée. Plusieurs valeurs de ζ sont possibles telles que $\zeta=1$ (Nei and Jin, 1989; Bulmer, 1991), $\zeta=2$ (Fitch and Margoliash, 1967; Kuhner and Felsenstein, 1994; Felsenstein, 1997) ou la récente estimation par bootstrap $\zeta=1.823$ (Sanjuán and Wróbel, 2005). Ainsi, si on considère les pondérations standards de SDM, *i.e.* $w_p=\ell_p/\mathcal{N}$, avec $\mathcal{N}=1$, $\mathcal{N}=\tilde{n}_p$ ou $\mathcal{N}=\tilde{n}_p(\tilde{n}_p-1)$, alors l'équation (6.11) se réécrit :

$$V_{ij}^{\text{SDM}} = \frac{1}{W_{ij}^2}\sum_{\substack{1\leq p\leq k\\ i,j\in\mathcal{L}_{\Delta^p}}} \frac{\ell_p}{\mathcal{N}^2}\alpha_p^2(\Delta_{ij}^p)^\zeta, \qquad \text{avec} \qquad W_{ij} = \sum_{\substack{1\leq p\leq k\\ i,j\in\mathcal{L}_{\Delta^p}}} \frac{\ell_p}{\mathcal{N}}.$$

Les algorithmes NJ*, UNJ*, BIONJ* et MVR* sont donc tous utilisables dans le cadre de l'inférence phylogénomique par combinaison moyenne lorsqu'on les emploie conjointement avec la méthode SDM.

25%				
k	NJ*	UNJ*	BIONJ*	MVR*
2	0.0928	0.0908	**0.0846**	0.0859
4	0.0594	0.0540	**0.0498**	0.0520
6	0.0448	0.0424	**0.0369**	0.0404
8	0.0388	0.0335	**0.0319**	***0.0325***
10	0.0318	0.0290	**0.0267**	***0.0284***
12	0.0354	0.0315	**0.0284**	0.0310
14	0.0282	0.0258	**0.0230**	0.0252
16	0.0326	0.0313	**0.0280**	0.0302
18	0.0280	0.0275	**0.0232**	0.0263
20	0.0277	0.0266	**0.0229**	***0.0247***

75%				
k	NJ*	UNJ*	BIONJ*	MVR*
2	0.2188	0.2194	**0.2124**	***0.2143***
4	0.1828	0.1733	**0.1705**	***0.1709***
6	0.1544	0.1422	**0.1399**	***0.1400***
8	0.1275	0.1168	0.1120	**0.1099**
10	0.1098	0.0948	0.0940	**0.0891**
12	0.1019	0.0917	0.0865	**0.0832**
14	0.0851	0.0711	0.0715	**0.0668**
16	0.0781	0.0683	0.0640	**0.0618**
18	0.0690	***0.0561***	**0.0545**	***0.0550***
20	0.0673	0.0538	0.0542	**0.0510**

TAB. 6.2 – **Performances de NJ*, UNJ*, BioNJ* et MVR***

Les deux tableaux représentent les distances d_{quad} moyennes entre l'arbre $\hat{T}$, inféré par les quatre scénarios SDM+NJ*, SDM+UNJ*, SDM+BIONJ* et SDM+MVR*, et l'arbre modèle T dans le cadre des 25% et 75% de délétion de taxons. La colonne k indique le nombre de matrices de distance utilisées pour construire chacune des 500 supermatrices de distance avec la méthode SDM. Les autres colonnes représentent la moyenne des 500 valeurs d_{quad} observées pour chaque valeur de k. Les valeurs écrites en caractères gras représentent la plus petite distance d_{quad} moyenne pour chaque ligne. Les valeurs écrites en caractères gras italiques représentent les distances d_{quad} moyennes dont la valeur est proche (± 0.002) de la plus petite valeur dans la ligne correspondante.

Résultats de simulation

Afin d'observer les performances respectives des scénarios d'inférence phylogénomique SDM+NJ*, SDM+UNJ*, SDM+BIONJ* et SDM+MVR*, les supermatrices de distance $(\Delta_{ij}^{\text{SDM}})$ générées lors des simulations décrites dans la partie 5.3.2 (cf page 108) ont été ré-analysées avec les algorithmes NJ*, UNJ*, BIONJ* et MVR*. Le paramètre $s = 20$ a été fixé lors de la combinaison des quatre critères d'agglomération. Les paramètres $\mathcal{N} = 1$ et $\zeta = 2$ ont été fixés pour le calcul de $(\Delta_{ij}^{\text{SDM}})$ et de (V_{ij}^{SDM}). Chaque arbre inféré $\hat{T}$ a été comparé avec son arbre modèle T à l'aide de la distance topologique d_{quad}, et la moyenne des 500 valeurs

d_{quad} observées, pour chaque valeur de $k = 2, 4, ..., 20$ et chaque taux de délétion de taxons $(25\%, 75\%)$, a été représentée dans le Tableau 6.2.

Lorsque l'on compare les résultats de la colonne NJ* pour 75% de délétion de taxons dans le Tableau 6.2 (cf page 133) avec les résultats du tableau 6.1, on observe que l'utilisation successive des quatre critères d'agglomération Q^*_{xy}, $\tilde{N}^*_{xy}/|\tilde{C}^*_{xy}|$, $|\varnothing_x| + |\varnothing_y|$ et $\tilde{N}'^*_{xy}/|\tilde{C}^*_{xy}|$ améliore efficacement les performances de l'algorithme NJ*.

Plus précisément, on observe que les algorithmes NJ* et UNJ* présentent de moins bons résultats que BIONJ* et MVR* dans tous les cas de figure. L'algorithme BIONJ* présente toujours les meilleures valeurs d_{quad} moyennes dans le cadre des 25% de délétion de taxons. Pour 75% de délétion de taxons, l'algorithme MVR* présente parfois de meilleures performances que BIONJ* suivant les valeurs de k. L'utilisation d'une estimation de la variance associée à $(\Delta^{\text{SDM}}_{ij})$ permet donc d'améliorer sensiblement les performances des algorithmes agglomératifs dans les scénarios d'inférence phylogénomique par combinaison moyenne.

En termes de temps d'exécution sur un PC Pentium IV 1.8GHz (1Go RAM), pour $k = 10$, des durées d'environ 18 secondes et 14 secondes ont été nécessaires pour les taux de délétion de taxons de 25% et 75%, respectivement, pour construire l'ensemble des 500 arbres phylogénétiques. Dans les mêmes conditions de simulation, le logiciel FITCH met plus de 20 secondes pour construire un seul arbre (cf page 110). Les algorithmes agglomératifs mettant ainsi moins d'un dizième de secondes pour construire un arbre, ils représentent une alternative de choix pour l'inférence phylogénomique de plusieurs centaines d'espèces.

Disposant d'une méthode efficace d'inférence de supermatrices de distance (SDM) et d'algorithmes rapides d'inférence d'arbres (NJ*, UNJ*, BIONJ* et MVR*), le chapitre suivant étudie leur utilisation conjointe dans divers scénarios de combinaison. Il décrit et compare également les performances des autres approches d'inférence phylogénomique par combinaison basse (*e.g.* approches ML), moyenne (*e.g.* ACS+MW*, SDM+FITCH) et haute (*e.g.* MRP).

Chapitre 7

Comparaison de différents scénarios d'inférence phylogénomique

La simplicité est souvent l'élément clé d'un algorithme efficace en pratique.

Jens Stoye

Sommaire

7.1 Protocole de simulation 136

7.2 Scénarios de combinaison basse 137

7.2.1 Scénarios de combinaison basse utilisant des méthodes de distance . . . 137

7.2.2 Résultats et discussion 139

7.3 Scénarios de combinaison moyenne 140

7.3.1 Inférence d'arbre et complétion de distances manquantes à partir des supermatrices de distance inférées par SDM 141

7.3.2 Résultats et discussion 143

7.4 Scénarios de combinaison haute 144

7.4.1 Les supermatrices de distance inférées à partir d'une collection d'arbres phylogénétiques 145

7.4.2 Les techniques de représentation matricielle binaire MRP et MRH . . . 146

7.4.3 Résultats et discussion 148

7.5 Les différents niveaux de combinaison en inférence phylogénomique . . . 151

Les techniques d'inférence phylogénomique peuvent être classifiées en trois grands types de combinaison de données génétiques (cf page 72). La combinaison basse effectue l'analyse simultanée d'une supermatrice de caractères obtenu par concaténation des alignements de

chaque donnée. La combinaison haute infère un arbre phylogénétique à partir de chaque donnée, puis amalgame ces différents arbres en un unique superarbre. La combinaison moyenne passe par une étape d'encodage de l'information phylogénétique de chaque donnée, puis amalgame ces nouvelles informations afin d'inférer un arbre.

Les chapitres précédents de ce manuscrit de thèse ont explicité plusieurs techniques et algorithmes existants pour chacun des trois types de combinaison. Plusieurs nouveaux scénarios de combinaison moyenne ont été également proposés en s'appuyant sur la méthode SDM et les algorithmes NJ*, UNJ*, BIONJ* et MVR*.

Ce chapitre introduit de nouveaux scénarios d'inférence phylogénomique utilisant les méthodes de distance SDM, UNJ*, BIONJ* et MVR*. Les performances de chacun de ces nouveaux scénarios sont observées et discutées en utilisant les données générées lors des simulations décrites dans la partie 5.3.2 (cf page 108).

7.1 Protocole de simulation

La génération des arbres modèles et des données nucléiques décrite dans la partie 5.3.2 (cf page 108) est résumée ci-dessous :

• *Génération des arbres modèles* — Un arbre modèle défini sur $n = 48$ taxons a été généré à l'aide du logiciel R8S suivant le processus de Yule-Harding (Yule, 1925; Harding, 1971). L'arbre modèle ultramétrique enraciné UT ainsi obtenu respecte l'hypothèse de l'horloge moléculaire en garantissant une distance de 1 entre la racine et chacune des feuilles. Un arbre T a été ensuite obtenu à partir de UT en créant une déviation par rapport à cette hypothèse, puis dupliqué afin d'obtenir k arbre T^p de même topologie que T. Une homothétie a été appliquée sur les longueurs de branche de chaque arbre T^p afin que chacun d'entre eux possède une vitesse d'évolution propre, variant uniformément (en valeur relative) entre 0.4 et 9.0.

• *Génération des données* — Une collection de k alignements de $n = 48$ séquences sur ℓ_p sites a été générée à l'aide du logiciel SEQ-GEN à partir de chaque arbre T^p suivant le modèle K2P. La valeur de ℓ_p est tirée uniformément entre 200 et 1000 sites et est propre à chaque arbre T^p. Pour chaque alignement, certains taxons ont été aléatoirement supprimés avec une certaine probabilité de suppression. Deux valeurs de probabilité ont été utilisées : 25% pour une suppression faible et 75% pour une suppression forte. Un recouvrement d'au moins quatre taxons entre chaque paire de matrices a été néanmoins préservé afin de conserver une histoire évolutive et une information topologique significatives communes entre les paires de k gènes.

Ce processus a été répété 500 fois pour chaque valeur de $k = 2, 4, 6, ..., 20$ et chacune des deux probabilités de suppression, aboutissant ainsi à $500 \times 10 \times 2$ collections $\mathcal{C}_G$ de données portant sur 48 taxons. A partir de chaque collection de gènes $\mathcal{C}_G$, différents autres jeux de données ont été inférés :

- la supermatrice de caractères $\mathcal{S}$, construite par concaténation des k gènes formant $\mathcal{C}_G$, afin d'y appliquer plusieurs méthodes de combinaison basse,
- la collection $\mathcal{C}_{(\Delta)}$ des k matrices de distance estimées à partir de chaque gène suivant le modèle K2P, afin d'utiliser la méthode SDM (modèle SSM) dans les scénarios de combinaison moyenne,
- la collection $\mathcal{C}_T$, contenant les arbres phylogénétiques inférés par PHYML suivant le modèle K2P à partir de chaque gène, afin de construire des superarbres dans le cadre de la combinaison haute.

Pour chaque arbre inféré $\hat{T}$, la distance topologique d_{quad} entre $\hat{T}$ et l'arbre modèle T a été utilisée pour observer les performances des différents scénarios d'inférence phylogénomique. Les temps d'exécution présentés ont été observés sur un PC Pentium IV 1.8GHz (1Go RAM).

7.2 Scénarios de combinaison basse

Les critères MP ayant été montrés inconsistants (dans certains cas) et les critère ML impliquant d'importants temps de calcul, cette partie décrit et discute plusieurs scénarios d'inférence phylogénomique par combinaison basse utilisant des critères de distance, telle que la méthode SDM.

7.2.1 Scénarios de combinaison basse utilisant des méthodes de distance

Estimation d'une matrice de distance à partir d'une supermatrice de caractères

A partir de chaque supermatrice de caractères $\mathcal{S}$, une matrice de distance (Δ_{ij}) a été directement estimée suivant le modèle K2P et la procédure MISSDIST=IGNORE (cf page 53). Un arbre $\hat{T}$ a été inféré par l'algorithme BIONJ à partir de (Δ_{ij}). Les valeurs d_{quad} moyennes observées sont représentées graphiquement dans la Figure 7.1.

Ce scénario d'inférence phylogénomique par combinaison basse, nommé SAD (*Simultaneous Analysis with Distance*), a été comparé avec le scénario d'inférence phylogénomique par combinaison moyenne SDM+FITCH introduit et simulé dans la partie 5.3.2 (cf page 108). Les graphiques de la Figure 7.1 permettent ainsi d'observer les performances des méthodes de distance par combinaison basse et moyenne.

En moyenne, le temps d'exécution de SAD est toujours inférieur ou égal à une seconde. Le temps d'exécution moyen de SDM varie entre environ 2 et 10 secondes selon les valeurs de k et les deux taux de suppression. Ainsi, pour $k = 10$, SDM nécessite une durée moyenne d'environ 2 secondes pour 25% de suppression de taxons et d'environ 1 seconde pour 75%. Pour $k = 20$, ces durées moyennes montent à environ 9 et 2 secondes, respectivement. Le temps d'exécution

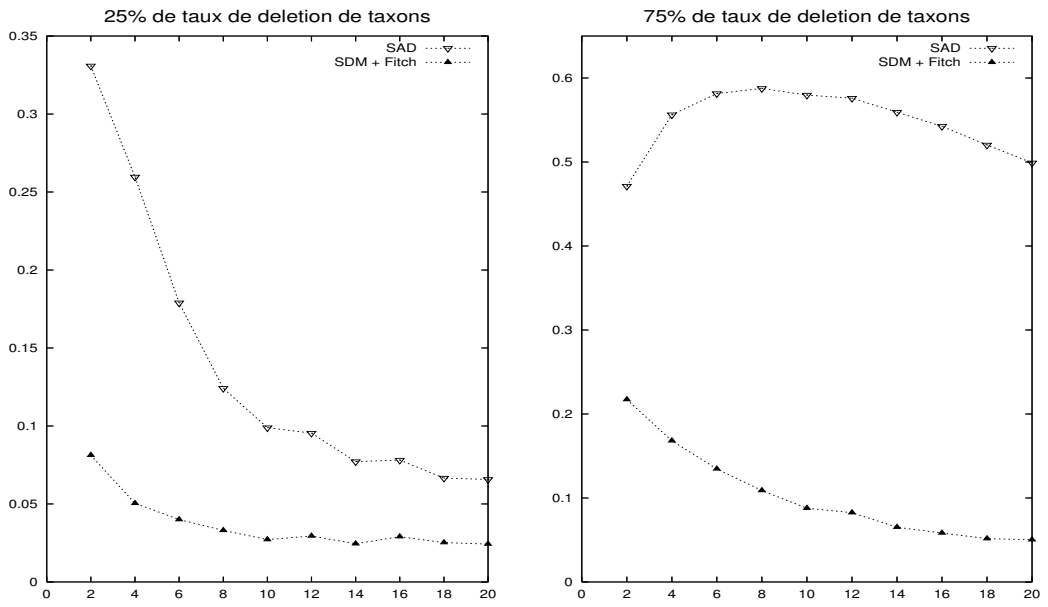

FIG. 7.1 – **Performances des méthodes de distance par combinaison basse et moyenne**

Les deux graphiques représentent les distances topologiques d_{quad} moyennes entre l'arbre $\hat{T}$, inféré par les deux scénarios SAD (combinaison basse) et SDM+FITCH (combinaison moyenne), et l'arbre modèle T dans le cadre des 25% et 75% de suppression de taxons. L'axe des abscisses représente les différentes valeurs de k. L'axe des ordonnées représente les distances topologiques d_{quad} moyennes observées pour les différents scénarios d'inférence phylogénomique.

moyen de FITCH est toujours largement supérieur à celui de SDM (*e.g.* pour $k = 20$, environ 23 secondes pour les deux taux de suppression de taxons).

Inférence ML à partir d'une supermatrice de caractères

Un arbre phylogénétique a été inféré à partir de chaque supermatrice de caractères $\mathcal{S}$ avec le logiciel PHYML suivant le modèle K2P. Par défaut, PHYML construit un arbre initial suivant le scénario SAD, puis effectue à partir de ce dernier une recherche locale par descente en utilisant des voisinages NNI augmentés. Ce scénario d'inférence phylogénomique est nommé SAD+PHYML.

Afin de mesurer la sensibilité de la recherche locale implémentée dans le logiciel PHYML, l'arbre inféré par le scénario SDM+FITCH a été utilisé comme nouveau point de départ. Ce scénario est nommé SDM+FITCH+PHYML[1].

[1] Le scénario SDM+FITCH+PHYML n'est pas un scénario de combinaison basse au sens strict du terme, car il utilise un arbre inféré par combinaison moyenne comme point de départ à une recherche locale employée dans une approche par combinaison basse. Toutefois, l'arbre renvoyé par ce scénario est celui qui maximise un critère ML à partir d'une supermatrice de caractères. Dans ce sens, c'est l'approche par combinaison basse qui "décide" de l'optimalité du critère choisi. Ainsi, l'utilisation de la recherche locale implémentée dans le logiciel PHYML sur une supermatrice de caractères

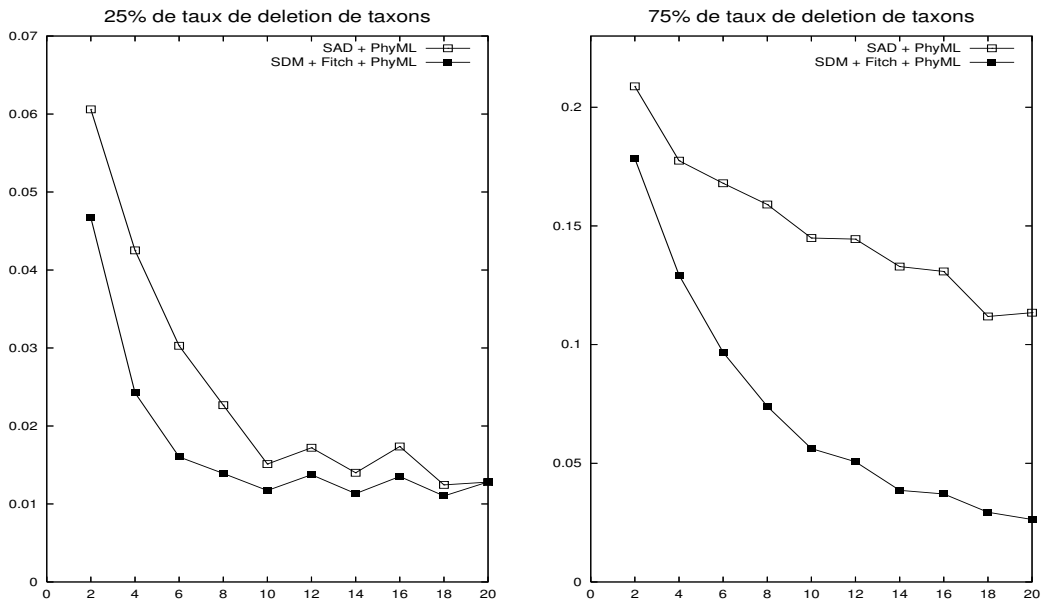

FIG. 7.2 – **Performances des scénarios ML d'inférence d'arbre par combinaison basse**

Les deux graphiques représentent les distances topologiques d_{quad} moyennes entre l'arbre $\hat{T}$, inféré par les deux scénarios SAD+PHYML et SDM+FITCH+PHYML, et l'arbre modèle T dans le cadre des 25% et 75% de suppression de taxons. L'axe des abscisses représente les différentes valeurs de k. L'axe des ordonnées représente les distances topologiques d_q moyennes observées pour les différents scénarios d'inférence phylogénomique.

