Méthodes et algorithmes pour l'approche statistique en phylogénie

Stéphane Guindon

U n i v e r s i t é M o n t p e l l i e r I I
— Sciences et Techniques du Languedoc —
U.F.R. Sciences de Montpellier

Thèse

pour obtenir le grade de
Docteur de l'Université Montpellier II

Discipline : Biologie
Formation Doctorale : Biologie Santé
École Doctorale : Sciences Chimiques et Biologiques pour la Santé

Méthodes et algorithmes pour l'approche statistique en phylogénie

par

Stéphane Guindon

présentée et soutenue publiquement le *7 juillet 2003* devant le jury :

M. Pierre Darlu, Directeur de recherche, CNRS/Université Denis Diderot, Paris ..... Examinateur
M. Olivier Gascuel, Directeur de recherche, CNRS/LIRMM, Montpellier ..... Directeur de thèse
M. Manolo Gouy, Directeur de recherche, CNRS/Université Claude Bernard, Lyon ..... Rapporteur
M. Alain Jean-Marie, Professeur, UMII, Montpellier ..... Examinateur
M. Hervé Philippe, Professeur, Université de Montréal ..... Rapporteur

Remerciements

Cette thèse est dédiée à mon oncle, Hugues, et ma grand-mère, Marie-Lou.

Je tiens évidemment à remercier Olivier Gascuel, mon directeur de thèse pour avoir guidé adroitement et calmement mes recherches. Sa disponibilité, sa rigueur scientifique et son intuition m'ont beaucoup aidé. Ces presque quatre années passées sous sa direction ont été très formatrices, tant au niveau scientifique que personnel.

Je remercie également les membres de mon jury et plus particulièrement les deux rapporteurs, Manolo Gouy et Hervé Philippe pour leurs judicieuses remarques à propos du mémoire, ainsi que Pierre Darlu pour la pertinence de ses questions. Être jugé par des chercheurs aussi brillants est un honneur. Un grand merci aussi à Alain Jean-Marie pour ses connaissances sans faille en statistique et son aide précieuse concernant les chaînes de Markov.

Merci aussi à Nicole pour sa gentillesse et son aide indispensable. À Marie et Julie pour leur patience inébranlable devant mon incapacité à gérer simplement la moindre démarche administrative.

Merci à Emmanuel Douzery pour avoir suivi mon travail régulièrement. À Frédéric Delsuc (j'arrive!), Marc Robinson-Rechavi, Vincent Berry, et Xavier Perrier pour m'avoir accueilli chaleureusement au sein du groupe de travail «super-arbre ». À tous les membres du laboratoire de Biométrie et Biologie Évolutive, de Lyon, et plus particulièrement à Loïc Ponger et Guy Perrière.

Merci surtout à tous les thésards et ex-thésards du département d'informatique fondamentale au LIRMM avec lesquels j'ai passé de très bons moments. Ils m'ont initié aux algorithmes, à linux et avant tout au BLAST. Dans un ordre quelconque : merci à Vincent pour sa gentillesse et pour m'avoir fait découvrir la magie du fameux «double parcours récursif ». À Hervé (Herv8 devrais-je écrire) : élu miiIiister grooOoovy à l'unanimité. À David pour avoir repris à zéro ma formation en C (je te dédie mon For à moi), m'avoir fait comprendre la différence entre un bon programmeur et un bon programmeur, et surtout, pour ses caricatures tordantes. À Pat pour ses connaissances pratiquement infinies sur linux et l'informatique en général. Merci surtout à toi, Pat, pour tout le temps consacré à ma formation d'apprenti root. À Laurent pour sa gentillesse, son humour, ses conseils précieux en statistique et son bon sens. Merci aussi pour l'algo de projection d'un arbre sur un autre. À Gilles pour ne pas m'avoir fait écouter sa musique. À Rico pour m'avoir fait mourir de rire en m'assaillant de fautes d'orthographes. À Yannic pour avoir ostenté pendant des mois une certaine affiche, honteusement pastichée par un certain D. Genest dont je ne citerai que le prénom : David, et présentant le très talentueux «Bouley» dans une situation fort inconfortable. Au P'tit Jeune pour sa simplicité et pour sa bonne humeur malgré toutes les tares qui

l'affligent. Et enfin, un grand merci à Sandra : 1 thèse = 50000 km à vélo (finalement, il est bouclé ce tour du monde, non ?), et à Sèverine qui m'a adressé, j'en suis sûr et certain, un bon millier de grimaces durant ces trois ans passés dans mon dos. À tous, merci beaucoup !

Merci à mes amis de toujours : Mary, Mathieu, Steph et bien sur ce vieux Alex. Merci aussi à Antoine, Magali, Dave et Anne-Laure ainsi que toutes les autres personnes rencontrées à Montpellier (Jean-Claude, Jean, Fabrice, ...).

Merci à ma douce et tendre, Elsa, pour m'avoir aimé et soutenu durant ces quelques années. Puissions-nous continuer notre chemin ensemble pendant encore longtemps.

Enfin, merci à mes parents pour m'avoir convaincu de persévérer dans les études. Merci d'avoir été fiers de moi depuis toujours, de m'avoir guidé sans jamais rien m'imposer.

Table des matières

Introduction

La phylogénie moléculaire est la discipline visant à reconstruire l'histoire évolutive de séquences génétiques et, dans la plupart des cas, à en déduire celle des espèces correspondantes. Elle constitue une branche majeure de la **systématique** et de la **génomique comparative**. Ces deux champs d'applications incluent en effet une dimension historique qui peut être décrite simplement grâce à l'outil phylogénétique. Ainsi, outre l'identification, la description, et l'inventaire des êtres vivants, une des principales tâches de la systématique est la classification des espèces à partir d'arbres phylogénétiques. Or, cette étape est capitale car elle seule permet de rendre intelligible l'immense diversité des espèces. De même, l'extraction d'informations pertinentes à partir de la comparaison de séquences génétiques repose fréquemment sur l'histoire des séquences analysées. Par exemple, l'identification de familles de gènes orthologues ou la détection de transferts horizontaux de matériel génétique repose sur l'inférence et la comparaison de phylogénies.

Aujourd'hui, la plupart des méthodes de construction d'arbres sont basées sur la **modélisation statistique** de l'évolution des macromolécules. Les approches proposées procèdent à partir d'un modèle de substitution entre les états des caractères considérés et d'une phylogénie. Cette dernière correspond à une topologie d'arbre, présentant les parentés évolutives entre organismes, munie de longueurs de branches. Ces deux ensembles de paramètres sont considérés conjointement et la fiabilité de la topologie inférée est généralement dépendante de la qualité de l'estimation des longueurs de branches et des paramètres du modèle de substitution. Une partie du travail réalisé pendant cette thèse porte sur cette interdépendance. Plus précisément, nous proposons une méthode d'estimation du paramètre mesurant la variabilité des vitesses entre sites, particulièrement adaptée à la construction de phylogénies à partir de distances évolutives. De manière surprenante, cette approche empirique, renforcée par des arguments théoriques, montre que la valeur optimale de ce paramètre n'est pas nécessairement la valeur sous-jacente au processus évolutif.

Outre une paramétrisation adéquate des modèles de substitution, un autre aspect important de la phylogénie moléculaire est la mise au point d'algorithmes efficaces de construction d'arbres. Parmi les différentes classes de méthodes, l'approche du **maximum de vraisemblance** est parmi les plus fiables. Malheureusement, celle-ci est pénalisée par la lourdeur des calculs impliqués et l'analyse de grands jeux de données est, de ce fait, difficile. Une solution à ce problème est présentée ici. Les performances de l'algorithme que nous avons développé sont très satisfaisantes. En effet, pour une précision dans l'estimation de topologies d'arbres égale, voire supérieure, à celles des approches les plus performantes de ce point de vue, les temps de calculs sont comparables à ceux des méthodes de distances, pourtant réputées pour leur

rapidité.

Le premier chapitre de ce mémoire est consacré à la description du cadre historique qui à vu naître la science de la classification des espèces. Les principaux courants de pensées ainsi que l'apparition des données moléculaires en phylogénie sont aussi présentés. Le second chapitre s'attache à la description des fondements et des outils de l'approche statistique en phylogénie moléculaire. Nous présentons notamment les procédures classiques pour le calcul de distances évolutives ou la vraisemblance d'un arbre. Le troisième chapitre décrit les algorithmes d'inférence d'arbres. Sans prétendre à une description exhaustive de ces méthodes, nous présentons, discutons et comparons les principales d'entre elles, et insistons plus particulièrement sur les avancées récentes dans ce domaine. Les deux derniers chapitres sont consacrés à la présentation des travaux effectués pendant cette thèse. Le quatrième chapitre porte sur l'influence du paramètre décrivant la variabilité des vitesses entre sites sur l'estimation de topologies d'arbres. Enfin, le cinquième chapitre concerne la description d'une méthode rapide et fiable pour l'estimation de phylogénies suivant le principe du maximum de vraisemblance.

Chapitre 1

Des premières classifications à l'avènement de la phylogénie moléculaire

Ce chapitre a pour objectif de replacer la classification des espèces et la phylogénie moléculaire dans leurs contextes historique et intellectuel respectifs. Pour débuter, nous présentons quelques points importants sur la genèse de la classification des espèces en tant que discipline scientifique ainsi que l'avènement des données moléculaires en systématique. Les principaux éléments méthodologiques caractérisant les différentes approches pour la reconstruction de phylogénies sont introduits ensuite.

1.1 Recherche d'une classification naturelle

Bien que le mot taxinomie –ou **taxonomie**– ait été proposé au XIX^{e} siècle, les botanistes du troisième siècle avant Jésus Christ inventoriaient et classaient déjà les plantes médicinales. Cependant, il faut attendre le XVI^{e} siècle pour observer un réel foisonnement de méthodes de classification, là-encore réservées à l'étude des végétaux. Ces tentatives étaient fondées sur différents critères : taille, forme générale, forme des feuilles, des racines, etc., et se partageaient entre deux types d'approches. La première, héritée d'Aristote, est une logique «**divisive**» où l'ensemble des organismes est partagé suivant des critères prédéfinis. Ce partage est réitéré jusqu'à aboutir aux seules espèces. L'autre approche, dite «**agglomérative**», consiste à rassembler les espèces sur des critères de similitudes et répéter cette opération en considérant comme nouvelles unités les groupes ainsi définis. Carl von Linné (1707-1778), combinant ces deux logiques, est le premier à faire apparaître la notion de **niveaux hiérarchiques**. Les sept rangs traditionnels que sont les règnes, les embranchements, les classes, les ordres, les familles, les genres et les espèces sont alors constitués. Pour la petite histoire, six niveaux avaient été proposés à l'origine, mais Linné a ramené à sept le nombre de rangs car, dans l'esprit de l'époque, le nombre sept était supposé parfait. Bien évidemment, cette règle des sept niveaux hiérarchiques s'est trouvée rapidement transgressée et des rangs intermédiaires n'ont pas tardé à apparaître.

À la fin du XVIII^{e} siècle un résultat particulièrement inattendu s'imposa aux systématiciens. Dans la très grande majorité des cas, la plupart des grandes familles (rosacées, papilionacées, graminées, etc.) étaient retrouvées, et ceci quel que soit la méthode de classification employée. Or, ne perdons pas de vue que, jusqu'à cette époque, la classification des espèces n'avait qu'un but pratique : celui d'organiser, de «ranger» les êtres vivants suivant une procédure prédéfinie. Dans cette optique, observer des ressemblances entre deux classifications bâties à partir de méthodes distinctes est inattendu. La mise en évidence de familles conservées quel que soit la méthode de classification utilisée, indiqua qu'il existait peut-être une **classification naturelle**. Cette éventuelle classification naturelle amenait à penser que la Nature tout entière était ordonnée. Cet ordre était évidemment celui de la création divine.

Dès lors, les botanistes s'efforcèrent d'établir une méthode qui permette d'accéder à la classification naturelle. Antoine-Laurent de Jussieu (1748-1839) introduisit alors le **principe de subordination** qui fut une véritable révolution. Ce principe est inspiré d'une logique divisive : les espèces sont classées suivant une hiérarchie de caractères préétablie, inspirée du développement (ontogénie). La nouvelle classification des plantes ainsi proposée fut considérée comme naturelle. Elle remplaça très rapidement celle de Linné et une grande partie des familles distinguées par Jussieu sont encore reconnues aujourd'hui. Dès la fin du XVIII^{e} siècle, cette méthode fut appliquée aux animaux par Georges Cuvier (1769-1832) et ceux-ci furent alors scindés en quatre embranchements (Vertebrata, Arthropoda, Mollusca et Radiata).

À la fin du siècle des Lumières, la classification naturelle n'était plus unanimement considérée comme un ordre d'origine divine. Dès lors, l'enjeu n'était plus de construire cette classification, mais plutôt de donner un sens à celle-ci. Le **transformisme**, présenté par Jean-Baptiste Lamarck (1744-1829) en 1809, est issu de ce questionnement. Lamarck esquissa un mécanisme de la transformation des êtres vivants, associé à ce que l'on appellera plus tard l'hérédité des caractères acquis. Il fut ainsi le premier à proposer une filiation des animaux et ébaucher le concept de phylogénie. Mais les naturalistes de l'époque tenaient les différences entre individus comme un désordre négligeable, et c'est Charles Darwin (1809-1882) qui, le premier, s'intéressa aux variations de certains caractères morphologiques au sein d'une même espèce. De ces travaux naquît le célèbre paradigme de la **sélection naturelle** qui structure désormais la pensée en biologie. Darwin postula qu'au cours de la transmission des caractères d'une génération à la suivante, les variations favorables sont maintenues et les variations défavorables sont éliminées. Une population évolue ainsi au cours des générations. À une échelle supérieure, la variation des caractères héréditaires conjuguée à la sélection implique que les ressemblances entre espèces sont dues aux caractères hérités d'une espèce ancestrale. Comme ce raisonnement peut être utilisé à rebours, il devient opérationnel pour bâtir une classification fondée sur une recherche de parentés. C'est tout l'enjeu de la phylogénie. Ainsi, établir la classification naturelle des espèces revient à décrire leur histoire. Il fallut attendre plus d'un siècle pour que cette avancée fondamentale aboutisse à un paradigme opérationnel.

Durant cette centaine d'années, aucun cadre formel convainquant n'est proposé pour la classification.

Ainsi, les relations d'apparentement entre espèces sont parfois mêlées à des considérations adaptatives et écologiques. Par exemple, la parenté évolutive entre oiseaux et reptiles, clairement mise en évidence par l'analyse comparative des orteils et du bassin, était «masquée» sous prétexte qu'avec le vol les oiseaux avaient franchi un «saut adaptatif» qui méritait une classe à part. Plus grave encore, les relations de descendances entre espèces (la généalogie) sont souvent confondues avec les relations d'apparentement (la phylogénie).

Les premiers véritables efforts de formalisation pour la classification des espèces voient le jour au milieu du XX^{e} siècle. En 1950, l'entomologiste allemand Willi Hennig fonde la systématique phylogénétique, ou **cladistique**. En réaction à cette approche apparaissent la **taxonomie numérique**, fondée par Sneath et Sokal en 1963, et l' **approche probabiliste**, dérivant des travaux de Fisher en 1920, et développée par Edwards et Cavalli-Sforza en 1963. Bien que ces trois «écoles» présentent des caractéristiques communes, tant au niveau du fond que de la forme, elles divergent sur des points fondamentaux. La troisième partie de ce chapitre est consacrée à ceux-ci.

1.2 L'introduction des données moléculaires

En 1955, Fred Sanger et ses collaborateurs décryptent la séquence en acides aminés de l'insuline chez trois mammifères. Leur comparaison montre une corrélation évidente avec les différences anatomiques observées. Cependant, Zuckerkandl et Pauling (1962) sont les premiers à véritablement prôner l'utilisation des séquences de gènes pour la construction de phylogénies. À la même période, Margaret Dayhoff initie l'application de l'informatique à l'analyse de séquences. Elle développe ainsi des programmes pour l'analyse de protéines et propose, en 1965, un atlas décrivant 65 séquences. Au début des années 1970, Carl Woese et ses collaborateurs créent une base de données regroupant des séquences d'ARN de la petite sous-unité ribosomique. En 1977, l'analyse phylogénétique de ces séquences conduit à la séparation des trois grands «domaines» du vivant : Archaebactéries, Eubactéries et Eucaryotes, et confirme l'hypothèse d'endosymbiose en série pour l'origine des mitochondries et chloroplastes. Parallèlement, des avancées spectaculaires sont réalisées dans les techniques de séquencages et bientôt la première banque de séquences nucléiques, EMBL, voit le jour à Heidelberg en 1980. Dès lors, l'accumulation exponentielle de séquences génétiques, couplée à la création de nouvelles bases de données telles que Genbank, NBRF, DDBJ ou SwissProt, et de logiciels pour interroger celles-ci : ACNUC (Gouy et al., 1984), ENTREZ (NCBI, National Center for Biotechnology Information) ou SRS (Etzold et Argos, 1993), a permis de populariser l'utilisation des données moléculaires en phylogénie. En Janvier 2003, GenBank contenait plus de 17 milliards de nucléotides issus d'environ 100,000 espèces distinctes.

Outre ces outils informatiques dédiés à l'analyse des génomes au sens large, des efforts pour trier et organiser l'information de manière plus spécifique ont été entrepris. Ainsi, la base de données RDP (Ribosomal Database Project, Maidak et al., 1994-2001) est spécialement dédiée à l'analyse phylogénétique

et poursuit ainsi, 20 ans plus tard, les travaux initiés par C. Woese. Signalons aussi le projet Deep Green (voir Brown, 1999), portant sur l'analyse des données moléculaires pour inférer des phylogénies de plantes. Le succès rencontré par ces travaux, justifié notamment par des résultats spectaculaires concernant l'origine des plantes à fleurs, est incontestable. Ces découvertes sont d'une telle importance que bon nombre de botanistes sont prêts à remettre en cause le système linnéen de classification des plantes, pourtant en place depuis environ 250 ans !

1.3 Trois écoles pour la systématique

La cladistique, la phénétique et l'approche probabiliste sont les trois cadres formels au sein desquels se sont développées l'ensemble des méthodes modernes de classification. Comme nous le verrons, ces approches divergent tant sur le fond que sur la forme. L'objectif ici n'est pas de détailler leurs oppositions mais plutôt de présenter les principaux repères méthodologiques.

1.3.1 La cladistique

La première étape d'une analyse cladistique classique est généralement la **polarisation des caractères** : quels sont les états ancestraux, ou **plésiomorphes** et les états dérivés, ou **apomorphes** ? Pour les données moléculaire, cette polarisation est généralement réalisée implicitement au sein de l'algorithme de construction de l'arbre. Pour des données morphologiques la polarisation constitue une étape à part entière. Les fossiles des caractères contemporains étudiés apportent généralement les réponses les moins ambiguës. Malheureusement, ce type de données n'est pas systématique disponible et des critères de polarisation indirects doivent être appliqués. L'**ontogénie** est un de ces critères. Il est fondé sur la loi d'Haeckel selon laquelle l'ontogénie récapitule la phylogénie, autrement dit, lors de son développement, l'embryon présente la succession des états ancestraux d'un caractère donné. Par exemple, à certains stades embryonnaires primitifs de la lamproie, du lézard et du chat, la majorité du squelette est constitué de cartilage. Au cours des étapes suivantes du développement, l'os remplace en grande partie le cartilage chez le lézard et le chat. D'après la loi d'Haeckel, le cartilage est l'état primitif du caractère squelette, l'os étant le caractère dérivé. Notons que, dès les années vingt, de nombreux contre-exemples ont mis en doute la validité de cette loi. Hennig, lui-même, rejette totalement l'idée que l'ontogénie puisse permettre de polariser les caractères.

La deuxième approche pour l'identification des caractères ancestraux et dérivés est l'utilisation d'un groupe externe ou **outgroup**. Un groupe externe est une espèce, ou un groupe d'espèces, dont on est sur qu'elle (il) est extérieur(e) au groupe d'étude. Une fois le groupe externe défini, la polarisation des caractères s'effectue simplement : si, chez une ou plusieurs espèces(s) du groupe étudié (par exemple, un crocodile et un alligator), l'état du caractère (écaille) est le même que dans le groupe extérieur (présence

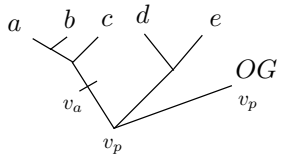
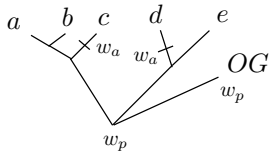
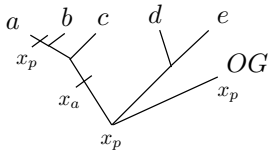
caractères	partitions	cladogramme	nombre de pas
v	a, b, c		1
w	c, d		2
x	b, c		2

FIG. 1.1 – **Différents partages de caractères ancestraux et dérivés.** Les caractères binaires analysés sont notés v, w, et x. v_p et v_a désignant les états plésiomorphes et apomorphes de v. Pour chaque caractère, les UEs partageant le même état forment une partition. Le nombre de pas est égal au nombre estimé de substitutions d'un état par un autre au sein du cladogramme.

d'écaille chez la tortue), cet état est considéré comme plésiomorphe au sein du groupe d'étude. Si, chez d'autres espèces (poulet, autruche), l'état (plume) du caractère est différent de celui présent au sein du groupe externe (écaille), cet état est considéré comme apomorphe. Considérer l'état du caractère observé chez l'outgroup comme étant «ancestral» repose sur l'hypothèse que ce dernier n'a pas divergé de l'ancêtre de toutes les espèces comparées. Or cette hypothèse n'est généralement pas vérifiable sauf lorsque des données fossiles sont disponibles. Cependant, les méthodes de construction d'arbres basées sur la cladistique n'impose pas de connaître les états des caractères analysés à la racine de l'arbre. Ainsi, la polarisation telle qu'elle est évoquée ici consiste véritablement à enraciner l'arbre plutôt qu'à identifier les états ancestraux des caractères.

Lorsqu'un outgroup est constitué, débute alors la construction d'un **cladogramme**. La Figure 1.1 donne le principe de l'évaluation d'un cladogramme. Afin de simplifier le propos, les caractères étudiés, v, w et x sont binaires. Les états plésiomorphes sont notés v_p, w_p et x_p ; les états apomorphes correspondants : v_a, w_a et x_a. Le groupe d'étude est constitué des espèces a, b, c, d, et e, et le groupe externe est noté OG. Dans le cadre de la phylogénie, les unités comparées (ici des espèces) sont des **unités évolutives** ou UEs. Les UEs a, b, et c présentent l'état dérivé du caractère v. Cette partition des espèces forme un **clade**. Le partage de l'état apomorphe d'un caractère par deux espèces appartenant à un même clade est une **synapomorphie** : la présence du caractère dérivé v_a chez a, b et c traduit une synapomorphie. Le partage de l'état plésiomorphe par deux espèces est une **symplésiomorphie** : la présence du caractère

ancestral v_p chez d et e traduit une symplésiomorphie. Synapomorphies et symplésiomorphies sont toutes deux des **homologies** : ce sont des similarités par filiations.

Les caractères w et x sont en désaccord avec les clades de l'arbre choisi. On considère ici que le passage de l'état w_p à l'état w_a est survenu dans deux lignées distinctes. Il s'agit d'une **convergence**. L'état x_p présent chez a, d et e n'est pas hérité d'un ancêtre commun à ces trois espèces : il est ancestral à d et e et secondairement transformé chez a. Il s'agit d'une **réversion**. Convergence et réversion sont des évènements faussant les relations entre espèces, ce sont des **homoplasies**.

Le cladogramme présenté dans notre exemple n'est évidemment pas le seul qu'il est possible de construire. Pour 6 UEs il existe en effet 105 topologies d'arbres distinctes. Il est donc nécessaire de choisir entre les différents scénarios évolutifs. Le **critère de parcimonie maximale** permet d'effectuer ce choix. Parmi l'ensemble des possibilités, le(s) cladogramme(s) retenu(s) est (sont) celui (ceux) qui requiert (requièrent) le plus petit nombre de substitutions d'un état par un autre. Dans la terminologie cladiste, chacune de ces substitutions correspond à un pas (voir colonne de droite de la Figure 1.1). Ainsi, pour reconstruire le ou les arbres les plus parcimonieux dans notre exemple, il est nécessaire d'examiner toutes les configurations topologiques qu'il est possible d'obtenir à partir d'un arbre enraciné à six UEs, de compter le nombre de pas associés à chacun, et de choisir ensuite la ou les configurations associées(s) au plus petit de ces nombres.

Il est généralement impossible d'examiner l'ensemble des topologies d'arbres car leur nombre croît exponentiellement avec le nombre d'UEs. Le recours à des **heuristiques** est donc indispensable. Ces dernières permettent de «parcourir» la fonction à minimiser (ou maximiser). Pour la parcimonie, l'objectif est de trouver la topologie qui minimise le nombre de pas. Cette recherche s'achève lorsqu'un extremum est atteint. La plupart des heuristiques ne garantissent pas le recouvrement d'un extremum global. Le ou les arbres obtenus correspondent alors à des extrema locaux de la fonction. Tout l'art dans l'établissement d'heuristiques performantes concerne l'évitement de ces minima locaux. Les chapitres 3 et 5 sont consacrés à ce sujet.

1.3.2 La taxonomie numérique

Les systématiciens partisans de cette approche infèrent des arbres –ou **phénogrammes**– à partir des similitudes globales entre UEs. La première partie du travail consiste donc à construire une matrice de distances entre UEs prises deux à deux. La fréquence des différences observées sur la suite de caractères considérée est une mesure de la distance entre deux UEs. Naturellement, il existe des méthodes plus sophistiquées de calcul de distances, faisant notamment appels à des **modèles stochastiques**, traduisant des hypothèses sur le mode de substitution entre les différents états des caractères. Nous reviendrons largement sur ces modèles par la suite.

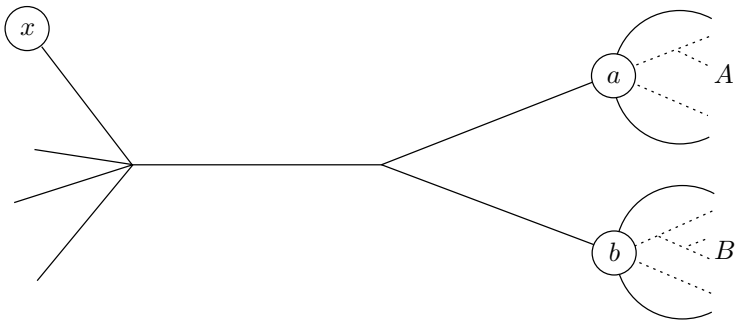


FIG. 1.2 – **Agglomération de a et b.** a et b sont les noeuds racines des sous-arbres A et B. x est un noeud externe ou un noeud interne à la racine d'un sous-arbre.

Les premières approches pour bâtir des phénogrammes sont basées sur une méthode de classification ascendante. Notons Δ la matrice de distances estimées. Δ_{ij} est la distance entre les éléments i et j. L'algorithme de construction d'un phénogramme comprend trois étapes qui sont répétées tant que la dimension de la matrice est supérieure à 1. (1) La distance minimale observée au sein de Δ désigne une paire de noeuds voisins dans l'arbre, notés a et b. (2) Ceux-ci sont joints et les longueurs des deux branches engendrées par cette agglomération sont toutes deux égales à $\frac{1}{2}\Delta_{ab}$. (3) Δ est ensuite réduite en remplaçant les entrées correspondant à a et b par une seule entrée, notée u.

Les différentes stratégies pour calculer les distances entre u et les autres entrées de la matrice définissent les variantes de cette méthode. L'étape de réduction repose généralement sur une variante de l'expression suivante :

$$\Delta_{ux} = \alpha_a \Delta_{ax} + \alpha_b \Delta_{bx}, \text{ avec } \alpha_a + \alpha_b = 1$$

a et b sont les racines des sous-arbres A et B présentant $|A|$ et $|B|$ UEs respectivement et x est une UE, ou un groupe d'UEs déjà constitué, exclue de A et B (Figure 1.2). Lorsque $\alpha_a = |A|/(|A| + |B|)$ et $\alpha_b = |B|/(|A| + |B|)$, les poids associés à A et B sont proportionnels à leurs effectifs. Par conséquent, un seul et même poids est associé à chacune des UEs présente au sein de ces sous-arbres. Cette approche non pondérée correspond alors à l'algorithme **UPGMA** («Unweighted Paired Group Method using Averages») (Sokal et Michener, 1958). Lorsque $\alpha_a = 1/2$ et $\alpha_b = 1/2$, le poids associé à chaque UE est inversement proportionnel à l'effectif du sous-arbre auquel il appartient. Cette approche correspond à l'algorithme **WPGMA** («Weighted Paired Group Method using Averages») (McQuitty, 1966). Notons ici que UPGMA et WPGMA sont fréquemment confondus dans la littérature : WPGMA associant le même poids aux sous-arbre A et B, cette méthode est souvent considérée comme non-pondérée. Cependant, les poids évoqués au sein des deux acronymes UPGMA et WPGMA portent sur les UEs et non les sous-arbres. Or, du point de vue des poids associés aux UEs, UPGMA et WPGMA sont bien des versions pondérées et non pondérées du même algorithme. La Figure 1.3 donne un exemple de construction d'un phénogramme par cette méthode. Enfin, la méthode du **lien simple** fixe $\alpha_a = 1$ et $\alpha_b = 0$ si $\Delta_{ax} < \Delta_{bx}$: seule la plus petite des deux distances parmi Δ_{ax} et Δ_{bx} intervient dans la réduction. La méthode du **lien complet** fixe $\alpha_a = 0$ et $\alpha_b = 1$ si $\Delta_{ax} < \Delta_{bx}$ et seule la plus grande des deux distances

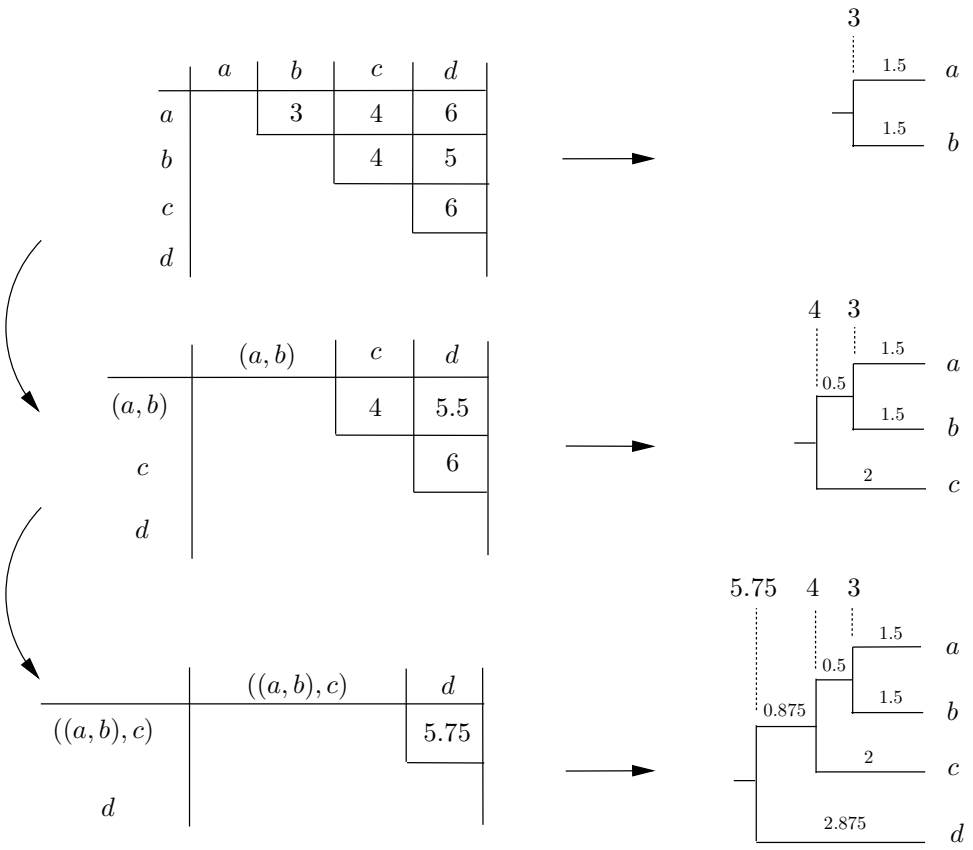


FIG. 1.3 – **Construction d'un phénogramme par WPGMA.** La matrice (à gauche) reporte les distances entre les noeuds agglomérés de l'arbre. Les positions de ces derniers sont indiqués par les valeurs situées au-dessus de l'arbre. Celles-ci correspondent aux distances minimales dans la matrice à chaque étape et désignent donc la suite de paires de noeuds agglomérés aboutissant au phénogramme complet.

intervient dans la réduction.