Les valeurs d_{quad} moyennes observées pour les deux scénarios SAD+PHYML et SDM+FITCH+PHYML sont représentées graphiquement dans la Figure 7.2.

Pour 25% de suppression de taxons et $k = 20$, le temps d'exécution moyen de PHYML est d'environ 20 minutes avec un point de départ SAD et d'environ 13 minutes avec un point de départ SDM+FITCH. Pour 75% de suppression de taxons, ces durées moyennes sont d'environ 68 minutes et 35 minutes, respectivement.

7.2.2 Résultats et discussion

Malgré un temps d'exécution quasi-immédiat, le scénario SAD infère des arbres phylogénétique de très mauvaise qualité (cf Figure 7.1). La distance d_{quad} moyenne observée entre deux arbres générés aléatoirement sur un même ensemble de taxon est toujours proche de $2/3$ (Steel and Penny, 1993). Or, dans le cadre des 75% de suppression de taxons, le scénario SAD présente des valeurs d_{quad} proche de cette limite asymptotique. Si i et j sont deux taxons correspondant chacun à un gène G^1 lent et un gène G^2 rapide, alors la distance Δ_{ij} estimée à partir de la concaténation de G^1 et G^2 est biaisée, car l'éloignement induit par G^1 est étouffé

peut être considérée comme un scénario de combinaison basse, quelle que soit la nature de l'arbre de départ (aléatoire, ou inféré par combinaison basse, moyenne ou haute)

par la grande distance induite par G^2. De plus, si il existe un troisième taxon x présent dans G^1 mais pas dans G^2, alors les distances Δ_{ix} et Δ_{jx} seront relativement faibles par rapport à la distance Δ_{ij}. Cette observation montre également la nécessité d'utiliser une technique prenant en compte la vitesse d'évolution de chaque gène. La méthode SDM effectuant cette estimation, on observe naturellement de meilleures performances en comparaison avec SAD (cf Figure 7.1).

Si SAD montre de si mauvaises performances, parfois proches d'un arbre aléatoire, alors son utilisation comme point de départ d'une recherche locale par descente implique l'exploration d'un grand nombre de voisinages avant d'atteindre un optimum local. Cette observation explique les différences de performance entre les scénarios SAD+PHYML et SDM+FITCH+PHYML. Premièrement, la recherche locale implémentée dans PHYML s'arrête beaucoup plus souvent dans un minimum local lorsque SAD est pris comme point de départ ; ce phénomène s'observe particulièrement dans la Figure 7.2 pour 75% de suppression de taxons. Deuxièmement, plus le nombre de voisinages explorés par la recherche locale est grand, plus le temps d'exécution de cette recherche locale est important ; conséquemment, pour $k = 20$ et 75% de suppression de taxons, les temps d'exécution moyens des scénarios SAD+PHYML et SDM+FITCH+PHYML ont été respectivement d'environ 68 et 36 minutes. L'utilisation du scénario SDM+FITCH comme point de départ d'une recherche locale minimisant un critère ML (le même raisonnement peut s'appliquer pour un critère MP) permet donc d'inférer des arbres de meilleure qualité en un temps d'exécution moindre. Le scénario d'inférence phylogénomique SDM+FITCH+PHYML peut donc se révéler extrèmement utile pour traiter des supermatrices de caractère de tailles importantes ou pour effectuer des études de bootstrap[2].

7.3 Scénarios de combinaison moyenne

La partie 6.3.2 a permis d'observer les performances en simulation des quatre scénarios d'inférence phylogénomique SDM+NJ*, SDM+UNJ*, SDM+BIONJ* et SDM+MVR* (cf page 133). Ces résultats ont montré que les scénarios utilisant une estimation de la variance de $(\Delta_{ij}^{\text{SDM}})$ (*i.e.* SDM+BIONJ* et SDM+MVR*) infèrent des arbres de meilleures qualités. Cette partie prolonge ces résultats en comparant les performances de ces algorithmes avec d'autres méthodes d'inférence d'arbre (*e.g.* FITCH, MW*), ainsi qu'en observant l'efficacité des algorithmes de complétion de distances incomplètes.

[2] Ces simulations, cherchant à mesurer l'impact d'un arbre inféré avec SDM lorsqu'il est utilisé comme arbre initial dans la recherche locale implémentée dans le logiciel PHYML, ont été effectuées à une époque antérieure au développement des algorithmes NJ*, UNJ*, BIONJ* et MVR*, d'où l'utilisation de FITCH. Ces simulations ayant été très longues (parfois plus de 15 jours pour un seul point dans les graphiques de la Figure 7.2), elles n'ont pas été recommencées en utilisant les algorithmes agglomératifs. Néanmoins, il est fort probable qu'un scénario tel que SDM+BIONJ*+PHYML présente des résultats très similaires à ceux du scénario SDM+BIONJ*+PHYML

25%				
k	SDM+FITCH	SDM+MW*	SDM+BIONJ*	SDM+MVR*
2	**0.0834**	0.0906	***0.0846***	0.0859
4	***0.0503***	0.0546	**0.0498**	0.0520
6	0.0399	0.0445	**0.0369**	0.0404
8	***0.0330***	0.0356	**0.0319**	***0.0325***
10	***0.0271***	0.0299	**0.0267**	0.0284
12	***0.0294***	0.0317	**0.0284**	0.0310
14	***0.0244***	0.0266	**0.0230**	0.0252
16	***0.0290***	0.0318	**0.0280**	0.0302
18	***0.0252***	0.0278	**0.0232**	0.0263
20	***0.0242***	0.0259	**0.0229**	***0.0247***

75%				
k	SDM+FITCH	SDM+MW*	SDM+BIONJ*	SDM+MVR*
2	0.2154	0.2174	**0.2124**	***0.2143***
4	**0.1682**	0.1778	0.1705	0.1709
6	**0.1347**	0.1443	0.1399	0.1400
8	**0.1089**	0.1253	0.1120	***0.1099***
10	**0.0878**	0.1039	0.0940	***0.0891***
12	**0.0825**	0.0968	0.0865	***0.0832***
14	**0.0652**	0.0749	0.0715	***0.0668***
16	**0.0583**	0.0731	0.0640	0.0618
18	**0.0515**	0.0617	0.0545	0.0550
20	**0.0503**	0.0600	0.0542	***0.0510***

TAB. 7.1 – **Performances des scénarios SDM+FITCH, SDM+MW*, SDM+BIONJ* et SDM+MVR***

Les deux tableaux représentent les distances d_{quad} moyennes entre l'arbre $\hat{T}$, inféré par les quatre scénarios de combinaison moyenne, et l'arbre modèle T dans le cadre des 25% et 75% de suppression de taxons. La colonne k indique la taille de la collection $\mathcal{C}_{(\Delta)}$. Les autres colonnes représentent la moyenne des 500 valeurs d_{quad} observées pour chaque valeur de k. Les valeurs écrites en caractères gras représentent la plus petite distance d_{quad} moyenne pour chaque ligne. Les valeurs écrites en caractères gras italiques représentent les distances d_{quad} moyennes dont la valeur est proche (± 0.002) de la plus petite valeur dans la ligne correspondante.

7.3.1 Inférence d'arbre et complétion de distances manquantes à partir des supermatrices de distance inférées par SDM

Cette partie décrit et discute des simulations effectuées pour observer et comprendre les causes des différences de performances entre les algorithmes agglomératifs BIONJ* et MVR*, et les méthodes standards FITCH et MW*.

75%									
k	θ			NJ*			UNJ*		
	-	ADD	QUAD	-	ADD	QUAD	-	ADD	QUAD
2	32%	0%	11%	**0.2188**	0.2240	0.2322	**0.2194**	0.2230	0.2285
4	42%	0%	13%	**0.1828**	0.2051	0.1993	**0.1733**	0.1921	0.1895
6	42%	0%	11%	**0.1544**	0.1831	0.1660	**0.1422**	0.1655	0.1519
8	37%	0%	9%	**0.1275**	0.1658	0.1367	**0.1168**	0.1444	0.1213
10	32%	0%	6%	**0.1098**	0.1535	***0.1115***	**0.0948**	0.1266	0.0982
12	27%	0%	4%	**0.1019**	0.1424	0.1047	**0.0917**	0.1221	***0.0924***
14	22%	0%	3%	**0.0851**	0.1285	***0.0869***	**0.0711**	0.0976	***0.0715***
16	18%	0%	2%	**0.0781**	0.1211	0.0808	***0.0683***	0.0991	**0.0676**
18	14%	0%	1%	**0.0690**	0.1068	0.0716	**0.0561**	0.0857	***0.0569***
20	11%	0%	1%	**0.0673**	0.1018	***0.0685***	**0.0538**	0.0815	***0.0552***

TAB. 7.2 – **Performances des scénarios SDM+NJ*, SDM+Add+NJ*, SDM+Quad+NJ*, SDM+UNJ*, SDM+Add+UNJ* et SDM+Quad+UNJ***

La colonne k indique la taille de la collection $\mathcal{C}_{(\Delta)}$. La colonne θ représente la proportion de distances manquantes dans les supermatrices de distance avant (colonne $-$) et après les complétions additive (colonne ADD) et par quadruplet (colonne QUAD). Les autres colonnes représentent la moyenne des 500 valeurs d_{quad} observées pour chaque valeur de k. Les valeurs écrites en caractères gras représentent la plus petite distance d_{quad} moyenne pour chaque ligne. Les valeurs écrites en caractères gras italiques représentent les distances d_{quad} moyennes dont la valeur est proche (± 0.002) de la plus petite valeur dans la ligne correspondante.

Les méthodes standards d'inférence d'arbres à partir d'une supermatrice de distance incomplète

Le logiciel FITCH et la méthode MW* ont été appliqués sur les supermatrices de distance $(\Delta_{ij}^{\text{SDM}})$ inférées à partir de chaque collection $\mathcal{C}_{(\Delta)}$. Les valeurs d_{quad} moyennes observées pour SDM+FITCH et SDM+MW*, ainsi que pour les deux scénarios SDM+BIONJ* et SDM+MVR*, sont représentées dans le Tableau 7.1.

Pour $k = 10$ et 25% de suppression de taxons, FITCH et MW* nécessitent une durée totale d'environ 3 heures 11 minutes et 4 heures 40 minutes, respectivement, pour inférer les 500 arbres. Pour $k = 10$ et 75% de suppression de taxons, FITCH et MW* nécessitent une durée totale d'environ 2 heures 21 minutes et 4 heures 1 minute, respectivement. Les algorithmes agglomératifs BIONJ* et MVR* ont nécessité au plus 30 secondes pour inférer les 500 arbres, pour chaque valeur de k et chaque taux de suppression de taxons.

Les algorithmes de complétion de distances incomplètes

Les algorithmes de complétion additive et par quadruplets présentés dans la partie 3.5.1 (cf page 55) ont été implémentés et appliqués sur chaque supermatrice de distance $(\Delta_{ij}^{\text{SDM}})$ inférée dans le cadre des 75% de suppression de taxons. Ces deux méthodes, nommées

SDM+ADD et SDM+QUAD, respectivement, ont permis d'obtenir des supermatrices de distance contenant relativement peu de distances manquantes (cf Tableau 7.2, colonne θ). Les algorithmes agglomératifs NJ* et UNJ* ont été appliqués sur chaque supermatrice de distance SDM+ADD et SDM+QUAD.

Les valeurs d_{quad} moyennes observées pour SDM+ADD+NJ*, SDM+QUAD+NJ*, SDM+ADD+UNJ* et SDM+QUAD+UNJ* sont représentées dans le Tableau 7.2.

Chaque algorithme de complétion a nécessité une durée moyenne de moins de 2 secondes par supermatrice de distance pour chaque valeur de k.

7.3.2 Résultats et discussion

Dans le cadre des 25% de suppression de taxons, le scénario d'inférence phylogénomique SDM+BIONJ* présente globalement les meilleures valeurs d_{quad} moyennes, suivi par les scénarios SDM+FITCH et SDM+MVR*. Dans le cadre des 75% de suppression de taxons, le scénario SDM+FITCH présente globalement de meilleures performances que les scénarios SDM+BIONJ* et SDM+MVR*. Nénamoins, les différentes valeurs d_{quad} demeurent très proches entre ces trois scénarios. Ainsi, malgré des temps d'exécution entre 200 et 1800 fois plus long que les algorithmes BIONJ* et MVR*, le logiciel FITCH n'infère pas toujours les meilleurs arbres phylogénétiques. Les excellentes performances de BIONJ* et MVR*, en terme de rapport entre qualité et rapidité, font de ces deux algorithmes des outils de choix pour l'inférence phylogénomique par combinaison moyenne en association avec la méthode SDM.

La méthode MW* présente des valeurs d_{quad} moyennes un peu plus élevées que les autres méthodes. Ce fait s'explique par l'application préalable d'un algorithme de complétion ultramétrique ou additive. Si $\theta = \mid \overline{\{i,j : \Delta_{ij}^{\text{SDM}} = \varnothing\}} \mid$ représente la proportion de distances manquantes dans $(\Delta_{ij}^{\text{SDM}})$, alors, suivant la Formule (3.3), il existe au plus $n-2$ estimations ultramétriques, chacune ayant une probabilité d'existence de $(1-\theta)^2$. Ainsi le nombre moyen d'estimateurs pour la complétion ultramétrique est de $(n-2)(1-\theta)^2$. Suivant raisonnement similaire à partir de la Formule (3.4) formalisant la complétion additive, il y a $(n-2)(n-3)(1-\theta)^5/2$ estimateurs en moyenne pour la complétion additive. Ainsi, la complétion additive est plus informative que la complétion ultramétrique si

$$\frac{(n-2)(n-3)(1-\theta)^5}{2} \; > \; (n-2)(1-\theta)^2,$$

ce qui conduit, après calculs, à la condition suivante (Makarenkov and Lapointe, 2004) :

$$n \; > \; 3 + \frac{2}{(1-\theta)^3}.$$

Si cette condition est vérifiée, MW* effectue une complétion additive des distances manquantes. Dans le cas contraire, MW* applique un algorithme de complétion ultramétrique. Comme l'ensemble des supermatrices de distance $(\Delta_{ij}^{\text{SDM}})$ inférées à partir de chaque $\mathcal{C}_\Delta$ vérifient cette condition, la méthode MW* infère un arbre à partir de la matrice de distance SDM+ADD. Or,

les résultats représentés dans le Tableau 7.2 montrent que SDM+ADD est une technique de complétion présentant de mauvais résultats. Si $\hat{T}$ est l'arbre inféré à partir de $(\Delta_{ij}^{\text{SDM}})$, ces mauvais résultats sont dûs au fait que la complétion additive tend à produire des multifurcations dans $\hat{T}$ lorsque $\Delta_{xy}^{\text{SDM}} = \varnothing$ et que x et y induisent une cerise dans $\hat{T}$ (Guénoche and Grandcolas, 1999).

Le Tableau 7.2 montre que la complétion par quadruplets, qui corrige le biais de la complétion additive en ne complétant pas toutes les distances manquantes, permet aux algorithmes NJ* et UNJ* d'inférer des arbres de meilleure qualité que ceux inférés à partir de SDM+ADD. Néanmoins, malgré les meilleures performances de la complétion par quadruplets SDM+QUAD, on n'observe pas de réelles améliorations en comparaison avec les résultats des scénarios SDM+NJ* et SDM+UNJ* n'utilisant pas de complétion (cf Tableau 6.2, page 133). Ainsi, les méthodes d'inférence d'arbre à partir de distances incomplètes ne nécessitent pas les algorithmes de complétion actuels pour renvoyer des arbres de meilleures qualités.

Seuls les algorithmes NJ* et UNJ* ont été utilisés pour observer les performances des algorithmes de complétion additive et par quadruplets. Il est montré dans l'Annexe C que le calcul de la variance V_{xy}^{SDM} d'une distance Δ_{xy}^{SDM} estimée par complétion additive s'effectue en $O(n)$ afin de pouvoir utiliser les algorithmes BIONJ* et MVR*. Néanmoins, sachant qu'il est préférable d'employer la complétion par quadruplets, il est également montré dans l'Annexe C que le calcul de la variance V_{xy}^{SDM} d'une distance Δ_{xy}^{SDM} estimée par complétion par quadruplets s'effectue en $O(n^5)$ pour employer l'algorithme BIONJ* et en $O(n^3)$ pour employer l'algorithme MVR*. Ces résultats impliquant des calculs de l'ordre de $O(n^5)$ et plus, ces cas de figure n'ont pas été jugés efficaces.

7.4 Scénarios de combinaison haute

Cette partie explicite l'adaptation à la combinaison haute des scénarios d'inférence phylogénomique utilisant la méthode SDM. La seule méthode actuelle de combinaison haute combinant des distances additives est la méthode ACS (Lapointe and Cucumel, 1997). La méthode de combinaison haute la plus couramment utilisée est la méthode MRP (Baum, 1992; Ragan, 1992). Cette partie explicite différentes utilisations de la méthode SDM et des algorithmes UNJ*, BIONJ* et MVR* dans le cadre de la combinaison haute, introduit une nouvelle méthode, nommée MRH (*Matrix Representation with Hamming distance*) et très proche de la méthode MRD (*Matrix Representation with Distance* ; Lapointe *et al.*, 2003), et discute de leurs performances relatives en comparaison avec ACS et MRP.

7.4.1 Les supermatrices de distance inférées à partir d'une collection d'arbres phylogénétiques

Observation des différentes méthodes d'inférence d'arbre

Si on considère les matrices de distance additive (T_{ij}^p) équivalentes à chacun des k arbres phylogénétiques formant la collection $\mathcal{C}_T$, alors la méthode SDM peut être appliquée sur $\mathcal{C}_T$. Pour chaque valeur de k et chaque taux de suppression de taxons, une supermatrice de distance $(\Delta_{ij}^{\text{SDM}})$ a été inférée à partir de $\mathcal{C}_T$ avec les mêmes paramètres utilisés dans les simulations de combinaison moyenne. Cette première partie de scénario d'inférence phylogénomique est nommée PHYML+SDM. Chacune des supermatrices de distance produite par PHYML+SDM a été analysée par les algorithmes FITCH, MW*, NJ*, UNJ*, BIONJ* et MVR*.

Pour chaque valeur de k et et chaque taux de suppression de taxons, les valeurs d_{quad} moyennes observées entre le superarbre inféré $\hat{T}$ et son arbre modèle T ont été reportées dans le Tableau 7.3.

Les temps d'exécution sont similaires à ceux constatés dans le cadre de la combinaison moyenne.