Le nombre réel de substitutions d'un état par un autre définit une **distance d'arbre** et cette dernière est représentée par une unique **phylogénie** décrivant correctement les liens de parentés entre les organismes comparés. Lorsque les vitesses d'évolution sont constantes, les distances d'arbres sont dites **ultramétriques**. Dans cette situation et pour des séquences suffisamment longues, les phénogrammes construits par UPGMA ou WPGMA correspondent aux vraies phylogénies. En effet, lorsque deux noeuds sont agglomérés, la divergence entre ceux-ci est distribuée de manière égale entre les deux branches créées à cette étape. Le calcul des longueurs de branches suit donc l'hypothèse d'égalité des vitesses d'évolution au sein des lignées. Lorsque les distances réelles ne sont pas ultramétriques, les phénogrammes établis par ces deux méthodes ne correspondent généralement plus aux vraies phylogénies.

Les méthodes ADDTREE (Sattah et Tversky, 1977) et NJ (Saitou et Nei, 1987) permettent cependant d'établir la phylogénie correcte à partir de distances d'arbres sans que celles-ci soient ultramétriques. Cette caractéristique correspond en fait à une divergence profonde de ces deux approches vis à vis des méthodes phénétiques au sens strict. Alors que l'objectif de ces dernières est d'obtenir une simple représentation des

divergences entre UEs, ADDTREE, NJ et la plupart des méthodes actuelles d'inférence d'arbres basées sur les distances évolutives, ont pour but de retracer les liens de parentés entre organismes. ADDTREE et NJ construisent des phylogénies, UPGMA et WPGMA construisent des phénogrammes qui, sous certaines conditions (les distances vraies sont ultramétriques), correspondent à des phylogénies. Une partie du chapitre 3 est consacrée à la description des principales méthodes de construction de phylogénies actuelles à partir de distances entre UEs.

1.3.3 Les méthodes probabilitstes

Les approches actuelles assimilées à ce courant de pensées sont fondées sur le principe statistique du **maximum de vraisemblance**. La vraisemblance d'une phylogénie est la probabilité d'occurrence des données sous cet arbre et pour le modèle d'évolution choisi.

Prenons l'exemple de la phylogénie présentée dans la Figure 1.4. Les UEs sont notées a, b et c et les noeuds ancestraux sont notés r et u. Comme précédemment, nous considérons ici que les caractères sont binaires : les deux états sont notés 0 et 1. Les données correspondent ici aux états 0, 1 et 1 observés chez les UEs a, b et c, respectivement.

La vraisemblance de la phylogénie, notée L, s'écrit ici :

$$L = \sum_{x \in \{1,0\}} P(r = x) P_{x0}(l_a) \sum_{y \in \{1,0\}} P_{xy}(l_u) P_{y1}(l_b) P_{y1}(l_c)$$

où $P(r = x)$ est la probabilité que l'état au noeud r soit x et $P_{x0}(l_a)$ est la probabilité de changement de l'état x au noeud r vers l'état 0 au noeud a, r et a étant joints par une branche de longueur l_a. Les autres termes de l'expression désignent également à des probabilités de changements et sont définis de façon similaire. Ces probabilités sont déduites de modèles stochastiques, identiques à ceux évoqués plus haut dans le cadre de l'estimation de distances évolutives. Comme nous le verrons par la suite, de tels modèles jouent un rôle central dans l'inférence de phylogénies à partir de données moléculaires.

Le principe du maximum de vraisemblance, bien connu en statistique, consiste à sélectionner l'hypothèse maximisant la probabilité d'occurrence des données. Dans le cadre de la phylogénie, cette hypothèse est de nature complexe. Il s'agit en effet d'une topologie d'arbre dont les longueurs de branches sont valuées, ainsi que certains paramètres du modèle de substitution. L'application de cette méthode à l'inférence de phylogénies a été proposée en 1964 par Edwards et Cavalli-Sforza mais c'est à partir de 1981 et des travaux de Felsenstein, que cette approche a commencé à être véritablement utilisée. Les différents aspects de l'estimation de phylogénies suivant ce principe sont abordés aux chapitres 2, 3 et 5.

Le calcul de la vraisemblance est aussi à la base de **l'inférence bayesienne**. L'objectif est ici de trouver l'hypothèse de probabilité maximale au vue des données étudiées. L'application de cette technique d'estimation à la phylogénie consiste à chercher l'arbre, ou la topologie d'arbre, de probabilité *a posteriori* maximale. Outre la vraisemblance, le calcul de cette quantité fait intervenir des probabilités *a priori* sur

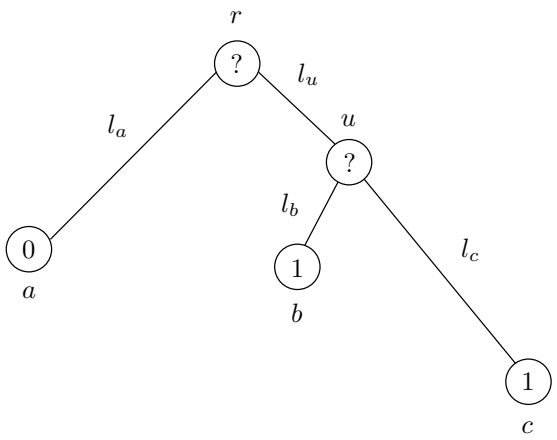


FIG. 1.4 – **Arbre modèle pour le calcul de vraisemblance d'une phylogénie.** 0 et 1 sont les deux états des caractères étudiés. ? : l'état du caractère est inconnu.

la phylogénie. L'inférence bayesienne a été introduite dans le domaine de la phylogénie en 1996 par Yang et Rannala. Elle s'est largement développée depuis et constitue, à l'heure actuelle, une stratégie très répandue pour l'estimation d'arbres à partir de données moléculaires. Le principe de cette approche est présenté au chapitre 2 et détaillé au chapitre 3.

1.4 Vers une approche statistique pour la phylogénie

Les modèles de substitutions jouent, à l'heure actuelle, un rôle central dans le domaine de la phylogénie. Ainsi, comme nous l'avons évoqué précédemment, ils sont à la base de l'estimation de distances évolutives ou du calcul de la vraisemblance d'un arbre. Cette place prépondérante s'explique en grande partie par les caractéristiques favorables des données moléculaires pour l'inférence statistique en phylogénie.

La première de ces caractéristiques réside dans la nature même des macromolécules analysées. Celles-ci sont constituées d'une suite de caractères pouvant prendre un nombre d'états fini (les quatre bases de l'ADN et les vingt acides aminés des protéines). Il est donc possible de proposer des hypothèses simples concernant les mécanismes de substitutions entre ces différents états. À l'opposé, déterminer la probabilité de substitution d'un caractère morphologique par un autre en un temps donné est généralement périlleux.

Le second avantage présenté par les données moléculaires est la possibilité de disposer d'importantes quantités de séquences. Par exemple, la base de données RDP propose quelques 16,000 séquences alignées d'ARN de la petite sous-unité ribosomique. Aussi, en Février 2002, HOBACGEN (Homologous Bacterial Genes Database) (Perrière et al., 2000) contenait 260,025 protéines organisées en près de 24,000 familles homologues potentiellement exploitables pour la construction de phylogénies. Or, dans le domaine des statistiques inférentielles, auquel se rattache l'estimation d'arbres fondée sur des modèles de substitutions, une méthode d'estimation **consistante** est d'autant plus fiable que la taille de l'échantillon analysé est importante.

Ainsi, dès 1969, Jukes et Cantor proposent un modèle de substitution s'appliquant aux séquences nucléiques et protéiques. De nombreuses approches, de plus en plus sophistiquées (voir Swofford et al., 1996 pour une revue) et spécialement adaptées aux données à analyser (voir Tamura et Nei, 1993 pour un exemple), ont été développées par la suite. La validation (et surtout l'invalidation) de ces modèles a permis d'affiner nos connaissances sur la manière dont évoluent les gènes et les protéines. La construction de phylogénies à partir de ceux-ci a aussi largement contribué à remettre en cause ou confirmer certains acquis en systématique (voir Graur et Li, 1991, D'Erchia et al., 1996 puis Murphy et al. 2001 pour l'exemple de la monophylie des rongeurs).

Notons cependant que l'utilisation de modèles stochastiques dans le cas des données moléculaires a été et reste sévèrement critiquée par les partisans de l'école cladiste. Le point le plus conflictuel concerne la validité biologique des modèles proposés. Il est en effet vraisemblable que ces derniers ne prennent pas en compte toute la complexité des processus évolutifs auxquels sont soumis les séquences nucléiques ou protéiques. Or, *a priori*, il semble difficile d'établir des phylogénies fiables à partir de méthodes d'inférence basées sur des hypothèses erronées. Cependant, comme le souligne Yang (1997a), en réponse à Purvis et Quicke (1997), la violation des hypothèses soutenues par le modèle évolutif n'engendre pas obligatoirement l'apparition d'erreurs dans l'estimation des phylogénies. Les méthodes basées sur une approche statistique sont généralement **robustes** vis à vis des écarts au modèle (Gaut et Lewis, 1995), à condition que ce dernier capture les grands traits du mode d'évolution des séquences.

Par conséquent, une telle critique ne justifie en aucun cas l'abandon de l'inférence statistique en phylogénie. En revanche, elle indique à juste titre la nécessité d'établir des modèles pertinents, capables de rendre compte du processus évolutif de manière réaliste, sans toutefois recourir à un nombre excessif de paramètres. Le chapitre suivant présente et discute les solutions proposées.

Chapitre 2

Modéliser l'évolution des séquences

La modélisation statistique est la traduction en termes mathématiques d'hypothèses concernant le processus de génération des données analysées. Dans le domaine de l'évolution moléculaire, les données correspondent à un alignement de séquences homologues et le processus se réfère aux modifications, au cours de l'évolution, de la suite de caractères composant ces séquences. Ce chapitre est consacré à la description des modèles markoviens de substitutions entre bases nucléiques ou acides aminés. Dans un premier temps, nous décrivons les données étudiées, puis les hypothèses sous lesquelles se placent la plupart des modèles sont discutées. Les principaux modèles de substitution sont présentés ensuite et leur utilisation dans le cadre des différentes méthodes d'inférence phylogénétique est détaillée.

2.1 Que modélise-t-on ?

L'évolution moléculaire est rendue possible par l'action conjointe de deux processus : le premier est la génération de nouveaux variants, ou allèles, le second est le maintien ou l'élimination de ceux-ci. Les nouveaux allèles sont engendrés par **mutations**, c'est à dire le remplacement d'un nucléotide par un autre, par **insertion** de nucléotides, ou par la **délétion** d'une partie de la séquence originale. Ces évènements sont, dans la majorité des cas, sans conséquences car ils affectent les cellules somatiques. Néanmoins, lorsque ceux-ci touchent les génomes de cellules germinales, les nouveaux allèles sont transmis à la génération suivante. La **dérive génétique**, c'est à dire l'échantillonnage aléatoire des allèles d'une génération à la suivante, peut alors provoquer la perte de certains variants, ou au contraire, leur **fixation** dans la population après quelques générations. La sélection naturelle intervient également à ce niveau. Lorsqu'un nouvel allèle procure un avantage aux individus porteurs, il est sélectionné **positivement** et éventuellement fixé dans la population. Si, au contraire, le nouvel allèle défavorise les individus porteurs, il est sélectionné **négativement** et, à terme, éliminé de la population. Dans de nombreux cas, les différents types de sélection couplés à la dérive génétique permettent de maintenir dans la population plusieurs allèles à différentes fréquences.

Une mutation maintenue au sein de la population est une **substitution**. C'est à ce niveau qu'intervient

la modélisation des processus évolutifs en phylogénie moléculaire. Les évènements de délétion ou insertion ne sont généralement pas intégrés au modèle proposé, bien qu'il existe quelques exceptions (Thorne et al., 1991, 1992 ; Mitchison et Durbin, 1995). Ainsi, les modèles actuels englobent deux éléments de natures différentes. Le premier, la mutation, est un processus biochimique, tandis que le second se rapporte aux forces agissant pour le maintien du nouvel allèle au sein de la population, et donc de l'espèce. Le processus de modélisation se situe ici au carrefour entre la biologie moléculaire et la génétique des populations. En pratique, les modèles sont construits à partir de l'observation des différents types de différences entre séquences et ne distinguent donc pas les processus biochimiques de l'action conjointe de la dérive génétique et la sélection naturelle.

2.2 Une approche probabiliste du processus de substitution

Nous décrivons ici brièvement les données exploitées pour l'inférence d'arbres. Les hypothèses sous-jacentes aux modèles de substitutions sont ensuite détaillées.

2.2.1 Les données

Les modèles de substitution permettent de décrire les processus évolutifs auxquels sont soumis des séquences nucléiques ou protéiques **homologues**, c'est à dire des séquences dérivant d'une même séquence ancestrale. Généralement, cette description concerne les bases nucléiques ou les acides aminés. Par conséquent, confronter le modèle aux données n'a de sens que lorsque l'homologie porte sur chaque position, ou site, des séquences étudiées (Figure 2.1). Aligner les séquences est donc une étape primordiale de la reconstruction phylogénétique.

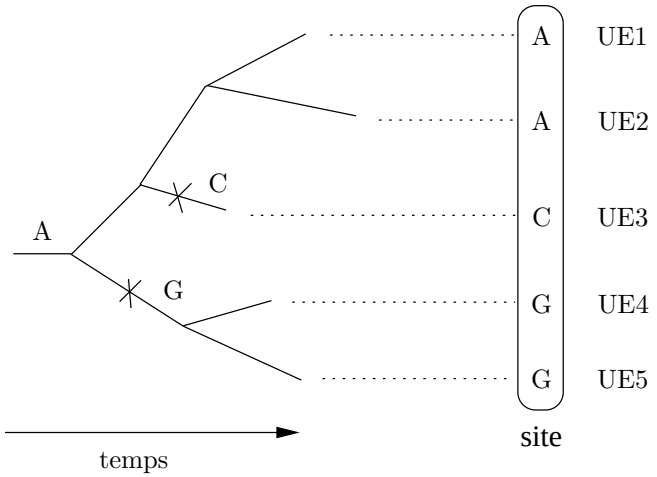


FIG. 2.1 – **Génération de séquences homologues.** Les croix portées sur la phylogénie indiquent des évènements de substitutions. À droite de l'arbre figurent les états homologues obtenus pour les UEs considérées.

L'alignement de deux séquences définit une suite de paires d'éléments. Pour l'ADN, chacune de ces paires est constituée d'un nucléotide appartenant à chaque séquence, ou d'un nucléotide et d'un caractère nul, correspondant à un **gap**, c'est à dire une insertion ou une délétion (Figure 2.2). La plupart des méthodes d'alignement de deux séquences fixent un coût à chaque couple d'états : par exemple, aligner X avec X coûte 0 (identité), aligner X avec Y coûte 2 (substitution) et aligner X avec le caractère nul coûte 3 (insertion/délétion). L'alignement choisi au final est celui qui minimise la somme des coûts associés à chaque couple de caractères. L'algorithme de Needleman et Wunsch (1970) permet d'obtenir cette solution optimale.

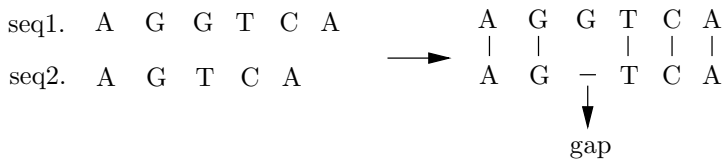


FIG. 2.2 – **Alignement de deux séquences nucléiques.** L'alignement optimal entre ces deux séquences engendre un gap, noté '-', indiquant un évènement d'insertion (au sein de la séquence 1 dans cet exemple) ou de délétion (affectant la séquence 2).

En pratique, reconstruire une phylogénie nécessite d'aligner plusieurs séquences simultanément. La procédure standard (Higgins et al., 1996), implémentée au sein du logiciel CLUSTAL, consiste à calculer les scores des meilleurs alignements des séquences deux à deux, d'en déduire un arbre phylogénétique, puis de «parcourir» cet arbre et d'en extraire les séquences et groupes de séquences à aligner. Contrairement au cas précédent, il n'existe pas d'algorithme efficace garantissant d'obtenir l'alignement multiple de coût minimum.

Les alignements de séquences provenant de différentes espèces présentent généralement des gaps. Ceux-ci sont traités de manière variable suivant les logiciels d'analyse phylogénétique. La solution la plus répandue est plutôt brutale : lorsqu'un gap (ou plus) apparaît au niveau d'un site, ce dernier est éliminé avant toute autre analyse. Pour le calcul de distances évolutives entre paires de séquences, il est possible d'utiliser une approche moins brutale. Les sites contenant des gaps ne sont pas supprimés globalement («global gap removal»), mais sont éliminés au sein de chaque couple de séquences comparées («pairwise gap removal»). Une autre manière d'appréhender cette singularité est de considérer le gap comme un emplacement libre où chaque base (acide aminé) peut prendre place. Nous verrons plus loin comment cette approche se traduit au niveau du calcul de la vraisemblance d'une phylogénie. Signalons enfin les travaux de Thorne et al. (1991, 1992) et Mitchison et Durbin (1995) permettant d'incorporer explicitement les insertions/délétions au sein du modèle décrivant l'évolution des séquences.

2.2.2 Hypothèses et outils mathématiques

Tout modélisation statistique repose sur des hypothèses concernant le processus générant les données. Nous présentons celles-ci dans le cadre des modèles d'évolution moléculaire.

H1. Les sites sont indépendants : les évènements évolutifs affectant un site ne sont pas influencés et n'influencent pas les évènements relatifs aux autres sites de la séquence. De manière plus formelle, on écrit :

$$P(S(t+dt)=W|S(t)=V)=\prod_{i=1}^{N}P(s_i(t+dt)=w_i|s_i(t)=v_i)$$

où S est la variable aléatoire «séquence», et V et W correspondent aux états de cette variable aux instants t et $t+dt$ respectivement. s_i est la variable aléatoire «séquence au site i», v_i et w_i sont les deux états de s_i aux instants t et $t+dt$ respectivement. $P(S(t+dt)=W|S(t)=V)$ est la probabilité pour que la séquence S, dans l'état V à l'instant t, prenne l'état W à l'instant $t+dt$. $P(s_i(t+dt)=w_i|s_i(t)=v_i)$ est la probabilité pour que la séquence S au site i, dans l'état v à l'instant t, prenne l'état w à l'instant $t+dt$. N est le nombre de sites de la séquence.

L'hypothèse d'indépendance est fréquemment violée. Par exemple, il est avéré que certaines substitutions sont compensées. Ainsi, le site 53 de la protéine α-synucleine est normalement occupé par une Alanine et la mutation Alanine→Thréonine prédispose à la maladie de Parkinson chez l'homme. Paradoxalement, les souris saines présentent une Thréonine à ce site. La différence homme-souris au site 53 est donc certainement compensée par une ou plusieurs autres différences à des sites distincts (Kondrashov et al., 2002). Plusieurs auteurs ont également proposé des modèles de substitutions spécialement adaptés aux couples de nucléotides en interaction au sein de boucles de nucléotides, abondantes chez les ARN de transferts notamment. Au lieu de considérer seulement quatre états (les quatre bases), le processus de substitution décrit ici l'évolution des seize états correspondant à l'ensemble des couples de bases qu'il est possible de former (Muse, 1995 ; Rzhetsky et al., 1995 ; Schöniger et von Haesler, 1994, 1995). Felsenstein et Churchill (1996) proposent aussi un modèle de Markov caché afin de prendre en compte une éventuelle corrélation des vitesses d'évolution entre sites voisins. Ces derniers travaux montrent qu'il est possible de relaxer la contrainte d'indépendance lorsque la corrélation entre sites est décrite précisément et intervient explicitement dans le modèle de substitution proposé.

H2. Les sites sont identiquement distribués : le même processus de substitution affecte l'ensemble des sites de la séquence. La probabilité d'un changement d'état ne dépend pas de la position du caractère considéré. On a :

$$P(S(t+dt)=W|S(t)=V)=\prod_{i=1}^{N_p}P(s_i(t+dt)=w_i|s_i(t)=v_i)^{n_i}$$

où N_p est le nombre de sites différents dans la séquence, ou «patterns», et n_i est le nombre de répétitions du site de type i. Ainsi, les N sites de l'alignement original sont ramenés à $N_p < N$ patterns distincts et pondérés. Chaque pattern correspond à un site de la séquence originale et sa pondération (n_i pour le i-ème pattern) est proportionelle à sa fréquence observée.

Là-encore, certains mécanismes biologiques vont à l'encontre de cette hypothèse. Par exemple, les substitutions responsables de la disparition des îlots CpG des zones régulatrices de la transcription font

intervenir des mécanismes biochimiques spécifiques, différents de ceux agissant sur d'autres régions. Si les sites ne sont effectivement pas identiquement distribués, le modèle peut alors être localement faux.

La variabilité des vitesses d'évolution entre sites, affectant la plupart des jeux de données, est un autre phénomène biologique réfutant clairement l'hypothèse d'homogénéité du processus d'évolution le long de la séquence. Cependant, comme nous le verrons en détail par la suite, il est possible de prendre en compte de manière efficace cette caractéristique et, de ce point de vue, relaxer la contrainte engendrée par l'hypothèse de distribution identique des sites.

H3. Le processus de substitution est markovien : la distribution de probabilité des états dans le futur est conditionnée par la distribution actuelle et non la distribution antérieure. Le processus est donc **sans mémoire**. On a :

$$\begin{aligned} P(s_i(t+dt) = w_i | s_i(t) = v_i, \{s_i(u) : 0 \leq u < t\}) &= P(s_i(t+dt) = w_i | s_i(t) = v_i) \\ &= P_{vw,t}(dt) \end{aligned}$$

$P_{vw,t}(dt)$ est donc la probabilité de changement de l'état v, présent à l'instant t, vers l'état w au cours d'un intervalle de temps dt. Cette hypothèse suggère que le remplacement de l'état v par w, dépend de v uniquement, et non des états précédents. Ceci est difficilement réfutable du point de vue de la biologie. Dans le cadre des processus markoviens en temps continu, les probabilités de transitions entre états sont généralement indépendantes des moments auxquels ces évènements interviennent. Ceci se traduit par l'hypothèse d'homogénéité présentée ci-dessous.

H4. Le processus de substitution est homogène dans le temps : la distribution de probabilité des états conditionnellement à une distribution antérieure dépend de la taille de l'intervalle de temps séparant les deux instants et non de leur position sur l'axe temporel. En d'autres termes, le même **mode évolutif**, ou pattern de substitution, affecte toutes les lignées de la phylogénie. Ainsi :

$$\begin{aligned} P(s_i(t+dt) = w_i | s_i(t) = v_i) &= P(s_i(t'+dt) = w_i | s_i(t') = v_i) \qquad \forall t', t, i \\ &= P_{vw}(dt) \end{aligned}$$

Certains modèles d'évolution permettent de s'affranchir de l'hypothèse d'homogénéité. À ce propos, signalons les travaux récents de Galtier (2001) et Huelsenbeck (2002) permettant de tenir compte d'une éventuelle variation dans l'arbre de la vitesse d'évolution à un site donné. Nous détaillons ce modèle à la fin du chapitre.

H5. Le processus de substitution est stationnaire : la distribution de probabilité des états est constante dans le temps. La probabilité d'observer un état donné ne dépend pas du moment auquel s'effectue l'observation. On note :

$$P(s_i(t) = v_i) = \pi_v$$

où π_v est la fréquence de l'état v lorsque le processus markovien de substitution a atteint l'équilibre.

Là encore, l'hypothèse de stationnarité ne correspond pas à une contrainte absolue. Par exemple, un modèle non-stationnaire a été proposé afin de prendre en compte les variations de composition en bases entre séquences (Galtier, 1997; Galtier et Gouy, 1995), et éviter ainsi le regroupement dans l'arbre de séquences sur la base de taux de G+C similaires. Son application à l'analyse d'ARN ribosomiques a permis d'estimer la composition en bases du génome du dernier ancêtre commun à l'ensemble des organismes vivants contemporains. Ces travaux apportent ainsi un nouvel éclairage sur la question de l'hyperthermophilie de cet ancêtre (Galtier et al., 1999).

H6. Le processus de substitution est réversible : pour un intervalle de temps donné, noté t, le nombre de substitutions de l'état v par l'état w est, en espérance, égal au nombre de substitutions de l'état w par l'état v. On a :

$$\pi_v P_{vw}(dt) = \pi_w P_{wv}(dt)$$

L'utilisation de modèles non-réversibles lors du calcul de la vraisemblance d'une phylogénie nécessite de connaître la position de la racine de l'arbre. Ce point est abordé plus loin dans ce chapitre.

2.3 Des taux de substitution aux probabilités de changements

Cette partie est consacrée à la description du rôle des matrices de taux de substitution instantanés et leur utilisation pour le calcul des probabilités de changements entre états sous les hypothèses précédentes.

2.3.1 Matrice des taux de substitutions instantanés

Le mode évolutif des caractères étudiés est décrit à partir des transformations des états de ces caractères au cours d'une unité de temps élémentaire, notée dt, lors de laquelle se produit une unique substitution. Prenons l'exemple de l'ADN. On a :

$$\begin{array}{rclcrcrcrcr}
A(t+dt) & = & A(t) & - & A(t)R_{A.}dt & + & C(t)R_{CA}dt & + & G(t)R_{GA}dt & + & T(t)R_{TA}dt \\
C(t+dt) & = & C(t) & + & A(t)R_{AC}dt & - & C(t)R_{C.}dt & + & G(t)R_{GC}dt & + & T(t)R_{TC}dt \\
G(t+dt) & = & G(t) & + & A(t)R_{AG}dt & + & C(t)R_{CG}dt & - & G(t)R_{G.}dt & + & T(t)R_{TG}dt \\
T(t+dt) & = & T(t) & + & A(t)R_{AT}dt & + & C(t)R_{CT}dt & + & G(t)R_{GT}dt & - & T(t)R_{T.}dt
\end{array}$$

où $A(t)$, $C(t)$, $G(t)$ et $T(t)$ sont les probabilités des quatre bases à l'instant t. R_{XY} est le taux de substitution instantané de l'état X par l'état Y et $R_{X.} = \sum_{Y \in \mathcal{A}, Y \neq X} R_{XY}$, est le taux de substitution instantané de l'état X par un quelconque autre état appartenant à l'alphabet des caractères, noté $\mathcal{A}$. Sous forme matricielle, on écrit :

$$\begin{aligned}
\mathbf{F}(t+dt) &= \mathbf{F}(t) + \mathbf{F}(t)\mathbf{R}dt \\
&= \mathbf{F}(t)(\mathbf{I} + \mathbf{R}dt) \qquad (2.1)
\end{aligned}$$

où $\mathbf{F}(t)$ est le vecteur ligne des probabilités des états à l'instant t. Dans la terminologie statistique, $\mathbf{R}$ est le générateur infinitésimal de la chaîne de Markov.

De l'expression 2.1, on déduit :

$$\frac{d\mathbf{F}(t)}{dt} = \mathbf{F}(t)\mathbf{R} \tag{2.2}$$

La résolution de cette équation différentielle donne :

$$\begin{aligned} \mathbf{F}(t) &= \mathbf{F}(0)e^{\mathbf{R}t} \\ &= \mathbf{F}(0)\mathbf{P}(t) \end{aligned} \tag{2.3}$$

où $e^{\mathbf{R}t} = \sum_{n=0}^{+\infty} \frac{\mathbf{R}^n t^n}{n!}$ est l'exponentielle de $\mathbf{R}t$ et $\mathbf{P}(t)$ est la matrice des probabilités de changements sur un temps écoulé égal à t. En d'autres termes, $P_{xy}(t)$ est la probabilité d'observer y à l'instant t sachant que x était présent à l'instant 0. Notons aussi que lorsque $t = dt \ll 1$, les termes de second ordre dans le développement limité de l'exponentielle matricielle sont négligeables. On a donc $\mathbf{P}(dt) = \mathbf{I} + \mathbf{R}dt$, ce qui est conforme à l'expresion 2.1. La Figure 2.3 détaille les éléments de $\mathbf{P}(t)$ et la relation entre cette dernière et $\mathbf{R}$.

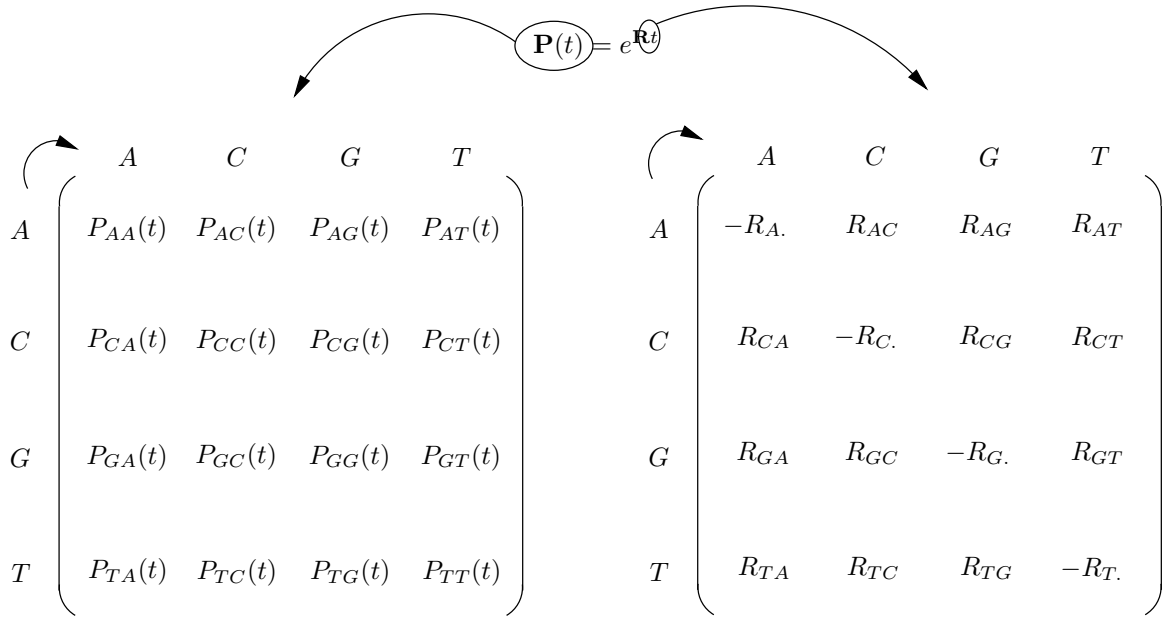


FIG. 2.3 – **Probabilités de substitutions** *vs.* **taux de substitutions instantanés**

L'expression 2.3 montre que les probabilités des états à un instant donné sont des fonctions du produit des taux de substitution instantanés par le temps écoulé. De part la nature des données analysées, ces deux dernières quantités sont généralement indissociables. La matrice $\mathbf{R}$ décrit les fréquences **relatives** des différents types de substitutions : seuls les rapports entre valeurs de R_{ij} sont informatifs. Tout facteur commun à l'ensemble des R_{ij} est lié au **rythme** global des substitutions. Il n'est pas indispensable de faire référence explicitement à ce dernier car il «disparaît» lorsque l'unité de temps considérée correspond au temps réel multiplié par ce paramètre. Ainsi, seuls les paramètres libres du modèle de substitution (comme

le ratio transition/transversion, les fréquences en bases à l'équilibre, etc.) participent à la description du modèle.

2.3.2 Les principaux modèles

Nous présentons ici les matrices de taux de substitutions instantanés correspondant aux principaux modèles appliqués aux séquences nucléiques. Chacune des matrices comprend des **taux relatifs** plus ou moins nombreux suivant la complexité du modèle, et des **paramètres de fréquences**, correspondant aux probabilités des états à l'équilibre.

Avant de présenter les différents modèles, il nous a semblé utile d'expliquer pourquoi les matrices de taux de substitutions instantanées ne sont pas décrites sous une seule et même forme à travers la littérature. Le premier point concerne la mise en évidence de la distribution stationnaire au sein de ces matrices. Lorsque celles-ci sont réversibles, il est en effet possible de réécrire la matrice $\mathbf{R}$ sous la forme :

$$\mathbf{R} = \begin{pmatrix} -R'_{A.} & \pi_C R'_{AC} & \pi_G R'_{AG} & \pi_T R'_{AT} \\ \pi_A R'_{CA} & -R'_{C.} & \pi_G R'_{CG} & \pi_T R'_{CT} \\ \pi_A R'_{GA} & \pi_C R'_{GC} & -R'_{G.} & \pi_T R'_{GT} \\ \pi_A R'_{TA} & \pi_C R'_{TC} & \pi_G R'_{TG} & -R'_{T.} \end{pmatrix}$$

où $R'_{X.} = \sum_{Y \neq X} \pi_Y R'_{XY}$ et $R'_{XY} = R'_{YX}$.

Soit $\mathbf{\Pi}$ le vecteur des fréquences stationnaires, décrivant la distribution des fréquences des états du processus markovien à l'équilibre. Plus formellement, on a $\mathbf{\Pi} = \mathbf{F}(0)\mathbf{P}(\infty)$, $\forall \mathbf{F}(0)$ ou encore $\mathbf{\Pi}(\mathbf{I}+\mathbf{R}dt) = \mathbf{\Pi}$, et donc $\mathbf{\Pi R} = 0$. Lorsque les π_X donnés dans la matrice ci-dessus vérifient $\mathbf{\Pi} = (\pi_X)$, alors $\mathbf{\Pi R} = 0$ lorsque $\mathbf{R}$ s'écrit sous la forme ci-dessus. Autrement dit, cette matrice des taux de substitution instantanés fait bien apparaître la distribution stationnaire, facilitant ainsi la compréhension des modèles. Remarquons que ceci s'applique uniquement pour des modèles réversibles, c'est à dire $R'_{XY} = R'_{YX}$, $\forall X, Y$.

Un autre point important concernant l'écriture des matrices de substitution porte sur la définition de l'unité de temps. En phylogénie, l'hypothèse d'homogénéité correspond généralement à la constance des ratios entre les valeurs de R'_{XY}, $\forall X, Y$. Cependant, au sens strict, cette hypothèse est plus contraignante : elle impose que les valeurs de R'_{XY} soient constantes, et ceci quelle que soit la branche considérée. Ceci se vérifie lorsque l'horloge moléculaire est respectée. Dans le cas général, les vitesses de substitutions varient entre lignées et R'_{XY} est constant au sein d'une branche mais varie de branches en branches. Soit l_k la longueur d'une branche k quelconque, mesurée au sens usuel, c'est à dire une espérance du nombre de substitutions ramenée à la taille des séquences. Pour un modèle homogène (au sens strict) le long de k, on a :

$$l_k = \left(\sum_X \pi_X \sum_{Y \neq X} \pi_Y R'_{XY,k} \right) t \tag{2.4}$$

où $R'_{XY,k}$ correspond au taux de substitution instantané de X par Y le long de k. On pose alors :

$$C_k = \sum_X \pi_X \sum_{Y \neq X} \pi_Y R'_{XY,k}, \tag{2.5}$$

on a alors $t = \frac{C_k}{l_k}$ et le temps t considéré ici est en fait un «pseudo-temps» proportionnel à la vitesse d'accumulation des substitutions le long de la branche k.

Le premier modèle proposé est celui de Jukes et Cantor (1969), noté **JC69** :

$$\mathbf{R} = \begin{pmatrix} -\frac{3}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & -\frac{3}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & -\frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & -\frac{3}{4} \end{pmatrix}$$

Pour ce modèle, l'application de l'expression 2.4 donne $t = \frac{4}{3} l_k$. Outre les hypothèses de stationnarité, d'homogénéité et de réversibilité, ce modèle suppose que toutes les substitutions se produisent au même taux et que les fréquences des bases à l'équilibre sont toutes égales à $\frac{1}{4}$.

Le modèle de Kimura (1980), noté **K80**, introduit un paramètre relatif. Il permet de tenir compte d'une éventuelle différence de fréquences entre les transitions, c'est à dire les substitutions entre purines ($A \leftrightarrow G$) ou entre pyrimidines ($C \leftrightarrow T$), et les transversions ($\{A, G\} \leftrightarrow \{C, T\}$). κ est le taux relatif décrivant l'équilibre entre ces deux types d'évènements, c'est à dire le ratio transition/transversion. Comme pour le modèle JC69, la matrice de taux de substitution instantané est symétrique. Les fréquences en bases à l'équilibre sont donc censées être toutes égales. On a :

$$\mathbf{R} = \begin{pmatrix} -\frac{1}{4}(\kappa+2) & \frac{1}{4} & \frac{1}{4}\kappa & \frac{1}{4} \\ \frac{1}{4} & -\frac{1}{4}(\kappa+2) & \frac{1}{4} & \frac{1}{4}\kappa \\ \frac{1}{4}\kappa & \frac{1}{4} & -\frac{1}{4}(\kappa+2) & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4}\kappa & \frac{1}{4} & -\frac{1}{4}(\kappa+2) \end{pmatrix}$$

et $t = 4l_k/(\kappa+2)$. Il existe plusieurs définitions du ratio transition/transversion dans la littérature. La plus répandue est celle utilisée dans le package PHYLIP (Felsenstein, 1991, 1993). Dans ce logiciel, le paramètre mesurant le biais entre transition et transversion, noté K, correspond à l'espérance du rapport entre le nombre de transitions et le nombre de transversions. Il existe un lien simple entre K et κ :

$$K = \frac{\pi_A R_{AG} + \pi_G R_{GA} + \pi_C R_{CT} + \pi_T R_{TC}}{\pi_A R_{AC} + \pi_C R_{CA} + \pi_A R_{AT} + \pi_T R_{TA} + \pi_C R_{CG} + \pi_G R_{GC} + \pi_G R_{GT} + \pi_T R_{TG}}$$

où les R_{xy} correspondant à des transitions sont les fonctions de κ données dans la matrice ci-dessus. Pour le modèle K2P, les π_x sont remplacés par $\frac{1}{4}$ et on obtient $K = \frac{\kappa}{2}$. L'analyse de différents jeux de données indique que $\kappa = 4.0$ est une valeur consensuelle.

Le modèle **F81** (Felsenstein, 1981) ne fait pas intervenir de taux relatifs mais autorise des fréquences en bases à l'équilibre différentes de $\frac{1}{4}$:

$$\mathbf{R} = \begin{pmatrix} -(\pi_Y + \pi_G) & \pi_C & \pi_G & \pi_T \\ \pi_A & -(\pi_R + \pi_T) & \pi_G & \pi_T \\ \pi_A & \pi_C & -(\pi_Y + \pi_A) & \pi_T \\ \pi_A & \pi_C & \pi_G & -(\pi_R + \pi_C) \end{pmatrix}$$

où $\pi_R = \pi_A + \pi_G$ et $\pi_Y = \pi_C + \pi_T$ sont les fréquences à l'équilibre des purines et des pyrimidines respectivement. Ici $t = l_k/(2(\pi_R\pi_Y + \pi_A\pi_G + \pi_C\pi_T))$. Notons que, même si cette matrice n'est pas symétrique lorsque les fréquences en bases à l'équilibre sont différentes de $\frac{1}{4}$, ce modèle est réversible. En effet, on a $\pi_x R_{xy} = \pi_y R_{yx}$, $\forall x \neq y$. Par exemple, $\pi_A R_{AC} = \pi_A\pi_C = \pi_C\pi_A = \pi_C R_{CA}$.

Le modèle **HKY85** (Hasegawa et al., 1985) est une synthèse des deux précédents. La matrice des taux instantanés s'écrit donc :

$$\mathbf{R} = \begin{pmatrix} -(\pi_Y + \kappa\pi_G) & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & -(\pi_R + \kappa\pi_T) & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & -(\pi_Y + \kappa\pi_A) & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & -(\pi_R + \kappa\pi_C) \end{pmatrix}$$

et $t = l_k/(2(\pi_R\pi_Y + \kappa\pi_A\pi_G + \kappa\pi_C\pi_T))$. Ce modèle est largement utilisé en phylogénie moléculaire car il représente un bon compromis entre la qualité de la description des données et le nombre de paramètres à estimer (trois paramètres de fréquences et un paramètre libre).

L'analyse de séquences homologues de la région de contrôle de l'ADN mitochondrial chez l'humain et le chimpanzé montre une différence entre les taux de transitions et transversions ainsi qu'un excès de transitions entre pyrimidines ($C \leftrightarrow T$) comparées aux transitions entre purines ($A \leftrightarrow G$). Le modèle de Tamura et Nei (1993), noté **TN93**, permet de tenir compte d'un tel déséquilibre au sein des transitions :

$$\mathbf{R} = \begin{pmatrix} -(\pi_Y + \kappa_R\pi_G) & \pi_C & \kappa_R\pi_G & \pi_T \\ \pi_A & -(\pi_R + \kappa_Y\pi_T) & \pi_G & \kappa_Y\pi_T \\ \kappa_R\pi_A & \pi_C & -(\pi_Y + \kappa_R\pi_A) & \pi_T \\ \pi_A & \kappa_Y\pi_C & \pi_G & -(\pi_R + \kappa_Y\pi_C) \end{pmatrix}$$

et $t = l_k/(2(\pi_R\pi_Y + \kappa_R\pi_A\pi_G + \kappa_Y\pi_C\pi_T))$. Ce modèle présente un paramètre supplémentaire comparé au précédent. κ est en effet décomposé en deux termes κ_R et κ_Y. Il est aussi possible d'exprimer ces deux paramètres en fonction de κ et d'un taux relatif , noté λ. On a :

$$\begin{aligned} \kappa_R &= \kappa\frac{2\lambda}{1+\lambda} \\ \kappa_Y &= \kappa\frac{2}{1+\lambda} \end{aligned}$$

et λ représente le paramètre supplémentaire du modèle. Lorsque sa valeur est égale à 1, $\kappa_R = \kappa_Y = \kappa$ et le modèle TN93 correspond exactement au modèle HKY85. Le modèle de Felsenstein (1993), noté **F84**,

est basé sur la même idée mais il n'inclut pas de paramètre supplémentaire λ. En effet, pour ce modèle, λ n'est pas ajusté, il est déduit des fréquences en bases à l'équilibre et du ratio transition/transversion :

$$\lambda = \frac{\pi_Y + (\pi_R - \pi_Y)/2\kappa}{\pi_R - (\pi_R - \pi_Y)/2\kappa}$$

Ce modèle est implémenté dans l'ensemble des programmes du package PHYLIP basés sur le maximum de vraisemblance.

Enfin, le modèle **GTR** (Lanave et al., 1984; Tavaré, 1986; Barry et Hartigan, 1987; Rodriguez et al., 1990), pour General Time Reversible, est le modèle réversible le plus riche en paramètres (trois paramètres de fréquences et six paramètres libres) :

$$\mathbf{R} = \begin{pmatrix} -(a\pi_C + b\pi_G + c\pi_T) & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & -(a\pi_A + d\pi_G + e\pi_T) & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & -(b\pi_A + d\pi_C + f\pi_T) & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & -(c\pi_A + e\pi_C + f\pi_G) \end{pmatrix}$$

et $t = l_k/(2(a\pi_A\pi_C + b\pi_A\pi_G + c\pi_A\pi_T + d\pi_C\pi_G + e\pi_C\pi_T + f\pi_G\pi_T))$.

Il existe d'autres modèles de substitution (Kimura, 1981; Zharkikh, 1994), traduisant différentes manières de prendre en compte les transitions, transversions ou fréquences en bases à l'équilibre, mais *a priori* aucun modèle n'est plus adapté qu'un autre à l'analyse d'un jeu de séquences donné. Le choix de celui-ci dépend évidemment des fréquences des types de différences observées mais aussi de la quantité et de la qualité des données disponibles. En effet, un modèle riche en paramètres nécessite un grand nombre de sites homologues pour estimer ses paramètres de manière satisfaisante. Des problèmes similaires d'estimation de paramètres apparaissent lorsque les divergences entre séquences sont très importantes, ou au contraire, très faibles.

Concernant les protéines, l'établissement de matrices de taux de substitution instantanés ne repose généralement pas sur l'estimation de paramètres de substitution pour chaque nouveau jeu de données analysé. En effet, Dayhoff et al. (1978) et Jones et al. (1992) ont construit ces matrices à partir de l'analyse de plusieurs protéines homologues peu divergentes. L'analyse par parcimonie de ces données permet d'estimer les valeurs de taux de substitution pour chaque paire d'acides aminés. L'approche développée par Adachi et Hasegawa (1990) suit la même idée mais se base sur le principe du maximum de vraisemblance afin d'éliminer les biais liés à la parcimonie.

2.3.3 Expression des probabilités de substitution

La construction des matrices de taux de substitutions instantanés est la traduction des hypothèses biologiques en termes mathématiques. Ces matrices décrivent de manière synthétique les hypothèses sur le processus de substitution mais elles n'expriment pas directement les probabilités des observations.

L'obtention de telles probabilités à partir des matrices de taux instantanés fait intervenir des calculs rarement décrits dans la littérature. Pour cette raison, nous présentons ici en détails les étapes du passage de la matrice des taux instantanés aux probabilités de substitutions correspondantes pour le modèle TN93.

L'exponentielle de $\mathbf{R}t$ est la matrice $\mathbf{P}(t)$ des probabilités de changements sur un intervalle de temps t. Si $\mathbf{R}t$ est diagonalisable, alors $\mathbf{R}t = \mathbf{V}_d\mathbf{D}t{\mathbf{V}_d}^{-1} \Leftrightarrow \mathbf{P}(t) = e^{\mathbf{R}t} = \mathbf{V}_d e^{\mathbf{D}t}{\mathbf{V}_d}^{-1}$, où $\mathbf{V}_d$ est la matrice des vecteurs propres à droite de $\mathbf{R}$. $\mathbf{D}$ est une matrice diagonale, dont les éléments sont les valeurs propres de $\mathbf{R}$. $e^{\mathbf{D}t}$ est une matrice diagonale dont les éléments sont les exponentielles des éléments diagonaux de $\mathbf{D}$, multipliés par t.

Les valeurs propres de $\mathbf{R}$ sont les solutions de l'équation $det(\mathbf{R} - \lambda\mathbf{I}) = 0$. On obtient facilement :

$$\mathbf{D} = \begin{pmatrix} \lambda_1 = 0 & 0 & 0 & 0 \\ 0 & \lambda_2 = -\kappa_R\pi_R - \pi_Y & 0 & 0 \\ 0 & 0 & \lambda_3 = -\kappa_Y\pi_Y - \pi_R & 0 \\ 0 & 0 & 0 & \lambda_4 = -\pi_R - \pi_Y \end{pmatrix}$$

La matrice $\mathbf{V}_d$ des vecteurs propres à droite est calculée à partir des équations $(\mathbf{R} - \lambda_i\mathbf{I})\mathbf{V}_{i,d} = 0$, où $\mathbf{V}_{i,d}$ est le vecteur (colonne) propre à droite associé à la valeur propre i. On obtient :

$$\mathbf{V}_d = \begin{pmatrix} 1 & \pi_G & 0 & 1/\pi_R \\ 1 & 0 & \pi_T & -1/\pi_Y \\ 1 & -\pi_A & 0 & 1/\pi_R \\ 1 & 0 & -\pi_C & -1/\pi_Y \end{pmatrix}$$

où la i-ème colonne de $\mathbf{V}_d$ correspond au i-ème vecteur propre de $\mathbf{R}$. Le système d'équations $(\mathbf{R} - \lambda_i\mathbf{I})\mathbf{V}_{i,d} = 0$ étant dégénéré, chaque colonne de $\mathbf{V}_d$ peut être multipliée par une constante notée α_i. Nous verrons ci-dessous comment ajuster ces valeurs.

Il faut à présent déterminer l'expression de ${\mathbf{V}_d}^{-1}$. Inverser $\mathbf{V}_d$ est un tâche fastidieuse en pratique. Il existe un moyen plus simple de parvenir à nos fins. En effet, $\mathbf{V}_d^{-1}$ est la matrice des vecteurs propres à gauche de $\mathbf{R}$, notée $\mathbf{V}_g$ dont l'expression est déterminée aisément à partir des équations $\mathbf{V}_{i,g}(\mathbf{R} - \lambda_i I) = 0$, où $\mathbf{V}_{i,g}$ est le vecteur (ligne) propre à gauche associé à la valeur propre i. À partir de calculs similaires aux précédents, on obtient :

$$\mathbf{V}_g = \begin{pmatrix} \pi_A & \pi_C & \pi_G & \pi_T \\ \pi_Y(1/\pi_R + 1/\pi_Y) & 0 & -\pi_Y(1/\pi_R + 1/\pi_Y) & 0 \\ 0 & \pi_R(1/\pi_R + 1/\pi_Y) & 0 & -\pi_R(1/\pi_R + 1/\pi_Y) \\ \pi_Y\pi_A & -\pi_R\pi_C & \pi_Y\pi_G & -\pi_R\pi_T \end{pmatrix}$$

où la i-ème ligne de $\mathbf{V}_g$ est le i-ème vecteur propre à gauche de $\mathbf{R}$. Signalons ici que le vecteur propre associé à la valeur propre égale à 0 fait apparaître la distribution stationnaire. Ceci est une caractéristique

générale des modèles de Markov qui correspond au fait que $\mathbf{\Pi}(\mathbf{I} + \mathbf{R}dt) = \mathbf{\Pi}$, donc $\mathbf{\Pi R} = 0$ et $\mathbf{\Pi}$ est, par conséquent, le vecteur propre à gauche de $\mathbf{R}$ associé à la valeur propre nulle. Ici encore, chaque ligne de $\mathbf{V}_g$ est multipliée par un facteur β_j. À partir de l'égalité $\mathbf{V}_d\mathbf{V}_g = \mathbf{I}$, on en déduit $\alpha_i\beta_i = 1$, $\forall i$. Il est aisément vérifiable que lorsque $\alpha_i = \beta_i = 1$, $\forall i$, le produit $\mathbf{V}_d\mathbf{V}_g$ est bien égal à $\mathbf{I}$.

L'expression de $\mathbf{P}(t) = e^{\mathbf{R}t}$ se déduit des calculs précédents. On a $\mathbf{P}(t) =$

$$\begin{pmatrix} \pi_A + e_3\frac{\pi_A\pi_Y}{\pi_R} + e_1\frac{\pi_G}{\pi_R} & \pi_C - e_3\pi_C & \pi_G + e_3\frac{\pi_G\pi_Y}{\pi_R} - e_1\frac{\pi_G}{\pi_R} & \pi_T - e_3\pi_T \\ \pi_A - \pi_A e_3 & \pi_C + e_3\frac{\pi_C\pi_R}{\pi_Y} + e_2\frac{\pi_T}{\pi_Y} & \pi_G - e_3\pi_G & \pi_T + e_3\frac{\pi_T\pi_R}{\pi_Y} - e_2\frac{\pi_T}{\pi_Y} \\ \pi_A + e_3\frac{\pi_A\pi_Y}{\pi_R} - e_1\frac{\pi_A}{\pi_R} & \pi_C - e_3\pi_C & \pi_G + e_3\frac{\pi_G\pi_Y}{\pi_R} + e_1\frac{\pi_A}{\pi_R} & \pi_T - e_3\pi_T \\ \pi_A - e_3\pi_A & \pi_C + e_3\frac{\pi_C\pi_R}{\pi_Y} - e_2\frac{\pi_C}{\pi_Y} & \pi_G - e_3\pi_G & \pi_T + e_3\frac{\pi_T\pi_R}{\pi_Y} + e_2\frac{\pi_C}{\pi_Y} \end{pmatrix} \quad (2.6)$$

où $e_1 = e^{(-\kappa_r r - y)t}$, $e_2 = e^{(-\kappa_y y - r)t}$, $e_3 = e^{(-r-y)t}$.

Il est aisé de déduire de cette matrice les probabilités de substitutions pour des modèles moins généraux. Par exemple, pour le modèle JC69, ces probabilités sont obtenues en fixant $\pi_A = \pi_C = \pi_G = \pi_T = \frac{1}{4}$ et $\kappa_R = \kappa_Y = 1.0$. En revanche, la complexité des calculs analytiques pour le modèle GTR ne permet pas d'exhiber les expressions analytiques des probabilités de substitutions. La plupart des logiciels implémentant ce modèle ont recours à des méthodes numériques pour les calculs de diagonalisation. De telles méthodes sont aussi appliquées pour diagonaliser les matrices de taux de substitutions instantanés ente acides aminés, et en déduire les probabilités de substitutions en fonction d'une longueur de branche donnée.

2.4 Applications à la phylogénie

Savoir calculer les probabilités de substitution permet d'estimer des distances évolutives entre séquences ainsi que la vraisemblance d'une phylogénie dont les longueurs de branches sont connues. Ces deux points sont détaillés ici.

2.4.1 Notations

Quelques notations et définitions portant sur les arbres phylogénétiques sont tout d'abord introduites. D'un point de vue mathématique, un arbre est un **graphe** connexe non-cyclique. Un graphe est un ensemble de sommets (les noeuds) reliés par des arêtes (les branches). Il est dit connexe lorsqu'il existe au moins un chemin entre chaque sommet, et non-cyclique si aucune des extrémités d'un quelconque chemin coïncident. Deux types de sommets se distinguent : les noeuds **internes** et les noeuds **externes** ou **feuilles**. Les noeuds internes correspondent aux ancêtres des UEs comparées et observées au niveau des feuilles de l'arbre. La topologie de l'arbre est complètement **résolue** lorsque les noeuds internes sont de degré trois (trois branches sont connectées à ce noeud). Lorsque le degré d'un des noeuds internes est supérieur à trois, la topologie est alors localement irrésolue.

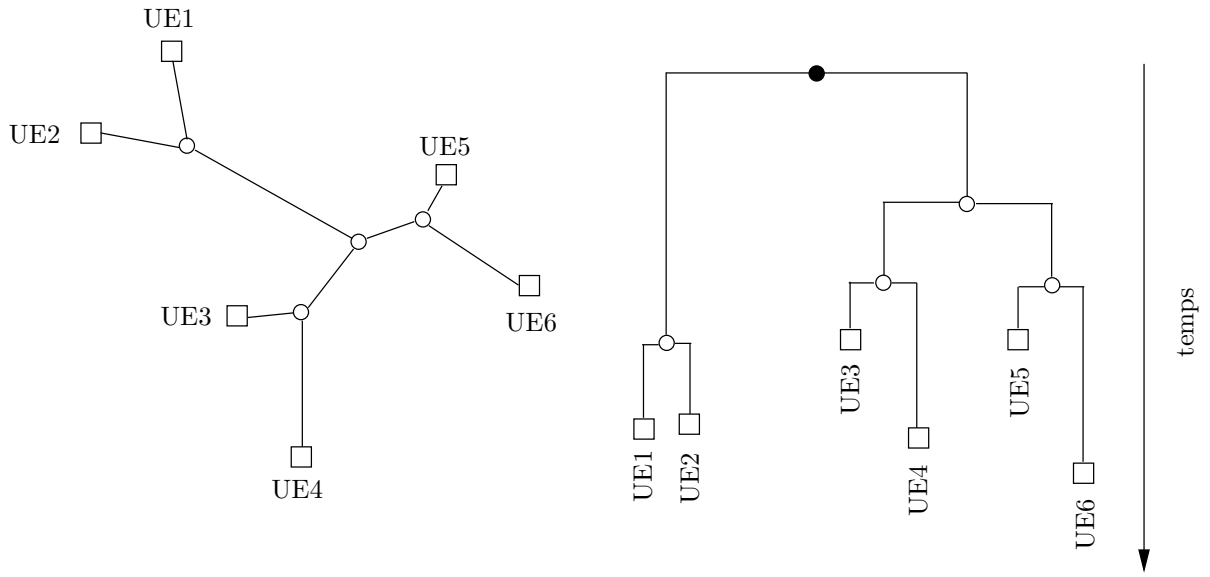


FIG. 2.4 – **Arbres non-enraciné (gauche) et enraciné (droite).** Les carrés désignent les noeuds externes, les cercles vides sont des noeuds internes et le cercle plein indique le noeud racine.

Une phylogénie peut être enracinée ou non. La racine d'une phylogénie est le noeud interne de degré deux, ancêtre de l'ensemble des autres noeuds. Situer ce point particulier sur l'arbre permet d'**orienter** celui-ci (Figure 2.4). *A priori*, la plupart des méthodes de construction d'arbres phylogénétique, UPGMA et dérivées mis à part, infèrent des phylogénies non-enracinées. L'introduction de groupes externes aux séquences analysées permet d'estimer la position de la racine (voir chapitre 1).

Représenter l'évolution des macromolécules par un arbre constitue en soi une hypothèse. Or, cette hypothèse est parfois violée : les transferts de matériel génétique entre génomes procaryotes sont relativement fréquents. Dans ce cas, un graphe acyclique n'est plus une description pertinente de l'histoire des séquences ayant subi un ou plusieurs transferts. Des travaux récents (Strimmer et Moulton, 2000) portent sur l'application du principe de vraisemblance maximale en phylogénie à des graphes cycliques, ou sur la reconstruction de tels graphes à partir de distances évolutives (Bandelt et Dress, 1992b). Outre la modélisation des recombinaisons et transferts horizontaux, nous verrons au chapitre 4 que ce type de graphe permet de mettre en évidence les zones d'une phylogénie peu supportées par les données.

2.4.2 Calculs de distances évolutives

La construction de phylogénies à partir de distances évolutives est une approche très répandue. La distance entre deux séquences est l'espérance du nombre de substitutions par site s'étant produit depuis leur divergence. Si ces séquences sont peu divergentes et que chaque différence observée correspond à un unique évènement de substitution, la proportion de différences, ou **distance de Hamming**, est une estimation parfaite de la distance évolutive. Cependant, il se peut qu'une différence observée soit la conséquence d'une succession de plusieurs substitutions. Dans ce cas, il est nécessaire de tenir compte de

ces substitutions masquées et la modélisation est alors un outil indispensable.

La distance entre deux séquences ayant divergé à un instant t antérieur à l'instant présent est notée d. Elle est définie de la façon suivante :

$$d = \int_0^{2t} \sum_{x \in \mathcal{A}} \pi_x(t) \sum_{y \in \mathcal{A}, y \neq x} R_{xy}(t) dt$$

L'hypothèse de stationnarité permet de simplifier cette expression :

$$d = \int_0^{2t} \sum_{x \in \mathcal{A}} \pi_x \sum_{y \in \mathcal{A}, y \neq x} R_{xy}(t) dt$$

et pour un modèle homogène, on écrit :

$$d = \sum_{x \in \mathcal{A}} \pi_x \sum_{y \in \mathcal{A}, y \neq x} R_{xy} 2t$$

Trois étapes sont nécessaires pour obtenir l'expression analytique de la distance estimée : (1) la distance entre deux séquences est tout d'abord exprimée comme une fonction des taux de substitution instantanés et du temps ; (2) les quantités mesurables, comme la proportion de différences observées entre deux séquences, sont aussi exprimées comme des fonctions de ces mêmes paramètres ; (3) la distance estimée est déduite des deux expressions précédentes. Ces trois étapes sont appliquées ici au modèle JC69.