Utilisation de la méthode ACS

La méthode ACS est très similaire à la méthode SDM, en ce sens qu'elle effectue une moyenne des distances additives T_{ij}^p disponibles pour chaque paire de taxons ij suivant la Formule (4.1) (cf page 85) :

$$\overline{T}_{ij} = \frac{1}{k_{ij}} \sum_{1 \leq p \leq k} T_{ij}^p,$$

où k_{ij} est le nombre de fois où la paire ij est incluse dans chaque arbre source T^p, *i.e.* $k_{ij} = \left| \{p : i, j \in \mathcal{L}_{T^p}\} \right|$. Comme chaque distance additive (T_{ij}^p) nécessite d'être normalisée par rapport aux autres, la formule suivante a été utilisée :

$$\Delta_{ij}^{\text{ACS}} = \frac{1}{W_{ij}} \sum_{\substack{1 \leq p \leq k \\ i,j \in \mathcal{L}_{T^p}}} \alpha_p T_{ij}^p,$$

où (Lapointe and Cucumel, 1997) :

$$\alpha_p = \left(\max_{i,j \in \mathcal{L}_{T^p}} [\, T_{ij}^p \,] \right)^{-1}.$$

Comme il a également été préconisé d'employer une méthode minimisant un critère LS pour inférer un arbre à partir de $(\Delta_{ij}^{\text{ACS}})$ (Lapointe and Cucumel, 1997), le logiciel FITCH a été utilisé dans ce but. Ce scénario d'inférence phylogénomique est nommé PHYML+ACS+FITCH.

Pour chaque valeur de k et chaque taux de suppression de taxons, les distance topologique d_{quad} moyennes entre l'arbre $\hat{T}$ inféré par les scénarios PHYML+ACS+FITCH et PHYML+SDM+FITCH et l'arbre modèle T sont représentées graphiquement dans la Figure 7.3.

Pour chaque valeur de k et chaque taux de suppression de taxons, le temps d'exécution de ACS est de moins d'une seconde par collection. Les temps d'exécution de SDM+FITCH sont similaires à ceux constatés dans le cadre de la combinaison moyenne.

7.4.2 Les techniques de représentation matricielle binaire MRP et MRH

La méthode MRH

La méthode MRP consiste à encoder l'information topologique contenue dans la collection $\mathcal{C}_T$ sous la forme d'une matrice de caractères binaires, puis d'appliquer une méthode d'inférence d'arbre minimisant un critère MP. La technique de représentation matricielle binaire standard (Brooks, 1981; Baum, 1992; Doyle, 1992; Ragan, 1992) consiste, pour chaque arbre $T^p \in \mathcal{C}_T$ et chaque bipartition $A|B$ des taxons définissant $\mathcal{L}_{T^p}$, à construire un vecteur binaire associé aux taxons de $\mathcal{L}_{\mathcal{C}_T}$ où chaque taxon dans A est codé par `1` et chaque taxon dans B est codé par `0`. Les éventuels taxons dans $\mathcal{L}_{\mathcal{C}_T}$ qui ne pas présents dans $\mathcal{L}_{T^p}$ sont codés par un état de caractère manquant). La représentation matricielle binaire de $\mathcal{C}_T$ est obtenue par concaténation de tous les vecteurs binaire construits à partir de chaque arbre T^p.

Si on considère la représentation matricielle $\mathbb{MR}^p$ d'un arbre T^p, on remarque qu'elle contient en son sein l'ensemble des bipartitions des taxons de T^p. Ainsi il existe une bijection entre la topologie de T^p et $\mathbb{MR}^p$. De plus, si on calcule la matrice de distance de Hamming Hm^p_{ij} à partir de chaque paire de séquences binaires dans $\mathbb{MR}^p$, on obtient une distance additive (Hm^p_{ij}) équivalente à $(\mathcal{T}^p_{ij})$. En effet, comme on a la relation $\mathsf{Hm}^p_{ij} = \mathcal{T}^p_{ij}/\ell_p$, où ℓ_p est le nombre de sites composant la matrice binaire $\mathbb{MR}^p$, la distance (Hm^p_{ij}) représente la distance additive $(\mathcal{T}^p_{ij})$ normalisée par le nombre ℓ_p de branches dans T^p. Ainsi, chaque représentation $\mathbb{MR}^p$ et (Hm^p_{ij}) est équivalente à la topologie de T^p (Farris et al., 1970; Barthélemy and Guénoche, 1988; Lapointe et al., 2003).

La représentation matricielle binaire de $\mathcal{C}_T$ étant constituée par la concaténation des différentes matrices binaires $\mathbb{MR}^p$, une approche simple pour construire un superarbre consiste à calculer la distance de Hamming Hm_{ij} entre chaque paire de séquences binaires concaténées, *i.e.*

$$\Delta^{\mathsf{MRH}}_{ij} = \mathsf{Hm}_{ij} = \frac{1}{W_{ij}} \sum_{\substack{1\leq p\leq k \\ i,j\in\mathcal{L}_{T^p}}} \mathcal{T}^p_{ij} \qquad \text{où} \qquad W_{ij} = \sum_{\substack{1\leq p\leq k \\ i,j\in\mathcal{L}_{T^p}}} \ell_p. \tag{7.1}$$

Observant que l'équation (7.1) peut se réécrire :

$$\Delta^{\mathsf{MRH}}_{ij} = \frac{1}{W_{ij}} \sum_{\substack{1\leq p\leq k \\ i,j\in\mathcal{L}_{T^p}}} \ell_p \mathsf{Hm}^p_{ij},$$

on constate qu'elle correspond à la moyenne des différentes distances additives Hm^p_{ij} pondérée par le nombre ℓ_p de branches de chaque arbre T^p. Comme pour les méthodes ACS et SDM

utilisées en combinaison haute, cette moyenne de distances additives permet d'obtenir une estimation de la distance additive équivalente à un superarbre de la collection $\mathcal{C}_T$. La pondération par les valeurs ℓ_p est très justifiée, car plus un arbre T^p contient de taxons (et conséquemment de branches), plus la distance $\mathcal{T}_{ij}^p$ est proche du nombre de branches séparant i et j dans le superarbre. Ainsi, un estimateur Hm_{ij}^p sera d'autant plus fiable que la valeur ℓ_p est élevée.

La méthode MRH (*Matrix Representation with Hamming distance*) consiste donc à calculer la matrice de distance $(\Delta_{ij}^{\mathsf{MRH}})$, puis d'inférer un superarbre en utilisant une méthode de distance. Le calcul de $(\Delta_{ij}^{\mathsf{MRH}})$ s'effectuant en $O(kn^2)$, cette méthode peut ne nécessiter qu'une complexité polynomiale de l'ordre de $O(kn^3)$ si elle utilise les algorithmes agglomératifs NJ*, UNJ*, BIONJ* et MVR* pour l'étape d'inférence d'arbre.

Comparaison des performances de MRP et MRH

Pour chaque collection $\mathcal{C}_T$, les techniques de combinaison haute MRP et MRH ont été appliquées avec la représentation matricielle standard (Brooks, 1981; Baum, 1992; Doyle, 1992; Ragan, 1992).

Chaque superarbre MRP a été inféré par le logiciel TNT (Goloboff et al., 2003) paramétré pour effectuer 25 recherches locales par descente à l'aide de voisinages TBR et complétées par la technique du *parsimony ratchet* (Nixon, 1999).

Chaque superarbre MRH a été inféré par l'algorithme UNJ*. Toutes les branches de longueur inférieure à

$$\left(\max_{i,j \in \mathcal{L}_{\mathcal{C}_T}} \left[\sum_{\substack{1 \leq p \leq k \\ i,j \in \mathcal{L}_{T^p}}} \ell_p \right] \right)^{-1}$$

ont été transformées en multifurcation dans le superabre inféré par MRH. En effet, la longueur d'une branche représentant une estimation du nombre moyen de différences entre deux séquences binaires, une longueur représentant moins d'une différence en moyenne correspond à une multifurcation.

Pour chaque valeur de k et chaque taux de suppression de taxons, les distance topologique d_{quad} moyennes entre l'arbre $\hat{T}$ inféré par les méthodes MRP et MRH, et l'arbre modèle T sont représentées dans le Tableau 7.4.

Pour un taux de suppression de taxons de 25%, le logiciel TNT a inféré un superarbre en un temps moyen de 12 secondes pour $k = 10$, et 23 secondes pour $k = 20$. Pour 75% de taux de suppression de taxons, des durées moyennes de 7 secondes et 15 secondes ont été observées pour les mêmes valeurs respectives de k. Dans tous les cas de figure, la méthode MRH a duré moins d'une seconde par superarbre inféré.

25%						
k	FITCH	MW*	NJ*	UNJ*	BIONJ*	MVR*
2	0.0558	0.0561	0.0588	0.0550	0.0558	**0.0526**
4	***0.0337***	0.0345	0.0364	0.0352	0.0348	**0.0317**
6	***0.0252***	0.0265	0.0273	***0.0245***	0.0264	**0.0237**
8	***0.0227***	***0.0227***	**0.0211**	***0.0226***	***0.0216***	***0.0213***
10	***0.0187***	***0.0188***	0.0193	***0.0183***	***0.0190***	**0.0171**
12	***0.0197***	***0.0207***	0.0216	***0.0209***	***0.0199***	**0.0190**
14	**0.0160**	***0.0164***	***0.0162***	***0.0162***	***0.0163***	***0.0161***
16	***0.0208***	**0.0204**	***0.0209***	***0.0209***	***0.0213***	***0.0209***
18	**0.0170**	***0.0177***	***0.0175***	***0.0179***	***0.0173***	***0.0175***
20	***0.0162***	***0.0168***	***0.0171***	***0.0165***	***0.0159***	**0.0158**

75%						
k	FITCH	MW*	NJ*	UNJ*	BIONJ*	MVR*
2	0.1869	0.1877	***0.1823***	***0.1833***	***0.1817***	**0.1815**
4	0.1395	0.1397	0.1374	0.1388	0.1375	**0.1336**
6	0.1094	0.1125	0.1128	0.1110	0.1112	**0.1060**
8	0.0865	0.0891	0.0919	0.0879	0.0859	**0.0811**
10	0.0690	0.0739	0.0755	0.0710	0.0717	**0.0663**
12	0.0641	0.0670	0.0712	0.0656	0.0670	**0.0605**
14	0.0508	0.0538	0.0569	0.0527	0.0523	**0.0486**
16	0.0504	0.0518	0.0558	0.0506	0.0514	**0.0462**
18	***0.0409***	***0.0416***	0.0479	***0.0411***	0.0424	**0.0398**
20	0.0403	0.0435	0.0453	0.0425	0.0425	**0.0369**

TAB. 7.3 – **Performances des scénarios PHYML+SDM+FITCH, PHYML+SDM+MW*, PHYML+SDM+NJ*, PHYML+SDM+UNJ*, PHYML+SDM+BIONJ* et PHYML+SDM+MVR***

Les deux tableaux représentent les distances topologiques d_{quad} moyennes entre l'arbre $\hat{T}$, inféré par les six scénarios de combinaison haute, et l'arbre modèle T dans le cadre des 25% et 75% de suppression de taxons. La colonne k indique la taille de la collection $\mathcal{C}_T$. Les autres colonnes représentent la moyenne des 500 valeurs d_{quad} observées pour chaque valeur de k. Les valeurs écrites en caractères gras représentent la plus petite distance d_{quad} moyenne pour chaque ligne. Les valeurs écrites en caractères gras italiques représentent les distances d_{quad} moyennes dont la valeur est proche (± 0.002) de la plus petite valeur dans la ligne correspondante.

7.4.3 Résultats et discussion

Les approches par supermatrice de distance

Comparée à l'approche ACS, la méthode SDM présente de très bonnes performances dans le cadre de la combinaison haute. Le scénario d'inférence phylogénomique PHYML+SDM+FITCH permet presque toujours d'inférer des superarbres $\hat{T}$ contenant deux fois plus de quadruplets communs avec les arbres modèles T que le scénario PHYML+ACS+FITCH (cf Figure 7.3). Or les algorithmes agglomératifs BIONJ* et MVR* ont montré des performances similaires, voire

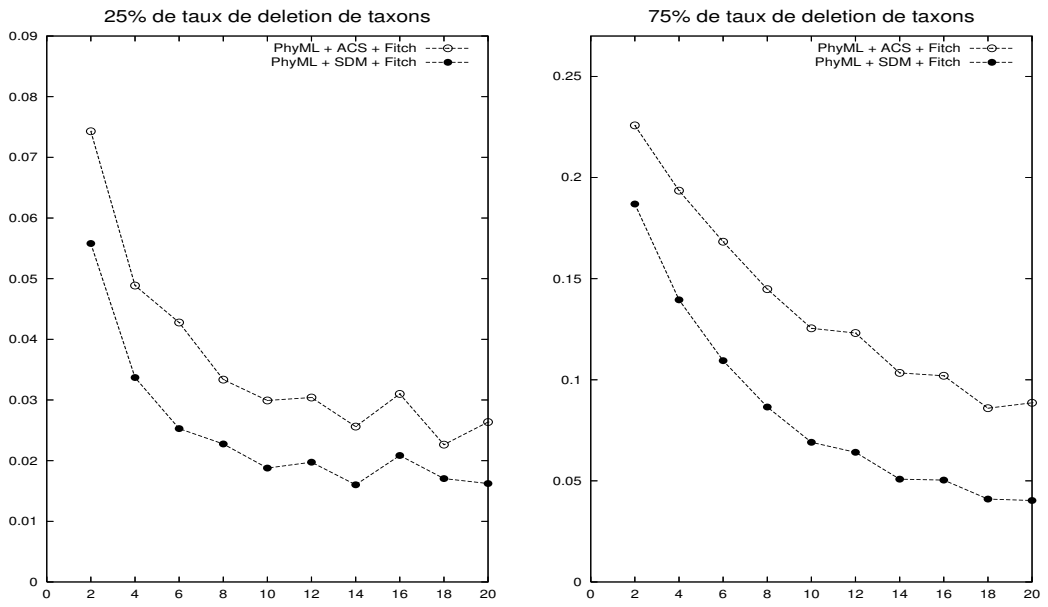

FIG. 7.3 – **Performances des scénarios PHYML+ACS+FITCH et PHYML+SDM+FITCH**

Les deux graphiques représentent les distances topologiques d_{quad} moyennes entre l'arbre $\hat{T}$, inféré par les deux scénarios de combinaison haute, et l'arbre modèle T dans le cadre des 25% et 75% de suppression de taxons. L'axe des abscisses représente les différentes valeurs de k. L'axe des ordonnées représente les distances topologiques d_{quad} moyennes observées pour les différents scénarios d'inférence phylogénomique.

meilleures, que celles de FITCH en combinaison moyenne (cf Tableau 7.1). Comme attendu, ces algorithmes agglomératifs montrent également des performances du même ordre lorsqu'ils sont appliqués sur les supermatrices de distance inférées suivant le scénario PHYML+SDM (cf Tableau 7.3), et ce avec des temps d'exécution bien plus rapides.

Plus précisément, on observe, à partir des simulations de cette section, que le scénario d'inférence phylogénomique PHYML+SDM+MVR* se révèle être l'approche la plus performante des techniques de combinaison haute utilisant les distances additives induites par les longueurs de branche des arbres composant $\mathcal{C}_T$. L'utilisation de $(\Delta_{ij}^{\text{SDM}})$, ainsi que de la variance associée aux supermatrices de distance, lors de l'étape d'inférence d'arbre, se révèle être le meilleur choix en comparaison avec tous les scénarios possibles utilisant ACS ou SDM, et NJ*, UNJ*, BIONJ*, MVR*, FITCH ou MW*.

Des alternatives à la technique standard MRP

Si on compare les résultats du scénario PHYML+SDM+MVR* dans le Tableau 7.3 et ceux de PHYML+MRP dans le Tableau 7.4, on constate que les deux approches présentent des performances proches dans le cadre des 25% de suppression de taxons. En moyenne, seul environ 1% de quadruplets incorrects sépare les superarbres inférés par PHYML+SDM+MVR* et PHYML+MRP. Pour un taux de 75% de suppression de taxons, le

25%			75%		
k	MRP	MRH	k	MRP	MRH
2	0.0660	**0.0649**	2	0.2539	**0.1928**
4	**0.0351**	0.0374	4	0.2229	**0.1487**
6	**0.0254**	0.0270	6	0.1800	**0.1212**
8	**0.0200**	***0.0214***	8	0.1438	**0.0999**
10	**0.0175**	***0.0193***	10	0.1152	**0.0826**
12	**0.0189**	***0.0208***	12	0.0920	**0.0726**
14	**0.0148**	***0.0164***	14	0.0716	**0.0604**
16	**0.0187**	***0.0205***	16	0.0666	**0.0579**
18	**0.0130**	***0.0147***	18	0.0487	**0.0460**
20	**0.0146**	***0.0163***	20	**0.0446**	***0.0460***

TAB. 7.4 – **Performances des méthodes de combinaison haute MRP et MRH sur des collections de *k* arbres de tailles similaires**

La colonne k indique le nombre d'arbres sources utilisés pour construire chacun des 500 superarbres avec les méthodes MRP et MRH. Les autres colonnes représentent la moyenne des 500 valeurs d_{quad} observées pour chaque valeur de k. Les valeurs écrites en caractères gras représentent la plus petite distance d_{quad} moyenne pour chaque ligne. Les valeurs écrites en caractères gras italiques représentent les distances d_{quad} moyennes dont la valeur est proche (± 0.002) de la plus petite valeur dans la ligne correspondante.

scénario PHYML+SDM+MVR* permet d'observer de meilleures valeurs d_{quad} moyennes que PHYML+MRP. Le taux de quadruplets erronés séparant les superarbres inférés par les scénario PHYML+SDM+MVR* et PHYML+MRP varie en moyenne entre 1% et 10%. Ainsi le scénario d'inférence phylogénomique basé sur les supermatrices de distance calculées par la méthode SDM représente une alternative méthodologique à la technique standard MRP de combinaison haute, notamment lorsque les données sont très incomplètes.

En termes de complexité algorithmique, la polynomialité de la procédure SDM+MVR*, en $O(k^3n^3)$, permet d'offir des temps d'exécution bien plus rapides que la procédure MRP. En effet, la technique standard d'inférence phylogénétique minimisant un critère MP consiste à effectuer une recherche locale par descente. Le logiciel TNT, utilisé dans le scénario PHYML+MRP pour effectuer cette recherche locale, est un des plus rapides existant actuellement. Malgré tout, la technique d'inférence de superarbre MRP demeure de deux à sept fois plus longue en pratique que la procédure SDM+MVR*. Ainsi, lorsque l'on dispose d'une collection d'arbres sources $\mathcal{C}_T$ dont les branches sont valuées suivant un modèle d'évolution, l'approche par combinaison haute par les supermatrices de distance SDM permet d'offrir un très bon rapport entre fiabilité et rapidité.

L'utilisation des méthodes de distance à partir de la représentation matricielle MR

Si les branches des arbres sources de $\mathcal{C}_T$ ne sont pas valués (absence d'un modèle d'évolution, données sources morphologiques, ...), la méthode SDM ne peut pas être utilisée. Néanmoins, la méthode MRH s'appuyant, comme MRP, sur la représentation matricielle binaire MR des topologies des arbres sources composant $\mathcal{C}_T$, elle constitue une méthode de distance pouvant traiter tout type d'arbres phylogénétiques sources. Leur complexité polynomiale, en $O(kn^3)$, en font une des approches de combinaison haute les plus rapides (moins d'une seconde pour $k = 20$ et 25% de suppression de taxons, contre 9 secondes pour SDM+MVR* et 23 secondes pour MRP). Les valeurs d_{quad} moyennes observées pour PHYML+MRH sont toujours légèrement moins bonnes que celle observées pour PHYML+SDM+MVR*, ce qui démontre l'intérêt de l'utilisation des longueurs des branches de chaque arbre source T^p. Néanmoins, les performances globales de PHYML+MRH sont toujours relativement proches de celles de PHYML+SDM+MVR*, en comparaison avec le scénario PHYML+MRP (cf Tableaux 7.3 et 7.4).

7.5 Les différents niveaux de combinaison en inférence phylogénomique

A la lumière des différents résultats de simulations présenté dans ce chapitre, un ordonnancement des différentes approches en inférence phylogénomique peut être proposé. Le Tableau 7.5 présente les scénarios ayant inféré les meilleurs arbres phylogénétiques en moyenne pour chaque taux de suppression de taxons et pour chaque niveau de combinaison.