[**Étape 1**] Pour un modèle réversible et homogène (ce qui est le cas ici), considérer deux séquences ayant divergé à un instant t antérieur à l'instant présent revient à comparer une même séquence aux temps 0 et $2t$. Pour le modèle JC69, l'expression 2.4 donne alors :

$$d_{JC69} = \frac{3}{2} t \tag{2.7}$$

Les probabilités de substitution de l'expression 2.6 donnent :

$$P_{xy}(2t) = \frac{1}{4} - \frac{1}{4} e^{-2t}, \forall x \neq y$$

Or, la probabilité d'observer une différence entre les deux séquences à un site pris au hasard, notée $p(2t)$, est donnée par :

$$p(2t) = \sum_{x \in \mathcal{A}} \pi_x \sum_{y \in \mathcal{A}, y \neq x} P_{xy}(2t)$$

[**Étape 2**] Pour le modèle JC69, on obtient :

$$p(2t) = \frac{3}{4} - \frac{3}{4} e^{-2t} \tag{2.8}$$

[**Étape 3**] On déduit des équations 2.8 et 2.7, l'expression la distance évolutive, d_{JC69} :

$$\begin{aligned} t &= -\frac{1}{2} ln(1 - \frac{4}{3} p(2t)) \\ d_{JC69} &= \frac{3}{2} t = -\frac{3}{4} ln(1 - \frac{4}{3} p(2t)) \end{aligned} \tag{2.9}$$

et l'estimateur usuel :

$$\widehat{d}_{JC69} = -\frac{3}{4} ln(1 - \frac{4}{3}\widehat{p})$$

où $\widehat{p}$ est l'estimation de $p(2t)$ et correspond à la proportion observée de différences entre les deux séquences comparées. Il est possible de montrer que l'estimateur obtenu est celui du maximum de vraisemblance.

[Étape 1] L'estimateur de la distance pour le modèle TN93 est obtenue en suivant une procédure similaire. La distance entre deux séquences s'écrit ici :

$$\begin{aligned} d &= \sum_{x \in \mathcal{A}} \pi_x \sum_{y \in \mathcal{A}, y \neq x} R_{xy} 2t \\ &= 4t(\pi_R \pi_Y + \kappa_R \pi_A \pi_G + \kappa_Y \pi_C \pi_T) \end{aligned} \tag{2.10}$$

[Étape 2] Les probabilités de substitutions correspondant à des changements entre purines et entre pyrimidines, notée $p_R(2t)$ et $p_Y(2t)$, ainsi que les probabilités de transversions, $q(2t)$, sur un temps écoulé égal à $2t$ se décomposent de la façon suivante :

$$\begin{aligned} p_R(2t) &= \pi_A P_{AG}(2t) + \pi_G P_{GA}(2t) \\ p_Y(2t) &= \pi_C P_{CT}(2t) + \pi_T P_{TC}(2t) \\ q(2t) &= \pi_A P_{AC}(2t) + \pi_C P_{CA}(2t) + \pi_A P_{AT}(2t) + \pi_T P_{TA}(2t) \\ &+ \pi_C P_{CG}(2t) + \pi_G P_{GC}(2t) + \pi_G P_{GT}(2t) + \pi_T P_{TG}(2t) \end{aligned}$$

En remplaçant dans ces trois équations les probabilités de substitutions par leurs expressions (voir équation 2.6), on obtient :

$$\begin{aligned} t &= -\frac{1}{2} ln(1 - \frac{q(2t)}{2\pi_R \pi_Y}) \\ \kappa_R &= -\frac{1}{2\pi_R t} ln(\pi_R + \pi_Y e^{-2t} - \frac{\pi_R p_R(2t)}{2\pi_A \pi_G}) - \pi_Y/\pi_R \\ \kappa_Y &= -\frac{1}{2\pi_Y t} ln(\pi_Y + \pi_R e^{-2t} - \frac{\pi_Y p_Y(2t)}{2\pi_C \pi_T}) - \pi_R/\pi_Y \end{aligned}$$

[Étape 3] En remplaçant t, κ_R et κ_Y par leurs estimations, l'expression 2.10 devient :

$$\begin{aligned} \widehat{d}_{TN93} = &- 2ln(1 - \frac{\widehat{q}}{2\widehat{\pi_R}\widehat{\pi_Y}})(\widehat{\pi_R}\widehat{\pi_Y} - \frac{\widehat{\pi_A}\widehat{\pi_G}\widehat{\pi_Y}}{\widehat{\pi_R}} - \frac{\widehat{\pi_C}\widehat{\pi_T}\widehat{\pi_R}}{\widehat{\pi_Y}}) \\ &- 2\frac{\widehat{\pi_A}\widehat{\pi_G}}{\widehat{\pi_R}} ln(1 - \frac{\widehat{q}}{2\widehat{\pi_R}} - \frac{\widehat{\pi_R}}{2\widehat{\pi_A}\widehat{\pi_G}}\widehat{p_R}) \\ &- 2\frac{\widehat{\pi_C}\widehat{\pi_T}}{\widehat{\pi_Y}} ln(1 - \frac{\widehat{q}}{2\widehat{\pi_Y}} - \frac{\widehat{\pi_Y}}{2\widehat{\pi_C}\widehat{\pi_T}}\widehat{p_Y}) \end{aligned}$$

où $\widehat{q}$, $\widehat{p_Y}$ et $\widehat{p_R}$ sont les fréquences observées de transversions, de transitions entre pyrimidines et entre purines respectivement. $\widehat{\pi_X}$ est la fréquence observée de l'état X.

2.4.3 Vraisemblance d'une phylogénie

Les méthodes d'estimation de phylogénies basées sur le principe de maximum de vraisemblance sont de plus en plus utilisées. Ceci est particulièrement vrai depuis 1981, date à laquelle Felsenstein a décrit

le premier algorithme efficace pour le calcul de la vraisemblance d'une phylogénie à partir de séquences nucléiques ou protéiques.

La vraisemblance d'une hypothèse H, notée $P(D|H)$, est la probabilité d'observer les données D sachant que H est correcte. En phylogénie, les données sont constituées par l'ensemble des séquences homologues. L'hypothèse est généralement composée de plusieurs éléments : la topologie de l'arbre, $\mathcal{T}$, à laquelle est associée des longueurs de branches, regroupées au sein du vecteur $\mathbf{l}$, ainsi que le vecteur des paramètres libres du modèle de substitution, $\mathbf{m}$. $\mathbf{l}$ et $\mathbf{m}$ sont des paramètres dits de «**nuisance**», c'est à dire des valeurs que l'on se doit de considérer si l'on veut estimer $\mathcal{T}$. $\boldsymbol{\Upsilon}$ est le vecteur désignant de tels paramètres. La vraisemblance de la phylogénie s'écrit donc $P(D|\mathcal{T}, \boldsymbol{\Upsilon})$.

Si l'hypothèse est constituée de la topologie seulement, la vraisemblance s'écrit alors :

$$P(D|\mathcal{T}) = \int_{\boldsymbol{\Upsilon}} P(D|\mathcal{T}, \boldsymbol{\Upsilon}) P(\boldsymbol{\Upsilon}|\mathcal{T}) d\boldsymbol{\Upsilon}$$

où $P(\boldsymbol{\Upsilon}|\mathcal{T})$ est probabilité des paramètres de nuisance conditionnellement à la topologie de l'arbre.

La fonction de distribution des paramètres de nuisance conditionnellement à la topologie de l'arbre est généralement inconnue mais des méthodes de Monte Carlo par chaînes de Markov permettent d'approximer cette intégrale.

Afin de simplifier les notations, la vraisemblance de la phylogénie étudiée est notée L. L'hypothèse d'indépendance des sites permet d'écrire :

$$L = \prod_{s=1}^{N} L_s \tag{2.11}$$

où L_s est la vraisemblance de la phylogénie au site s et N est le nombre total de sites. Calculer le logarithme de la vraisemblance, noté lnL, permet d'éviter les dérives numériques liées au produit de l'équation 2.11. De plus, deux sites identiques ont la même vraisemblance. En pratique, on applique donc l'équation suivante :

$$lnL = \sum_{s=1}^{N_p} n_s lnL_s$$

où N_p est le nombre de patterns et n_s est le nombre de répétitions (ou poids) du pattern s, et lnL_s est le logarithme de la vraisemblance au site s. Cette factorisation permet d'accélérer les temps de calculs comparé à l'équation 2.11. Le facteur d'accélération est d'autant plus important que les séquences sont peu divergentes et peu nombreuses.

Considérons l'arbre de la Figure 2.5a. La racine est notée r. u et v sont les deux noeuds fils de r. Dans la suite de ce mémoire, les références aux sous-arbres s'expriment par l'intermédiaire des noeuds racines de ces derniers. Ainsi, dans cet exemple, les références aux sous-arbres U et V sont faites par

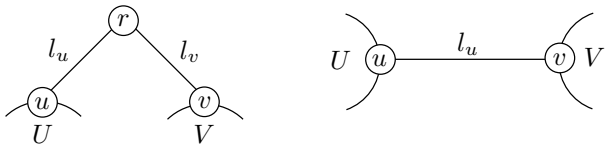


FIG. 2.5 – **Arbre modèle raciné (a) et non-enraciné (b).** U et V sont deux sous-arbres dont les noeuds u et v sont les racines. Les topologies de ces sous-arbres restent inchangées lorsque l'arbre est sous forme enracinée ou non-enracinée.

l'intermédiaire de leurs racines, u et v respectivement. On a :

$$\begin{aligned} L_s &= \sum_{x\in\mathcal{A}} \pi_x L_s(r=x) \\ &= \sum_{x\in\mathcal{A}} \pi_x \left[\sum_{y\in\mathcal{A}} P_{xy}(l_u) L_s(u=y)\right] \left[\sum_{y\in\mathcal{A}} P_{xy}(l_v) L_s(v=y)\right] \end{aligned} \tag{2.12}$$

$L_s(r=x)$ est la vraisemblance au site s du sous-arbre dont la racine est le noeud r, sachant que x est l'état observé à ce noeud. Il s'agit donc de la **vraisemblance conditionnelle** du sous-arbre en question. Si r est un noeud externe alors $L_s(r=x)=1$ si x est effectivement la base observée au site s chez l'UE correspondant à ce noeud ; $L_s(r=x)=0$ sinon.

Si l'état observé est un gap, ou un quelconque autre caractère inconnu, alors $L_s(r=x)=1.0$ quel que soit x. Ce choix peut se traduire en termes mathématiques de la façon suivante. Considérons le cas simple où le site étudié présente un unique gap ou un quelconque caractère inconnu. Soit D_s le vecteur des données au site s. La vraisemblance de la phylogénie T s'écrit $L_s = P(D_s|T) = \sum_{x\in\mathcal{A}} P(D'_s \cap x|T)$, où $D'_s \cap x$ correspond au vecteur D lorsque le caractère inconnu est x. Si le caractère inconnu en question s'observe chez l'UE r, appliquer ce calcul de la vraisemblance revient à fixer $L_s(r=x)=1.0$, $\forall x$. En d'autres termes, nous calculons ici la probabilité de générer des données conformes aux séquences analysées. Cette solution est utilisée au sein des programmes de PHYLIP (Felsenstein, 1993) et PAML (Yang, 1997c) basés sur la vraisemblance.

2.4.4 Calcul de la vraisemblance lorsque le modèle est réversible et homogène

Lorsque le modèle de substitution est réversible et homogène, la vraisemblance de la phylogénie est indépendante de la position de la racine. En effet, pour un tel modèle, on peut écrire :

$$\begin{aligned} L_s &= \sum_{x\in\mathcal{A}} \pi_x \left[\sum_{y\in\mathcal{A}} P_{xy}(l_u) L_s(u=y)\right] \left[\sum_{z\in\mathcal{A}} P_{xz}(l_v) L_s(v=z)\right] \\ &= \sum_{x\in\mathcal{A}}\sum_{y\in\mathcal{A}}\sum_{z\in\mathcal{A}} \pi_x P_{xy}(l_u) P_{xz}(l_v) L_s(u=y) L_s(v=z) \qquad (2.13) \\ &= \sum_{x\in\mathcal{A}}\sum_{y\in\mathcal{A}}\sum_{z\in\mathcal{A}} \pi_y P_{yx}(l_u) P_{xz}(l_v) L_s(u=y) L_s(v=z) \qquad (2.14) \\ &= \sum_{y\in\mathcal{A}}\sum_{z\in\mathcal{A}} \pi_y P_{yz}(l_u+l_v) L_s(u=y) L_s(v=z) \qquad (2.15) \end{aligned}$$

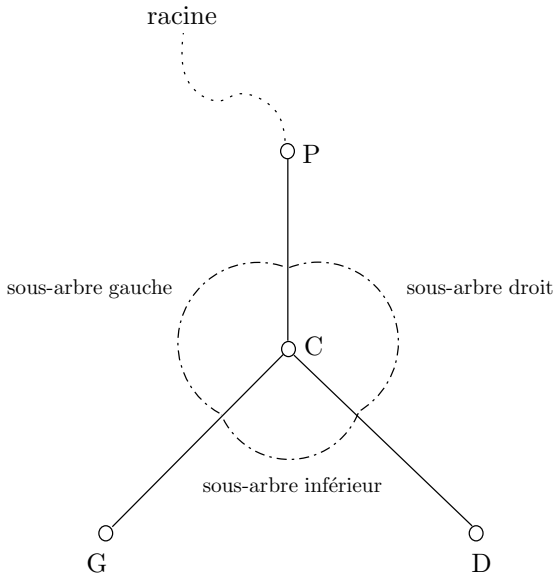


FIG. 2.6 – **Structure d'arbre pour le calcul des vraisemblances conditionnelles.** C est le noeud courant. G et D sont les deux «fils» de ce dernier. C est situé aux racines de trois sous-arbres : les sous-arbres inférieur (G+D), gauche (P+G) et droit (P+D).

L'hypothèse de réversibilité est nécessaire pour le passage de l'équation 2.13 à l'équation 2.14 et celle d'homogénéité du processus de substitution dans l'arbre est nécessaire au passage de l'équation 2.14 à l'équation 2.15. Dans ces trois expressions, la position de la racine passe de la branche joignant v et u (équation 2.13) au noeud u (équation 2.15). Il est alors possible de poursuivre ce déplacement sur une quelconque autre branche adjacente à u. La racine peut donc être placée à un point quelconque de l'arbre. Ces calculs démontrent le fameux «principe de poulie» décrit (et prouvé) par Felsenstein (1981).

2.4.5 Calcul récursif de la vraisemblance

L'expression 2.12 montre que le calcul de la vraisemblance d'une phylogénie repose sur une **procédure récursive** : les vraisemblances conditionnelles au noeud r sont obtenues à partir des vraisemblances conditionnelles aux deux noeuds fils u et v. L'information véhiculée par les vecteurs de vraisemblances conditionnelles «remonte» donc des feuilles vers une racine correspondant au point où est calculée la vraisemblance de l'arbre entier. Décrivons plus en détail ce calcul essentiel.

L'arbre est tout d'abord enraciné par un point choisi arbitrairement. En pratique, la racine correspond souvent à une des feuilles de l'arbre. La structure de données utilisée lors du parcours de la phylogénie est présentée dans la Figure 2.6. Le noeud courant, noté C, a un père, noté P, et deux descendants : les noeuds G et D. C se situe à la racine de trois sous-arbres : les sous-arbres inférieur (G+D), gauche (G+P) et droit (D+P).

Le parcours d'arbre décrit dans l'algorithme 1 permet de calculer les vraisemblances conditionnelles de tous les sous-arbres inférieurs. Nous considérons ici que les vraisemblances conditionnelles au niveau des

feuilles ont été préalablement initialisées. Les vecteurs de vraisemblances conditionnelles sont ici calculés pour chaque noeud fils (lignes 1 et 2) avant de «remonter» vers la racine (ligne 3). La vraisemblance de la phylogénie est obtenue à l'issue du calcul des vraisemblances conditionnelles à la racine. Cette procédure correspond au fameux algorithme «**pruning**» de Felsenstein (1981).

Algorithme 1: Calcul des vraisemblances conditionnelles des sous-arbres inférieurs

```
   PostOrder
   Données : C
   début
       si C est un noeud interne alors
1          PostOrder G;
2          PostOrder D;
3          Calculer les vraisemblances conditionnelles du sous-arbre (G+D) (expression 2.12);
       fin
       sinon retourner ;
   fin
```

2.4.6 Calculer efficacement les vraisemblances conditionnelles de tous les sous-arbres

Considérons à présent la version non-enracinée de l'arbre modèle (Figure 2.5b). Lorsque la longueur l_u est modifiée, la vraisemblance de la phylogénie varie mais les vraisemblances conditionnelles $L_s(u = y)$ et $L_s(v = z)$ restent inchangées. En effet, les vraisemblances conditionnelles d'un sous-arbre sont fonctions de la topologie de ce dernier et des longueurs de branches dans ce sous-arbre uniquement. Obtenir la nouvelle valeur de vraisemblance ne nécessite donc pas de parcourir à nouveau entièrement l'arbre. Calculer et stocker les valeurs de vraisemblances conditionnelles est donc très utile. Une manière efficace d'effectuer ce calcul est présentée ici.

Dans un premier temps, l'algorithme 1 est appliqué. À l'issue de cette première étape, toutes les vraisemblances conditionnelles des sous-arbres inférieurs sont connues. Les vraisemblances conditionnelles des sous-arbres gauches et droits sont ensuite obtenues grâce au parcours d'arbre de l'algorithme 2. Cette méthode est fréquemment utilisée dans le domaine de la phylogénie (V. Ranwez, communication personnelle) et s'applique parfaitement à notre problème : seulement deux parcours d'arbres sont nécessaires pour le calcul de trois vecteurs de vraisemblances conditionnelles à chaque noeud interne. Dans le premier parcours, les calculs de vraisemblances conditionnelles (ligne 3) interviennent après l'appel récursif à la fonction (lignes 1 et 2). Il s'agit d'un parcours «**post-order**». Dans le second, les calculs de vraisemblances conditionnelles (lignes 1 et 2) interviennent avant l'appel récursif (lignes 3 et 4). Il s'agit d'un parcours «**pre-order**».

Le stockage des vecteurs de vraisemblances conditionnelles est relativement coûteux en terme d'espace mémoire. En effet, le nombre total d'octets utilisés pour stocker les vraisemblances conditionnelles d'un arbre à n UEs, pour des séquences de longueurs N sachant que chaque valeur décimale est codée sur o

Algorithme 2: Calcul des vraisemblances conditionnelles des sous-arbres gauches et droits (les vraisemblances conditionnelles des sous-arbres inférieurs ont été calculées au préalable)

```
    PreOrder
    Données : C
    début
        si C est un noeud interne alors
1           Calculer les vraisemblances conditionnelles du sous-arbre (G+P);
2           Calculer les vraisemblances conditionnelles du sous-arbre (D+P);
3           PreOrder G;
4           PreOrder D;
        fin
        sinon retourner ;
    fin
```

octets, est égal à $|\mathcal{A}| \times (n-3) \times 3 \times N \times o$. $|\mathcal{A}|$ est la taille de l'alphabet des caractères considérés (4 pour l'ADN et 20 pour les protéines), $n-3$ est le nombre total de noeuds internes dans une arbre à n UEs et 3 est le nombre de sous-arbres par noeud interne. Ainsi, l'espace mémoire nécessaire est de l'ordre de grandeur de la matrice de séquences alignées ($n \times N$). En général, une valeur décimale est codée sur huit octets. Pour des séquences de 1,000 pb et 500 UEs, les vecteurs de vraisemblances conditionnelles occupent environ 48 Mo. Précisons qu'il s'agit ici d'une borne supérieure car nous ne tenons pas compte de la factorisation des sites (pour 500 UEs cette factorisation est quasiment inexistante). L'espace utilisé n'est donc pas rédhibitoire ici, mais il peut le devenir lorsque le nombre d'états que peuvent prendre les caractères augmente. Ainsi, pour les protéines, les calculs de vraisemblance portent sur vingt états. L'espace mémoire utilisé pour l'exemple précédent dépasse alors 238 Mo. De même, pour l'ADN, lorsque le modèle de substitution inclut la variabilité des vitesses entre sites à partir d'une loi gamma discrétisée, l'espace nécessaire est multiplié par le nombre de catégories de cette loi. Pour quatre catégories, la valeur la plus courante, l'espace mémoire utilisé approche alors 192 Mo.

2.4.7 Problème des probabilités faibles

Pour de grands jeux de données, l'espace mémoire occupé par les vecteurs de vraisemblances conditionnelles n'est pas le seul facteur posant problème. Ainsi, le calcul des vraisemblances conditionnelles sur des phylogénies présentant plus de 100 UEs, peut être entravé de problèmes numériques importants. Il arrive en effet que les valeurs de ces vraisemblances soient trop petites pour être représentées par un ordinateur standard. Nous présentons ici une solution à ce problème similaire à celle décrite par Yang (2000).

Reprenons l'arbre de la Figure 2.5a. Considérons à présent que r est un noeud quelconque de l'arbre et u et v sont ses deux fils. On a :

$$L_s^*(r=x) = \left[\sum_{y\in\mathcal{A}} P_{xy}(l_u)\frac{L_s(u=y)}{\rho_s(u)}\right]\left[\sum_{y\in\mathcal{A}} P_{xy}(l_v)\frac{L_s(v=y)}{\rho_s(v)}\right]$$

où $\rho_s(u) = \max_{y\in\mathcal{A}}(L_s(u=y))$ et $\rho_s(v) = \max_{y\in\mathcal{A}}(L_s(v=y))$. On choisit de calculer $L_s^*(r=x)$, la

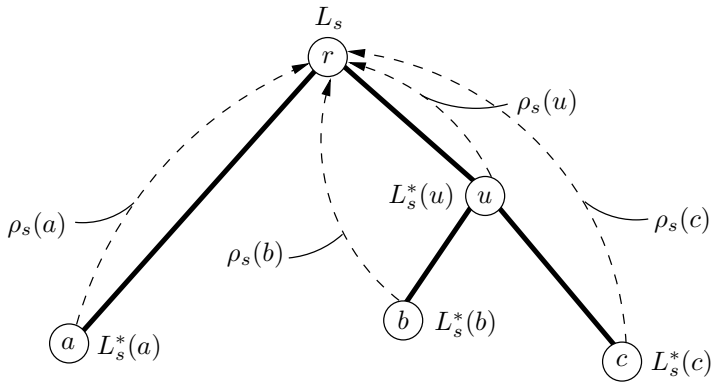


FIG. 2.7 – **Utilisation des facteurs d'adaptation pour le calcul de la vraisemblance.** Au site s, les vraisemblances conditionnelles des sous-arbres inférieurs q sont associées à des facteurs d'adaptation (voir texte), $\rho_s(q)$. Ces derniers sont utilisés pour le calcul de la vraisemblance de la phylogénie au noeud racine r.

vraisemblance conditionnelle **adaptée**, au lieu de $L_s(r = x)$, la vraisemblance conditionnelle originale, lorsque cette dernière valeur est inférieure à un seuil numérique en deçà duquel un nombre est difficilement représentable par un ordinateur. Diviser par la valeur maximale (elle-même proche de 0), permet d'augmenter la valeur de chacune des vraisemblances conditionnelles, et les ramène à une échelle raisonnable. Lorsqu'une des vraisemblances conditionnelles originale d'un sous-arbre est inférieure au seuil, alors le facteur d'adaptation est appliqué à chaque état. À partir de l'expression précédente, on obtient :

$$L_s^* = \frac{1}{\rho_s(u)\rho_s(v)} \sum_{x \in \mathcal{A}} \pi_x \left[\sum_{y \in \mathcal{A}} P_{xy}(l_u) L_s(u = y) \right] \left[\sum_{y \in \mathcal{A}} P_{xy}(l_v) L_s(v = y) \right]$$

et :

$$lnL_s^* = -ln(\rho_s(u)) - ln(\rho_s(v)) + lnL_s$$

On obtient donc :

$$lnL_s = ln(\rho_s(u)) + ln(\rho_s(v)) + lnL_s^* \tag{2.16}$$

Ainsi, les vraisemblances conditionnelles sont rendues calculables grâce à des facteurs multiplicatifs dont les valeurs sont conservées puis réinjectées ensuite (équation 2.16). Les calculs détaillés ci-dessous correspondent au cas simple où l'adaptation des vraisemblances conditionnelles est réalisée aux deux noeuds descendants du noeud racine auquel est calculée la vraisemblance de l'arbre. En pratique, il est possible que l'adaptation s'applique à différents sous-arbres, notés w. Les valeurs de ρ_s sont donc stockées pour chaque sous-arbre et le logarithme de la vraisemblance au site s s'écrit alors :

$$lnL_s = \sum_w ln(\rho_s(w)) + lnL_s^*$$

où la somme porte sur tous les sous-arbres inférieurs w au sein de la phylogénie enracinée au noeud r. La valeur de ρ_s pour les vecteurs de vraisemblance partielles non adaptées est égale à 1.0.

2.5 Avancées récentes

Nous présentons ici deux approches récentes, permettant de prendre en compte des caractéristiques importantes de l'évolution des séquences génétiques, et offrant ainsi la possibilité d'améliorer l'ajustement des modèles de substitution aux données analysées.

2.5.1 Modélisation de la variabilité des vitesses entre sites

L'hétérogénéité des vitesses d'évolution est une caractéristique évolutive très répandue au sein des séquences nucléiques (Uzzell et Corbin, 1971; Wakeley, 1993; Sullivan et al., 1995; Yang et Kumar, 1996; Yang, 1996). La structure même du code génétique favorise cette variabilité : la probabilité qu'une mutation sur la troisième position d'un codon modifie la nature de l'acide aminé traduit, est bien plus faible que si la mutation se produit sur la première ou sur la seconde position. Il en résulte une vitesse d'évolution des sites en troisième position bien supérieure à celles des sites en première ou seconde position.

Ignorer cette variabilité alors qu'elle affecte effectivement les séquences peut causer divers problèmes. Par exemple, les distances évolutives sont sous-estimées lorsque ce facteur n'est pas pris en compte (Jin et Nei, 1990; Tateno et al., 1994). Il en résulte une sous-estimation des longueurs de branches et du ratio transition/transversion dans le meilleur des cas (Wakeley, 1994), et des erreurs dans la topologie de la phylogénie inférée dans le pire des cas (Yang, 1993; Yang et al., 1994; Tateno et al., 1994).

L'analyse de différents groupes de gènes homologues montre que le nombre de substitutions par site est approximativement distribué suivant une loi binomiale négative (Uzzel et Corbin, 1971 mais voir aussi Golding, 1983 pour une revue plus complète). Ceci indique qu'une loi gamma représente convenablement la distribution des taux de substitutions par site. Cette loi présente un paramètre de forme, noté a, dont la valeur est une fonction décroissante de l'intensité de la variabilité des taux. Les trois distributions présentées dans la Figure 2.8 correspondent à trois valeurs distinctes de a. Dans cet exemple, les espérances de ces distributions sont toutes trois égales à 1.0.

La loi gamma a été introduite dans le cadre de l'estimation de distances évolutives (Golding, 1983; Jin et Nei, 1990; Tamura et Nei, 1993; Rzhetsky et Nei, 1994). La probabilité que deux séquences ayant divergé à un instant t antérieur à l'instant présent, différent à un site pris au hasard est alors donnée par :

$$p(2t) = \int_0^\infty p(2t\lambda)g(\lambda)d\lambda$$

où λ est un facteur multiplicatif distribué selon une loi gamma d'espérance égale à 1.0 et dont la fonction de répartition est notée $g(\lambda)$. La résolution de cette équation pour le modèle JC69 donne :

$$p(2t) = \frac{3}{4} - \frac{3}{4}\left(\frac{a}{a+8t}\right)^a$$

où a est le paramètre de forme de la loi gamma. En combinant cette expression avec l'équation 2.7, on

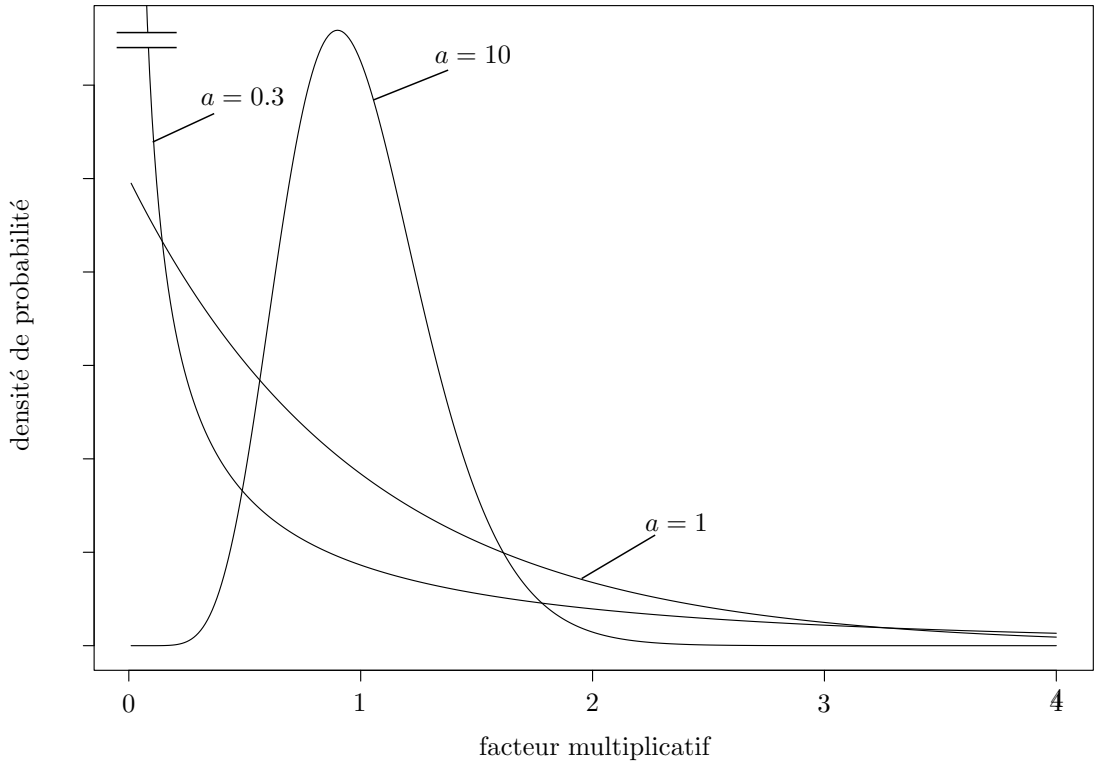


FIG. 2.8 – **Loi gamma.** L'axe des abscisses reporte la valeur du facteur multiplicatif de la vitesse moyenne d'évolution. Les espérances de ces trois distributions sont égales à 1.0 et leurs variances sont des fonctions décroissantes de a, le paramètre de forme de la loi.

obtient :

$$d_{JC69} = \frac{3}{4}a\left((1-\frac{4}{3}p(2t))^{-1/a}-1\right)$$

La distance estimée est donc :

$$\widehat{d}_{JC69} = \frac{3}{4}\alpha\left((1-\frac{4}{3}\widehat{p})^{-1/\alpha}-1\right) \tag{2.17}$$

L'estimateur de la distance évolutive est donc une fonction de α : la valeur estimée du paramètre de forme de la loi gamma. Il existe plusieurs méthodes pour estimer ce paramètre (Sullivan et al., 1995; Yang et Kumar, 1996; Tourasse et Gouy, 1997; Gu et Zhang, 1997). L'objectif de ces dernières est d'approcher la valeur du paramètre décrivant au mieux la variabilité réelle des vitesses d'évolution entre sites. Une nouvelle approche (Guindon et Gascuel, 2002 ; Annexe A) est présentée au chapitre 4. Contrairement aux autres travaux, l'estimation vise ici à définir la valeur de α la plus adaptée pour la construction de topologies d'arbres à partir de distances estimées.