Utilisation des méthodes de distance en inférence phylogénomique

On constate que la méthode SDM est présente dans les six cas de figure du Tableau 7.5 (*i.e.* trois niveaux de combinaison et deux taux de suppression). Les algorithmes agglomératifs BIONJ* et MVR* sont présents dans la moitié des cas de figure . Néanmoins, les arbres inférés par les scénarios de combinaison moyenne SDM+BIONJ* et SDM+FITCH étant globalement très proches, l'un est très probablement présent dans le voisinage NNI augmenté de l'autre. Les NNI augmentés étant les mouvements topologiques utlisés dans la recherche locale par descente implémentée dans le logiciel PHYML, on peut postuler que le scénario de combinaison basse SDM+BIONJ*+PHYML permet d'inférer des arbres très similaires à ceux renvoyés par le scénario SDM+FITCH+PHYML.

Ainsi, les techniques de distance SDM, BIONJ* et MVR* permettent d'améliorer les performances des trois niveaux de combinaison, quel que soit le taux d'incomplétude des données génétiques sources. Ces améliorations sont significatives aussi bien en termes de qualité des arbres inférés, qu'en termes de temps d'exécution. En effet, dans le cas de performances similaires (*e.g.* SDM+BIONJ* et SDM+FITCH dans la combinaison moyenne de données faiblement

combinaison	25%	75%
basse	**SDM+FITCH+PHYML**	**SDM+FITCH+PHYML**
moyenne	**SDM+BIONJ*** SDM+FITCH	**SDM+FITCH**
haute	**PHYML+SDM+MVR*** PHYML+MRP	**PHYML+SDM+MVR***

TAB. 7.5 – **Meilleurs scénarios d'inférence phylogénomique en simulation suivant le taux de suppression de taxons et le niveau de combinaison**

Un scénario a été sélectionné s'il présente au moins huit fois les meilleures valeurs $d_{\text{quad}} \pm 0.002$ moyennes sur les dix valeurs de k. Les scénarios représentés en caractères gras sont les plus rapides en moyenne par taux de suppression et par niveau de combinaison. On rappelle que le scénario de combinaison basse SDM+FITCH+PHYML est sujet à différentes remarques (cf notes de bas des pages 138 et 140).

délétées, ou PHYML+SDM+MVR* et PHYML+MRP dans la combinaison haute de données très incomplètes), ce sont toujours les nouvelles méthodes SDM ou BIONJ* qui permettent d'obtenir ces résultats en un temps d'exécution moindre.

Performances respectives des différents niveaux de combinaison

Pour chaque niveau de combinaison, les valeurs d_{quad} moyennes des scénarios présentant les meilleures performances en termes de fiabilité de l'arbre inféré et de temps d'exécution (*i.e.* en caractères gras dans le Tableau 7.5) ont été représentées graphiquement dans la Figure 7.4.

On observe que la combinaison moyenne infère les arbres de moins bonne qualité dans l'ensemble des simulations. Néanmoins, les temps d'exécution des méthodes SDM et BIONJ*, parmi les plus rapides, en font des points de départ de choix pour les recherches locales cherchant à minimier des critères impliquant de grandes quantités de calcul (*e.g.* ML). Conséquemment, on observe que le scénario de combinaison basse SDM+FITCH+PHYML, utilisant un arbre de combinaison moyenne comme point de départ, présente les meilleures valeurs d_{quad} moyennes dans l'ensemble des simulations. La combinaison haute se situe à un niveau intermédiaire entre les performances des méthodes de combinaison basse et moyenne.

Il est intéressant de constater que ces différents scénarios s'inter-alimentent pour offrir les meilleurs arbres. Ainsi la méthode SDM, initialement conçue comme technique de combinaison moyenne à partir de matrices de distance, permet d'offrir de meilleurs résultats en combinaison haute, grâce à l'inférence phylogénétique d'arbres sources par minimisation d'un critère ML. D'un autre coté, la recherche locale minimisant ce critère ML en combinaison basse est fortement améliorée par l'utilisation de SDM dans le cadre de la combinaison moyenne. Ainsi, une

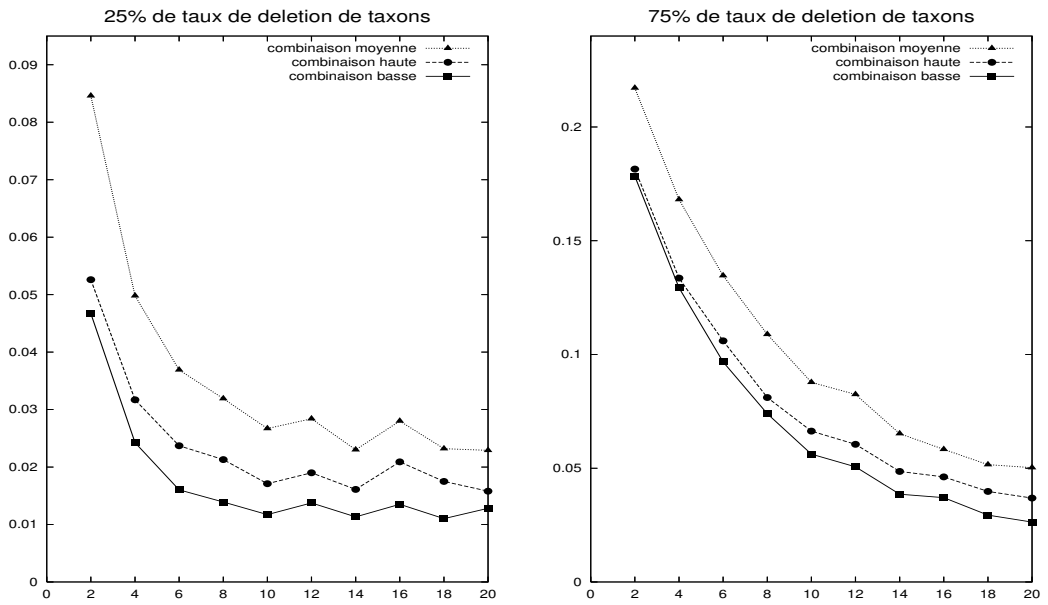

FIG. 7.4 – **Performances des meilleurs scénarios de combinaison moyenne, haute et basse**
Les deux graphiques représentent les distances topologiques d_{quad} moyennes entre l'arbre $\hat{T}$ inféré par combinaison moyenne, haute ou basse, et l'arbre modèle T dans le cadre des 25% et 75% de suppression de taxons. L'arbre $\hat{T}$ a été inféré par les scénarios SDM+BIONJ*, PHYML+SDM+MVR* et SDM+FITCH+PHYML dans le cadre des 25% de suppression de taxons, et par les scénarios SDM+FITCH, PHYML+SDM+MVR* et SDM+FITCH+PHYML dans le cadre des 75% de suppression de taxons. L'axe des abscisses représente les différentes valeurs de k. L'axe des ordonnées représente les distances topologiques d_{quad} moyennes observées pour les différents scénarios d'inférence phylogénomique. On rappelle que le scénario de combinaison basse SDM+FITCH+PHYML est sujet à différentes remarques (cf notes de bas des pages 138 et 140).

utilisation pertinente et complémentaire de différents scénarios de combinaison, chacun s'appuyant sur différents critères, permet une inférence phylogénomique de qualité, et ce, avec des temps d'exécution relativement raisonables.

Néanmoins, si la taille des données sources le permet ou si les temps d'exécution ne représentent pas une contrainte qualitative, une lecture attentive de l'ensemble des simulations de ce chapitre laisse présager qu'un scénario assemblant une méthode de combinaison haute et une inférence ML par combinaison basse doit conduire aux meilleurs arbres possibles.

Chapitre 8

Conclusions et perspectives

> Toutes choses sont dites déjà ; mais comme personne n'écoute, il faut toujours recommencer.
>
> *André Gide*

Cette thèse a permis de développer plusieurs méthodes s'inscrivant dans la thématique récente de l'inférence phylogénomique. Ce type d'inférence, cherchant à construire des arbres phylogénétiques à partir de vastes collections de gènes, est souvent limité en pratique par les problèmes de temps de calcul et d'espace mémoire importants impliqués par les très grands jeux de données considérés.

Initialement développée pour effectuer la combinaison moyenne d'une collection de gènes, la méthode SDM (*Super Distance Matrix* ; Criscuolo *et al.*, 2006) permet d'estimer la vitesse d'évolution relative de chaque gène avec une complexité polynomiale rendant ce calcul très rapide en pratique. La méthode SDM permet également de calculer une supermatrice de distance contenant l'information phylogénétique induite par chaque gène source. L'application d'un algorithme d'inférence d'arbre sur cette supermatrice de distance permet ainsi de produire un arbre relativement fiable représentant l'histoire évolutive de l'ensemble des taxons contenus dans la collection de gènes sources. Cette technique de distance permet d'utiliser la méthode SDM dans le cadre de la combinaison moyenne (en utilisant les matrices de distance directement estimées à partir de chaque gène source), mais également dans des approches par combinaison haute (en utilisant des arbres phylogénétiques inférés à partir de chaque gène source).

L'apparition d'importantes zones de données manquantes représente un autre problème pratique de l'inférence phylogénomique. Ce problème se traduit par l'apparition de distances manquantes dans les supermatrices de distance calculées par la méthode SDM, en particulier lorsque les gènes sources présentent peu de taxons en communs. Différents algorithmes, nommés NJ*, UNJ*, BIONJ* et MVR* (Criscuolo, 2006), ont été développés afin de pouvoir construire des arbres de bonne qualité malgré la présence de ces distances manquantes. Ces algorithmes ont en commun une faible complexité algorithmique, mais diffèrent par les critères qu'ils cherchent à optimiser (*e.g.* considération de la variance associée à chaque valeur existante

dans la supermatrice de distance).

D'importants protocoles de simulations ont été mis en oeuvre afin d'observer les performances des différents scénarios d'inférence phylogénomique envisageables par l'utilisation conjointe de la méthode SDM et des algorithmes NJ*, UNJ*, BIONJ* et MVR*. D'autres méthodes et algorithmes existants ont également été utilisés durant ces simulations et ont permis d'observer les très bonnes performances relatives des techniques développées durant cette thèse, aussi bien en termes de temps d'exécution que de fiabilité des arbres inférés. Ces simulations ont utilisé plusieurs approches phylogénomiques par combinaison basse, moyenne et haute, et ont permis de démontrer l'utilité des techniques de distance dans les trois catégories. Elles ont montré entre autre que les méthodes de distance constituaient des alternatives efficaces à certaines techniques standards (*e.g.* la méthode de combinaison haute MRP), ou des approches complémentaires permettant d'optimiser la fiabilité et les temps d'exécution d'approches plus lourdes en temps de calcul (*e.g.* combinaison basse avec critère ML). Enfin, ces simulations représentent un premier support pratique dans le grand débat scientifique opposant les partisans de l'approche par combinaison basse (*supermatrix*) et ceux de la combinaison haute (*supertree*).

Les travaux menés durant cette thèse peuvent être poursuivis dans plusieurs directions complémentaires.

L'inférence phylogénomique par combinaison basse en utilisant le critère ML a permis de construire les arbres les plus fiables dans les simulations présentées. Ce résultat a pu être obtenu en utilisant le logiciel PHYML et en utilisant comme point de départ un arbre inféré par combinaison moyenne grâce à la méthode SDM. Or il a été montré qu'associer un paramètre modélisant la vitesse d'évolution de chaque gène, lors de l'analyse simultanée d'une supermatrice de caractères, permettait d'améliorer la qualité des arbres inférés (Pupko et al., 2002; Bevan et al., 2006). La méthode SDM permettant d'estimer très rapidement chacun de ces paramètres, leur utilisation dans les calculs de vraisemblance, associée à un arbre de départ inféré par le scénario SDM+BIONJ* ou SDM+MVR*, ainsi qu'à la recherche locale implémentée dans PHYML permettrait d'inférer des arbres de très bonne qualité, sans augmenter significativement le temps d'exécution et l'espace mémoire nécessaire.

L'inférence phylogénomique est très souvent liée à la notion d'incongruence de gènes. Utilisée en combinaison moyenne, la méthode SDM permet d'inférer rapidement une supermatrice de distance représentant l'information phylogénétique induite par chaque matrice de distance source. Ainsi, en comparant la supermatrice de distance et chaque matrice source, plusieurs tests d'incongruence de gènes ont été envisagés. Ces tests sont d'autant plus utiles en pratique qu'ils bénéficient de la rapidité d'exécution de la méthode SDM.

Les algorithmes agglomératifs présentés dans cette thèse s'appuient sur l'idée d'un filtrage des paires potentielles de taxons à agglomérer. Ainsi chacune des s meilleures paires de taxons

sélectionnées par le critère Q^*_{xy} sont ensuite testées par le critère plus précis $\tilde{N}^*_{xy}/|\tilde{C}^*_{xy}|$. Or ce dernier critère peut sélectionner plusieurs paires potentielles de taxons à agglomérer. A cette étape, plusieurs nouveaux critères peuvent être utilisés aussi bien en inférence phylogénomique que phylogénétique, comme par exemple sélectionner la paire xy qui, agglomérée au noeud interne u, permettra d'obtenir l'estimation T_{xu} la plus fiable (*e.g.* minimisant la variance des estimateurs de T_{xu}). Cette idée peut également être exploitée dans la méthode MRH, où, parmi les paires xy potentielles à agglomérer, on choisirait celle qui contredit le moins les arbres sources.

Les algorithmes d'inférence d'arbre à partir d'une matrice de distance sont toujours définis par plusieurs équations mathématiques. Dans cette thèse, les algorithmes inférant un arbre à partir d'une matrice de distance incomplète s'appuient sur l'interprétation de chaque équation sous la forme d'une moyenne. Suivant ce principe, plusieurs autres algorithmes de distance (*e.g.* FASTME) pourraient également être transformés afin de les rendre utilisables sur des matrices de distance complètes ou non.

Annexe A

Annexes du Chapitre 5

Sommaire

A.1 Preuve d'invariance topologique des algorithmes ADDTREE, NJ, UNJ, BIONJ et MVR à certaines déformations de la distance (Δ_{ij}) **159**

A.1.1 Invariance topologique à la multiplication par un facteur 159

A.1.2 Invariance topologique à l'ajout d'une constante 161

A.2 Calcul des dérivées partielles de $f(v)$ **163**

A.1 Preuve d'invariance topologique des algorithmes ADDTREE, NJ, UNJ, BIONJ et MVR à certaines déformations de la distance (Δ_{ij})

Cette partie démontre que la classe des algorithmes agglomératifs de la Figure 3.3 renvoie la même topologie d'arbre lorsqu'elle est appliquée sur les distances (Δ_{ij}), $(\alpha\Delta_{ij})$ et $(\Delta_{ij} + a_i + a_j)$.

A.1.1 Invariance topologique à la multiplication par un facteur

Soit $(\tilde{\Delta}_{ij}) = (\alpha\Delta_{ij})$ la matrice de distance (Δ_{ij}) déformée par la multiplication par un facteur $\alpha > 0$.

Critères d'agglomération

Le critère d'agglomération N_{xy} défini par la Formule (3.13) et appliqué à la matrice de distance $(\tilde{\Delta}_{ij})$:

$$\tilde{N}_{xy} = \sum_{i,j \in \mathcal{L}_r} H\big(\tilde{\Delta}_{ix} + \tilde{\Delta}_{jy} - \tilde{\Delta}_{xy} - \tilde{\Delta}_{ij}\big) H\big(\tilde{\Delta}_{iy} + \tilde{\Delta}_{jx} - \tilde{\Delta}_{xy} - \tilde{\Delta}_{ij}\big),$$

se réécrit plus simplement :

$$\tilde{N}_{xy} = \sum_{i,j \in \mathcal{L}_r} H\big(\alpha(\Delta_{ix} + \Delta_{jy} - \Delta_{xy} - \Delta_{ij})\big) H\big(\alpha(\Delta_{iy} + \Delta_{jx} - \Delta_{xy} - \Delta_{ij})\big).$$

Or, si $t \geq 0$, alors $\alpha t \geq 0$, car $\alpha > 0$; de même, si $t < 0$, alors $\alpha t < 0$. Ainsi,

$$H\big(\alpha(\Delta_{ix} + \Delta_{jy} - \Delta_{xy} - \Delta_{ij})\big) = H\big(\Delta_{ix} + \Delta_{jy} - \Delta_{xy} - \Delta_{ij}\big),$$

et, par conséquent, $\tilde{N}_{xy} = N_{xy}$, pour tout $x, y \in \mathcal{L}_r$. La maximisation des critères $\tilde{N}_{xy}$ et N_{xy} est donc équivalente.

Le critère d'agglomération Q_{xy} défini par la Formule (3.14) et appliqué à la matrice de distance $(\tilde{\Delta}_{ij})$:

$$\tilde{Q}_{xy} \;=\; \tilde{R}_x \;+\; \tilde{R}_y \;-\; (r-2)\tilde{\Delta}_{xy},$$

s'appuie sur les fonctions $\tilde{R}_z = \sum_{i \in \mathcal{L}_r} \tilde{\Delta}_{zi} = \sum_{i \in \mathcal{L}_r} \alpha \Delta_{zi} = \alpha \sum_{i \in \mathcal{L}_r} \Delta_{zi}$, pour tout $z \in \mathcal{L}_r$. Ainsi, on a

$$\tilde{Q}_{xy} \;=\; \alpha\Big(R_x \;+\; R_y \;-\; (r-2)\Delta_{xy}\Big) \;=\; \alpha Q_{xy}.$$

Comme $\alpha > 0$ est une constante, maximiser $\tilde{Q}_{xy}$ revient à maximiser Q_{xy}.

Valuation des branches externes

La classe des fonctions de valuation des branches externes définie par la Formule (3.17) et appliquée à la matrice de distance $(\tilde{\Delta}_{ij})$:

$$\tilde{T}_{xu} = \frac{1}{2}\tilde{\Delta}_{xy} + \sum_{i \in \mathcal{L}_r - \{x,y\}} \tilde{w}_i(\tilde{\Delta}_{xi} - \tilde{\Delta}_{yi}) = \alpha \left(\frac{1}{2}\Delta_{xy} + \sum_{i \in \mathcal{L}_r - \{x,y\}} w_i(\Delta_{xi} - \Delta_{yi}) \right) = \alpha T_{xu}$$

permet de multiplier la branche externe de longueur T_{xu} par le facteur α si et seulement si on a $\tilde{w}_i = w_i$. Or, $\tilde{w}_i = w$ est une constante pour NJ et BIONJ, et $\tilde{w}_i$ ne dépend pas de (Δ_{ij}) pour UNJ. Quant à MVR, on a, d'après la Formule (3.20) :

$$\tilde{w}_i = \frac{\tilde{\mu}}{\tilde{V}_{xi} + \tilde{V}_{yi}} \qquad \text{où} \qquad \tilde{\mu} = \frac{1}{2} \left(\sum_{j \in \mathcal{L}_r - \{x,y\}} \frac{1}{\tilde{V}_{xj} + \tilde{V}_{yj}} \right)^{-1}.$$

Sachant que $\tilde{V}_{ij} = \alpha^2 V_{ij}$, on a $\tilde{\mu} = \alpha^2 \mu$ et donc $\tilde{w}_i = \alpha^2 \mu / (\alpha^2 V_{xi} + \alpha^2 V_{yi}) = w_i$. Ainsi, les algorithmes NJ, UNJ, BIONJ et MVR multiplient la branche externe de longueur T_{xu} par le facteur α. Ce résultat s'applique par symétrie à $\tilde{T}_{yu} = \tilde{\Delta}_{xy} - \tilde{T}_{xu}$.