Yang (1993) est le premier à décrire une implémentation de la loi gamma dans le cadre du maximum de vraisemblance. L'expression de la vraisemblance d'une phylogénie est la suivante :

$$L = \prod_{s=1}^{N_p}\int_0^\infty g(\lambda)L_s(\lambda)d\lambda$$

où $L_s(\lambda)$ est la vraisemblance de la phylogénie au site s lorsque toutes les longueurs de branches sont

multipliées par λ. Intégrer la vraisemblance sur une distribution continue des valeurs de λ est très coûteux en temps de calcul et utiliser une approximation discrète de cette distribution est une solution efficace. La vraisemblance s'écrit alors :

$$L = \prod_{s=1}^{N_p} \sum_{c=1}^{C} h(\lambda_c) L_s(\lambda_c)$$

où C est le nombre de catégories de la loi gamma discrétisée, λ_c est le facteur multiplicatif correspondant à la catégorie c et $h(\lambda_c)$ est la probabilité de la catégorie c. Yang (1994) propose une méthode de discrétisation de la loi gamma aboutissant à une distribution uniforme pour $h(\lambda_c)$. On a donc :

$$L = \prod_{s=1}^{N_p} \frac{1}{C} \sum_{c=1}^{C} L_s(\lambda_c) \tag{2.18}$$

En pratique, quatre catégories suffisent pour approximer de manière convenable la distribution continue (Yang et al., 1994). Le temps de calcul de la vraisemblance d'une phylogénie est donc, dans ce cas, multiplié par quatre par rapport au cas où la variabilité des vitesses entre sites n'est pas prise en compte.

La formule de calcul récursif de la vraisemblance (équation 2.12) s'écrit alors :

$$L_s = \sum_{c=1}^{C} \frac{1}{C} \sum_{x \in \mathcal{A}} \pi_x \left[\sum_{y \in \mathcal{A}} P_{xy}(l_u \lambda_c) L_s(u = y, \lambda_c) \right] \left[\sum_{y \in \mathcal{A}} P_{xy}(l_v \lambda_c) L_s(v = y, \lambda_c) \right]$$

où $L_s(u = y, \lambda_c)$ est la vraisemblance du sous-arbre dont u est la racine, conditionnellement à l'état y et lorsque les longueurs de l'ensemble des branches de la phylogénie sont multipliées par λ_c. Pour chacune des valeurs de λ_c, le calcul de la vraisemblance s'effectue de manière standard et les algorithmes 1 et 2 peuvent être utilisés pour l'obtention des vraisemblances conditionnelles.

2.5.2 Modélisation de la variabilité des vitesses entre lignées

Les modèles décrits précédemment reposent sur l'hypothèse de constance de la vitesse d'évolution de chaque site au cours du temps. Or, des fluctuations environnementales peuvent engendrer des variations de pression sélective et ainsi affecter la vitesse d'évolution des macromolécules. Aussi, à la suite de la duplication d'un gène, il est très fréquent d'observer une augmentation de la vitesse d'évolution au sein d'une des deux copies puis un ralentissement lorsqu'un nouvel équilibre fonctionnel et structural est atteint.

Très récemment, Galtier (2001), a proposé un modèle permettant de décrire les variations de vitesses le long d'une phylogénie. Les taux multiplicatifs de la loi gamma discrétisée, correspondant aux facteurs d'accélération de la vitesse d'évolution d'un site dans l'exemple précédent, sont, là-encore, des facteurs d'accélération, mais ceux-ci s'appliquent à présent à chaque position dans la phylogénie. Dans ce modèle,

les substitutions concernent à la fois les changements en bases mais aussi les transitions entre vitesses d'évolution.

Reprenons l'arbre modèle de la Figure 2.5a. La vraisemblance d'une phylogénie à un site donné s'écrit ici :

$$L_s = \sum_{x \in \mathcal{A}} \pi_x \sum_{r=1}^{C} \quad \frac{1}{C} \quad \left[\sum_{y \in \mathcal{A}} \sum_{k=1}^{C} P_{xy,\lambda_r\lambda_k}(l_u) L_s(u = y, \lambda_k) \right] \left[\sum_{y \in \mathcal{A}} \sum_{k=1}^{C} P_{xy,\lambda_r\lambda_k}(l_v) L_s(v = y, \lambda_k) \right]$$

$P_{xy,\lambda_r\lambda_k}(l_u)$ est la probabilité de changement de l'état x vers l'état y et du facteur d'accélération λ_r par λ_k sur une branche de taille l_u. $P_{xy,\lambda_r\lambda_k}(l_v)$ est défini de manière similaire. Les probabilités de changements concernent donc ici les états des caractères ainsi que les vitesses d'évolution. $L_s(u = y, \lambda_k)$ est la vraisemblance du sous-arbre dont u est la racine conditionnellement à l'état y et au facteur d'accélération λ_k à ce noeud. Les probabilités de changements s'obtiennent à partir de l'expression :

$$P_{xy,\lambda_r\lambda_k}(l) = P_{xy|\lambda_r\lambda_k}(l) P_{\lambda_r\lambda_k}(l)$$

Le terme $P_{\lambda_r\lambda_k}(l)$ est une fonction de λ_r, λ_k et l ainsi que d'un paramètre, noté ν, représentant la fréquence des variations de vitesses d'évolution. ν est le seul paramètre supplémentaire comparé au modèle précédent et le temps de calcul de la vraisemblance de la phylogénie à un site est à nouveau multiplié par C.

L'analyse de séquences homologues d'ARN ribosomique suggère que la variation des vitesses d'évolution au sein de la phylogénie est ici une caractéristique importante du mode d'évolution. En effet, tenir compte de ce phénomène améliore significativement l'ajustement du modèle aux données. Cette conclusion s'applique très certainement à la plupart des jeux de données pour lesquels les divergences entre séquences sont très anciennes. L'utilisation de ce type de modèle devrait donc être profitable à l'estimation d'une phylogénie universelle à partir de la comparaison de séquences homologues (Lopez et al., 1999; Philippe et al., 2000).

2.6 Conclusions

Les modèles d'évolution permettent de traduire en termes mathématiques des hypothèses sur les mécanismes évolutifs auxquels sont soumises les séquences. Comme nous l'avons vu dans ce chapitre, les premiers efforts de modélisation ont porté exclusivement sur le processus de substitution d'un état par un autre. Ainsi, pour l'ADN, le modèle le plus simple ne fait pas de distinction entre les états se substituant les uns aux autres tandis que chaque type de substitution est considéré de manière spécifique dans le modèle le plus complexe. Cette sophistication dans la description du processus de substitution est avantageusement complétée par la modélisation de traits évolutifs agissant à des échelles supérieures.

Par exemple, prendre en compte la variabilité des vitesses d'évolution entre sites permet d'améliorer l'ajustement du modèle à un grand nombre de jeux de séquences homologues. La variabilité des vitesses d'évolution au sein de la phylogénie semble aussi être un facteur important.

Les modèles statistiques actuels autorisent ainsi une description de plus en plus fine des processus évolutifs. Cependant, leur degré de complexité est contraint par la quantité limitée d'information qu'il est possible d'extraire des données. Du point de vue du biais affectant l'estimation des paramètres, un modèle riche en paramètres est supérieur à un modèle plus pauvre. Malheureusement, cet avantage est contrebalancé par une augmentation de la variance lorsque le nombre de paramètres augmente. L'estimation de la topologie d'arbre est très certainement concernée par ce phénomène, même si caractériser précisément cette relation est une tâche ardue car il est difficile de définir la variance et le biais d'un estimateur de topologie d'arbre. Ce problème est cependant central pour la question du choix d'un modèle. En effet, si la topologie d'arbre est un paramètre peu sensible aux biais engendrés par des modèles pauvres en paramètres, il peut alors être avantageux d'utiliser un modèle simple si l'objectif est d'obtenir une topologie d'arbre fiable. En d'autres termes, il est probable que la question du choix du modèle décrivant le mieux les données et celle du modèle le plus adapté pour la reconstruction d'une topologie d'arbre admettent fréquemment des réponses distinctes. Cette hypothèse est discutée plus précisément au chapitre 4, dans le cadre de l'estimation de la valeur du paramètre de forme de la loi gamma, modélisant la variabilité des vitesses entre sites.

Chapitre 3

Algorithmes de reconstruction d'arbres phylogénétiques

Le chapitre précédent décrit les principaux modèles d'évolution moléculaire et leur utilisation dans le cadre de l'inférence phylogénétique. Comme nous l'avons vu, l'approche statistique permet d'estimer des distances évolutives et calculer la probabilité d'occurrence des données sous une certaine phylogénie. Nous présentons ici les principaux critères constituant le cœur des méthodes d'inférence d'arbres. Les outils algorithmiques employés pour optimiser ces critères, et ainsi construire des phylogénies, sont aussi aussi largement évoqués. La première partie de ce chapitre est consacrée à la description des approches exploitant les distances estimées entre séquences prises deux à deux, ainsi que celles basées sur le calcul des vraisemblances de phylogénies. Dans les deux cas, nous insistons sur les stratégies utilisées pour optimiser les différents critères à la base de ces méthodes. Les approches ne reposant pas explicitement sur un modèle statistique, comme celle du maximum de parcimonie, ne sont pas discutées ici. Le second volet du chapitre détaille les stratégies mises en place pour explorer efficacement l'espace des topologies d'arbre. Cette partie du processus d'estimation jouent un rôle primordial dans l'ensemble des méthodes récentes d'inférence phylogénétique.

3.1 Méthodes basées sur les distances évolutives

L'utilisation de distances entre séquences est une approche très largement répandue pour l'estimation de phylogénies. Pour s'en persuader, nous noterons que l'article de Saitou et Nei (1987), décrivant le fameux algorithme **Neighbor-Joining** (NJ), a été cité plus de 7,000 fois à travers la littérature scientifique. Le succès des méthodes de distances s'explique notamment par leur rapidité, mais aussi et surtout par de bonnes propriétés asymptotiques. Ainsi, lorsque les distances estimées sont les distances réelles, l'arbre estimé par NJ, mais aussi par la majorité des méthodes de distances, est exactement l'arbre vrai.

Deux grandes familles de méthodes se distinguent : les approches **locales** et **globales**. Cette distinction est déterminée par la nature du processus de construction de l'arbre. Ainsi, certains algorithmes procèdent à partir d'une phylogénie localement résolue. Les méthodes d'**agglomérations** sont fondées

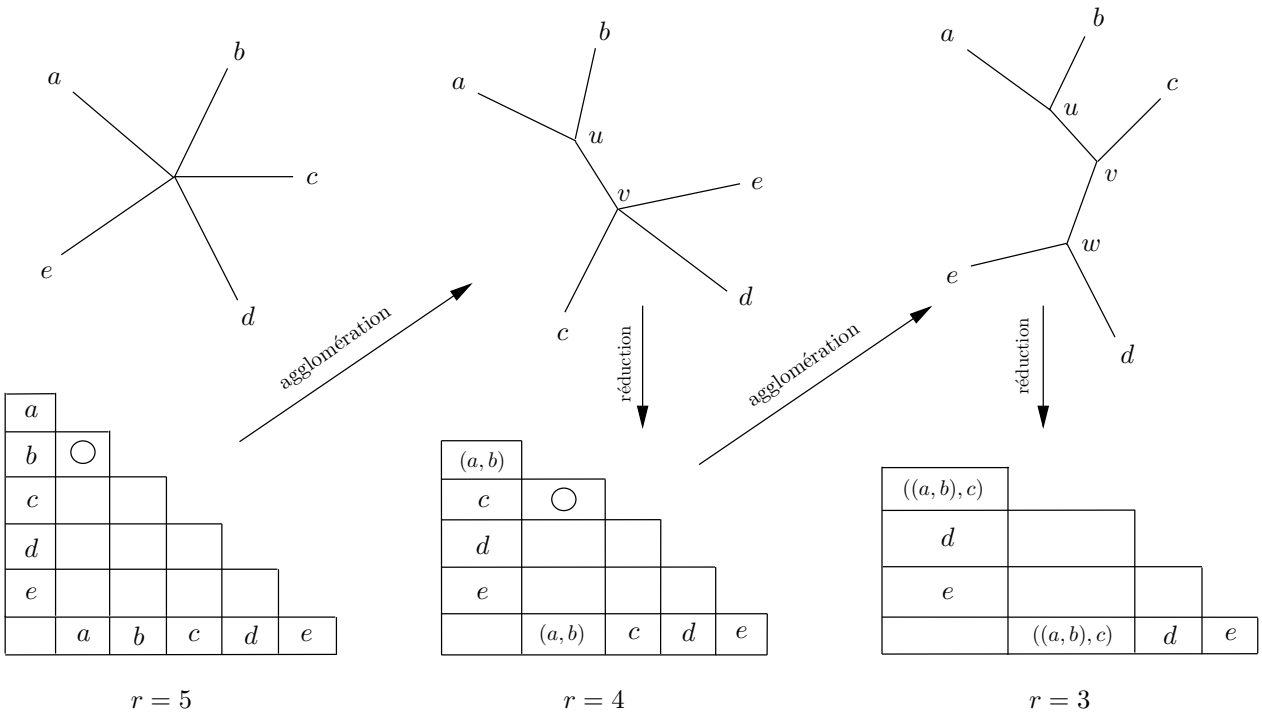


FIG. 3.1 – **Construction d'une phylogénie par agglomérations.** Le cercle situé au sein d'une des cellules de la matrice de distance (en bas) désigne la paire de noeuds à agglomérer. Chaque agglomération, c'est à dire le regroupement de deux noeuds dans l'arbre, entraîne une réduction de la dimension de la matrice, notée r.

sur ce principe : les noeuds agglomérés précédemment constituent une phylogénie «partielle» tandis que les noeuds non-agglomérés forment une «étoile». En revanche, les algorithmes basés sur une approche globale procèdent à partir d'un graphe correspondant globalement à une phylogénie à chaque étape de la construction. Les méthodes d'**insertion** font partie de cette famille. Nous verrons, dans la seconde partie de ce chapitre, que ces deux types d'approches n'offrent pas les mêmes possibilités en termes d'exploration des topologies d'arbres en cours de construction.

3.1.1 Approches locales

Le schéma agglomératif évoqué ci-dessus comprend trois étapes : (1) la sélection d'une paire de noeuds à partir de la valeur d'un critère (local) de voisinage, (2) l'estimation des longueurs des branches joignant chacun des deux noeuds de la paire, au nouveau noeud engendré par l'agglomération, et (3) la réduction de la matrice de distances, c'est à dire le calcul des distances entre le nouveau noeud et les noeuds non-agglomérés. Les distinctions entre méthodes agglomératives correspondent à différentes définitions données à chacun de ces trois points. Nous détaillons tout d'abord l'algorithme NJ dans sa version modifiée par Studier et Keppler (1988) et évoquons ensuite certaines de ses variantes.

Les deux noeuds agglomérés minimisent :

$$\mathcal{Q}_{ab} = (r-2)\Delta_{ab} - S_a - S_b \tag{3.1}$$

où a et b sont deux noeuds non-agglomérés, $S_x = \sum_{y=1}^{r} \Delta_{xy}$, et r est le nombre de noeuds non-agglomérés (Figure 3.1). La valeur du critère correspond, à une constante près, à la taille de l'arbre à cette étape, c'est à dire la somme de ses longueurs de branches (Gascuel, 1994). Ce critère de sélection des paires voisines est une version locale du critère d'évolution minimum. Ce dernier est détaillé plus loin, dans le cadre d'une approche globale pour l'inférence d'arbres à partir de distances évolutives. Lorsque a et b sont agglomérés, les longueurs des deux branches reliant a et u et b et u sont estimées par :

$$l_a = \frac{1}{2}(\Delta_{ab} + \frac{S_a - S_b}{r-2}) \tag{3.2}$$

$$l_b = \frac{1}{2}(\Delta_{ab} + \frac{S_b - S_a}{r-2}) \tag{3.3}$$

Ces estimations sont celles des moindres carrés ordinaires lorsque a et b sont des feuilles. Tel n'est plus le cas lorsque ces deux noeuds sont des racines de sous-arbres présentant plusieurs UEs (Gascuel, 1997b). Les expressions exactes des estimations au sens des moindres carrés pour toutes les longueurs de branches dans l'arbre sont présentées plus bas.

Enfin, l'étape de réduction consiste à calculer les distances entre le nouveau noeud u, considéré à présent comme un noeud externe, et tout autre noeud externe, noté k. On a :

$$\Delta_{uk} = \frac{1}{2}(\Delta_{ak} - l_a) + \frac{1}{2}(\Delta_{bk} - l_b) \tag{3.4}$$

Les étapes de sélection des paires à agglomérer (équation 3.1), de calcul des longueurs de branches (équations 3.2 et 3.3) et de réduction (équation 3.4), sont répétées tant que $r > 3$ (Figure 3.1).

Cette version de l'algorithme NJ est celle proposée par Studier et Kepler (1988). Bien que les phylogénies reconstruites par cette approche soient identiques à celles obtenues en appliquant la version originale de l'algorithme (Gascuel, 1994), ces deux approches divergent par le nombre d'opérations effectuées. Ce dernier est de l'ordre de n^5 pour l'approche décrite par Saitou et Nei (1987) tandis que la complexité de l'algorithme de Studier et Kepler est de $\mathcal{O}(n^3)$, rendant ainsi possible l'analyse de jeux de données présentant plusieurs milliers de séquences.

BIONJ (Gascuel, 1997a) et **Weighbor** (Bruno et al., 2000) sont des variantes de NJ largement utilisées à l'heure actuelle. BIONJ est fondée sur une expression mathématique plus générale pour l'étape de réduction, prenant en compte les variances des distances évolutives. La fiabilité des topologies d'arbres inférés par cette approche est supérieure à celle obtenue par NJ pour des temps de calculs identiques. Weighbor délaisse le principe d'évolution minimum pour l'étape d'agglomération et remplace celui-ci par un critère basé sur la vraisemblance. Les performances de cette méthode, toujours en termes de fiabilité des topologies inférées, sont supérieures à celles de NJ et BIONJ. Malheureusement, cette amélioration est contrebalancée par une nette augmentation des temps de calculs.

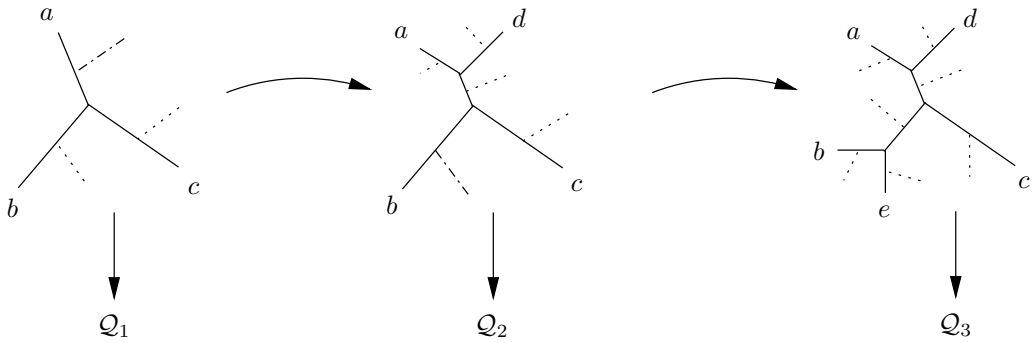


FIG. 3.2 – **Construction d'une phylogénie par insertions.** La phylogénie est construite par ajouts successifs d'UEs. À chaque étape du processus est calculée la valeur d'un critère $\mathcal{Q}$, déterminant le point d'insertion optimal pour le nouveau noeud.

3.1.2 Approches globales

Ici, l'agglomération de noeuds, constituant la base des approches locales, est remplacée par l'insertion d'UEs. Comme nous l'avons indiqué précédemment, cette stratégie permet de manipuler une phylogénie entièrement résolue à chaque étape de la construction. Le principe de construction d'un arbre par insertion est illustré dans la Figure 3.2. Pour chaque point d'insertion potentiel, les longueurs de branches sont ajustées et la valeur du critère est calculée. Le point d'insertion choisi correspond à la valeur minimum du critère à cette étape. Ce type d'algorithme est dit «**glouton**» : l'insertion d'une UE n'est jamais remise en cause au cours des insertions suivantes (sauf si les insertions sont couplées à des remaniements topologiques).

Les **moindres carrés** et le **minimum d'évolution** sont les deux principaux critères exploités par ce type d'approche. Leurs définitions et leurs utilisations dans le cadre de l'inférence de phylogénies sont détaillés ci-dessous.

Critère des moindres carrés

En 1967, Fitch et Margoliash proposent d'appliquer le critère des moindres carrés, très fréquemment utilisé en statistique, à l'inférence phylogénétique à partir de distances évolutives. Soit d_{ab}, la distance réelle entre a et b, c'est à dire le nombre exact de substitutions par site séparant les séquences a et b. Δ_{ab} est l'estimation de cette distance et Δ^T_{ab} est la distance estimée entre a et b au sein de la phylogénie inférée. Cette dernière correspond à la somme des longueurs des branches séparant les UEs a et b. Le critère à minimiser est noté $\mathcal{Q}_{mc}(T)$. On a :

$$\mathcal{Q}_{mc}(T) = \sum_{a \neq b} w_{ab}(\Delta_{ab} - \Delta^T_{ab})^2 \tag{3.5}$$

où n est le nombre d'UEs. w_{ab} correspond au poids associé à la différence entre Δ_{ab} et Δ^T_{ab}. Lorsque $w_{ab} = 1$, $\mathcal{Q}_{mc}$ est le critère des moindres carrés ordinaires et lorsque $w_{ab} = 1/\sigma^2_{ab}$, où σ^2_{ab} est la variance estimée de Δ_{ab}, $\mathcal{Q}_{mc}$ est le critère des moindres carrés pondérés. Des expressions analytiques approchées

des σ^2_{ab} peuvent être obtenues pour la plupart des modèles d'évolution, à partir de l'inverse de la dérivée seconde de la fonction de vraisemblance (voir Andrieu, 1997, p. 60 pour un exemple). Les variances sont inversement proportionnelles à la taille des séquences considérées et, au premier ordre, proportionnelles aux distances entre séquences. Fitch et Margoliash (1967) proposent ainsi d'utiliser $w_{ab} = 1/{\Delta_{ab}}^2$ comme pondération et $w_{ab} = 1/\Delta_{ab}$ est aussi très fréquemment employé. Signalons enfin que l'approche des moindres carrés généralisés introduit des pondérations prenant en compte les covariances entre distances. Théoriquement, cette approche est supérieure aux autres du point de vue de l'estimation des longueurs de branches (Bulmer, 1991).

Comme le montre l'expression 3.5, le but des méthodes basées sur les moindres carrés est de définir des distances d'arbres (estimées) les plus proches possible des distances observées entre séquences prises deux à deux. Or, les distances d'arbres sont des sommes de longueurs de branches. La minimisation de $\mathcal{Q}_{mc}$ repose donc entièrement sur la manière dont sont calculées de telles longueurs. Voyons l'approche classique pour ce calcul dans le cadre des moindres carrés ordinaires. Soit $\mathbf{A}$ la matrice décrivant la topologie de l'arbre. $A_{(i,j)b}$ est le terme de $\mathbf{A}$ situé sur la ligne (i,j), à la colonne b. i et j sont deux UEs et (i,j) est le rang de ce couple, tandis que b désigne le rang de la branche correspondante. $A_{(i,j)b} = 1$ si le chemin menant de i à j dans l'arbre passe par b et $A_{(i,j)b} = 0$ sinon. Pour l'arbre modèle de la Figure 3.3a on a :

$$\begin{pmatrix} \Delta_{ab} \\ \Delta_{ac} \\ \Delta_{ad} \\ \Delta_{bc} \\ \Delta_{bd} \\ \Delta_{cd} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix} \times \begin{pmatrix} l_a \\ l_b \\ l_c \\ l_d \\ l_e \end{pmatrix} + \begin{pmatrix} \epsilon_a \\ \epsilon_b \\ \epsilon_c \\ \epsilon_d \\ \epsilon_e \end{pmatrix}$$

Sous forme matricielle, cette expression s'écrit $\mathbf{\Delta} = \mathbf{A} \times \mathbf{l} + \epsilon$, où $\mathbf{\Delta}$ est le vecteur des distances estimées, correspondant ici à la variable à expliquer, et $\mathbf{l}$ est le vecteur des longueurs de branches, correspondant à la variable explicative, dont on veut estimer la valeur. L'ajustement des longueurs de branches au sens des moindres carrés consiste donc à trouver $\mathbf{l}$, tel que $\mathbf{\Delta}^T$ soit la plus proche possible de $\mathbf{\Delta}$. La résolution de ce système au sens des moindres carrés donne (il s'agit d'une régression linéaire) :

$$\mathbf{l} = (\mathbf{A}^t\mathbf{A})^{-1}\mathbf{A}^t\mathbf{\Delta}$$

Pour l'estimation des longueurs de branches au sens des moindres carrés pondérés et généralisés, on a $\mathbf{l} = (\mathbf{A}^t\mathbf{W}^{-1}\mathbf{A})^{-1}\mathbf{A}^t\mathbf{W}^{-1}\mathbf{\Delta}$, où $\mathbf{W}$ est la matrice de variances et de variances-covariances respectivement.

Cette méthode d'estimation des longueurs de branches est applicable lorsque le nombre de séquences est relativement restreint. Vach (1989) et Rzhetsky et Nei (1993) proposent des formules analytiques plus directes que le précédent calcul matriciel. Supposons que les noeuds a, b, c, et d de l'arbre de la Figure 3.3 sont les racines de sous-arbres A, B, C et D présentant respectivement $|A|$, $|B|$, $|C|$ et $|D|$ UEs. L'estimation au moindre carrés ordinaires de l_e est alors donnée par :

$$l_e = \frac{1}{2}[\lambda(\Delta_{AC} + \Delta_{BD}) + (1-\lambda)(\Delta_{AD} + \Delta_{BC}) - (\Delta_{AB} + \Delta_{CD})] \tag{3.6}$$

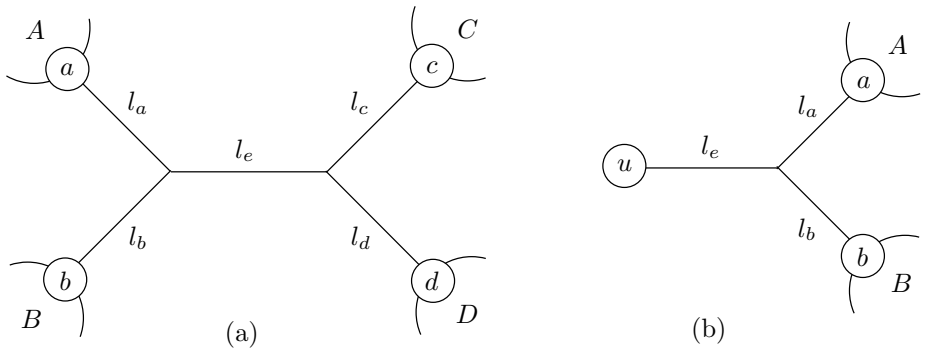


FIG. 3.3 – **Arbres modèles.** A, B, C, et D sont trois sous-arbres dont a, b, c et d sont les racines respectives. u est un noeud externe.

avec

$$\lambda = \frac{|A||D| + |B||C|}{(|A| + |B|)(|C| + |D|)}$$

et la distance entre deux sous-arbres V et W est la moyenne des distances entre UEs appartenant à ceux-ci :

$$\Delta_{VW} = \frac{1}{|V||W|} \sum_{v \in V, w \in W} \Delta_{vw} \tag{3.7}$$

L'estimation de la longueur de la branche externe menant à l'UE u sur l'arbre modèle de la Figure 3.3b s'écrit :

$$l_e = \frac{1}{2}(\Delta_{AU} + \Delta_{BU} - \Delta_{AB}) \tag{3.8}$$

Les équations 3.6, 3.7 et 3.8 montrent que les longueurs de branches estimées ne dépendent pas de la topologie des sous-arbres a, b, c, et d et a et b respectivement. Cette propriété explique en grande partie la rapidité des algorithmes d'inférence phylogénétique basés sur l'estimation des longueurs de branches aux moindres carrés ordinaires.

Critère d'évolution minimum

En 1971, Kidd et Sgaramella-Zonta proposent un nouveau critère pour l'inférence phylogénétique. Celui-ci repose sur l'idée que le scénario évolutif proposant un nombre de substitutions minimum est le plus probable. Cette idée est inspirée du principe du rasoir d'Ockham, stipulant que les hypothèses les plus simples doivent être préférées aux plus complexes. Notons $\mathcal{Q}_{em}$ ce critère. Pour un arbre non-enraciné, on a :

$$\mathcal{Q}_{em} = \sum_{i=1}^{2n-3} l_i$$

où $2n - 3$ est le nombre total de branches dans un arbre non-enraciné à n UEs. Optimiser le critère d'évolution minimum revient donc à chercher la phylogénie de taille minimale, c'est à dire l'arbre dont la somme des longueurs de branches est minimale. Là encore, le cœur de cette approche repose sur la façon dont sont estimées les longueurs de branches.