Réduction matricielle

La classe des fonctions de réduction matricielle définie par la Formule (3.18) et appliquée à la matrice de distance $(\tilde{\Delta}_{ij})$, s'écrit :

$$\tilde{\Delta}_{ui} = \tilde{\lambda}_i\tilde{\Delta}_{xi} + (1-\tilde{\lambda}_i)\tilde{\Delta}_{yi} - \tilde{\lambda}_i\tilde{T}_{xu} - (1-\lambda_i)\tilde{T}_{yu}.$$

Or, sachant que $(\tilde{\Delta}_{ij}) = \alpha(\Delta_{ij})$ et que, pour les algorithmes NJ, UNJ, BIONJ et MVR, on a $\tilde{T}_{xu} = \alpha T_{xu}$ et $\tilde{T}_{yu} = \alpha T_{yu}$, alors

$$\tilde{\Delta}_{ui} = \alpha\Big(\tilde{\lambda}_i\Delta_{xi} + (1-\tilde{\lambda}_i)\Delta_{yi} - \tilde{\lambda}_i T_{xu} - (1-\tilde{\lambda}_i)T_{yu} \Big).$$

Comme on montre facilement que $\tilde{\lambda}_i = \lambda_i$ pour NJ, UNJ, BIONJ et MVR, on a donc bien $\tilde{\Delta}_{ui} = \alpha\Delta_{ui}$ pour ces quatre algorithmes. Les formules de réduction des matrices de variance renvoient $\tilde{V}_{ui} = \alpha^2 V_{ij}$ pour les algorithmes BIONJ et MVR.

Sachant que

- le critère d'agglomération $\tilde{Q}_{xy}$ (resp. $\tilde{N}_{xy}$) sélectionne la même paire de taxons que Q_{xy} (resp. N_{xy}),
- les formules de valuation de branches externes calculent $\tilde{T}_{xu} = \alpha T_{xu}$ et $\tilde{T}_{yu} = \alpha T_{yu}$, si $\tilde{w}_i = w_i$, et
- les formules de réduction calculent $\tilde{\Delta}_{ui} = \alpha\Delta_{ui}$, si $\tilde{\lambda}_i = \lambda_i$, pour tout $i \neq x, y$,

alors les algorithmes NJ, UNJ, BIONJ et MVR appliqués sur (Δ_{ij}) et $(\tilde{\Delta}_{ij})$ renvoient les arbres T et $\tilde{T}$, respectivement, avec $T = \tilde{T}$ et $(T_{ij}) = (\alpha\tilde{T}_{ij})$.

A.1.2 Invariance topologique à l'ajout d'une constante

Soit $(\tilde{\Delta}_{ij}) = (\Delta_{ij} + a_{\hat{i}})$ la matrice de distance (Δ_{ij}) déformée par l'ajout d'une constante $a_{\hat{i}}$ à chaque valeur non diagonale correspondant au taxon $\hat{i}$ dans (Δ_{ij}).

Critères d'agglomération

Comme $\tilde{\Delta}_{\hat{i}x} - \tilde{\Delta}_{\hat{i}j} = \Delta_{\hat{i}x} + a_{\hat{i}} - \Delta_{\hat{i}j} - a_{\hat{i}} = \Delta_{\hat{i}x} - \Delta_{\hat{i}j}$, alors on a $\tilde{N}_{xy} = N_{xy}$, pour tout $x, y \in \mathcal{L}_r$. La maximisation des critères $\tilde{N}_{xy}$ et N_{xy} est donc équivalente.

Sachant que

$$\begin{cases} \tilde{R}_z = \sum_{i\in\mathcal{L}_r}\tilde{\Delta}_{zi} = a_{\hat{i}} + \sum_{i\in\mathcal{L}_r}\Delta_{zi} = a_{\hat{i}} + R_z & \text{si } z \neq \hat{i} \\ \tilde{R}_{\hat{i}} = \sum_{i\in\mathcal{L}_r}\tilde{\Delta}_{\hat{i}i} = (r-1)a_{\hat{i}} + \sum_{i\in\mathcal{L}_r}\Delta_{\hat{i}i} = (r-1)a_{\hat{i}} + R_{\hat{i}i}, \end{cases}$$

alors on a

$$\begin{cases} \tilde{Q}_{xy} = Q_{xy} + 2a_{\hat{i}} & \text{si } x, y \neq \hat{i} \\ \tilde{Q}_{x\hat{i}} = a_{\hat{i}} + R_x + (r-1)a_{\hat{i}} + R_{\hat{i}} - (r-2)(\Delta_{x\hat{i}} + a_{\hat{i}}) = Q_{x\hat{i}} + 2a_{\hat{i}}. \end{cases}$$

Ainsi $\tilde{Q}_{xy} = Q_{xy} + 2a_{\hat{\imath}}$, pour tout $x, y \in \mathcal{L}_r$, et la maximisation des critères $\tilde{Q}_{xy}$ et Q_{xy} est équivalente.

Valuation des branches externes

La classe des fonctions de valuation des branches externes définie par la Formule (3.17) et appliquée à la matrice de distance $(\tilde{\Delta}_{ij})$ se réécrit :

$$\tilde{T}_{xu} \;=\; \frac{1}{2}\Delta_{xy} + \sum_{i \in \mathcal{L}_r - \{x,y\}} \tilde{w}_i(\Delta_{xi} - \Delta_{yi}) + \tilde{w}_{\hat{\imath}}(a_{\hat{\imath}} - a_{\hat{\imath}}) \;=\; T_{xu},$$

si $x \neq \hat{\imath}$, et $\tilde{T}_{\hat{\imath}u} \;=\; T_{\hat{\imath}u} + a_{\hat{\imath}}$, si et seulement si on a $\tilde{w}_i = w_i$, ce qui est le cas pour NJ, UNJ et BIONJ. L'algorithme MVR vérifie également cette propriété, sachant que $\tilde{V}_{ij} = V_{ij}$, si $i, j \neq \hat{\imath}$, et que $\tilde{V}_{\hat{\imath}j} = \mathrm{Var}(\Delta_{\hat{\imath}j} + a_{\hat{\imath}}) = V_{\hat{\imath}j}$. Ainsi, les algorithmes NJ, UNJ, BIONJ et MVR ajoutent $a_{\hat{\imath}}$ à la branche externe de longueur $T_{\hat{\imath}u}$. Ce résultat s'applique par symétrie à $\tilde{T}_{yu} = \tilde{\Delta}_{xy} - \tilde{T}_{xu}$.

Réduction matricielle

La classe des fonctions de réduction matricielle définie par la Formule (3.18) et appliquée à la matrice de distance $(\tilde{\Delta}_{ij})$, se réécrit :

$$\tilde{\Delta}_{ui} = \tilde{\lambda}_i\Delta_{xi} + (1 - \tilde{\lambda}_i)\Delta_{yi} - \tilde{\lambda}_i T_{xu} - (1 - \tilde{\lambda}_i)T_{yu},$$

si $x, y \neq \hat{\imath}$ est la paire de taxons agglomérée, et

$$\tilde{\Delta}_{ui} = \tilde{\lambda}_i(\Delta_{\hat{\imath}i} + a_{\hat{\imath}}) + (1 - \tilde{\lambda}_i)\Delta_{xi} - \tilde{\lambda}_i(T_{\hat{\imath}u} + a_{\hat{\imath}}) - (1 - \tilde{\lambda}_i)T_{xu},$$

si $\hat{\imath}, x$ est la paire de taxons agglomérée. Ainsi, $\tilde{\Delta}_{ui} = \Delta_{ui}$ si et seulement si $\tilde{\lambda}_i = \lambda_i$. Cette propriété est trivialement vérifiée par les algorithmes NJ, UNJ, BIONJ et MVR.

Sachant que

- le critère d'agglomération $\tilde{Q}_{xy}$ (resp. $\tilde{N}_{xy}$) sélectionne la même paire de taxons que Q_{xy} (resp. N_{xy}),
- les formules de valuation de branches externes calculent $\tilde{T}_{xu} = T_{xu}$ et $\tilde{T}_{\hat{\imath}u} = T_{\hat{\imath}u} + a_{\hat{\imath}}$, si $\tilde{w}_i = w_i$, et
- les formules de réduction calculent $\tilde{\Delta}_{ui} = \Delta_{ui}$, si $\tilde{\lambda}_i = \lambda_i$, pour tout $i \neq x, y$,

alors les algorithmes NJ, UNJ, BIONJ et MVR appliqués sur (Δ_{ij}) et $(\tilde{\Delta}_{ij})$ renvoient les arbres T et $\tilde{T}$, respectivement, avec $T = \tilde{T}$ et $\tilde{T}_{\hat{\imath}i} = T_{\hat{\imath}i} + a_{\hat{\imath}}$.

A.2 Calcul des dérivées partielles de $f(v)$

Si

$$f(v) = \sum_{\substack{i,j \in \mathcal{L}_{\mathcal{C}(\Delta)} \\ i \neq j \\ k_{ij} \geq 2}} \sum_{\substack{1 \leq p \leq k \\ i,j \in \tilde{\mathcal{L}}_{\Delta^p}}} w_p \Big(\alpha_p \Delta_{ij}^p + a_{ip} + a_{jp} - \overline{\Delta}_{ij} \Big)^2,$$

avec $v = (\alpha_1, ..., \alpha_p, ..., \alpha_k, ..., a_{ip}, ...)$ et

$$\overline{\Delta}_{ij} = \frac{1}{W_{ij}} \sum_{\substack{1 \leq p \leq k \\ i,j \in \tilde{\mathcal{L}}_{\Delta^p}}} w_p (\alpha_p \Delta_{ij}^p + a_{ip} + a_{jp}) \quad \text{où} \quad W_{ij} = \sum_{\substack{1 \leq p \leq k \\ i,j \in \tilde{\mathcal{L}}_{\Delta^p}}} w_p,$$

alors

$$\frac{\partial}{\partial \alpha_m} f(v) = 2 \sum_{\substack{i,j \in \tilde{\mathcal{L}}_{\Delta^m} \\ i \neq j}} \Bigg[w_m \left(\Delta_{ij}^m - \frac{w_m}{W_{ij}} \Delta_{ij}^m \right) \left(\alpha_m \Delta_{ij}^m + a_{im} + a_{jm} - \overline{\Delta}_{ij} \right) + \sum_{\substack{1 \leq p \leq k \\ p \neq m \\ i,j \in \tilde{\mathcal{L}}_{\Delta^p}}} w_p \left(-\frac{w_m}{W_{ij}} \Delta_{ij}^m \right) \left(\alpha_p \Delta_{ij}^p + a_{ip} + a_{jp} - \overline{\Delta}_{ij} \right) \Bigg]$$

$$= 2 w_m \sum_{\substack{i,j \in \tilde{\mathcal{L}}_{\Delta^m} \\ i \neq j}} \Delta_{ij}^m \Bigg[\alpha_m \Delta_{ij}^m + a_{im} + a_{jm} - \overline{\Delta}_{ij} - \frac{1}{W_{ij}} \sum_{\substack{1 \leq p \leq k \\ i,j \in \tilde{\mathcal{L}}_{\Delta^p}}} w_p \Big(\alpha_p \Delta_{ij}^p + a_{ip} + a_{jp} - \overline{\Delta}_{ij} \Big) \Bigg]$$

$$= 2 w_m \sum_{\substack{i,j \in \tilde{\mathcal{L}}_{\Delta^m} \\ i \neq j}} \Delta_{ij}^m \Big(\alpha_m \Delta_{ij}^m + a_{im} + a_{jm} - \overline{\Delta}_{ij} \Big),$$

et

$$\frac{\partial}{\partial a_{im}} f(v) = 4 \sum_{\substack{i,j \in \tilde{\mathcal{L}}_{\Delta^m} \\ j \neq i}} \Bigg[w_m \left(1 - \frac{w_m}{W_{ij}}\right) \left(\alpha_m \Delta_{ij}^m + a_{im} + a_{jm} - \overline{\Delta}_{ij}\right) + \sum_{\substack{1 \leq p \leq k \\ p \neq m \\ i,j \in \tilde{\mathcal{L}}_{\Delta^p}}} w_p \left(-\frac{w_m}{W_{ij}}\right) \left(\alpha_p \Delta_{ij}^p + a_{ip} + a_{jp} - \overline{\Delta}_{ij}\right) \Bigg]$$

$$= 4 w_m \sum_{j \in \tilde{\mathcal{L}}_{\Delta^m} - \{i\}} \Bigg[\alpha_m \Delta_{ij}^m + a_{im} + a_{jm} \; - \; \overline{\Delta}_{ij} - \frac{1}{W_{ij}} \sum_{\substack{1 \leq p \leq k \\ i,j \in \tilde{\mathcal{L}}_{\Delta^p}}} w_p \Big(\alpha_p \Delta_{ij}^p + a_{ip} + a_{jp} - \overline{\Delta}_{ij}\Big) \Bigg]$$

$$= 4 w_m \sum_{j \in \tilde{\mathcal{L}}_{\Delta^m} - \{i\}} \Big(\alpha_m \Delta_{ij}^m + a_{im} + a_{jm} \; - \; \overline{\Delta}_{ij}\Big).$$

Annexe B

Annexes du Chapitre 6

Réécriture des critères Q^*_{xy} et $\tilde{Q}^*_{xy}$

Le critère d'agglomération Q^*_{xy}

Si on considère les ensembles

$$S^*_{xy} = \big\{ i \in \mathcal{L}_r : \Delta_{xi}, \Delta_{yi} \neq \varnothing \big\} \qquad \text{et} \qquad \hat{S}^*_{xy} = \big\{ i \in \mathcal{L}_r - \{x, y\} : \Delta_{xi}, \Delta_{yi} \neq \varnothing \big\},$$

alors on observe que

$$S^*_{xy} = \hat{S}^*_{xy} \sqcup \{x, y\} \qquad \text{et} \qquad |S^*_{xy}| = |\hat{S}^*_{xy}| + 2.$$

Ainsi on obtient

$$R^*_{xy} \;=\; \sum_{i \in S^*_{xy}} \Big(\Delta_{xi} + \Delta_{yi} \Big) \;=\; \sum_{i \in \hat{S}^*_{xy}} \Big(\Delta_{xi} + \Delta_{yi} \Big) + 2\Delta_{xy},$$

et la fonction

$$Q^*_{xy} \;=\; \frac{2}{|\hat{S}^*_{xy}|} \Delta_{xy} \;+\; \frac{1}{|\hat{S}^*_{xy}|} \sum_{i \in \hat{S}^*_{xy}} \Big(\Delta_{xi} + \Delta_{yi} - \Delta_{xy} \Big)$$

qui se factorise de la manière suivante :

$$Q^*_{xy} \;=\; \left(\frac{2}{|\hat{S}^*_{xy}|} - 1 \right) \Delta_{xy} \;+\; \frac{1}{|\hat{S}^*_{xy}|} \sum_{i \in \hat{S}^*_{xy}} \Big(\Delta_{xi} + \Delta_{yi} \Big),$$

peut se réécrire

$$Q^*_{xy} \;=\; \left(\frac{2}{|S^*_{xy}| - 2} - 1 - \frac{2}{|S^*_{xy}| - 2} \right) \Delta_{xy} \;+\; \frac{1}{|S^*_{xy}| - 2} R^*_{xy},$$

ce qui donne

$$Q^*_{xy} \;=\; \frac{R^*_{xy}}{|S^*_{xy}| - 2} \;-\; \Delta_{xy}.$$

Le critère d'agglomération $\tilde{Q}^*_{xy}$

D'une manière similaire à Q^*_{xy}, la fonction

$$\tilde{Q}^*_{xy} \;=\; \frac{2}{r-2}\Delta_{xy} \;+\; \frac{1}{|\hat{S}^*_{xy}|}\sum_{i\in\hat{S}^*_{xy}}\Big(\Delta_{xi}+\Delta_{yi}-\Delta_{xy}\Big).$$

se réécrit

$$\tilde{Q}^*_{xy} \;=\; \left(\frac{2}{r-2}-1-\frac{2}{|S^*_{xy}|-2}\right)\Delta_{xy} \;+\; \frac{R^*_{xy}}{|S^*_{xy}|-2},$$

ce qui se simplifie en

$$\tilde{Q}^*_{xy} \;=\; \frac{R^*_{xy}}{|S^*_{xy}|-2} \;-\; \left(\frac{r-4}{r-2}+\frac{2}{|S^*_{xy}|-2}\right)\Delta_{xy}.$$

Annexe C

Annexes du Chapitre 7

Sommaire

C.1 Calcul de la variance d'une distance estimée par complétion 167

C.1.1 Complétion additive 167

C.1.2 Complétion par quadruplets 169

C.1 Calcul de la variance d'une distance estimée par complétion

C.1.1 Complétion additive

Si la distance inconnue $\Delta_{xy} = \varnothing$ est estimée par complétion additive avec la Formule (3.4) (cf page 56), alors il existe une paire de taxons $i, j \in \tilde{C}^*_{xy}$ avec

$$\tilde{C}^*_{xy} = \big\{(i,j) \in \mathcal{L}_r \times \mathcal{L}_r : i \neq x,y,j\ ,\ j \neq x,y,i\ ,\ \Delta_{ix}, \Delta_{jy}, \Delta_{xy}, \Delta_{ij} \neq \varnothing\big\},$$

telle que

$$\Delta_{xy} \ = \ \Delta_{xi} + \Delta_{yj} - \Delta_{ij}.$$

Ainsi, on a

$$V_{xy} \ = \ \Big(V_{xi} + V_{yj} + 2\mathsf{Cov}(\Delta_{xi}, \Delta_{yj})\Big) + V_{ij} - 2\mathsf{Cov}\big(\Delta_{xi} + \Delta_{yj}, \Delta_{ij}\big).$$

Sachant que $\mathsf{Cov}(X+Y, Z) = \mathsf{Cov}(X,Z) + \mathsf{Cov}(Y,Z)$, on a alors :

$$V_{xy} \ = \ V_{xi} + V_{yj} + V_{ij} + 2\Big(\mathsf{Cov}(\Delta_{xi}, \Delta_{yj}) - \mathsf{Cov}(\Delta_{xi}, \Delta_{ij}) - \mathsf{Cov}(\Delta_{yj}, \Delta_{ij})\Big). \quad \text{(C.1)}$$

Le modèle de variance-covariance de BIONJ et BIONJ*

Si on considère le modèle de variance-covariance de BIONJ, alors, pour toute paire de taxons distincts ij, on a $V_{ij} = \Delta_{ij}$. Quant à $\mathrm{Cov}(\Delta_{xi}, \Delta_{yj})$, pour tout quadruplet de taxons $ijxy$, sa valeur correspond à la longueur du chemin formant l'intersection entre les chemins $\{x, i\}$ et $\{y, j\}$, si celui-ci existe ; sinon on a $\mathrm{Cov}(\Delta_{xi}, \Delta_{yj}) = 0$ (Nei and Jin, 1989; Bulmer, 1991; Gascuel, 1997a). Suivant ce modèle, on a donc

$$\mathrm{Cov}(\Delta_{xi}, \Delta_{ij}) = \frac{1}{2}\Big(\Delta_{xi} + \Delta_{ij} - \Delta_{xj}\Big), \;\; \mathrm{Cov}(\Delta_{yj}, \Delta_{ij}) = \frac{1}{2}\Big(\Delta_{yj} + \Delta_{ij} - \Delta_{yi}\Big)$$

et

$$\mathrm{Cov}(\Delta_{xi}, \Delta_{yj}) = \frac{1}{8}\Bigg(\Delta_{xy}+\Delta_{xi}+\Delta_{yj}+\Delta_{ij}-2\Delta_{xj}-2\Delta_{yi} + \Big|\Delta_{xy} + \Delta_{xi} + \Delta_{yj} + \Delta_{ij} - 2\Delta_{xj} - 2\Delta_{yi}\Big|\Bigg).$$

Ce dernier calcul de covariance utilise l'inégalité quadrangulaire pour estimer la longueur du chemin formant l'intersection entre les chemins $\{x, i\}$ et $\{y, j\}$. La valeur absolue sert à rendre la covariance nulle si l'estimation de la longueur de ce chemin est négative (*i.e.* l'intersection entre les chemins $\{x, i\}$ et $\{y, j\}$ est vide). On obtient ainsi, après calculs, à partir de la Formule (C.1) :

$$V_{xy} \;=\; \frac{1}{4}\Bigg[\Delta_{xy} + \Delta_{xi} - 3\Delta_{yj} - 3\Delta_{ij} + 6\Delta_{xj} + 2\Delta_{yi} + \Big|\Delta_{xy} + \Delta_{xi} + \Delta_{yj} + \Delta_{ij} - 2\Delta_{xj} - 2\Delta_{yi}\Big|\Bigg].$$

Le modèle de variance de MVR et MVR*

Si on considère le modèle de variance de MVR, alors, pour tout quadruplet de taxons distincts $ijxy$, on a $\mathrm{Cov}(\Delta_{xi}, \Delta_{yj}) = \mathrm{Cov}(\Delta_{xi}, \Delta_{ij}) = \mathrm{Cov}(\Delta_{yj}, \Delta_{ij}) = 0$. Ainsi, on obtient, à partir de la Formule (C.1) :

$$V_{xy} \;=\; V_{xi} + V_{yj} + V_{ij}.$$

Le cas particulier de SDM

On cherche à calculer

$$V_{xy}^{\mathrm{SDM}} \;=\; V_{xi}^{\mathrm{SDM}} + V_{yj}^{\mathrm{SDM}} + V_{ij}^{\mathrm{SDM}} + 2\Big(\mathrm{Cov}(\Delta_{xi}^{\mathrm{SDM}}, \Delta_{yj}^{\mathrm{SDM}}) - \mathrm{Cov}(\Delta_{xi}^{\mathrm{SDM}}, \Delta_{ij}^{\mathrm{SDM}}) - \mathrm{Cov}(\Delta_{yj}^{\mathrm{SDM}}, \Delta_{ij}^{\mathrm{SDM}})\Big),$$

sachant la Formule (5.9) (cf page 102) :

$$\Delta_{ij}^{\mathsf{SDM}} \;=\; \frac{1}{W_{ij}} \sum_{\substack{1\leq p\leq k \\ i,j\in\mathcal{L}_{\Delta^p}}} w_p\big(\alpha_p\Delta_{ij}^p + a_{ip} + a_{jp}\big),$$

ainsi que la Formule (6.11) (cf page 132) :

$$V_{ij}^{\mathsf{SDM}} = \frac{1}{W_{ij}^2} \sum_{\substack{1\leq p\leq k \\ i,j\in\mathcal{L}_{\Delta^p}}} w_p^2\alpha_p^2\mathsf{Var}(\Delta_{ij}^p).$$

Si on cherche d'abord à calculer

$\mathsf{Cov}(\Delta_{xi}^{\mathsf{SDM}}, \Delta_{yj}^{\mathsf{SDM}}) =$

$$\mathsf{Cov}\left(\frac{1}{W_{xi}} \sum_{\substack{1\leq p\leq k \\ x,i\in\mathcal{L}_{\Delta^p}}} w_p\big(\alpha_p\Delta_{xi}^p + a_{xp} + a_{ip}\big), \frac{1}{W_{yj}} \sum_{\substack{1\leq q\leq k \\ y,j\in\mathcal{L}_{\Delta^q}}} w_q\big(\alpha_q\Delta_{yj}^q + a_{yq} + a_{jq}\big)\right),$$

sachant que $\mathsf{Cov}\big(\sum_p X_p, \sum_q Y_q\big) = \sum_{p,q}\mathsf{Cov}(X_p, Y_q)$, on obtient :

$$\mathsf{Cov}(\Delta_{xi}^{\mathsf{SDM}}, \Delta_{yj}^{\mathsf{SDM}}) \;=\; \frac{1}{W_{xi}W_{yj}} \sum_{\substack{1\leq p\leq k \\ 1\leq q\leq k \\ x,y,i,j\in\mathcal{L}_{\Delta^p}\cap\mathcal{L}_{\Delta^q}}} w_p w_q\alpha_p\alpha_q\mathsf{Cov}(\Delta_{xi}^p, \Delta_{yj}^q).$$

Or, comme $\mathsf{Cov}(\Delta_{xi}^p, \Delta_{yj}^q) = 0$, si $p \neq q$, on obtient l'écriture simplifiée :

$$\mathsf{Cov}(\Delta_{xi}^{\mathsf{SDM}}, \Delta_{yj}^{\mathsf{SDM}}) \;=\; \frac{1}{W_{xi}W_{yj}} \sum_{\substack{1\leq p\leq k \\ x,y,i,j\in\mathcal{L}_{\Delta^p}}} w_p^2\alpha_p^2\mathsf{Cov}(\Delta_{xi}^p, \Delta_{yj}^p).$$

Ainsi, en utilisant les modèles employés de BIONJ ou MVR, on peut estimer la variance V_{xy}^{SDM} d'une complétion additive dans une supermatrice de distance $(\Delta_{ij}^{\mathsf{SDM}})$ avec une complexité de l'ordre de $O(n)$.

C.1.2 Complétion par quadruplets

Si la distance inconnue $\Delta_{xy} = \varnothing$ est estimée par complétion par quadruplets, alors il existe un ensemble $\mathbb{S}_{xy}$ formé par toutes les paires de taxons i, j vérifiant le critère (3.5) (cf page 56), et la complétion est effectuée par la formule

$$\Delta_{xy} \;=\; \frac{1}{|\mathbb{S}_{xy}|} \sum_{\substack{i,j\in\mathbb{S}_{xy} \\ i<j}} (\Delta_{xi} + \Delta_{yj} - \Delta_{ij}).$$

On obtient alors la formule :

$$V_{xy} \;=\; \frac{1}{|\mathbb{S}_{xy}|^2}\left(\sum_{\substack{i,j\in\mathbb{S}_{xy}\\ i<j}} \mathsf{Var}(\Delta_{xi}+\Delta_{yj}-\Delta_{ij}) + 2\sum_{\substack{i,j,u,v\in\mathbb{S}_{xy}\\ i<j\\ u<v\\ (i,j)\neq(u,v)}} C_{ijuv}\right),$$

où

$$C_{ijuv} \;=\; \mathsf{Cov}\Big(\Delta_{xi}+\Delta_{yj}-\Delta_{ij}\;,\;\Delta_{xu}+\Delta_{yv}-\Delta_{uv}\Big).$$

Les $O(n^2)$ variances $\mathsf{Var}(\Delta_{xi}+\Delta_{yj}-\Delta_{ij})$ sont calculées suivant la Formule (C.1), et les neuf covariances issues de chacune des $O(n^4)$ valeurs C_{ijuv} peuvent se calculer à l'aide du modèle de variance-covariance employé par BIONJ. Le modèle de variance employé par MVR impliquant que $C_{ijuv}=0$, la valeur V_{xy} se calcule alors en $O(n^2)$ par l'équation suivante :

$$V_{xy} \;=\; \frac{1}{|\mathbb{S}_{xy}|^2}\sum_{\substack{i,j\in\mathbb{S}_{xy}\\ i<j}} \big(V_{xi}+V_{yj}+V_{ij}\big).$$

Pour le cas particulier d'une supermatrice de distance $(\Delta_{ij}^{\mathsf{SDM}})$, alors la formule

$$V_{xy}^{\mathsf{SDM}} \;=\; \frac{1}{|\mathbb{S}_{xy}|^2}\left(\sum_{\substack{i,j\in\mathbb{S}_{xy}\\ i<j}} \mathsf{Var}(\Delta_{xi}^{\mathsf{SDM}}+\Delta_{yj}^{\mathsf{SDM}}-\Delta_{ij}^{\mathsf{SDM}}) + 2\sum_{\substack{i,j,u,v\in\mathbb{S}_{xy}\\ i<j\\ u<v\\ (i,j)\neq(u,v)}} C_{ijuv}^{\mathsf{SDM}}\right),$$

se calcule en $O(n^5)$ avec le modèle de variance-covariance issu de BIONJ, et en $O(n^3)$ avec le modèle de variance-covariance simplifié issu de MVR.

Index

α_p, 93–95
a_{ip}, 95
ACS, 83
C^*_{xy}, 112, 114
$\mathcal{C}_T$, 73
$\mathcal{C}_{(T)}$, 83
$\mathcal{C}_{(\Delta)}$, 95
$\tilde{C}^*_{xy}$, 112, 114
Δ_{ij}, 51
$\Delta^{\mathsf{SDM}}_{ij}$, 100
d_{RF}, 66
d_{path}, 67
d_{quad}, 67
GLS, 56
H, 60
Hm_{ij}, 51
k_{ij}, 83, 96
ℓ, 43
ℓ_p, 96
λ_i, 63
λ_i, 62
λ^*_i, 116
$\mathcal{L}_T$, 25
$\mathcal{L}_r$, 60
$\mathcal{L}_{\Delta^p}$, 95
$\mathcal{L}_{\mathcal{C}_T}$, 75
$\tilde{\mathcal{L}}_{\Delta^p}$, 96
$\tilde{\mathcal{L}}_{\mathcal{C}_{(\Delta)}}$, 97
LS, 55
ME, 56
ML, 48
MP, 42
MR, 78
$\mathsf{MR}_{\mathcal{C}_T}$, 81
MRH, 143
MRP, 80
N^*_{xy}, 112
N'_{xy}, 61
N_{xy}, 60
$\tilde{N}^*_{xy}$, 112
$\tilde{n}$, 100
$\tilde{n}_p$, 96
n_p, 95
OLS, 55
ϕ_T, 58
PM, 93, 95
Q^*_{xy}, 114
Q_{xy}, 61, 62
$\tilde{Q}^*_{xy}$, 114
R_z, 61
SM, 93
SSM, 95
T_{ij}, 27
$\mathcal{T}_{ij}$, 27
TE, 71
w_i, 62, 63
w^*_i, 115
WLS, 55

AddTree, 59, 60
ADN, 41
algorithme
    — agglomératif, 28, 29, 60, 117
    — d'insertion, 30
    — de complétion, 53, 140

allèle, 20
analyse simultanée, 71
anthropocentrisme, 15
arbre, 25
  $X-$ —, 25
  — en étoile, 28
  — phylogénétique, 20, 25
    — binaire, 25
    — enraciné, 25
    — racine, 24
    — valué, 24, 27
    cerise, 25
    diamètre, 58
    feuille, 25
    noeud interne, 25
  chenille, 67
attraction des longues branches, 45, 72

BIONJ*, 129
BIONJ, 63
bipartition, 26
bootstrap, 67
BUILD, 75

clade, 24
  — trivial, 26
combinaison
  — basse, 70, 71, 135
  — haute, 70, 73, 142
  — moyenne, 70, 84, 138
compatible
  collection d'arbres —, 74
congruence
  in —, 72, 84, 92
consensus, 73, 74
  — majoritaire, 74
  — strict, 74
consistant, 57, 62
  in —, 45

créationnisme, 14
critère
  — d'agglomération, 60, 112
  — de distance, 55
  — de moindres-carrés, 55
  — du maximum de parcimonie, 42
  — du maximum de vraisemblance, 48
  Minimum Evolution, 56

darwinisme, 18
distance, 27
  — à centre, 27, 94
  — additive, 27
  — de Hamming, 51
  — évolutive, 51
  — incomplète, 53
  — topologique, 66
    — d'agrément, 66
    — de bipartition, 66
    — de quadruplets, 67
    — par chemin, 67
  — ultramétrique, 27
  matrice de —, 27
duplication, 22

état de caractère, 40
évènement mutationnel, 41
  — par délétion, 42
  — par substitution, 42
  transition, 47
  transversion, 47

FASTME, 57
FITCH, 57, 58, 65
fixisme, 16

gène, 19, 41
gap, 42
génération spontanée, 14
gradisme, 16

graphe, 25
— connexe, 25
— sans cycle, 25
groupe
— monophylétique, 24
— paraphylétique, 24

homologie, 16, 41
— de position, 41
homoplasie, 72
convergence, 72
réversion, 72

inégalité
— quadrangulaire, 27
— triangulaire, 27
— ultramétrique, 27

loi gamma, 48

modèle d'évolution, 46
JC, 47
K2P, 48
mouvement topologique, 35
LPR, 66
NNI, 35
NNI augmenté, 50, 149
SPR, 37
TBR, 37
MVR, 64
MVR*, 129
MW, 59
MW*, 59

NJ, 59, 61, 62
NJ*, 129
nucléotides, 41

parcimonie, 42
— de Fitch, 43
— de Sankoff, 43
PAUP*, 44, 51, 53, 57
PHYML, 50

recherche locale, 32
— par descente, 33
GRASP, 34
Recuit Simulé, 34
RVV, 34
Tabou, 34
voisinage, 33
représentation matricielle binaire, 78

sélection naturelle, 18, 19
site, 42
spéciation, 22, 24
superarbre, 73
Average Consensus Supertree, 83
Matrix Representation with Parsimony, 80
Matrix Representation with Hamming distance, 143
supermatrice
— de caractères, 71, 72
— de distance, 83, 100

T-REX, 59
taxon, 24
taxonomie, 24
— numérique, 20
TNT, 44
total evidence, 71
transformisme, 16

uniformitarisme, 17
UNJ, 62
UNJ*, 129

vraisemblance, 48

Table des figures

1.1 **Modélisation de l'évolution par Lamarck (1809)** 19
1.2 **Modélisation de l'évolution par Darwin (1837)** 21
1.3 **Arbre phylogénétique inféré par Fitch et Margoliash (1967)** 23

2.1 **Exemple d'arbres phylogénétiques** 28
2.2 **Le schéma agglomératif pour construire un arbre phylogénétique non enraciné** 31
2.3 **Le schéma divisif pour construire un arbre phylogénétique enraciné** 32
2.4 **La procédure d'insertion pour construire un arbre phylogénétique non enraciné** 33
2.5 **Exemple d'une recherche locale** 34
2.6 **Réarrangement de type NNI** 36
2.7 **Réarrangement de type SPR** 37
2.8 **Réarrangement de type TBR** 38

3.1 **Exemple d'alignement de séquences d'ADN** 43
3.2 **Association d'un vecteur de vraisemblance à un noeud interne ainsi qu'à ses deux noeuds fils** 51
3.3 **Schéma générique des algorithmes agglomératifs** 65

4.1 **Exemple de supermatrice de caractères** 74
4.2 **Construction du graphe représentant l'information topologique d'une collection de deux arbres** 78
4.3 **Construction de la représentation matricielle binaire de l'information topologique d'une collection de deux arbres** 81
4.4 **Exemple de représentation CGR d'une signature génomique** 89

5.1 **Estimation de la complexité en pratique de la résolution du système linéaire de la méthode SDM** 103
5.2 **Observation de la stabilité des estimations des vitesses d'évolution relatives par SDM** 107

5.3 **Observation du recouvrement topologique des scénarios de combinaison moyenne utilisant SDM** 110

6.1 **Schéma générique des algorithmes agglomératifs adaptés aux distances incomplètes** 119
6.2 **Distribution des valeurs des quatre critères d'agglomération** Q^*_{xy}, $\tilde{Q}^*_{xy}$, $\tilde{N}^*_{xy}/|\tilde{C}^*_{xy}|$ **et** $\tilde{N}'^*_{xy}/|\tilde{C}^*_{xy}|$ 126
6.3 **Performances de la combinaison de plusieurs critères d'agglomération** . . . 129

7.1 **Performances des méthodes de distance par combinaison basse et moyenne** 138
7.2 **Performances des scénarios ML d'inférence d'arbre par combinaison basse** 139
7.3 **Performances des scénarios PhyML+ACS+Fitch et PhyML+SDM+Fitch** . . 149
7.4 **Performances des meilleurs scénarios de combinaison moyenne, haute et basse** 153

Bibliographie

Adams III, E. (1972). Consensus techniques and the comparison of taxonomic trees. *Systematic Zoology*, 21 :390–397.

Aho, A. V., Sagiv, T. G., Szymanski, T. G., and Ullman, J. D. (1981). Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM Journal of Computing*, 10(3) :405–421.

Allen, B. and Steel, M. (2001). Subtree transfer operations and their induced metrics on evolutionary trees. *Annals of Combinatorics*, 5 :1–13.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment saerch tool. *Journal of Molecular Biology*, 215 :403–410.

Atteson, K. (1999). The performance of neighbor-joining methods of phylogenetic reconstruction. *Algorithmica*, 25 :251–278.

Auch, A. F., Henz, S. R., Holland, B. R., and Göker, M. (2006). Genome BLAST distance phylogenies inferred from whole plastid and whole mitochondrion genome sequences. *BMC Bioinformatics*, 7 :350.

Barrett, M., Donoghue, M., and Sober, E. (1991). Against consensus. *Systematic Zoology*, 4 :486–493.

Barthélemy, J.-P. and Guénoche, A. (1988). *Les arbres et les représentations des proximités*. Masson, Paris.

Baum, B. R. (1992). Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon*, 41 :3–10.

Ben-Dor, A., Chor, B., Graur, D., Ophir, R., and Pelleg, D. (1998). Constructing phylogenies from quartets : Elucidation of eutherian superordinal relationships. *Journal of Computational Biology*, 5(3) :377–390.

Berry, V. and Gascuel, O. (1996). Interpretation of bootstrap trees : threeshold of clade selection and induced gain. *Molecular Biology and Evolution*, 13(7) :999–1011.

Berry, V., Gascuel, O., and Caraux, G. (2000). Chosing the tree which actually best explains the data : another look at the bootstrap in phylogenetic reconstruction. *Computational Statistics and Data Analysis*, 38 :273–283.

Berry, V. and Semple, C. (2006). Fast computation of supertrees for comptatible phylogenies with nested taxa. *Systematic Biology*, 55(2) :270–288.

Bevan, R. B., Lang, B. F., and Bryant, D. (2006). Calculating the evolutionary rates of different genes : a fast, accurate estimator with applications to maximum likelihood phylogenetic analysis. *Systematic Biology*, 54 :900–915.

Bininda-Emonds, O. R. P. and Bryant, H. N. (1998). Properties of matrix representation with parsimony analysis. *Systematic Biology*, 47(3) :497–508.

Bond, J. E. and Hedin, M. (2006). A total evidence assessment of the phylogeny of north american euctenizine trapdoor spiders (Araneae, Mygalomorphae, Cyrtaucheniidae) using bayesian inference. *Molecular Phylogenetics and Evolution*, 41 :70–85.

Bourque, M. (1978). *Arbres de Steiner et réseaux dont varie l'emplagement de certains sommets*. PhD thesis, Département d'informatique et de recherche opérationnelle, Université de Montréal.

Brooks, D. R. (1981). Hennig's parasitological method : a proposed solution. *Systematic Zoology*, 30 :229–249.

Brossier, G. (1985). Approximation des dissimilarités par des arbres additifs. *Mathématiques et Sciences Humaines*, 91 :5–21.

Bryant, D. (2001). Optimal agreement supertrees. In *First International Conference on Biology, Informatics, and Mathematics (JOBIM 2000) Lecture notes in computer science 2066*, pages 24–31. Springer-verlag.

Bryant, D. (2003). A classification of consensus methods for phylogenetics. In Janowitz, M., Lapointe, F. J., McMorris, F., Mirkin, B., and Roberts, F., editors, *Bioconsensus*, pages 163–184. DIMACS, AMS.

Bryant, D. (2005). On the uniqueness of the selection criterion in Neighbor-joining. *Journal of Classification*, 22 :3 – 15.

Bryant, D., McKenzie, A., and Steel, M. (2003). The size of a maximum agreement subtree for random binary trees. *DIMACS series in Discrete Mathematics and Theoretical Computer Science*, 61 :55–65.