Un résultat théorique important a été apporté en 1993 par Rzhetsky et Nei. Ces auteurs ont montré que la somme des longueurs de branches estimées par la méthode des moindres carrés à partir des vraies distances, et donc de distances d'arbres, est la plus petite pour l'arbre vrai. Aucun autre arbre, de topologie ou de longueurs de branches différentes, n'a une taille estimée inférieure à celle de l'arbre vrai. Le critère d'évolution minimum combiné à l'estimation des longueurs de branches par les moindres carrés ordinaires est donc consistant : l'arbre estimé est l'arbre vrai lorsque les distances estimées sont suffisamment proches des vraies distances. Cette avancée théorique est décisive car, jusqu'à ces travaux, l'hypothèse que l'arbre vrai est celui qui possède une somme de longueurs de branches minimale, n'était pas fondée mathématiquement. Il existe d'autres résultats intéressants à ce sujet. Ainsi, en 2000, Gascuel, Denis et Bryant ont montré que le principe d'évolution minimum combiné à l'estimation des longueurs de branches par les moindres carrés pondérés et généralisés (voir Bulmer, 1991) n'est pas consistant. À partir des distances vraies, il est parfois possible de trouver un arbre de taille estimée inférieure à celle de l'arbre vrai. Il est plutôt inattendu que le principe du minimum d'évolution ne «fonctionne» pas correctement avec les moindres carrés pondérés et généralisés, alors qu'il se comporte très bien avec les moindres carrés ordinaires. Mais, comme le note David Bryant (communication personnelle), la vraie question n'est-elle pas plutôt : « Pourquoi le principe d'évolution minimum fonctionne si bien avec les moindres carrés ordinaires ? ». Ces considérations semblent, certes, plutôt techniques, mais elles pointent sans aucun doute sur des problèmes de portées plus générales, et, même si les travaux de Rzhetsky et Nei, et d'autres par la suite, ont montrés la consistance du principe d'évolution minimum sous certaines conditions, ils n'en expliquent pas vraiment la cause.

La méthode d'inférence d'arbres décrite par Rzhetsky et Nei (1992) est basée sur l'exploration de topologies d'arbres voisines de celles estimées par NJ. Pour chacune de ces topologies, les longueurs de branches sont estimées au sens des moindres carrés ordinaires et l'arbre retenu au final est celui dont la somme des longueurs de branches est la plus faible. Les résultats obtenus montrent que les phylogénies inférées par cette approche sont légèrement moins fiables que celles estimées par NJ (Gascuel, 2000).

Très récemment, Pauplin (2000) puis Desper et Gascuel (2002) ont proposé une approche originale basée sur une nouvelle définition des distances entre sous-arbres, distances intervenant dans l'estimation des longueurs de branches au sens des moindres carrés (voir équations 3.6 et 3.8). L'approche standard est non-pondérée : pour le calcul des distances entre sous-arbres Δ_{AC}, Δ_{BD}, Δ_{AD}, Δ_{BC}, Δ_{AB} et Δ_{CD}, le même poids est attribué aux UEs présentes au sein de A, B, C et D. La distance entre sous-arbres proposée ici est dite «balancée» car elle prend en compte le nombre d'UEs de chacun des ces sous-arbres. La construction de la phylogénie est basée sur un processus d'insertion complété de réarangements topologiques locaux. La complexité de l'algorithme proposé par Desper et Gascuel est de $\mathcal{O}(diam(T) \times n^2)$, où $diam(T)$ est le diamètre de l'arbre inféré, c'est à dire le plus grand nombre de branches séparant deux de ses feuilles. Notons que cette complexité est quasiment optimale car un minimum de $\mathcal{O}(n^2)$ opérations sont nécessaires pour l'analyse d'une matrice $n \times n$. En pratique, les temps de calculs sont inférieurs

à ceux de NJ (ou BIONJ) tandis que l'inférence de topologies est beaucoup plus fiable. Sur de grands jeux de données (100 UEs), les performances obtenues par cette approche en termes de fiabilité sont nettement supérieures à celles de Weighbor et FITCH (Felsenstein, 1993), une méthode fondée sur le critère des moindres carrés pondérés et procédant pas l'insertion d'UEs (Desper et Gascuel, en cours de publication).

3.2 Méthodes basées sur la vraisemblance

Comme nous l'avons vu au chapitre précédent, la vraisemblance d'une hypothèse est la probabilité d'occurrence des données sous celle-ci. Une méthode d'inférence intuitive consiste donc à chercher l'hypothèse maximisant cette probabilité. C'est l'approche du **maximum de vraisemblance**. Une deuxième stratégie consiste à calculer la probabilité des hypothèses conditionnellement aux données. Il s'agit de l'approche **bayesienne**. La différence fondamentale entre ces deux méthodes réside dans l'introduction de probabilités *a priori* sur les hypothèses dans le cadre de l'inférence bayesienne. Ces deux critères sont présentés ci-dessous.

3.2.1 Critère du maximum de vraisemblance

Reprenons les notations du chapitre précédent. T est la phylogénie considérée et $\mathcal{T}$, sa topologie. $\boldsymbol{\Upsilon}$ est le vecteur des paramètres de nuisance. Ce dernier se décompose en $\mathbf{l}$, le vecteur de longueurs de branches, et $\mathbf{m}$ le vecteur des paramètres libres du modèle de substitution (par exemple, le ratio transition/transversion). La phylogénie la plus vraisemblable, notée $\widehat{T}$, est définie de la façon suivante :

$$\widehat{T} = \underset{\{\mathcal{T},\mathbf{l},\mathbf{m}\}}{\operatorname{argmax}} (L(\mathcal{T},\mathbf{l},\mathbf{m},D))$$

Ici, l'approche exhaustive consiste à trouver, pour chaque topologie, les valeurs des longueurs de branches et des paramètres libres du modèle de substitution qui maximisent la vraisemblance de la phylogénie. Ainsi, les méthodes fondées sur le principe du maximum de vraisemblance, tout comme les méthodes de distances, reposent sur un algorithme d'exploration des topologies et le calcul d'un critère pour chacune de celles-ci.

Une première estimation de la phylogénie est généralement obtenue par insertions successives d'UEs suivant le même principe que celui utilisé dans le cadre des approches globales pour les méthodes de distances. Le critère des moindres carrés est simplement remplacé ici par celui de la vraisemblance. L'ordre d'insertion des UEs est généralement dicté par celui d'apparition des séquences homologues dans le fichier analysé. Or, deux permutations de ce classement n'aboutissent pas nécessairement à la même phylogénie. La plupart des programmes implémentant cette approche proposent donc la possibilité d'itérer le processus à partir de différents ordres d'insertions.

Une approche agglomérative est aussi envisageable : les deux UEs sélectionnées à chaque étape correspondent au couple qui maximise la vraisemblance de la phylogénie en cours de construction. Mais, il semble que le principe du maximum de vraisemblance s'accorde mal à une approche de type locale. En effet, les performances de cette méthode, en termes de fiabilité des topologies inférées, sont inférieures à la précédente (Adachi et Hasegawa, 1996, p. 48).

Les méthodes basées sur les **quartets** permettent aussi d'estimer une première phylogénie en s'inspirant du principe de vraisemblance maximale (Strimmer et von Haeseler, 1996). Un quartet est un arbre à quatre UEs. Le calcul de la vraisemblance d'une telle structure est très rapide et peut être réalisé pour l'ensemble des quartets déduits des n séquences homologues analysées. Pour résumer, les UEs sont insérées successivement et le point d'insertion correspond à la position la moins conflictuelle vis à vis des quartets inférés, pondérés par leurs vraisemblances respectives. La fiabilité des topologies inférées par cette approche est similaire, voir inférieure, à celle de NJ, pour des temps de calculs bien supérieurs (Ranwez et Gascuel, 2001). Ces performances décevantes comparées à celles de l'approche d'insertion classique s'expliquent probablement par la simple constatation que les fondements de cette approche ne sont pas ceux du maximum de vraisemblance. Bien que les arbres à quatre UEs soient inférés suivant ce principe, la procédure d'insertion utilisée ne vise pas à maximiser la probabilité des données conditionnellement à la phylogénie. De plus, l'inférence d'arbres à quatre UEs est particulièrement sensible à l'attraction des longues branches, un artefact bien connu en phylogénie, pouvant engendrer des erreurs dans l'estimation de la topologie (voir Felsenstein, 1978 et Philippe, 2000 pour une description du problème et de ses conséquences).

Les méthodes d'exploration de topologies d'arbres venant généralement compléter celles décrites ci-dessus, sont discutées plus loin dans ce chapitre. Nous insistons à présent sur l'ajustement des longueurs de branches à une topologie donnée ainsi que l'optimisation des paramètres du modèle.

En pratique, la vraisemblance d'une phylogénie est une fonction trop complexe pour exhiber les expressions analytiques des valeurs optimales des paramètres de nuisance. Ces valeurs sont donc obtenues par **optimisation numérique**. Les méthodes d'optimisation sont nombreuses (Press et al., 1988) et les performances de celles-ci, mesurées par leur rapidité et leurs capacités d'ajustement, varient suivant la nature de la fonction. Pour notre exemple, la fonction à maximiser décrit assurément un paysage complexe (la surface de vraisemblance) et présente généralement de multiples optima locaux (Steel, 1994), rendant difficile la recherche d'optima globaux. Les simulations réalisées par Rogers et Swofford (1999) indiquent cependant que de tels optima sont pratiquement absents lorsque les phylogénies considérées sont proches de l'arbre ayant servi à engendrer les données. Ceci signifie que l'ajustement des paramètres libres du modèle de substitution et des longueurs de branches pour une topologie proche ou identique à celle de l'arbre vrai ne nécessite généralement pas de recourir à des méthodes d'optimisation permettant d'échapper à des optima locaux de la fonction.

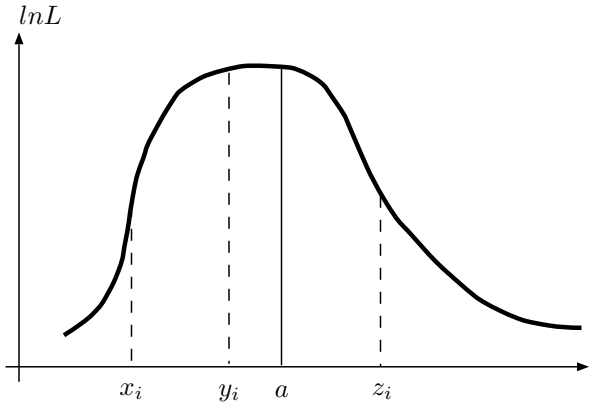


FIG. 3.4 – **Encadrement d'un maximum par le triplet** (x_i, y_i, z_i)**.** L'axe des abscisses reporte les valeurs du paramètre à ajuster. L'axe des ordonnées indique les valeurs du logarithme de la vraisemblance, lnL. Le triplet de points (x_i, y_i, z_i) encadre un maximum de la fonction. a est un nouveau point, et constitue un élément du triplet suivant, $(x_{i+1}, y_{i+1}, z_{i+1})$, dans le processus d'ajustement (voir texte). À chaque étape de l'optimisation, la distance sur l'axe des abscisses entre a et x_i est égale à la distance entre y_i et z_i.

3.2.2 Optimisation des paramètres libres du modèle de substitutions

Les paramètres libres du modèle de substitution sont généralement ajustés indépendamment des longueurs de branches. Les algorithmes de la **«section d'or»** et de **Brent** (1973) peuvent être appliqués pour ce problème (voir Annexe B pour un exemple). Ceux-ci sont décrits ci-dessous.

Méthode de section du nombre d'or

La méthode de la section d'or permet d'ajuster la valeur d'un unique paramètre. x_i, y_i et z_i sont trois valeurs de ce paramètre, telles que $x_i < y_i < z_i$. Ces trois valeurs constituent un triplet noté (x_i, y_i, z_i). L'indice i est le nombre d'itérations effectuées. Les vraisemblances en ces trois points sont notées $L(x_i)$, $L(y_i)$ et $L(z_i)$, et on a $L(y_i) > L(z_i)$ et $L(y_i) > L(x_i)$ quel que soit i (Figure 3.4).

Considérons que y_i est une fraction R du chemin entre x_i et z_i :

$$R = \frac{y_i - x_i}{z_i - x_i} \text{ et } 1 - R = \frac{z_i - y_i}{z_i - x_i}$$

Supposons à présent qu'un point a est placé entre y_i et z_i, définissant ainsi une nouvelle fraction S :

$$S = \frac{a - y_i}{z_i - x_i}$$

Si $L(a) > L(y_i)$, le nouveau triplet $(x_{i+1}, y_{i+1}, z_{i+1})$ correspond à (y_i, a, z_i). La distance entre x_{i+1} et z_{i+1} est alors égale à $1 - R$. Si $L(a) < L(y_i)$, le triplet $(x_{i+1}, y_{i+1}, z_{i+1})$ correspond à (x_i, y_i, a). La distance entre x_{i+1} et z_{i+1} est alors égale à $R + S$. Ainsi, à l'étape $i + 1$, les valeurs du paramètre encadrant le maximum de la fonction définissent soit un intervalle de taille $1 - R$ (si $L(a) > L(y_i)$), soit un intervalle de taille $R + S$ (si $L(a) < L(y_i)$). Afin de converger vers un maximum en un nombre d'étapes minimum,

il est souhaitable que la diminution de la taille de l'intervalle encadrant le maximum soit la même lorsque celui-ci mesure $1 - R$ ou $R + S$ à l'étape $i + 1$. Le choix le plus raisonnable pour placer a est donc de définir S tel que $1 - R = R + S$. On a alors :

$$S = 1 - 2R \tag{3.9}$$

S est positif seulement si $R < 0.5$ et le point a est donc placé dans le plus grand des deux segments. Cependant, il reste encore à déterminer la position précise de a sur ce segment. Pour cela, il suffit de remarquer qu'à l'étape i, le point y_i a été placé de façon optimale. Cette similarité d'échelle permet d'écrire :

$$\frac{S}{1 - R} = R \tag{3.10}$$

D'après les équations 3.9 et 3.10, on en déduit $R = \frac{3-\sqrt{5}}{2} \simeq 0.382$. Ainsi, à chaque étape i de l'algorithme, le rapport $\frac{y_i - x_i}{z_i - x_i}$ est égal à 0.382 et le rapport $\frac{z_i - y_i}{z_i - x_i}$ est égal à 0.618. Ces deux fractions correspondent aux fameuses «sections d'or».

Ces étapes de mises à jour du triplet encadrant un maximum sont répétées jusqu'à ce que la distance entre les deux extrema de l'intervalle passe en deçà d'un certain seuil. La convergence vers le maximum est **linéaire** car, à chaque étape, la taille de l'intervalle encadrant l'extrema est multipliée par un facteur 0.618.

Méthode de Brent (1973)

La section d'or est une approche prudente de l'optimisation d'une variable. Le nombre d'itérations avant d'atteindre un maximum est relativement élevé mais les hypothèses sur la forme de la fonction sont réduite. La méthode de Brent (1973) repose sur l'hypothèse que la fonction est de forme parabolique à proximité du maximum. Si une parabole s'ajuste parfaitement à cette zone, il existe une expression analytique simple donnant l'abscisse correspondant au maximum de la parabole.

La Figure 3.5 décrit deux étapes successives de ce processus d'ajustement. Reprenons les notations précédentes. À l'étape i, le triplet de points (x_i, y_i, z_i) encadrant le maximum sont les abscisses des points 1, 2, et 3 figurant sur la courbe. Par ces trois points passe la parabole A. Les coordonnées du point 4, dont l'abscisse correspond au maximum de cette première parabole, sont obtenues par interpolation parabolique inverse. Le nouveau triplet $(x_{i+1}, y_{i+1}, z_{i+1})$ est alors constitué des abscisses des points 1, 2 et 4 par lesquels passe la nouvelle parabole B. Pour cet exemple, l'abscisse correspondant au maximum de cette dernière parabole est une bonne approximation de la valeur du paramètre maximisant la fonction.

Lorsque la parabole s'ajuste à la fonction de manière satisfaisante, la procédure est rapide. En effet, dans cette situation la convergence est **quadratique**. Ceci signifie que, près d'un zéro de la fonction, le nombre de chiffres significatifs de la solution approximée par cette méthode, double à chaque étape. En revanche, dans certains cas, l'ajustement parabolique peut être bien plus lent, et surtout moins sur, que

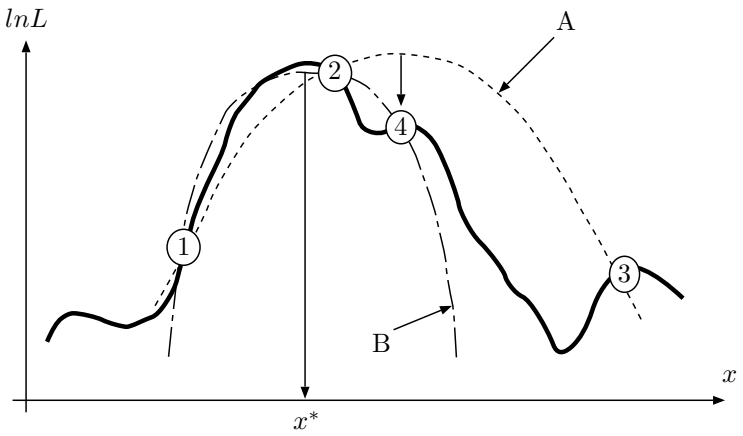


FIG. 3.5 – **Ajustement parabolique itéré.** Les points 1, 2 et 3 définissent une première parabole A. Le point 4 est une projection du maximum de cette parabole sur la fonction. La parabole B passe par les points 1, 2 et 4. L'abscisse de son maximum est une bonne approximation du maximum de lnL.

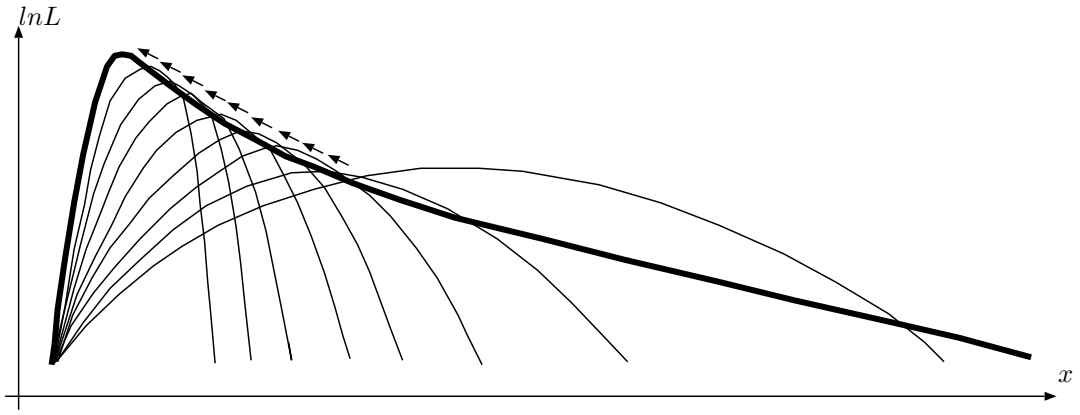


FIG. 3.6 – **Ajustement parabolique dans un cas défavorable.** La convergence vers le maximum de lnL est ici très lente.

l'algorithme du nombre d'or. Le Figure 3.6 illustre une telle situation. Ainsi, en pratique, la méthode du nombre d'or est appliquée jusqu'à ce qu'un critère mesurant l'ajustement d'une parabole à la fonction soit satisfait et l'ajustement parabolique prend alors le relais.

Application à la phylogénie

Ces deux méthodes d'optimisation reposent sur le calcul de la vraisemblance de la phylogénie pour différentes valeurs du paramètre à ajuster. En pratique, ni la topologie, ni les longueurs de branches de l'arbre vrai ne sont connues. La surface de vraisemblance déduite de l'arbre estimé est donc très fréquemment distincte de celle de l'arbre vrai. Cependant, pour une phylogénie inférée réaliste, cet écart entre les deux fonctions n'est généralement pas pénalisant et les valeurs optimales des paramètres peuvent être approximées de manière très satisfaisante. Par exemple, Yang (1996) signale que la valeur du paramètre de forme de la loi gamma, mesurant la variabilité des vitesses d'évolution entre sites, n'est sous-estimée

que lorsque l'estimation est effectuée à partir d'une phylogénie aberrante. Une phylogénie estimée rapidement par une méthode de distances constitue un support efficace pour l'ajustement de ce paramètre et d'autres, comme le ratio transition/transversion. Cette approche est généralement satisfaisante lorsque la finalité est l'inférence d'un arbre par maximum de vraisemblance. Lorsque l'objectif est plutôt d'obtenir une estimation très précise des paramètres du modèle, les valeurs ajustées peuvent être utilisées pour la construction d'un arbre à partir duquel sont à nouveau optimisées les valeurs des paramètres. Ces deux étapes sont itérées jusqu'à l'obtention d'une phylogénie et des valeurs de paramètres stables. La méthode d'inférence d'arbre décrite au chapitre 5 suit cette stratégie.

3.2.3 Optimisation des longueurs de branches

Les méthodes utilisées pour l'optimisation des **longueurs de branches** d'une phylogénie sont nombreuses. Distinguons ici les approches permettant d'ajuster **simultanément** l'ensemble des longueurs de branches de l'arbre, de celles procédant branche par branche, **successivement**. Considérer les branches une à une est un problème d'optimisation unidimensionnel qui peut être abordé par les deux méthodes décrites ci-dessus. Néanmoins, contrairement aux exemples précédents, les paramètres à ajuster sont ici fortement interdépendants : la longueur optimale d'une branche est liée aux longueurs des autres branches dans l'arbre. Afin de tenir compte de cette corrélation, les longueurs de chacune des branches sont optimisées une à une et cette série d'ajustements est réitérée tant que les longueurs optimales ne sont pas stables. Cependant, il est très probable que la longueur optimale d'une branche obtenue à la première itération soit différente de cette même longueur à la dernière itération. Par conséquent, optimiser «complètement» les longueurs de branches à chaque série d'ajustement est inutile. Concrètement, le nombre d'itérations dans l'algorithme d'ajustement d'une longueur de branche est limité *a priori*. Ainsi, tant que les longueurs de branches ne sont pas stables, celles-ci sont successivement optimisées «incomplètement». Cette approche est nettement plus rapide que la précédente car le temps passé à optimiser chaque branche est plus court. Ainsi, la nette diminution des temps de calculs de fastDNAml (Olsen et al., 1994) comparés à ceux de DNAML dans la version 3.2 du package PHYLIP est, en partie, due à cette stratégie d'optimisation.

Méthode de Newton-Raphson

Cette approche permet d'approximer les racines d'un système d'équations. Or, la dérivée d'une fonction est nulle à ses extrema. L'idée est donc d'utiliser la méthode de Newton-Raphson pour approximer les valeurs des paramètres annulant la dérivée du logarithme de la vraisemblance. Cet outil est utilisé pour l'ajustement des longueurs de branches au sein de logiciels largement diffusés, tels que fastDNAml (Olsen et al., 1994) ainsi que les différents programmes de PHYLIP (Felsenstein, 1993) basés sur la vraisemblance.

Nous détaillons tout d'abord le principe de l'algorithme lorsque la fonction présente un seul paramètre. Puis nous présentons la procédure générique, s'appliquant à un problème à plusieurs paramètres. Nous

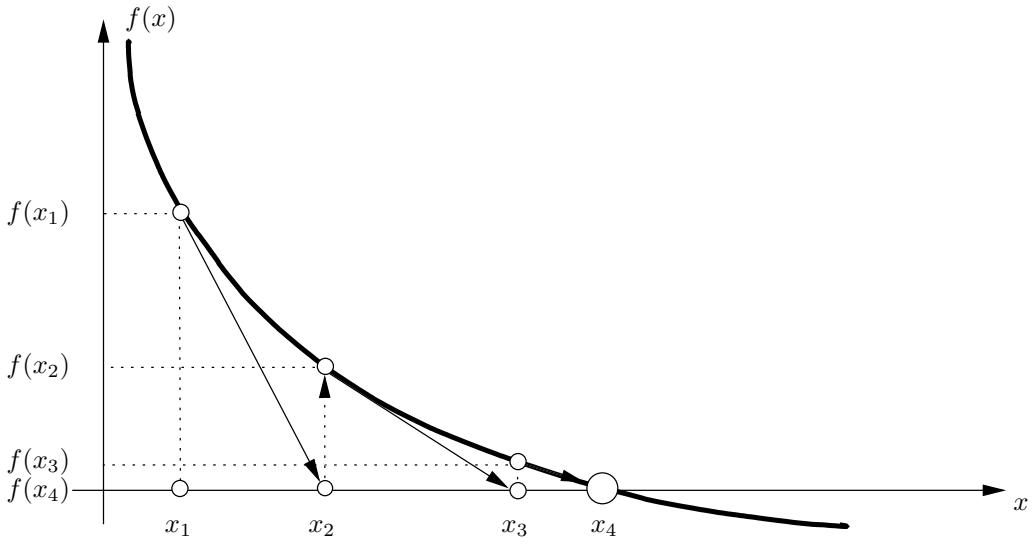


FIG. 3.7 – **Approximation de zéros d'une fonction par la méthode de Newton-Raphson.** Le principe de cet algorithme consiste à approximer successivement la fonction par ses tangentes (voir texte). La tangente au point d'abscisse x_1 désigne x_2 comme point d'abscisse suivant dans le processus d'optimisation. La valeur $f(x_2)$ représente l'erreur dans l'approximation de la fonction par la tangente d'abscisse x_1.

verrons enfin comment ces algorithmes sont utilisés, en pratique, pour l'optimisation des longueurs de branches de manière à maximiser le logarithme de la vraisemblance d'une phylogénie.

D'un point de vue géométrique (Figure 3.7), l'algorithme de Newton-Raphson consiste à (1) prolonger la tangente tracée au niveau du point d'abscisse courant, x_i. (2) L'intersection entre cette droite et l'axe des abscisses correspond au point x_{i+1}. (3) x_{i+1} devient le point d'abscisse courant. (4) Les étapes (1)-(3) sont itérées jusqu'à la convergence vers un zéro de la fonction.

D'un point de vue algébrique, le développement limité du premier ordre d'une fonction $f(x+\delta)$ au voisinage de x s'écrit :

$$f(x+\delta) \simeq f(x) + \delta f'(x) \tag{3.11}$$

où $f'(x) = \frac{d}{dx} f(x)$. On a donc :

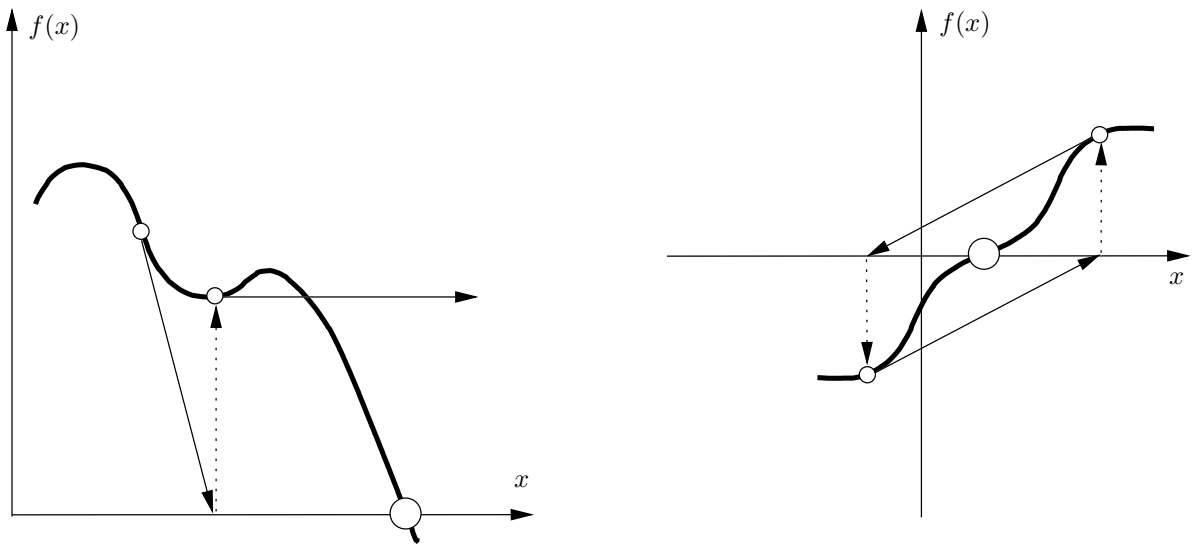
$$\delta \simeq \frac{f(x+\delta) - f(x)}{f'(x)}$$

Si $x+\delta$ est un zéro de f, alors :

$$\delta \simeq -\frac{f(x)}{f'(x)}$$

Cette dernière expression suggère une approche itérative permettant d'approximer le ou les zéro(s) de la fonction. La valeur de x_{i+1}, c'est à dire x à l'étape $i+1$, est obtenue à partir de celle de x_i à laquelle est ajoutée le pas $-\frac{f(x_i)}{f'(x_i)}$, ce qui correspond très précisément à l'interprétation de la Figure 3.7. Si les itérations procèdent au voisinage d'un zéro, cette approche permet généralement de converger vers ce dernier.

Lorsque le point de départ est éloigné d'un zéro, les termes de second ordre dans le développement



(a) (b)

FIG. 3.8 – **Deux situations dans lesquelles la méthode de Newton-Raphson échoue.** (exemples tirés de l'ouvrage «Numerical Recipes in C» (Press et al., 1988), p. 271, 272)

limité ne sont plus négligeables et les approximations réalisées ci-dessus ne sont plus valables. Ceci aboutit à des situations dans lesquelles la convergence n'est pas garantie. Par exemple, il se peut qu'un extremum de la fonction, dont l'abscisse est située entre l'abscisse initiale et un zéro, conduise l'algorithme vers des valeurs aberrantes (Figure 3.8a). Une autre situation empêchant la convergence est décrite dans la Figure 3.8b. Dans le cadre de l'optimisation d'une longueur de branche, il pourrait être intéressant de caractériser l'évolution des valeurs de la dérivée première du logarithme de la vraisemblance, et de rapprocher éventuellement les courbes obtenues aux cas «pathologiques» pour lesquels la méthode de Newton-Raphson est inefficace. Ceci permettrait peut-être d'avoir une idée plus précise de l'effort à fournir pour obtenir des valeurs initiales des longueurs de branches conduisant l'algorithme à converger.

Malgré ces inconvénients, dans des conditions d'applications favorables, l'algorithme de Newton-Raphson présente l'avantage non négligeable de converger à une vitesse **quadratique**. Il existe donc ici un compromis à trouver entre la rapidité de convergence et le temps nécessaire au calcul de la dérivée. La recherche d'un extremum d'une fonction nécessite un calcul supplémentaire : celui de la dérivée seconde de la fonction. Par conséquent, il est fortement recommandé d'appliquer cet algorithme à partir d'une valeur initiale la plus proche possible d'un extremum.