Bryant, D., Semple, C., and Steel, M. (2004). *Phylogenetic supertrees : Combining information to reveal the tree of life*, chapter Combining evolutionary trees with ancestral divergence dates. Kluwer Academic, Dordrecht, The Nederlands, O.R.P. Bininda-Emonds edition.

Bryant, D. and Waddell, P. (1998). Rapid evaluation of least-squares and minimum evolution criteria on phylogenetic trees. *Molecular Biology and Evolution*, 15 :1346–1359.

Bulmer, M. (1991). Use of the method of the generalized least-squares in reconstructing phylogenies from sequence data. *Molecular Biology and Evolution*, 8 :868–883.

Buneman, P. (1971). The recovery of trees from measures of dissimilarity. In Hudson, F., Kendall, D., and Tautu, P., editors, *Mathematics in archeological and historical sciences*, pages 387–395. Edinburgh University Press, Edinburgh.

Caraux, G., Gascuel, O., Andrieux, G., and Levy, D. (1995). Approches informatiques de la reconstruction phylogénétique. *Techniques et Sciences Informatiques*, 14(2) :113–139.

Carnap, R. (1962). *The Logical Foundations of Probability*. University of Chicago Press, Chicago.

Carnap, R. (1997). Replies and systematic expositions. In Schlipp, P. A., editor, *The philosophy of Rudolph Carnap*, pages 859–1013. Open Court, LaSalle, IL.

Carroll, J. D. and Pruzansky, S. (1980). Discrete and hybrid scaling models. In Lantermann, E. B. and Feger, H., editors, *Similarity and Choice*, pages 108–139. Hans Huber, Bern.

Cavalli-Sforza, L. L. and Edwards, A. W. F. (1967). Phylogenetic analysis : models and estimation procedures. *American Journal of Human Genetics*, 19 :223–257.

Cavender, J. (1978). Taxonomy with confidence. *Mathematical Biosciences*, 40 :271–280.

Chapus, C., Dufraigne, C., Edwards, S., Giron, A., Fertil, B., and Deschavanne, P. J. (2005). Exploration of phylogenetic data using a global sequence analysis method. *BMC Evolutionary Biology*, 5 :63.

Charon, I. and Hudry, O. (1993). The noising method : a new method for combinatorial optimization. *Operations Research Letters*, 14 :133–137.

Chen, D., Diao, L., Eulenstein, O., Fernández-Baca, D., and Sanderson, M. (2003). Flipping : a supertree construction method. In Janowitz, M., Lapointe, F. J., McMorris, F., Mirkin, B., and Roberts, F., editors, *Bioconsensus*, pages 135–160. DIMACS, AMS.

Chen, D., Eulenstein, O., Fernández-Baca, D., and Sanderson, M. (2006). Minimum-flip supertree : complexity and algorithms. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3(2) :165–173.

Chippindale, P. T. and Wiens, J. J. (1994). Weighting, partitioning, and combining characters in phylogenetic analysis. *Systematic Biology*, 43 :278–287.

Constantinescu, M. and Sankoff, D. (1995). An efficient algorithm for supertrees. *Journal of Classification*, 12 :101–112.

Cosner, M. E., Jansen, R. K., Moret, B. M. E., Raubeson, L. A., Wang, L.-S., Warnow, T., and Wyman, S. (2000). A new fast heuristic for computing the breakpoint phylogeny and experimental analyses of real and synthetic data. In Bourne, P., Gribskov, M., Altman, R., Jensen, N., Hope, D., Lengauer, T., Mitchell, J., Scheeff, E., Smith, C. Strande, S., and Weissig, H., editors, *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*, La Jolla, CA, USA.

Cox, D. R. and Miller, H. D. (1965). *The Theory of Stochastic Processes*. Chapman & Hall, London, UK.

Creevey, C. J. and McInerney, J. O. (2005). CLANN : investigating phylogenetic information through supertree analyses. *Bioinformatics*, 21(3) :390–392.

Criscuolo, A. (2006). Adaptation de algorithmes agglomératifs nj, unj, BIONJ et mvr pour l'inférence très rapide de superarbres. In Denise, A., Durrens, P., Robin, S., Rocha, E., de Daruvar, A., and Groppi, A., editors, *JOBIM06*, pages 227–240.

Criscuolo, A., Berry, V., Douzery, E. J. P., and Gascuel, O. (2006). SDM : A fast distance-based approach for (super)tree building in phylogenomics. *Systematic Biology*. (*in press*).

Cunningham, J. P. (1978). Free trees as representations of psychological distances. *Journal of Mathematical Psychology*, 17 :165–188.

Daniel, P. and Semple, C. (2004). *Phylogenetic supertrees : Combining information to reveal the tree of life*, chapter Supertree algorithms for nested taxa. Kluwer Academic, Dordrecht, The Nederlands, O.R.P. Bininda-Emonds edition.

Darlu, P. and Tassy, P. (1993). *La reconstruction phylogénétique. Concepts et méthodes.* Masson, Paris, France.

Day, W. H. E. (1987). Computational complexity of inferring phylogenies by dissimilarity matrices. *Bulletin of Mathematical Biology*, 49 :461–467.

Day, W. H. E. (1996). Complexity theory : an introduction for practitionners of classification. In Arabie, P., Hubert, L. J., and De Soete, G., editors, *Clustering and Classification*, pages 199–233. World Scientific, London.

Day, W. H. E. and Sankoff, D. (1986). Computational complexity of inferring phylogenies by compatibility. *Systematic Zoology*, 35(2) :224–229.

de Queiroz, A., Donoghue, M. J., and Kim, J. (1995). Separate versus combined analysis of phylogenetic evidence. *Annual review of ecology and systematics*, 26 :657–681.

De Soete, G. (1983). A least squares algorithm for fitting additive trees to proximity data. *Psychometrika*, 48 :621–626.

De Soete, G. (1984a). Additive tree representations of incomplete dissimilarity data. *Quality and Quantity*, 18 :387–393.

De Soete, G. (1984b). Ultrametric tree representations of incomplete dissimilarity data. *Journal of classification*, 1 :235–242.

Denis, F. and Gascuel, O. (2003). On the consistency of the minimum evolution principle of phylogenetic inference. *Discrete Applied Mathematics*, 127 :63–77.

Deschavanne, P. J., Giron, A., Vilain, J., Fagot, G., and Fertil, B. (1999). Genomic signature : characterization and classification of species assessd by chaos game representation of sequences. *Molecular Biology and Evolution*, 16 :1391–1399.

Desper, R. and Gascuel, O. (2002). Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *Journal of Computational Biology*, 19(5) :687–705.

Desper, R. and Gascuel, O. (2004). Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. *Molecular Biology and Evolution*, 21(3) :587–598.

Diaconis, P. and Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American*, 248 :116–130.

Dicks, J. (2000). *Comparative Genomics : Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families*, chapter CHROMTREE : Maximum likelihood estimation of chromosomal phylogenies. Kluwer Academic, Dordrecht, The Nederlands, D. Sankoff and J. H. Nadeau edition.

Doolittle, W. F. (1999). Phylogenetic classification and the universal tree. *Science*, 284 :2124–2129.

Doyle, J. J. (1992). Gene trees and species trees : molecular systematics as one-character taxonomy. *Systematic Botany*, 17 :144–163.

Driskell, A. C., Ané, C., Burleigh, J. G., McMahon, M. M., O'Meara, B. C., and Sanderson, M. J. (2004). Pospects for building the tree of life from large sequence databases. *Science*, 306 :1172–1174.

Edwards, A. W. F. and Cavalli-Sforza, L. L. (1967). Phylogenetic analysis : models and estimation procedures. *Evolution*, 21 :550–570.

Eernisse, D. J. and Kluge, A. G. (1993). Taxonomic congruence versus total evidence, and amniote phylogeny inferred from fossils, molecules and morphology. *Molecular Biology and Evolution*, 10(6) :1170–1195.

Efron, B. (1979). Bootstrap methods : another look at the jacknife. *Annals of Statistics*, 7(1) :1–26.

Efron, B. (1981). Nonparametric estimates of standard error : the jackknife, the bootstrap and other methods. *Biometrika*, 68 :589–599.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*. Society of Industrial and Applied Mathematics CBMS-NSF Monographs.

Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap*. Chapman & Hall, New York.

Elemento, O. and Gascuel, O. (2002). A fast and accurate distance algorithm to reconstruct tandem duplication trees. *Bioinformatics*, 18 :92–99.

Estabrook, G. F., McMorris, F. R., and Meacham, C. A. (1985). Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. *Systematic Zoology*, 34 :193–200.

Eulenstein, O., Chen, D., Burleigh, J. G., Fernandez-Baca, D., and Sanderson, M. J. (2004). Performance of flip supertree construction with a heuristic algorithm. *Systematic Biology*, 53(2) :299–308.

Farris, J. S. (1970). Methods for computing Wagner trees. *Systematic Zoology*, 19 :83–92.

Farris, J. S., Källersjö, M., Kluge, A. G., and Bult, C. (1995). Testing significance of incongruence. *Cladistics*, 1 :315–319.

Farris, J. S., Kluge, A. G., and Eckardt, M. J. (1970). A numerical approach to phylogenetic systematics. *Systematic Zoology*, 19 :172–191.

Felsenstein, J. (1973). Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Zoology*, 22 :240–249.

Felsenstein, J. (1978a). Cases in which parsimony and compatibility methods will be positively misleading. *Systematic Zoology*, 27 :401–410.

Felsenstein, J. (1978b). The number of evolutionary trees. *Systematic Zoology*, 27 :27–33.

Felsenstein, J. (1981). Evolutionary trees from dna sequences : a maximum likelihood approach. *Journal of Molecular Evolution*, 17 :368–376.

Felsenstein, J. (1985). Confidence limits on phylogenies : an approach using the bootstrap. *Evolution*, 39 :783–791.

Felsenstein, J. (1993). PHYLIP : Phylogeny inference package, version 3.6. *Distributed by the author. University of Washington, Seattle, Washington*.

Felsenstein, J. (1997). An alternating least-squares approach to inferring phylogenies. *Systematic Biology*, 46 :101–111.

Felsenstein, J. and Kishino, H. (1993). Is there something wrong with the bootstrap ? a reply to Hillis and Bull. *Systematic Biology*, 42 :193–200.

Feo, T. and Resende, M. (1995). Greedy randomized adaptive search procedures. *Journal of Global Optimization*, 6 :109 – 133.

Finden, C. R. and Gordon, A. D. (1985). Obtaining common pruned trees. *Journal of Classification*, 2 :255–276.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Systematic Zoology*, 19 :99–113.

Fitch, W. M. (1971). Toward defining the course of evolution : minimum change for a specific tree topology. *Systematic Zoology*, 20 :406–416.

Fitch, W. M. and Margoliash, E. (1967). The construction of phylogenetic trees - a generally applicable method utilizing estimates of the mutation distance obtained from cytochrome c sequences. *Science*, 155 :279–284.

Foulds, L. R. and Graham, R. L. (1982). The steiner problem in phylogeny is NP-complete. *Advances in Applied Mathematics*, 3 :43–49.

Gascuel, O. (1994). A note on Sattath and Tversky's, Saitou ans Nei's and Studier and Keppler's algorithms for inferring phylogenies from evolutionary distances. *Molecular Biology Evolution*, 11(6) :961–963.

Gascuel, O. (1997a). BIONJ : an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, 14 :685–695.

Gascuel, O. (1997b). Concerning the NJ algorithm and its unweighted version, UNJ. In Mirkin, B., McMorris, F., Roberts, F., and Rzhetsky, A., editors, *Mathematical Hierarchies and Biology*, pages 149–170. DIMACS Series in Discrete Mathematics and Theoretical Computer Science. American Mathematical Society, Providence.

Gascuel, O. (2000). Data model and classification by trees : the minimum variance reduction (MVR) method. *Journal of Classification*, 17(1) :67–99.

Gascuel, O., Bryant, D., and Denis, F. (2001). Strengths and limitations of the minimum evolution principle. *Systematic Biology*, 50 :621–627.

Gascuel, O. and Levy, D. (1996). A reduction algorithm for approximating a (non-metric) dissimilarity by a tree distance. *Journal of Classification*, 13 :129–155.

Gascuel, O. and Steel, M. (2006). Neighbor joining revealed. -. (submitted).

Gatesy, J., Matthee, C., DeSalle, R., and Hayashi, C. (2002). Resolution of a supertree/supermatrix paradox. *Systematic Biology*, 51(4) :652–664.

Gatesy, J. and Springer, M. S. (2004). *Phylogenetic supertrees : Combining information to reveal the tree of life*, chapter A critique of matrix representation with parsimony supertrees. Kluwer Academic, Dordrecht, The Nederlands, O.R.P. Bininda-Emonds edition.

Gauss, C. F. (1876). Beiträge zur theorie der algebraischen gleichungen. In *Werke*, pages 71–102. K. Gesellschaft Wissenschaft, Göttingen.

Glover, F. (1986). Future paths for integer programming and links to artificial intelligence. *Computers and Operations Research*, 13 :533 – 549.

Goddard, W., Kubicka, E., Kubicki, G., and McMorris, F. R. (1994). The agreement metric for labeled binary trees. *Mathematical Bioscience*, 123 :215–226.

Goloboff, P., Farris, J., and Nixon, K. (2003). TNT : Tree analysis using new technology. *Distributed by the authors*.

Goloboff, P. and Pol, D. (2002). Semi-strict supertrees. *Cladistics*, 18 :514–525.

Gomory, R. E. and Hu, T. C. (1961). Multi-terminal network flows. *Journal of the Society for Industrial and Applied Mathematics*, 9(4) :551–570.

Gordon, A. D. (1980). On the assessment and comparison of classifications. In Tomassine, R., editor, *Analyse de Données et Informatique*, pages 149–160, Le Chesnay, INRIA, France.

Gordon, A. G. (1986). Consensus supertrees : the synthesis of rooted trees containing overlapping sets of labelled leaves. *Journal of Classification*, 3 :335–348.

Graybeal, A. (1998). Is it better to add taxa or characters to a difficult phylogenetic problem ? *Systematic Biology*, 47(1) :9–17.

Guindon, S. and Gascuel, O. (2003). A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52(5) :696–704.

Guénoche, A. and Garreta, H. (2001). Can we have confidence in a tree representation ? In *Proceedings of JOBIM'2000, Lecture Notes in Computer Sciences*, volume 2066, pages 43–53.

Guénoche, A. and Grandcolas, S. (1999). Approximations par arbre d'une distance partielle. *Mathématiques, Informatique et Sciences Humaines*, 146 :51–64.

Guénoche, A. and Leclerc, B. (2001). The triangle method to build x-trees from incomplete distance matrices. *RAIRO Operations Research*, 35 :283–300.

Hansen, P. (1986). The steepest ascent mildest descent heuristic for combinatorial programming. In *Congress on Numerical Methods in Combinatorial Optimization*, Capri, Italy.

Harding, E. (1971). The probabilities of rooted tree-shapes generated by random bifurcation. *Advanced Applied Probabilities*, 3 :44–77.

Hartigan, J. A. (1967). Representation of similarity matrices by trees. *Journal of American Statistics*, 62 :1140–1158.

Hartigan, J. A. (1973). Minimum mutation fits to a given tree. *Biometrics*, 29 :53–65.

Hasegawa, M., Adachi, J., and Milinkovitch, M. C. (1997). Novel phylogeny of whales supported by total molecular evidence. *Journal of Molecular Evolution*, 44 :117–120.

Heaviside, O. (1893). On operators in physical mathematics, part 1. In *Proceeding of the Royal Society*, volume 52, pages 504–529, London.

Hendy, M. D. and Penny, D. (1989). A framework for the quantitative study of evolutionary trees. *Systematic Zoology*, 38 :297–309.

Henz, S. R., Huson, D. H., Auch, A. F., Nieselt-Struwe, K., and Schuster, S. C. (2005). Whole-genome prokaryotic phylogeny. *Bioinformatics*, 21(10) :2329–2335.

Henzinger, M. R., King, V., and Warnow, T. (1999). Constructing a tree from homeomorphic subtrees, with applications to computational evolutionary biology. *Algorithmica*, 24 :1–13.

Hillis, D. M. (1996). Inferring complex phylogenies. *Nature*, 383 :130–131.

Hillis, D. M. (1998). Taxonomic sampling, phylogenetic accuracy, and investigator bias. *Systematic Biology*, 47 :3–8.

Hillis, D. M. and Bull, J. J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology*, 42 :192–200.

Hillis, D. M., Bull, J. J., White, M., Badgett, M. R., and Molineux, I. J. (1992). Experimental phylogenetics : generation of a known phylogeny. *Sience*, 255 :589–592.

Hillis, D. M., Huelsenbeck, J. P., and Cunningham, C. W. (1994). Application and accuracy of molecular phylogenies. *Sience*, 264 :671–677.

Hordijk, W. and Gascuel, O. (2005). Improving the efficiency of spr moves in phylogenetic tree search algorithms based on maximum-likelihood. *Bioinformatics*, 21(24) :4338–4347.

Hubert, L. J. and Arabie, P. (1995). Iterative projection strategies for the least-squares fitting of tree structures to proximity data. *British Journal of Mathematical and Statistical Psychology*, 48 :281–317.

Huelsenbeck, J. P. (1995). Performance of phylogenetic methods in simulation. *Systematic Biology*, 44 :17–48.

Huelsenbeck, J. P., Bull, J. J., and Cunningham, C. W. (1996). Combining data in phylogenetic analysis. *Tree*, 11(4) :152–158.

Huelsenbeck, J. P. and Hillis, D. M. (1993). Success of phylogenetic methods in the four-taxon case. *Systematic Biology*, 42 :247–264.

Huelsenbeck, J. P. and Ronquist, F. (2001). MrBayes : Bayesian inference of phylogeny. *Bioinformatics*, 17 :754–755.

Jardine, C. J., Jardine, N., and Sibson (1967). The structure and construction of taxonomic hierarchies. *Mathematical Bioscience*, 1 :173–179.

Jeffrey, H. J. (1990). Chaos game representation of gene structure. *Nucleic Acids Research*, 18 :2163–2170.

Johnson, J. L., Anderson, R. S., and Ordal, E. J. (1970). Nucleic acid homologies among oxidase-negative Moraxella species. *Journal of Bacteriology*, 101 :568–573.

Jukes, T. H. and Cantor, C. R. (1969). Evolution of protein molecules. In Munro, H. N., editor, *Mammalian Protein Metabolism*, volume III, chapter 24, pages 21–132. Academic Press, New York.

Kidd, K. and Sgaramella-Zonta, L. (1971). Phylogenetic analysis : concepts and methods. *American Journal of Human Genetics*, 23 :235–252.

Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16 :111–120.

Kimura, M. and Ohta, T. (1972). On the stochastic model for estimation of mutational distance between homologous proteins. *Journal of Molecular Evolution*, 2 :87–90.

Kirkpatrick, S., Gelatt Jr., G., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598) :671–680.

Kishino, H. and Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *Journal of Molecular Evolution*, 29 :170–179.

Kluge, A. G. (1989). A concern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae, Serpentes). *Systematic Zoology*, 38 :7–25.

Kuhner, M. K. and Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under inequal evolutionary rates. *Molecular Biology and Evolution*, 11 :459–468.

Lafay, B., Smith, A. B., and Christen, R. (1995). A combined morphological and molecular approach to the phylogeny of asteroids (Asteroidea : Echinodermata). *Systematic Biology*, 44(2) :190–208.

Lanave, C., Preparata, G., Saccone, C., and Serio, G. (1984). A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution*, 20(1) :86–93.

Landry, P.-A. and Lapointe, F.-J. (1997). Estimation of missing distances in path-length matrices : problems and solutions. In *DIMACS*, volume 37 of *Discrete mathematics and theoretical computer science*, pages 209–218.