La méthode de Newton-Raphson permet aussi d'annuler simultanément un ensemble de fonctions F_k de plusieurs paramètres. Soit $\mathbf{x} = (x_1, \ldots, x_k, \ldots, x_q)$ le vecteur des q paramètres. On cherche :

$$F_k(x_1, \ldots, x_k, \ldots, x_q) = 0 \qquad \forall k = 1, 2, \ldots, q$$

$\mathbf{F}$ est le vecteur des fonctions F_k. Le développement limité au premier ordre de $\mathbf{F}$ s'écrit :

$$\begin{pmatrix} F_1(\mathbf{x}+\delta\mathbf{x}) \\ \vdots \\ F_k(\mathbf{x}+\delta\mathbf{x}) \\ \vdots \\ F_N(\mathbf{x}+\delta\mathbf{x}) \end{pmatrix} \simeq \begin{pmatrix} F_1(\mathbf{x}) \\ \vdots \\ F_k(\mathbf{x}) \\ \vdots \\ F_N(\mathbf{x}) \end{pmatrix} + \begin{pmatrix} \frac{\partial}{\partial x_1}F_1(\mathbf{x}) & \dots & \frac{\partial}{\partial x_k}F_1(\mathbf{x}) & \dots & \frac{\partial}{\partial x_q}F_1(\mathbf{x}) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_1}F_k(\mathbf{x}) & \dots & \frac{\partial}{\partial x_k}F_k(\mathbf{x}) & \dots & \frac{\partial}{\partial x_q}F_k(\mathbf{x}) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_1}F_q(\mathbf{x}) & \dots & \frac{\partial}{\partial x_k}F_q(\mathbf{x}) & \dots & \frac{\partial}{\partial x_q}F_q(\mathbf{x}) \end{pmatrix} \begin{pmatrix} \delta x_1 \\ \vdots \\ \delta x_k \\ \vdots \\ \delta x_N \end{pmatrix}$$

où encore :

$$\mathbf{F}(\mathbf{x}+\delta\mathbf{x}) \simeq \mathbf{F}(\mathbf{x}) + \mathbf{J}\delta\mathbf{x}$$

où $\mathbf{J}$ est la matrice jacobienne. Le pas d'ajustement est donc ici :

$$\delta\mathbf{x} \simeq -\mathbf{J}^{-1}\mathbf{F}(\mathbf{x})$$

Ainsi, à chaque étape de l'ajustement l'ensemble des valeurs composant le vecteur $\mathbf{x}$ sont mises à jour. Les différents inconvénients et avantages de l'approche multidimensionnelle de la méthode de Newton-Raphson sont identiques à ceux décrits dans la cas unidimensionnel.

Voyons à présent comment utiliser ces deux approches pour l'ajustement des longueurs de branches d'une phylogénie de manière à maximiser le logarithme de sa vraisemblance. Optimiser des longueurs de branches une à une est un problème d'optimisation à une dimension. Notons $lnL'(l_x) = \frac{d}{dl_x}lnL(l_x)$, $lnL''(l_x) = \frac{d^2}{dl_x^2}lnL(l_x)$, où $lnL(l_x)$ est le logarithme de la vraisemblance et l_x est la longueur de branche ajustée. On a :

$$lnL'(l_x+\delta_x) \simeq lnL'(l_x) + \delta_x lnL''(l_x)$$

et au voisinage d'un zéro de la dérivée du logarithme de la vraisemblance :

$$\delta_x \simeq -\frac{lnL'(l_x)}{lnL''(l_x)}$$

L'ajustement d'une seule longueur de branche consiste donc à ajouter à l_x le pas $-\frac{lnL'(l_x)}{lnL''(l_x)}$ à chaque étape. Obtenir la valeur de ce dernier nécessite le calcul des dérivées premières et secondes du logarithme de la vraisemblance par rapport à chaque longueur de branche. Nous verrons ci-dessous comment calculer ces quantités.

Pour l'ajustement simultané, le vecteur $\mathbf{l}$, composé de l'ensemble des longueurs de branches, est mis à jour à chaque étape. La matrice jacobienne dans le cas précédent est ici la matrice hessienne de $lnL(\mathbf{l})$, notée $\mathbf{H}$:

$$\mathbf{H} = \begin{pmatrix} \frac{\partial^2}{\partial l_1^2}lnL(\mathbf{l}) & \dots & \frac{\partial^2}{\partial l_1 \partial l_k}lnL(\mathbf{l}) & \dots & \frac{\partial^2}{\partial l_1 \partial l_q}lnL(\mathbf{l}) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial l_k \partial l_1}lnL(\mathbf{l}) & \dots & \frac{\partial^2}{\partial l_k^2}lnL(\mathbf{l}) & \dots & \frac{\partial^2}{\partial l_k \partial l_q}lnL(\mathbf{l}) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial l_q \partial l_1}lnL(\mathbf{l}) & \dots & \frac{\partial^2}{\partial l_q \partial l_k}lnL(\mathbf{l}) & \dots & \frac{\partial^2}{\partial l_q^2}lnL(\mathbf{l}) \end{pmatrix}$$

où q est le nombre de branches dans l'arbre. Le pas d'ajustement est ici :

$$\delta \mathbf{l} = -\mathbf{H}^{-1} \frac{\partial}{\partial l} lnL(\mathbf{l})$$

où $\frac{\partial}{\partial l} lnL(\mathbf{l})$ est le vecteur colonne des dérivées premières du logarithme de la vraisemblance par rapport à chaque longueur de branche l. D'une étape à la suivante, l'ensemble des longueurs de branches est mis à jour. Afin de simplifier les calculs, les éléments de la matrice hessienne situés hors de la diagonale sont considérés comme nuls. Par conséquent, le pas d'ajustement des longueurs de chaque branche est ici exactement le même que celui appliqué dans le cas unidimensionnel. La vraie distinction entre ces deux approches réside donc plutôt dans l'opposition entre la modification successive *vs.* simultanée des longueurs de branches, plutôt qu'ajustement unidimensionnel *vs.* multidimensionnel. Notons aussi que l'inverse de la matrice hessienne correspond à la matrice de variance-covariance entre longueurs de branches (au signe près). Considérer comme nuls les éléments non-diagonaux revient à négliger les covariances entre longueurs de branches. Les conséquences de cette approximation sont généralement minimes. Elles sont discutées plus en détails dans la thèse de N. Galtier (1997, p. 64-67).

Voyons à présent comment calculer les dérivées premières et secondes du logarithme de la vraisemblance par rapport aux longueurs de branches. Pour l'arbre modèle de la Figure 2.5b, les dérivées première et seconde de la vraisemblance au site s sont obtenues à partir des expressions suivantes :

$$\begin{aligned}
\frac{d}{dl_u} L(l_u) &= \sum_{y \in \mathcal{A}} \pi_y \sum_{z \in \mathcal{A}} L_s(u = y) \frac{d}{dl_u} P_{yz}(l_u) L_s(v = z) \\
\frac{d^2}{dl_u^2} L(l_u) &= \sum_{y \in \mathcal{A}} \pi_y \sum_{z \in \mathcal{A}} L_s(u = y) \frac{d^2}{dl_u^2} P_{yz}(l_u) L_s(v = z)
\end{aligned}$$

On en déduit les valeurs des dérivées du logarithme de la vraisemblance en appliquant :

$$\begin{aligned}
\frac{d}{dl_u} lnL_s(l_u) &= \frac{d}{dl_u} L_s(l_u) \frac{1}{L_s} \\
\frac{d^2}{dl_u^2} lnL_s(l_u) &= -\frac{1}{{L_s}^2} \left(\frac{d}{dl_u} L_s(l_u) \right)^2 + \frac{1}{L_s} \frac{d^2}{dl_u^2} L_s(l_u)
\end{aligned}$$

Les dérivées du logarithme de la vraisemblance pour le jeu de données complet s'obtiennent à partir d'expressions similaires à celui du calcul de la vraisemblance :

$$\begin{aligned}
\frac{d}{dl_u} lnL(l_u) &= \sum_{s=1}^{N_p} n_s \frac{d}{dl_u} lnL_s(l_u) \\
\frac{d^2}{dl_u^2} lnL(l_u) &= \sum_{s=1}^{N_p} n_s \frac{d^2}{dl_u^2} lnL_s(l_u)
\end{aligned}$$

n_s est le nombre de répétitions du pattern s et N_p est le nombre total de patterns.

Ainsi, pour chaque site de l'alignement, les dérivées premières et secondes de la vraisemblance sont calculées par rapport à chacune des branches de la phylogénie. Ce calcul est analogue à celui de la vraisemblance. On en déduit les dérivées du logarithme de la vraisemblance et le pas d'ajustement pour

chacune de ces branches. Les longueurs de branches sont ensuite modifiées puis les vraisemblances conditionnelles de chaque sous-arbre sont mises à jours. Enfin, la vraisemblance de la phylogénie est évaluée. Ces ajustements simultanés cessent lorsque la différence de vraisemblance entre deux étapes est en deçà d'une valeur seuil fixée *a priori*.

Lorsque la phylogénie initiale présente une topologie ou des longueurs de branches aberrantes, il se peut que la vraisemblance diminue entre deux étapes de la procédure d'optimisation. Ce phénomène est aussi parfois observé lors des dernières étapes de l'ajustement. Felsenstein et Churchill (1996) proposent une modification de la méthode Newton-Raphson permettant de garantir la convergence vers un maximum. Cette modification suit le schéma suivant : lorsque la vraisemblance de la phylogénie décroît entre les étapes i et $i+1$, les valeurs des pas sont divisées par deux puis appliquées aux longueurs de branches obtenues à l'étape i. Si cette modification ne conduit pas à un accroissement de la vraisemblance, la division des valeurs de pas et la mise à jour des longueurs de branches sont à nouveau effectuées. Ce processus est poursuivi tant que la vraisemblance est inférieure ou égale à sa valeur à l'étape i.

Ajustements simultanés *vs.* successifs

L'ajustement des longueurs de branches une à une peut être réalisé de manière efficace en appliquant la méthode décrite par Adachi et Hasegawa (1996). Cet algorithme astucieux de parcours d'arbre est présenté en détails dans la thèse de V. Ranwez (2002). Il autorise l'optimisation des longueurs de branches lors d'un unique parcours d'arbre nécessitant la mise à jour d'un nombre réduit de vecteurs de vraisemblances partielles. En revanche, l'ajustement simultané requiert une mise à jour de l'ensemble des vraisemblances conditionnelles après chaque application d'un pas d'ajustement.

À partir de l'analyse de jeux de données simulés, nous avons pu observer un léger avantage en faveur de l'approche successive du point de vue de l'optimisation de la vraisemblance. Yang (2000) aboutit à des conclusions similaires concernant l'efficacité des deux approches. Il signale cependant qu'ajuster les longueurs de branches une à une peut être inefficace lorsque celles-ci sont fortement corrélées à différents paramètres libres du modèle de substitution, en particulier la variabilité des vitesses entre sites.

3.2.4 L'approche bayesienne

L'inférence bayesienne est apparue récemment dans le domaine de la phylogénie moléculaire (Rannala et Yang, 1996; Mau, 1996; Li, 1996; Simon et Larget, 2000; Huelsenbeck et Ronquist, 2001). L'objectif est ici d'estimer les probabilités conjointes des différentes topologies ($\mathcal{T}$) et paramètres de nuisance ($\boldsymbol{\Upsilon}$) en fonction des données, ou probabilité *a posteriori*.

Soit $T = (\mathcal{T}, \boldsymbol{\Upsilon})$ une phylogénie de topologie $\mathcal{T}$, associée au vecteur $\boldsymbol{\Upsilon}$ des paramètres de nuisance.

D'après la formule de Bayes, on a :

$$P(T|D) = \frac{P(D|T)P(T)}{P(D)} \tag{3.12}$$

Cette dernière expression peut s'écrire sous la forme :

$$P(\mathcal{T}, \boldsymbol{\Upsilon}|D) = \frac{P(D|\mathcal{T}, \boldsymbol{\Upsilon})P(\mathcal{T}, \boldsymbol{\Upsilon})}{\sum_{j=1}^{B(n)} \int_{\boldsymbol{\Upsilon}_j} P(D|\mathcal{T}_j, \boldsymbol{\Upsilon}_j)P(\mathcal{T}_j, \boldsymbol{\Upsilon}_j)d\boldsymbol{\Upsilon}_j} \tag{3.13}$$

ou encore :

$$P(\mathcal{T}, \boldsymbol{\Upsilon}|D) \quad \propto \quad P(D|\mathcal{T}, \boldsymbol{\Upsilon})P(\mathcal{T}, \boldsymbol{\Upsilon}) \tag{3.14}$$

car la valeur du dénominateur dans l'expression 3.13 est indépendante de $\mathcal{T}$ et $\boldsymbol{\Upsilon}$. $B(n)$ est le nombre total de topologies d'arbres à n UEs. $P(D|T)$ est la vraisemblance de la phylogénie T et $P(T)$ est la probabilité *a priori* de celle-ci. L'équation 3.14 montre que, pour une distribution uniforme des probabilités *a priori* des phylogénies, chercher l'arbre le plus probable revient à chercher l'arbre le plus vraisemblable. Ainsi, du point de vue des topologies d'arbres, le point central des méthodes bayesiennes, résidant dans l'introduction de connaissances *a priori*, n'est véritablement exploité qu'à condition d'introduire des probabilités *a priori* informatives, et donc non-uniformes.

Si les paramètres de nuisance sont indépendants de la topologie de l'arbre, $P(\mathcal{T}, \boldsymbol{\Upsilon})$ est le produit de la probabilité *a priori* de la topologie à laquelle sont associées des longueurs de branches, par la probabilité *a priori* des paramètres libres du modèle de substitution. Le choix des distributions de ces probabilités est subjectif. C'est d'ailleurs à ce niveau que se situe l'avantage offert par l'inférence bayesienne.

Larget et Simon (1999) utilisent une distribution uniforme sur les **histoires labellées**, ou distribution de Yule. Une histoire labellée est une phylogénie racinée pour laquelle les évènements de spéciation sont ordonnés. Il est connu qu'une telle distribution favorise une certaine forme de topologie d'arbres. Or, il est difficile de savoir si cette tendance correspond à une réalité biologique.

Yang et Rannala (1997) proposent une méthode permettant d'établir les probabilités *a priori* des histoires labellées associées à des longueurs de branches. Ces deux auteurs se basent sur un processus de naissance et extinction de lignées modulé notamment par deux paramètres dont les espérances sont estimées au sens du maximum de vraisemblance. La probabilité conjointe *a priori* d'une l'histoire labellée et des longueurs de ses branches est déduite analytiquement à partir des valeurs de ces deux paramètres générées aléatoirement et uniformément dans des intervalles centrés sur leurs moyennes respectives.

Les distributions des probabilités *a priori* des paramètres libres du modèle de substitution sont généralement uniformes dans des intervalles excluant les valeurs aberrantes (Huelsenbeck et Bollback, 2001). Lorsqu'un modèle pertinent, noté $\mathcal{G}$, permet d'obtenir des expressions analytiques des probabilités *a priori*, deux stratégies sont alors envisageables. La première est d'utiliser les valeurs les plus vraisemblables des paramètres de $\mathcal{G}$ pour le calcul analytique. C'est l'approche **empirique**. La deuxième solution est d'intégrer sur les valeurs des paramètres de $\mathcal{G}$. C'est l'approche **hiérarchique**.

Revenons sur l'expression de la probabilité d'une phylogénie conditionnellement aux données (équation 3.13). Le dénominateur fait intervenir une somme sur l'ensemble des topologies d'arbres à n UEs. En pratique, ce nombre est bien trop grand pour espérer calculer la probabilité *a posteriori* d'une phylogénie sur des durées raisonnables. Les méthodes de Monte Carlo par Chaîne de Markov permettent de contourner ce problème. L'idée est de construire une chaîne de Markov permettant d'échantillonner au sein de la distribution *a posteriori* des phylogénies. Ce point est détaillé dans la partie qui suit.

3.3 Exploration de l'espace des topologies d'arbres

Outre l'estimation des valeurs des paramètres de nuisance, l'exploration de l'espace des topologies d'arbre est un aspect central de l'inférence phylogénétique. La nature discrète des topologies inscrit cette tâche dans le domaine de l'**optimisation combinatoire**. Comme précédemment, l'objectif est d'ajuster ce paramètre particulier afin de minimiser (ou maximiser) une fonction, telle que la taille de la phylogénie, sa vraisemblance ou sa probabilité *a posteriori*. Dans la suite de ce chapitre, une telle fonction est appelée fonction de coût et l'objectif est de minimiser ce dernier.

Le nombre de topologies d'arbres non-enracinés à n UEs est égal à $\prod_{i=3}^{n}(2i-5)$ et, comme le soulignent Lemmon et Milinkovitch (2002), il existe donc environ 3×10^{84} topologies distinctes pour une phylogénie à 55 UEs, soit plus que le nombre d'atomes dans l'univers ! Par conséquent, il n'est pas envisageable d'utiliser une approche exhaustive pour analyser la plupart des jeux de données. Ceci est d'autant plus vrai que les grandes quantités de séquences disponibles à l'heure actuelle constituent fréquemment des familles de gènes homologues présentant plusieurs centaines, voire plusieurs milliers de séquences. Il n'est donc pas surprenant que la mise au point de méthodes efficaces pour l'optimisation de topologies soit un thème de recherche privilégié.

Deux grandes classes de stratégies se distinguent. Les premières sont **déterministes** : la topologie de l'arbre est perturbée et ce changement est accepté s'il conduit à une diminution du coût. Il est systématiquement refusé si le coût augmente. Les approches appartenant à cette classe sont des méthodes d'améliorations itératives. Les secondes sont **stochastiques** : la topologie est perturbée et ce changement est accepté s'il conduit à une diminution du coût, ou refusé avec une certaine probabilité lorsque le coût augmente.

Nous décrivons à présent les différentes opérations permettant de modifier une topologie. Leurs utilisations dans le cadre de méthodes d'améliorations itératives sont ensuite discutées. Les principes des approches stochastiques, de plus en plus utilisées à l'heure actuelle, sont détaillés par la suite.

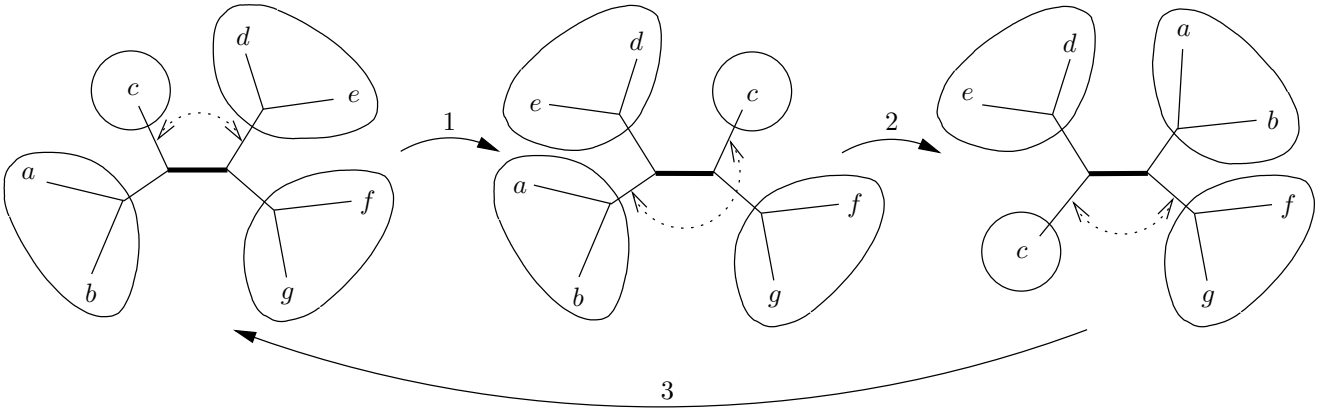


FIG. 3.9 – **NNI.** Les flèches en pointillés indiquent les échanges de sous-arbres. Les flèches pleines soulignent les transitions entre topologies.

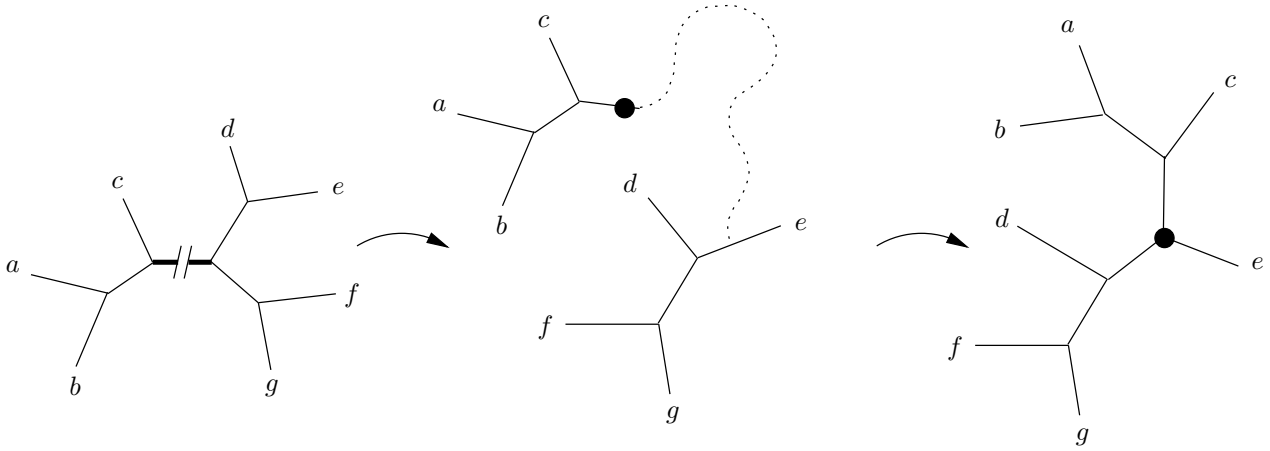


FIG. 3.10 – **SPR.** Après section (à gauche), le sous-arbre constitué des UEs a, b et c est «donneur» (voir texte). Le cercle plein indique le noeud de liaison à ce sous-arbre (au centre).

3.3.1 Perturbation d'une topologie d'arbre

Les trois «mouvements élémentaires» les plus répandus sont : «Nearest Neighbor Interchange» (**NNI**), «Subtree Pruning and Regrafting» (**SPR**), et «Tree Bisection and Reconnection» (**TBR**) (Figures 3.9, 3.10 et 3.11). Appliqué de manière adéquate, chacun de ces trois mouvements autorise une exploration exhaustive de l'espace des topologies.

NNI correspond à des modifications topologiques **locales** dans l'arbre. Chaque branche interne d'une phylogénie non-enracinée définit quatre sous-arbres. Dans l'exemple de la Figure 3.9, la branche interne considérée apparaît en trait épais et les quatre sous-arbres sont constitués par les UEs $\{a, b\}$, $\{c\}$, $\{d, e\}$ et $\{f, g\}$, numérotés 1, 2, 3 et 4, respectivement. Les trois configurations topologiques qu'il est possible d'engendrer à partir de ces quatre groupes sont : $\{1, 2\}/\{3, 4\}$ (à gauche sur la figure), $\{1, 3\}/\{2, 4\}$ (au centre), et $\{1, 4\}/\{2, 3\}$ (à droite).

SPR engendre des changements de topologies plus **globaux** que ceux définis pas NNI. En effet, la

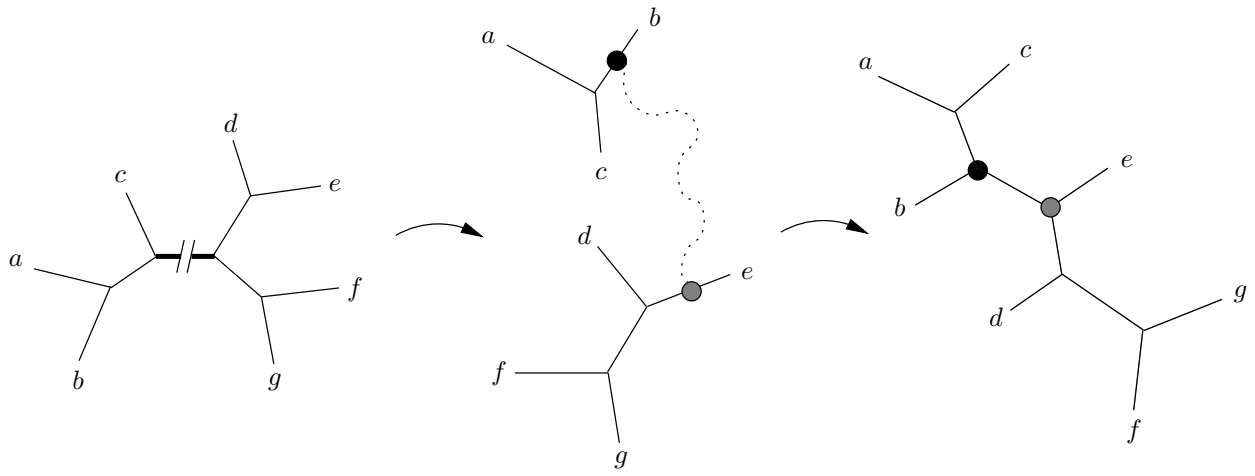


FIG. 3.11 – **TBR.** Après section (à gauche), un point de liaison est définit sur chaque sous-arbre (au centre). Ces deux points constituent les extrémités de la nouvelle branche joignant les deux sous-arbres (à droite).

plupart des modifications topologiques issues d'un SPR sont équivalents à une suite de plusieurs NNIs : le mouvement décrit dans la Figure 3.10 ne peut pas être accompli à l'issue d'un unique NNI (deux sont nécessaires). Dans une premier temps, la topologie est sectionnée au niveau d'une branche interne (à gauche sur la figure). Deux graphes sont définis à l'issue de cette étape. Le premier, constitué des UEs a, b et c dans notre exemple, est une topologie non-enracinée «donneuse», caractérisée par un noeud externe ne correspondant pas à une séquence contemporaine. La seconde est une topologie «receveuse» : chacune de ses branches est un point d'insertion potentiel pour la topologie donneuse (au centre). Dans la Figure 3.10, le point d'insertion est situé sur la branche externe menant à l'UE e (à droite).

TBR induit des mouvements plus globaux que ceux engendrés par SPR : la modification topologique présentée dans la Figure 3.11 ne peut pas être le fruit d'un unique SPR. De la même manière que précédemment, la topologie initiale est tout d'abord sectionnée au niveau d'une branche interne (à gauche sur la figure). En revanche, il n'y a plus de distinction de type donneur/receveur entre ces deux graphes. Une branche est identifiée sur chacune des deux topologies comme étant un point de liaison (au centre) et les deux topologies sont jointes par ces deux noeuds (à droite).

Signalons enfin que SPRs et TBRs peuvent être utilisés pour «recombiner» différentes topologies. Par exemple, certaines méthodes d'inférence phylogénétique procèdent à partir de plusieurs topologies en parallèle (Lewis, 1998; Lemmon et Milinkovitch, 2002). Des sous-arbres extraits de ces différentes phylogénies peuvent alors être agglomérés de manière à constituer une nouvelle phylogénie. Ce type d'opération est notamment utilisé au sein d'algorithmes génétiques, dont le principe général est décrit plus loin.

3.3.2 Les approches déterministes

Les procédures déterministes de réarangements topologiques correspondent à des suites de mouvements élémentaires. Ces suites sont plus ou moins longues suivant le mouvement considéré, et l'espace des topologies explorées est donc plus ou moins vaste. Ainsi, comme nous allons le montrer ci-dessous, les modifications topologiques engendrées par NNIs correspondent à une succession de modifications moins nombreuses que celles engendrées par SPR, qui, elles-mêmes, sont moins nombreuses que celles induites par TBR.

Chacune des $n-3$ branches internes de la phylogénie est soumise à un NNI. Lorsqu'un de ces mouvements conduit à une diminution de la fonction de coût, la topologie est modifiée en conséquence. Deux stratégies peuvent alors être adoptées. Soit l'algorithme est réinitialisé et les NNIs s'appliquent à l'ensemble des branches internes. Soit l'algorithme n'est pas réinitialisé et les NNIs s'appliquent sur les branches internes qui n'ont pas encore été analysées. Pour chacune des $n-3$ branches internes, deux nouvelles configurations sont testées. Le nombre total de topologies explorées lors d'un parcours en profondeur de la phylogénie est donc de l'ordre de n.

L'algorithme de réarangements topologique basé sur des SPRs est, lui aussi, très simple. Chacune des $n-3$ branches internes est sectionnée et définit deux sous-topologies. L'ensemble des points d'insertions sont ensuite testés, en considérant alternativement chacune de ces deux sous-topologies comme donneuse et receveuse. Le nombre de points d'insertions sur une topologie receveuse étant de l'ordre de n, le nombre total de topologies explorées lors d'un parcours de la phylogénie est donc de l'ordre de n^2.

L'algorithme d'exploration de l'espace des topologies par l'application de TBRs est similaire au précédent. Là encore, chacune des $n-3$ branches internes est sectionnée et les deux graphes engendrés correspondent exactement à deux topologies d'arbres. Les branches des deux topologies ainsi créées sont ensuite utilisées comme points de liaisons. Le nombre de ces branches est de l'ordre de n et le nombre de combinaisons testées est donc de l'ordre de n^2. Par conséquent, le nombre total de topologies explorées lors d'un parcours de la phylogénie est de l'ordre de n^3.

Comme nous l'avons vu précédemment, les deux principales stratégies pour l'obtention d'une première phylogénie sont l'insertion et l'agglomération d'UEs. Les réarangements topologiques décrits ci-dessus peuvent être combinés à ces procédures. Ainsi, l'algorithme fastDNAml (Olsen et al., 1994) couple les insertions d'UEs à des réarangements locaux de type NNIs. À l'issue d'une insertion, la suite de NNIs décrite précédemment s'applique à la phylogénie encore incomplète. Les réarangements globaux basés sur des SPRs ou des TBRs et bien plus coûteux en temps de calcul que le précédent, sont utilisés uniquement lorsque toutes les UEs sont insérées.

Combiner la construction d'arbres par agglomérations à des réarangements topologiques semble plus délicat. À notre connaissance, seuls les programmes MOLPHY (Adachi et Hasegawa, 1996) et PAML

(Yang, 1997c) implémentent une approche agglomérative pour l'inférence de phylogénies suivant le principe du maximum de vraisemblance. Or, aucun de ces deux programmes ne propose de coupler réarangements et agglomérations. Il est probable que cette stratégie n'offre pas d'amélioration significative des résultats comparé au cas où les réarangements sont appliqués sur l'arbre complet.

3.3.3 Les approches stochastiques

Considérons les valeurs du critère (vraisemblance ou taille de la phylogénie, par exemple) connues pour l'ensemble des phylogénies qu'il est possible de construire pour un nombre d'UEs fixé. À partir d'un quelconque point de départ, toutes les topologies explorées par la suite peuvent être identifiées par avance lorsque l'algorithme d'exploration des topologies est déterministe. Identifier *a priori* quelles sont les topologies visitées par un algorithme stochastique est impossible. Cette différence est la conséquence d'une appréciation divergente des phylogénies proposées à l'issue de perturbations topologiques. Pour un algorithme d'améliorations itératives, seules les modifications topologiques offrant une diminution de coût sont acceptées. En revanche, dans le cadre d'un algorithme stochastique, il est envisageable d'accepter un mouvement topologique conduisant à une augmentation du coût.