Landry, P.-A., Lapointe, F.-J., and Kirsch, J. A. W. (1996). Estimating phylogenies from lacunose distance matrices : Additive is superior to ultrametric estimation. *Molecular Biology and Evolution*, 13 :818–823.

Lanyon, S. (1993). Phylogenetic frameworks : towards a firmer foundation of the comparative approach. *Biological Journal of the Linnean Society*, 49 :45–61.

Lapointe, F.-J. and Cucumel, G. (1997). The average consensus procedure : combination of weighted trees containing identical or overlapping sets of taxa. *Systematic Biology*, 46(2) :306–312.

Lapointe, F.-J. and Kirsch, J. A. W. (1995). Estimating phylogenies from lacunose distance matrices, with special reference to DNA hybridization data. *Molecular Biology and Evolution*, 12 :266–284.

Lapointe, F.-J. and Levasseur, C. (2004). *Phylogenetic supertrees : Combining information to reveal the tree of life*, chapter Everything you always wanted to know about the average consensus, and more. Kluwer Academic, Dordrecht, The Nederlands, O.R.P. Bininda-Emonds edition.

Lapointe, F.-J., Wilkinson, M., and Bryant, D. (2003). Matrix representation with parsimony or with distances : two sides of the same coin ? *Systematic Biology*, 52(6) :865–868.

Lawson, C. M. and Hanson, R. J. (1974). *Solving Least Squares Problems*. Prentice Hall, Englewood Cliffs, NJ.

Lecointre, G. and Deleporte, P. (2005). Total evidence requires exclusion of phylogenetically misleading data. *Zoologica Scripta*, 34(1) :101–117.

Lecointre, G., Philippe, H., Van Le, H., and Le Guyader, H. (1993). Species sampling has a major impact on phylogenetic inference. *Molecular Phylogeny and Evolution*, 2 :205–224.

MacStewart, W. (1941). A note on the power of the sign test. *Annals of Mathematics and Statistics*, 12 :236–239.

Maddison, D. R. (1991). The discovery and importance of multiple islands of most-parsimonious trees. *Systematic Zoology*, 43(3) :315–328.

Makarenkov, V. (2001). T-REX : reconstructing and visualizing phylogenetic trees and reticulation networks. *Bioinformatics*, 17(7) :664–668.

Makarenkov, V. (2002). Comparison of four methods for inferring phylogenetic trees from incomplete dissimilarity matrices. In Jajuga, K., Sokolowski, A., and Bock, H.-H., editors, *Classification, Clustering, and Data Analysis*, pages 371–378, Cracow, Poland. Springer.

Makarenkov, V. and Lapointe, F.-J. (2004). A weighted least-squares approach for inferring phylogenies from incomplete distance matrices. *Bioinformatics*, 20 :2113–2121.

Makarenkov, V. and Leclerc, B. (1999). An algorithm for the fitting of a phylogenetic tree according to a weighted least-squares criterion. *Journal of classification*, 16(1) :3–26.

Mirkin, B. (1996). *Mathematical classification and clustering*. Kluwer Academic, London.

Miyamoto, M. (1985). Consensus cladograms and general classifications. *Cladistics*, 1 :186–189.

Mlanedovic, N. and Hansen, P. (1997). Variable neighborhood search. *Computers and Operations Research*, 24 :1097–1100.

Nadeau, J. and Taylor, B. (1984). Lengths of chromosomal segments conserved since divergence of man and mouse. *Proceedings of the National Academy of Science of USA*, 81 :814–818.

Nei, M. and Jin, L. (1989). Variances of the average numbers of nucleotide substitutions within and between populations. *Molecular Biology and Evolution*, 6 :290 – 300.

Nelson, G. and Ladiges, P. Y. (1992). Information-content and fractional weight of 3-item statements. *Systematic Biology*, 41 :490 – 494.

Nelson, G. and Ladiges, P. Y. (1994). *Models in Phylogeny Reconstruction*, chapter Three-item consensus : empirical test of fractional weighting. Clarendon, Oxford, R. W. Scotland, Seibert, D. J. and Williams, D. M. edition.

Ng, M. P., Steel, M. A., and Wormald, N. C. (2000). The difficulty of constructing a leaf-labelled tree including or avoiding given subtrees. *Discrete Applied Mathematics*, 98(3) :227 – 235.

Nixon, K. C. (1999). The parsimony ratchet, a new method for rapid parsimony analysis. *Cladistics*, 15(4) :407–414.

Nixon, K. C. and Carpenter, J. M. (1996). On simultaneous analysis. *Cladistics*, 12 :221–241.

Olsen, G. J., Matsuda, H., Hagstrom, R., and Overbeek, R. (1994). fastdnaml : A tool for construction of phylogenetic trees of dna sequences using maximum likelihood. *Computational Applied Bioscience*, 10 :41–48.

Page, R. (2002). Modified mincut supertrees. In Guigo, R. and Gusfield, D., editors, *WABI 2002*.

Page, R. D. M. (2004). *Phylogenetic supertrees : Combining information to reveal the tree of life*, chapter Taxonomy, supertrees, and the tree of life. Kluwer Academic, Dordrecht, The Nederlands, O.R.P. Bininda-Emonds edition.

Pauplin, Y. (2000). Direct calculation of a tree length using a distance matrix. *Journal of Molecular Evolution*, 51 :41–47.

Philippe, H. and Douzery, E. J. P. (1994). The pitfalls of molecular phylogeny based on four species, as illustrated by the Cetacea/Artiodactyla relationships. *Journal of Mammalian Evolution*, 2(2) :133–152.

Piaggio-Talice, R., Burleigh, G. J., and Eulenstein, O. (2004). *Phylogenetic supertrees : Combining information to reveal the tree of life*, chapter Quartets supertrees. Kluwer Academic, Dordrecht, The Nederlands, O.R.P. Bininda-Emonds edition.

Pisani, D. and Wilkinson, M. (2002). Matrix representation with parsimony, taxonomic congruence, and total evidence. *Systematic Biology*, 51 :151–155.

Poincaré, H. (1901). Second complément à l'analysis situs. *Proceedings of the London Mathematical Society*, 32 :277–308.

Ponstein, J. (1966). *Matrices in graphs and network theory*. Van Gorcum, Assen, Netherlands.

Prager, E. M. and Wilson, A. C. (1988). Ancient origin of lactalbumin from lysozyme : Analysis of dna and amino acid sequences. *Journal of Molecular Evolution*, 27 :326–335.

Pupko, T., Huchon, D., Cao, Y., Okada, N., and Hasegawa, M. (2002). Combining multiple data sets in a likelihood analysis : which models are the best ? *Molecular Biology and Evolution*, 19(12) :2294–2307.

Purvis, A. (1995). A modification to Baum and Ragan's meyhod for combining phylogenetic trees. *Systematic Biology*, 44 :251–255.

Ragan, M. A. (1992). Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution*, 1 :53–58.

Rambaut, A. and Grassly, N. C. (1997). SEQ-GEN : an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences*, 13 :235–238.

Rannala, B., Huelsenbeck, J. P., Yang, Z., and Nielsen, R. (1998). Taxon sampling and the accuracy of large phylogenies. *Systematic Biology*, 47 :702–710.

Ranwez, V. and Gascuel, O. (2001). Quartet based phylogenetic inference : improvements and limits. *Molecular Phylogenetics and Evolution*, 18(6) :1103–1116.

Rieppel, O. (2005). The philosophy of total evidence and its relevance for phylogenetic inference. *Papéis Avulsos de Zoologia*, 45(8) :77–89.

Robinson, D. F. (1971). Comparison of labeled trees with valency three. *Journal of Combinatorial Theory*, 11 :105–119.

Robinson, D. F. and Foulds, L. (1979). Comparison of weighted labeled trees. *Lectures Notes in Mathematics*, 748 :119–126.

Robinson-Rechavi, M. and Graur, D. (2001). Usage optimization of unevenly sampled data through the combination of quartet trees : An eutherian draft phylogeny based on 640 nuclear and mitochondrial proteins. *Israelian Journal of Zoology*, 47 :259–270.

Rodrìguez, R., Oliver, J. L., Marin, A., and Medina, J. R. (1990). The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology*, 142 :485–501.

Ronquist, F. (1996). Matrix representation of trees, redundancy, and weighting. *Systematic Biology*, 45 :247–253.

Ronquist, F. and Huelsenbeck, J. P. (2003). MrBayes 3 : Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19 :1572–1574.

Roux, M. (1988). Techniques of approximation for building two tree structures. In Hayashi, C., Diday, E., Jambu, M., and Ohsumi, N., editors, *Recent Developments in Clustering and Data Analysis*, pages 151–170. Academis Press, New York.

Rzhetsky, A. and Nei, M. (1992). A simple method for estimating and testing minimum-evolution trees. *Molecular Biology and Evolution*, 9 :945–967.

Rzhetsky, A. and Nei, M. (1993). Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Molecular Biology and Evolution*, 10 :1073–1095.

Saitou, N. and Nei, M. (1987). The neighbor-joining method : a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4 :406–425.

Sanderson, M. (2003). Inferring absolute rates of molecular evolution and divergence times in the absence of molecular clock. *Bioinformatics*, 19 :301– 302.

Sanjuán, R. and Wróbel, B. (2005). Weighted least-squares likelihood ratio test for branch testing in phylogenies reconstructed from distance measures. *Systematic Biology*, 54(2) :218–229.

Sankoff, D. and Cedergreen, R. (1983). Simultaneous comparison of three or more sequences related by a tree. In Sankoff, D. and Kruskal, J., editors, *Time warps, String edit, and Macromolecules : the theory and practice of sequence comparison*, pages 253–284. Addison-Wesley, Reading, MA.

Sankoff, D., Deneault, M., Bryant, D., Lemieux, C., and Turmel, M. (2000). Chloroplast gene order and the divergence of plants and algae, from the normalized number of induced breakpoints. In Sankoff, D. and Nadeau, J., editors, *Comparative Genomics*, pages 89–98. Kluwer Academic.

Sankoff, D., Leduc, G., Antoine, N., Paquin, B., Lang, B. F., and Cedergren, R. (1992). Gene order comparisons for phylogenetic inference : evolution of the mitochondrial genome. *Proceedings of the National Academy of Science USA*, 89 :6575–6579.

Sankoff, D. and Rousseau, P. (1975). Locating the vertices of a Steiner tree in an arbitrary space. *Mathematical Programming*, 9 :240–246.

Sattath, S. and Tversky, A. (1977). Additive similarity trees. *Psychometrika*, 42 :319–345.

Savva, G., Dicks, J., and Roberts, I. N. (2003). Current approaches to whole genome phylogenetic analysis. *Briefings in Bioinformatics*, 4(1) :63–74.

Schmidt, H. A. (2003). *Phylogenetic Trees from Large Datasets*. PhD thesis, Dusseldorf, Germany.

Schmidt, H. A., Strimmer, K., Vingron, M., and von Haeseler, A. (2002). TREE-PUZZLE : maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, 18 :502–504.

Semple, C. (2002). Reconstructing minimal rooted trees. *Discrete Applied Mathematics*, 127 :489–503.

Semple, C., Daniel, P., Hordijk, W., Page, R. D. M., and Steel, M. (2004). Supertree algorithms for ancestral divergence dates and nested taxa. *Bioinformatics*, 20(15) :2355–2360.

Semple, C. and Steel, M. (2000). A supertree method for rooted trees. *Discrete Applied Mathematics*, 105 :147–158.

Semple, C. and Steel, M. (2002). Cyclic permutations and evolutionary trees. Research Report 2002/15, University of Canterbury, Department of Mathematics and Statistics, Canterbury, New Zealand.

Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. *Systematic Biology*, 51 :492–508.

Shimodaira, H. and Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution*, 16 :1114–1116.

Sibson, R. (1972). Order invariant methods for data analysis. *Journal of Royal Statistician Society*, 34 :311–349.

Simões-Pereira, J. M. S. (1969). A note on the tree realizability of a distance matrix. *Journal of Combinatorial Theory*, 6 :303–310.

Smith, A. B., Littlewood, D. T. J., and Wray, G. A. (1995). Comparing patterns of evolution : larval and adult life history stages and ribosomal rna of post-palaeozoic echinoids. *Philosophical Transactions : Biological Sciences*, 349 :11–18.

Smolenskii, Y. A. (1969). A method for linear recoding of graphs. *Computational Mathematics and Mathematical Physics*, 2 :396–397.

Sober, E. (1988). *Reconstructing the past : Parsimony, evolution, and inference*. MIT Press, Cambridge, MA.

Sokal, R. R. and Rohlf, F. J. (1981). Taxonomic congruence in the leptopodomorpha re-examined. *Systematic Zoology*, 30 :309–325.

Springer, M. S. and de Jong, W. W. (2001). Phylogenetics. which mammalian supertree to bark up ? *Science*, 291 :1709–1711.

Stamatakis, A., Ludwig, T., and Meier, H. (2005). RAxML-III : a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics*, 21(4) :456–463.

Steel, M. and Penny, D. (1993). Distribution of tree comparison metrics – Some new results. *Systematic Biology*, 42(2) :126–141.

Steel, M. A. (1992). The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of Classification*, 9 :91–116.

Strimmer, K., Goldman, N., and von Haeseler, A. (1997). Bayesian probabilities and quartet puzzling. *Molecular Biology and Evolution*, 14 :210–213.

Strimmer, K. and von Haeseler, A. (1996). Quartet puzzling : A quartet maximum likelihood method for reconstructing tree topologies. *Molecular Biology and Evolution*, 13 :964–969.

Strimmer, K. and von Haeseler, A. (1997). Likelihood-mapping : A simple method to visualize phylogenetic content of a sequence alignment. *Proceedings of the National Academy of Science of USA*, 94 :6815–6819.

Studier, J. A. and Keppler, K. J. (1988). A note on the neighbor-joining method of Saitou and Nei. *Molecular Biology and Evolution*, 5 :729–731.

Susko, E. (2003). Confidence regions and hypothesis tests for topologies using generalized least squares. *Molecular Biology and Evolution*, 20 :862–868.

Swofford, D. L. (2002). *PAUP* :Phylogenetic Analysis using Parsimony (*and other methods), version 10*, Sinauer, Sunderland, Massachussetts edition.

Swofford, D. L., Olsen, G. J., Waddell, P. J., and Hillis, D. M. (1996). Phylogenetic inference. In Hillis, D. M., Moritz, C., and Mable, B. K., editors, *Molecular Systematics*. Sinauer Associates, Massachussets, Sinauer Associates.

Tassy, P. (1986). *L'ordre et la diversité du vivant. Quel statut scientifique pour les classifications biologiques ?* Fondation Diderot / Fayard, Paris.

Templeton, A. (1983). Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. *Evolution*, 37 :221–224.

Thorley, J. L. (2000). *Cladistic information, leaf stability, and supertree construction*. PhD thesis, University of Bristol, UK.

Uno, T. and Yagiura, M. (2000). Fast algorithms to enumerate all common intervals of two permutations. *Algorithmica*, 26(2) :290–309.

Vach, W. and Degens, P. O. (1991). Least-squares approximation of additive trees to dissimilarities characterizations and algorithms. *Computational Statistics Quaterly*, 3 :203–218.

Voudouris, C. (1997). *Guided local search for combinatorial optimisation problems*. PhD thesis, Department of Computer Science, University of Essex, Colchester, UK.

Wiens, J. J. (1998a). The accuracy of methods for coding and sampling higher-level taxa for phylogenetic analysis : A simulation study. *Systematic Biology*, 47 :381–397.

Wiens, J. J. (1998b). Combining data sets with different phylogenetic histories. *Systematic Biology*, 47 :568–581.

Wiens, J. J. (1998c). Does adding characters with missing data increase or decrease phylogenetic accuracy ? *Systematic Biology*, 47 :625–640.

Wiens, J. J. and Servedio, M. R. (1998). Phylogenetic analysis and intraspecific variation : Performance of parsimony, likelihood, and distance methods. *Systematic Biology*, 47 :228–253.

Wilkinson, M., Cotton, J. A., and Thorley, J. L. (2004a). The information content of trees and their matrix representations. *Systematic Biology*, 53 :989–1001.

Wilkinson, M., Thorley, J. L., Littlewood, D. T. J., and Bray, R. A. (2001). *Interrelationships of the Platyhelminthes*, chapter Towards a phylogenetic supertree of Platyhelminthes, pages 292–301. Taylor & Francis, London, Littlewood, D. T. J. and Bray, R. A. edition.

Wilkinson, M., Thorley, J. L., Pisani, D., Lapointe, F.-J., and McInerney, J. O. (2004b). *Phylogenetic supertrees : Combining information to reveal the tree of life*, chapter Some desiderata for liberal supertrees. Kluwer Academic, Dordrecht, The Nederlands, O.R.P. Bininda-Emonds edition.

Williams, T. L. and Moret, B. M. E. (2003). An investigation of phylogenetic likelihood methods. In *Third IEEE Symposium on BioInformatics and BioEngineering (BIBE'03)*, page 79.

Williams, W. T. and Clifford, H. T. (1971). On the comparison of two classifications of the same set of elements. *Taxon*, 20 :519–522.

Willson, S. J. (1999). Building phylogenetic trees from quartets by using local inconsistency measures. *Molecular Biology and Evolution*, 16 :685–693.

Yang, Z. (1994). Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution*, 39 :105–111.

Yang, Z. (1996a). Among-site rate variation and its impact on phylogenetic analysis. *Trends in Ecology and Evolution*, 11 :367–372.

Yang, Z. (1996b). Maximum-likelihood models for combined analysis of multiple sequence data. *Journal of Molecular Evolution*, 42 :587–596.

Yang, Z. (1997). PAML : a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics*, 13(5) :55–556.

Yule, G. U. (1925). A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis. *Philosophical Transactions of the Royal Society of London, Ser. B,*, 213 :21–87.

Zaretskii, K. (1965). Postroenie dereva po naboru rasstoianii mezhdu visiacimi vershinami. *Uspekhi Mathematicekih Nauk*, 20 :90–92.

Résumé

L'inférence phylogénomique cherche à combiner le signal évolutif induit par un ensemble de gènes dans le but de construire un unique arbre phylogénétique. Elle peut être décomposée en trois grandes familles méthodologiques : la combinaison basse, qui s'appuie sur la concaténation des différents gènes, la combinaison haute, qui considère l'ensemble des arbres inférés à partir de chaque gène, et la combinaison moyenne, qui encode les différents signaux phylogénétiques puis combine ces différents encodages. Une méthode d'inférence d'arbre est ensuite appliquée sur le résultat de la combinaison.

Cette thèse développe de nouveaux scénarios d'inférence phylogénomique, principalement basés sur l'estimation de distances évolutives entre chaque paire de taxons. Elle propose une nouvelle méthode de combinaison moyenne, nommée SDM, qui considère les matrices de distance estimées à partir de chaque gène et qui les combine en une unique supermatrice de distance. Cette dernière pouvant parfois contenir des distances manquantes, cette thèse décrit également de nouveaux algorithmes, nommés NJ*, UNJ*, BIONJ* et MVR*, permettant d'inférer très rapidement un arbre à partir d'une matrice de distance complète ou incomplète. De nombreuses simulations ont permis d'observer les bonnes performances de ces nouvelles méthodes de distance. Initialement développées pour la combinaison moyenne, elles permettent toutefois d'améliorer significativement les résultats de certaines approches standards en combinaison basse, et représentent une alternative efficace à MRP, la plus utilisée des techniques de combinaison haute, en termes de fiabilité et de rapidité. La taille des jeux de données phylogénomiques étant de plus en plus importante, les méthodes développées dans cette thèse constituent ainsi des outils de choix pour construire l'Arbre de la Vie.

Mots-clés Phylogénie - Phylogénomique - Superarbre - Distance Evolutive - Neighbor Joining

Keywords Phylogenetics - Phylogenomics - Supertree - Evolutionary Distance - Neighbor Joining