Recuit simulé

Le principe du recuit simulé a été introduit en 1953 par Metropolis et al.. Il a fallu ensuite attendre 30 ans (Kirkpatrick et al., 1983) pour une première application de cette technique à la résolution de problèmes combinatoire de grandes dimensions (le fameux problème du voyageur de commerce, par exemple). Dès lors, l'utilisation d'algorithmes basés sur le recuit simulé s'est largement répandue. Dans le domaine de la génomique, de nombreux logiciels utilisent cette approche pour l'ordonnancement de séquences d'ADN sur un génome ou l'inférence de la structure tridimensionnelle d'une protéine.

Le terme recuit simulé dérive de l'analogie avec le processus physique de chauffer puis refroidir lentement une substance de manière à obtenir une structure cristalline, d'énergie minimale. Le refroidissement s'effectue par paliers successifs : à chaque étape, la température doit être maintenue suffisamment longtemps pour que le système atteigne un état stable. Il s'agit de la phase de **thermalisation**. Lorsque la température est proche de zéro, le système se stabilise dans l'état d'énergie minimale.

L'algorithme de recuit simulé présente cinq étapes. (1) Une première solution, S_0, est générée aléatoirement. q_0 est la valeur initial du paramètre analogue à la température, définissant la probabilité d'acceptation d'une nouvelle solution. S_0 est ici la solution courante, notée S_c dans le cas général, et q_0 est la valeur courante du paramètre q, notée q_t dans le cas général. (2) Une nouvelle solution, est proposée à partir d'une modification de la solution à l'étape précédente. (3) Si le coût de la nouvelle solution, notée S_n, est inférieure au coût de la solution courante, cette dernière est remplacée par la nouvelle. Sinon, la

probabilité d'accepter la nouvelle solution est :

$$R = \exp(\frac{f(S_c) - f(S_n)}{q_t}),$$

Une itération de cette procédure de sélection d'un nouvel état à partir d'un critère stochastique est l'étape dite de **Metropolis**. L'étape de Metropolis est répétée tant qu'une condition d'arrêt n'est pas satisfaite. Cette condition correspond généralement à un nombre minimum de nouvelles solutions acceptées. La succession des étapes de Metropolis pour une valeur fixée de q_t est analogue au processus de thermalisation. (4) La valeur de q_t est mise à jour en utilisant, par exemple, l'expression :

$$q_{t+1} = \alpha q_t$$

où α est un réel dont la valeur est généralement comprise entre 0.80 et 0.99 (Aarts et Lenstra, 1997). (5) Les étapes (2)-(4) sont répétées jusqu'à atteindre le seuil pour la valeur de q_t, analogue à une température proche de zéro. Ce seuil est généralement déterminé en fonction du problème étudié. La solution la plus simple est de fixer une borne inférieure pour q_t en deçà de laquelle l'exécution de l'algorithme est suspendue. Cette valeur est déterminée de manière à ce que, pour des valeurs de q_t proches du seuil, aucune nouvelle solution n'est acceptée.

Salter et Pearl (2001) ont récemment appliqué le recuit simulé pour l'estimation de phylogénies de vraisemblances maximales. Contrairement à l'algorithme standard, décrit ci-dessus, q_t décroît systématiquement lorsqu'un nouvel état est proposé. Cette approche correspond donc à un cas particulier du recuit simulé. Les performances obtenues sont satisfaisantes, tant en termes de vitesses que de vraisemblances des phylogénies estimées. Ainsi, pour des phylogénies de vraisemblances égales, les temps de calculs sont divisés par un facteur proche de quatre comparé à l'approche d'améliorations itératives standard, basée sur des mouvements topologiques de type NNIs et implémentée au sein du logiciel PAUP* (Swofford, 1999). Cependant, la stratégie proposée par ces deux auteurs ne s'applique que lorsque l'hypothèse de l'horloge moléculaire est respectée.

Lorsque la décroissance de q est suffisamment lente, l'algorithme converge vers la solution optimale (Aarts et Lenstra, 1997). En pratique, l'ajustement des différents paramètres de réglage de l'algorithme (principalement la valeur initiale de q, et sa fonction de décroissance ainsi que le nombre d'étapes de Metropolis) correspond à un compromis entre le temps d'exécution et la fiabilité de la procédure d'optimisation. Cette contrainte offre cependant un choix intéressant (Salter et Pearl, 2001). Ainsi, au lieu de décider entre une exploration relativement rapide de l'espace des topologies par NNIs, et une exploration par SPRs ou TBRs, nettement plus coûteuse en temps de calculs (voire même souvent impossible), le recuit simulé offre la possibilité de régler précisément le temps à accorder à un problème donné. Cette propriété est particulièrement souhaitable dans le cadre de l'analyse de grands jeux de données, présentant plusieurs milliers de séquences.

Algorithmes génétiques

Cette stratégie d'optimisation a été introduite par Holland en 1975 (cité par Aarts et Lenstra, 1997). Son nom provient des analogies entre certains de ses mécanismes et différents concepts en génétique des populations et théorie de l'évolution. Longtemps utilisée pour la résolution de problèmes d'optimisation complexes dans divers domaines, elle n'a été appliquée à la biologie que récemment. Des thèmes de recherches standards, tels que la comparaison de structures de protéines, l'assemblage de segments d'ADN, ou l'inférence de structures secondaires d'ARN ont ainsi pu être abordés par des méthodes fondées sur ce type d'algorithmes.

Les algorithmes génétiques procèdent à partir d'un ensemble de solutions soumises successivement à trois processus. Les deux premiers sont la perturbation et la recombinaison de ces solutions de manière à en générer de nouvelles. Le troisième est la sélection des solutions de coûts minima. Lewis (1998) a adapté un tel algorithme pour l'inférence de phylogénies de vraisemblances maximales. La première étape est (1) la génération aléatoire d'un ensemble de q phylogénies distinctes. (2) Ces arbres sont ensuite «clonés», et le nombre de clones de chaque arbre est une fonction croissante de sa vraisemblance. (3) Les nouvelles phylogénies ainsi engendrées sont soumises à diverses perturbations topologiques et des valeurs des paramètres de nuisance. Seule la phylogénie de vraisemblance maximale est laissée intacte de manière à conserver le meilleur arbre trouvé jusqu'ici. Les modifications de topologies sont basées sur des mouvements standards de types NNIs, SPRs ou TBRs. Les arbres peuvent aussi être combinés par paires (Figure 3.12). Les modifications des longueurs de branches sont locales. Ainsi, une proportion aléatoire de celles-ci sont multipliées par un facteur distribué selon une loi gamma d'espérance 1.0. (4) q phylogénies sont choisies aléatoirement parmi les clones modifiés. Les étapes (2)-(4) sont itérées jusqu'à ce qu'une condition d'arrêt soit satisfaite. Par exemple, le processus prend fin lorsque la vraisemblance maximale observée parmi les q valeurs de vraisemblance reste inchangée pendant un nombre prédéfini d'itérations.

Récemment, Lemmon et Milinkovitch (2002) ont proposé un algorithme génétique différent de celui de Lewis (1998). La principale originalité est ici l'utilisation en parallèle de plusieurs ensembles, ou populations, de phylogénies. Les perturbations de chacun des arbres sont alors guidées par des comparaisons avec les phylogénies de vraisemblances maximales dans chacune de ces populations. De telles comparaisons visent à identifier les régions fiables de la topologie de l'arbre, qui ne doivent donc pas être remises en cause.

Cette méthode est relativement rapide : elle autorise l'analyse d'une centaine de séquences homologues sur des durées raisonnables. De plus, il existe une forte corrélation entre la fréquence des branches internes au sein des phylogénies les plus vraisemblables de chaque population, et les estimations de probabilités *a posteriori* des clades correspondants. Il est donc possible d'avoir une idée assez précise du support des données pour chaque branche de la phylogénie obtenue.

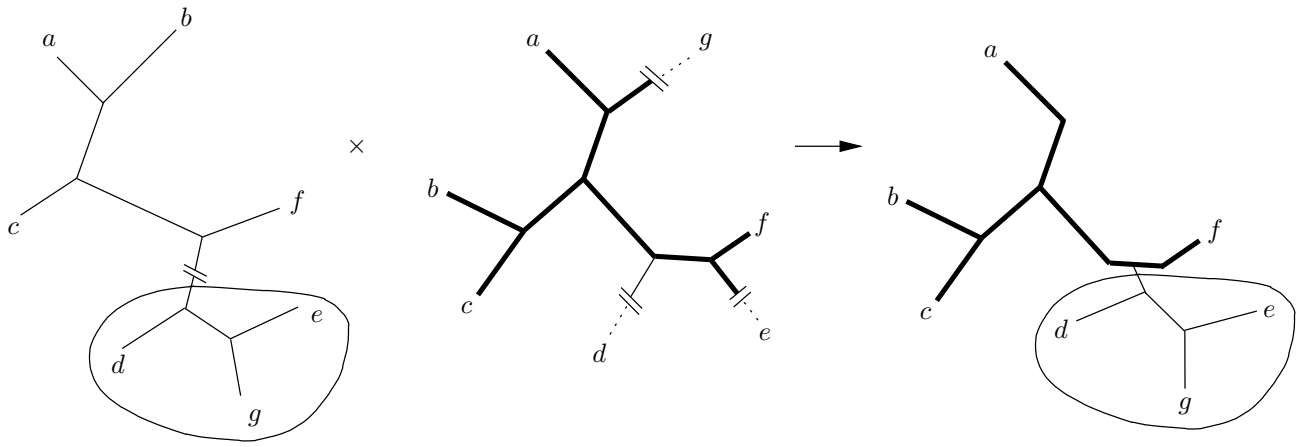


FIG. 3.12 – **Recombinaison de deux phylogénies.** Les deux topologies à recombiner sont présentées à gauche et au centre de la figure. Le sous-arbre présentant les UEs d, e et g (à gauche) est séparé du reste de cette phylogénie. Ces UEs sont aussi retirées du second arbre. Le sous-arbre est ensuite inséré sur une branche quelconque de cette seconde phylogénie.

Malheureusement, Lemmon et Milinkovitch (2002) n'ont pas testé les performances de leur méthode du point de vue de la fiabilité des topologies inférées. Ceci manque d'autant plus que, lorsque la condition d'arrêt de l'algorithme est satisfaite, longueurs de branches et paramètres libres du modèle sont alors ajustés par des méthodes d'optimisation numérique standard. Par conséquent, le processus de modifications stochastiques des valeurs paramètres de nuisance couplée à la sélection des phylogénies les plus vraisemblables ne permet pas d'obtenir des estimations précises de ces paramètres. Il se peut qu'une tendance similaire s'observe pour la topologie de l'arbre (voir Guindon et Gascuel, 2003 pour un exemple)

D'une manière plus générale, le point fort des méthodes d'optimisation stochastique, résidant dans la possibilité d'étudier les comportements asymptotiques des algorithmes, est difficilement exploitable ici. Déterminer, par exemple, l'équilibre optimal entre les différents opérateurs de perturbations des solutions courantes n'est pas aisé. Il en résulte une grande variété d'implémentations (voir Whitley, 1994 pour une revue), chacune de celles-ci correspondant à différentes manières d'évaluer les solutions courantes, de modifier et recombiner celles-ci. Cependant, ce constat illustre aussi le point fort de ces méthodes, résidant dans leur flexibilité. Ainsi, cette approche est généralement efficace pour l'optimisation combinatoire et peut être facilement adaptée à la résolution de problèmes variés. Son application à la phylogénie moléculaire est donc une avancée indéniable, même s'il est probable que les algorithmes proposés jusqu'à aujourd'hui puissent être améliorés de manière significative.

Monte Carlo par chaînes de Markov

L'objectif n'est plus ici de minimiser une fonction de coût, mais plutôt d'échantillonner les solutions de manière à approximer leur distribution de probabilités *a posteriori*. La solution choisie au final est généralement celle qui est le plus fréquemment visitée durant cet échantillonnage. Comme précédemment,

nous présentons tout d'abord brièvement l'algorithme dans le cas général puis discutons de son application à l'inférence phylogénétique.

Le principe de cette approche est simple. L'idée est de construire une chaîne de Markov dont la fréquence stationnaire est égale à la distribution de probabilité des solutions conditionnellement aux données, ou probabilités *a posteriori*. Cette construction s'effectue de la façon suivante : (1) un premier état de la chaîne, noté S_0, correspondant à une des solutions, est généré aléatoirement. S_0 est alors la solution courante. Elle est notée S_c dans le cas général. (2) Un nouvel état de la chaîne, noté S_n est ensuite proposé à partir de S_c. La probabilité *a priori* de transition de S_c vers S_n est $P(S_c \rightarrow S_n)$, et la probabilité d'effectuer le mouvement inverse est $P(S_n \rightarrow S_c)$. (3) La probabilité d'accepter la nouvelle solution est :

$$\begin{aligned} R &= \min(1, \frac{P(S_n|D)}{P(S_c|D)} \times \frac{P(S_n \rightarrow S_c)}{P(S_c \rightarrow S_n)}) \\ &= \min(1, \frac{P(D|S_n)P(S_n)/P(D)}{P(D|S_c)P(S_c)/P(D)} \times \frac{P(S_n \rightarrow S_c)}{P(S_c \rightarrow S_n)}) \\ &= \min(1, \frac{P(D|S_n)}{P(D|S_c)} \times \frac{P(S_n)}{P(S_c)} \times \frac{P(S_n \rightarrow S_c)}{P(S_c \rightarrow S_n)}) \end{aligned}$$

Ces trois expressions équivalentes montrent que, malgré l'impossibilité de calculer indépendamment les probabilités *a posteriori* $P(S_c|D)$ ou $P(S_n|D)$ (car $P(D)$ est généralement incalculable, voir équation 3.13), le rapport de ces deux termes se ramène à une quantité mesurable. Cette quantité est le produit du rapport des vraisemblances de S_c et S_n par le rapport des probabilités *a priori* de ces deux états. Le rapport $P(S_n \rightarrow S_c)/P(S_c \rightarrow S_n)$ est le **ratio de Hastings**. (4) Les étapes (1)-(3) sont répétées un grand nombre de fois et la fréquence de visites d'une solution quelconque est une bonne approximation de sa probabilité *a posteriori*.

Dans le cadre de l'inférence phylogénétique, les états de la chaîne de Markov correspondent à des phylogénies (Huelsenbeck et al., 2001; Larget et Simon, 1999), ou à des histoires labellées (Yang et Rannala, 1997) sur les UEs considérées. Bien que ces deux approches diffèrent sur la nature des solutions explorées, l'objectif principal reste le même : estimer les probabilités *a posteriori* des différentes topologies d'arbres.

Le point le plus délicat concernant l'efficacité de l'échantillonnage réside essentiellement dans l'ajustement des mécanismes de transitions entre états de la chaîne de Markov. Deux excès sont à éviter. Le premier correspond au cas où les différences entre deux états successivement explorés sont trop faibles. La conséquence est une corrélation entre la valeur de l'état et le moment où il est atteint. Le second excès correspond à la situation inverse : les états successivement échantillonnés sont très différents et ceux présentant des vraisemblances élevées sont alors rarement visités.

Les transitions entre phylogénies sont généralement effectuées à partir d'un processus stochastiques entraînant une modification des longueurs de branches et, fréquemment, une perturbation topologique de

type NNI. À l'instar de la vraisemblance, la surface de la fonction de densité de probabilité *a posteriori* des phylogénies est certainement «accidentée» et présente probablement plusieurs pics (Huelsenbeck et Ronquist, 2001). Or, suivant le paramétrage de l'algorithme, des mouvements de types NNIs successifs peuvent être efficaces pour l'échantillonnage au sein d'un pic, mais moins adaptés une exploration de pics en pics, ou l'inverse. Une solution à ce problème consiste à construire plusieurs chaînes de Markov en plus de la chaîne originale, chaîne pour laquelle une nouvelle phylogénie est acceptée si sa vraisemblance est relativement proche de celle la phylogénie courante (Huelsenbeck et al., 2001). Les ratios de Hastings dans les chaînes supplémentaires sont supérieurs à celui de la chaîne originale et permettent ainsi d'accepter plus fréquemment des mouvements entre phylogénies relativement dissemblables. Les différentes chaînes sont construites en parallèle et périodiquement combinées. Cette stratégie empirique permet d'explorer plus efficacement la surface de la fonction de densité, mais au prix d'une multiplication linéaire des temps de calculs, proportionnelle au nombre de chaînes construites (Huelsenbeck et Ronquist, 2001).

Bien que la fiabilité des topologies inférées par une approche bayesienne n'ait pas été testée par simulations sur un grand nombre de jeux de données, les performances de la méthode proposée par Huelsenbeck, Ronquist, Nielsen et Bollack (2001) semblent satisfaisantes sur ce point (Guindon et Gascuel, 2003). De plus, la fréquence d'échantillonnage d'un clade donné est une estimation de sa probabilité *a posteriori*. Les méthodes de Monte Carlo par chaînes de Markov permettent donc d'évaluer le support des données pour chaque branches interne d'une phylogénie. Un autre avantage important fourni par ce type d'algorithme est la possibilité d'intégrer sur les valeurs des paramètres de nuisance. La topologie est ainsi inférée en prenant en compte les incertitudes liées aux estimations des paramètres libres du modèle de substitution et des longueurs de branches.

3.4 Conclusions

Ce chapitre aborde les principaux aspects méthodologiques pour l'estimation de topologies d'arbres, des longueurs de branches et des paramètres libres du modèle de substitution. Cette estimation repose sur la maximisation (ou la minimisation) d'un critère constituant véritablement le cœur des différentes méthodes. Les critères utilisés actuellement correspondent à des canons de la statistique (par exemple, le maximum de vraisemblance, la probabilité *a posteriori*, les moindres carrés). Le minimum d'évolution est beaucoup plus récent. Ce critère est d'«inspiration biologique» mais pas toujours mathématiquement fondé car il s'applique dans des conditions bien particulières (estimation des longueurs de branches aux moindres carrés) et ne se justifie pas dans le cadre général.

Une propriété importante des méthodes de distances est la bijection entre l'ensemble des phylogénies et l'ensemble des distances d'arbres (Barthélemy et Guénoche, 1988) (à une matrice de distances d'arbres correspond une unique phylogénie). Cependant, il n'existe pas de bijection entre l'ensemble des distances estimées et l'ensemble des jeux de séquences homologues (à différents jeux de séquences peuvent corres-

pondre une seule matrice de distances estimées). Par conséquent, une perte d'information accompagne nécessairement le calcul des distances évolutives, et la reconstruction de phylogénies à partir de celles-ci n'est pas optimale de ce point de vue. Néanmoins, l'information contenue au sein de distances exactes est suffisante pour construire une phylogénie sans erreurs. Par conséquent, cette critique (Page et Holmes, 1998) admet certaines limites.

L'approche du maximum de vraisemblance est une méthode statistique très répandue pour l'estimation de paramètres numériques. Lorsque le modèle utilisé décrit exactement les données, les estimateurs sont asymptotiquement sans biais et de variance minimale. Cependant, l'utilisation de ce principe en phylogénie pose certains problèmes. Le principal concerne l'estimation de la topologie de l'arbre : aucun paramètre se rattachant à la topologie n'apparaît explicitement dans l'expression de la vraisemblance d'une phylogénie. La topologie de l'arbre n'est donc pas estimée au sens du maximum de vraisemblance comme peuvent l'être les paramètres de nuisance.

Néanmoins, l'estimation de phylogénies suivant ce critère statistique rencontre un large succès. Celui-ci est sans nul doute lié aux bonnes performances de l'approche du maximum de vraisemblance sur le plan de la fiabilité des topologies estimées. De nombreuses études comparatives basées sur des simulations (Huelsenbeck et Hillis, 1993; Kuhner et Felsenstein, 1994; Huelsenbeck, 1995; Rosenberg et Kumar, 2001; Ranwez et Gascuel, 2002) indiquent la supériorité du maximum de vraisemblance comparée aux méthodes de distances ou parcimonie. De plus, Yang (1994b) propose une preuve pour la consistance de cette approche. Par conséquent, les temps de calculs importants constituent, à l'heure actuelle, le principal point faible des méthodes d'inférence d'arbre basées sur le principe du maximum de vraisemblance. Les solutions récentes apportées à ce problème reposent sur l'exploration de l'espace des topologies d'arbres par des méthodes d'optimisation stochastique. Les approches basées sur des algorithmes génétiques et le recuit simulé offrent ainsi la possibilité d'analyser des jeux de données présentant plusieurs centaines de séquences. Une nouvelle méthode d'inférence basée sur un algorithme d'améliorations itératives de la vraisemblance et offrant des performances supérieures à ces dernières, est présentée au chapitre 5.

Enfin, l'inférence bayesienne est apparue récemment dans le domaine de la phylogénie moléculaire. Telle qu'elle est appliquée à l'heure actuelle, cette méthode permet de construire relativement rapidement des intervalles de confiance autour des valeurs des paramètres de nuisance et estimer les probabilités *a posteriori* de chacun des sous-arbres. Cependant, cette avancée ne doit pas tant à la théorie statistique composant l'approche bayesienne qu'à l'utilisation de chaînes de Markov par Monte Carlo. C'est en effet grâce à cette stratégie algorithmique qu'il est possible d'intégrer sur les valeurs des paramètres de nuisance et obtenir ainsi une mesure de la fiabilité d'une topologie indépendante de tout autre paramètre. À nos yeux, ce dernier point est nettement plus important que l'introduction de probabilités *a priori* pour l'inférence phylogénétique.

Chapitre 4

Paramètres de nuisance et inférence de topologies d'arbres

Comme nous l'avons indiqué précédemment (chapitres 2 et 3), il est souvent nécessaire d'estimer les longueurs de branches et les valeurs des paramètres libres du modèle de substitution pour inférer une topologie d'arbre. Généralement, les estimateurs de ces paramètres visent à approximer le plus précisément possible les valeurs sous-jacentes au processus de génération des données. Cependant, lorsque le but final est d'inférer une topologie d'arbre, il peut être souhaitable d'utiliser des estimateurs explicitement adaptés à cette tâche.

Nous nous sommes intéressés à ce problème dans le cadre de l'estimation du paramètre de forme de la loi gamma, mesurant la variabilité des vitesses d'évolution entre sites, avec pour objectif l'inférence de phylogénies à partir de distances évolutives. Dans la première partie de ce chapitre, nous présentons les principales méthodes d'estimation de ce paramètre. Puis, à partir de simulations, nous montrons que la vraie valeur de ce dernier n'est généralement pas la valeur la plus adaptée à l'inférence de topologies d'arbres. Une nouvelle méthode d'estimation est alors présentée. Cette approche est basée sur l'optimisation du paramètre de forme de la loi gamma de manière à minimiser un critère d'ajustement entre les distances estimées et la phylogénie inférée à partir de ces dernières. De nouveaux résultats de simulations montrent que la fiabilité des arbres construits à partir de cette approche est supérieure à celle qu'il serait possible d'obtenir si la valeur sous-jacente au processus d'évolution était connue. Ces travaux ont été publiés en 2002 dans la revue «Molecular Biology and Evolution» (voir Annexe A).

4.1 Estimation du paramètre de forme de la loi gamma

Deux grands types d'approches se distinguent. Les premières procèdent par l'ajustement du paramètre de manière à maximiser la vraisemblance d'une phylogénie dont la topologie et les longueurs de branches sont fixées (Yang, 1994; Gu et al., 1995). Les méthodes de section du nombre d'or ou celle de Brent peuvent être appliquées ici (voir chapitre 3). Il est aussi possible d'utiliser des algorithmes d'optimisation numérique basés sur les dérivées du logarithme de la vraisemblance (Yang, 1997c, par exemple). Contrairement aux

longueurs de branches, la valeur du paramètre de forme de la loi gamma n'apparaît pas explicitement dans l'expression de la vraisemblance (voir équation 2.18). Par conséquent, il n'est pas possible d'exhiber les expressions analytiques des dérivées dans ce cas. Celles-ci doivent alors être approximées au prix de calculs relativement lourds. Ceci rend sans doute préférable les approches ne nécessitant pas de connaître les valeurs des dérivées, comme celle de Brent ou de la section d'or.

Si le taux de substitution instantané est constant dans l'arbre (hypothèse d'homogénéité), mais varie entre sites selon une loi gamma de paramètre de forme α et d'espérance égale à μ, alors le nombre de substitutions dans l'arbre et par site, noté N_s, suit une loi binomiale négative d'espérance égale à μ et de variance $\mu(1+\mu/\alpha)$. Bon nombre de méthodes d'estimation de α s'inspirent de cette observation (Sullivan et al., 1995; Tourasse et Gouy, 1997; Gu et Zhang, 1997). Les approches développées comprennent deux étapes : (1) l'estimation du nombre de substitutions dans l'arbre pour chaque site, et (2) l'optimisation de α de manière à ajuster au mieux le modèle (la distribution binomiale négative) aux données (la distribution des N_s). Pour l'étape (1), une phylogénie est inférée par maximum de parcimonie ou à partir de distances évolutives, et les états des caractères aux noeuds internes de cet arbre sont estimés par une approche bayesienne (Gu et Zhang, 1997) ou par maximum de parcimonie (Sullivan et al., 1995; Tourasse et Gouy, 1997). Les valeurs de N_s sont ensuite déduites à partir de la comparaison des états aux deux extrémités de chaque branche de l'arbre. Appliquer ici le principe du maximum de parcimonie revient à considérer qu'il ne se produit au plus qu'une substitution par branche. Or, cette hypothèse peut conduire à sous-estimer le nombre de substitutions par site. Afin de corriger ce biais, Gu et Zhang (1997) proposent d'estimer la distribution du nombre de substitutions par branche et par site au sens du maximum de vraisemblance. Plusieurs substitutions peuvent alors survenir sur une branche. Enfin, lorsque la distribution des valeurs de N_s est connue, la valeur de α peut être déterminée par la méthode des moments, mais c'est généralement au sens du maximum de vraisemblance qu'est réalisée l'estimation.

D'une manière générale, la plupart de ces méthodes permettent d'estimer très précisément la valeur du paramètre de forme de la loi gamma (voir Tableaux 3-5, Gu et Zhang, 1997). Cependant, la sous-estimation du nombre de substitutions par site entraîne fréquemment une sur-estimation du paramètre, c'est à dire la sous-estimation de l'intensité de la variabilité entre sites. Ainsi, lorsque les divergences entre séquences sont fortes, les performances du maximum de parcimonie sont relativement faibles (Gu et Zhang, 1997). L'approche du maximum de vraisemblance (au sens de Yang, 1994a) est la plus précise et semble peu sensible à la topologie de la phylogénie utilisée, à condition que cette dernière soit relativement proche de celle de l'arbre vrai (Yang, 1996).

Comme nous l'avons souligné précédemment, l'objectif de ces différentes approches est d'estimer une valeur du paramètre la plus proche possible de celle sous-jacente au jeu de données. Dans ce qui suit, nous analysons la pertinence de ces estimations du point de vue de la fiabilité des topologies inférées. Pour ce faire, nous nous restreignons au cas où la valeur estimée du paramètre de forme de la loi gamma

est utilisée pour le calcul de distances évolutives, à partir desquelles est déduite une phylogénie.

4.2 Résultats préliminaires

Des simulations (voir Annexe A) ont été entreprises afin d'analyser l'évolution de la distance topologique moyenne entre la vraie phylogénie et l'arbre inféré, construit à partir de distances estimées pour différentes valeurs du paramètre de forme de la loi gamma. La vraie valeur de ce dernier, c'est à dire la valeur à partir de laquelle sont générées les données, est notée a. α est l'estimation utilisée pour le calcul des distances évolutives. $\widehat{\mathcal{T}^\alpha}$ est la topologie de l'arbre déduite de ces distances. Celle-ci est comparée à $\mathcal{T}_v$, la topologie de l'arbre vrai. Nous utilisons ici la distance de Robinson et Foulds (1979), notée RF, définie comme la proportion de branches internes (ou bipartitions) présentes au sein d'un arbre et absente dans l'autre. Ses valeurs sont comprises entre 0.0 (les deux topologies sont identiques) et 1.0 (les deux topologies ne partagent aucune branche en commun).

La Figure 4.1 montre l'évolution de $\overline{RF}(\mathcal{T}_v, \widehat{\mathcal{T}^\alpha})$ en fonction de α, pour des écarts à l'horloge moléculaire dans l'arbre vrai fort ou modéré et deux intensités de la variabilité des vitesses d'évolution entre sites au sein des séquences générées ($a = 0.1$ et $a = 0.7$ correspondent à une variabilité forte et modérée respectivement).

Le point le plus important mis en évidence par ces courbes est le décalage systématique entre la vraie valeur du paramètre, a, et la valeur optimale, notée α^{opt}, minimisant les écarts moyens entre la topologie de l'arbre vrai et celle de l'arbre estimé. α^{opt} est supérieur à a, signifiant que les topologies d'arbres les plus fiables sont obtenues en sous-estimant l'hétérogénéité des vitesses entre sites et les distances estimées par la même occasion. L'écart entre a et α^{opt} croît lorsque l'écart à l'horloge diminue. Lorsque l'écart à l'horloge est nul (résultats non présentés), les courbes obtenues sont monotones et décroissantes ($\alpha^{opt} = +\infty$). Remarquons aussi l'aplanissement autour de α^{opt} lorsque la variabilité des vitesses entre sites diminue. Pour $a = 0.1$, les minima sont situés dans des «puits» étroits, tandis que pour $a = 0.7$, ces puits sont nettement plus larges. Cette dernière observation s'explique par les propriétés de la loi gamma. Pour de faibles valeurs du paramètre de forme (par exemple $a = 0.1$), les variations de α autour de a induisent une grande variabilité des distances estimées, et, par conséquent, de fortes perturbations de la topologie inférée.

L'effet de l'écart à l'horloge moléculaire s'explique à partir d'arguments inspirés de ceux proposés par Steel et Penny (2000). Soit d_{xy} la distance entre x et y. Si l'on suppose que le mécanisme de substitution respecte les hypothèses décrites au chapitre 2 (voir partie «Hypothèses et outils mathématiques»), et si celui-ci est couplé à une distribution des vitesses d'évolution entre sites décrite par une loi gamma de paramètre a, alors $d_{xy} = f(\mathbf{p}_{xy}, a)$, où $\mathbf{p}_{xy}$ est le vecteur regroupant les fréquences des différents types de différences entre x et y, en fonction du modèle choisi. Par exemple, pour le modèle K2P, $\mathbf{p}_{xy}$

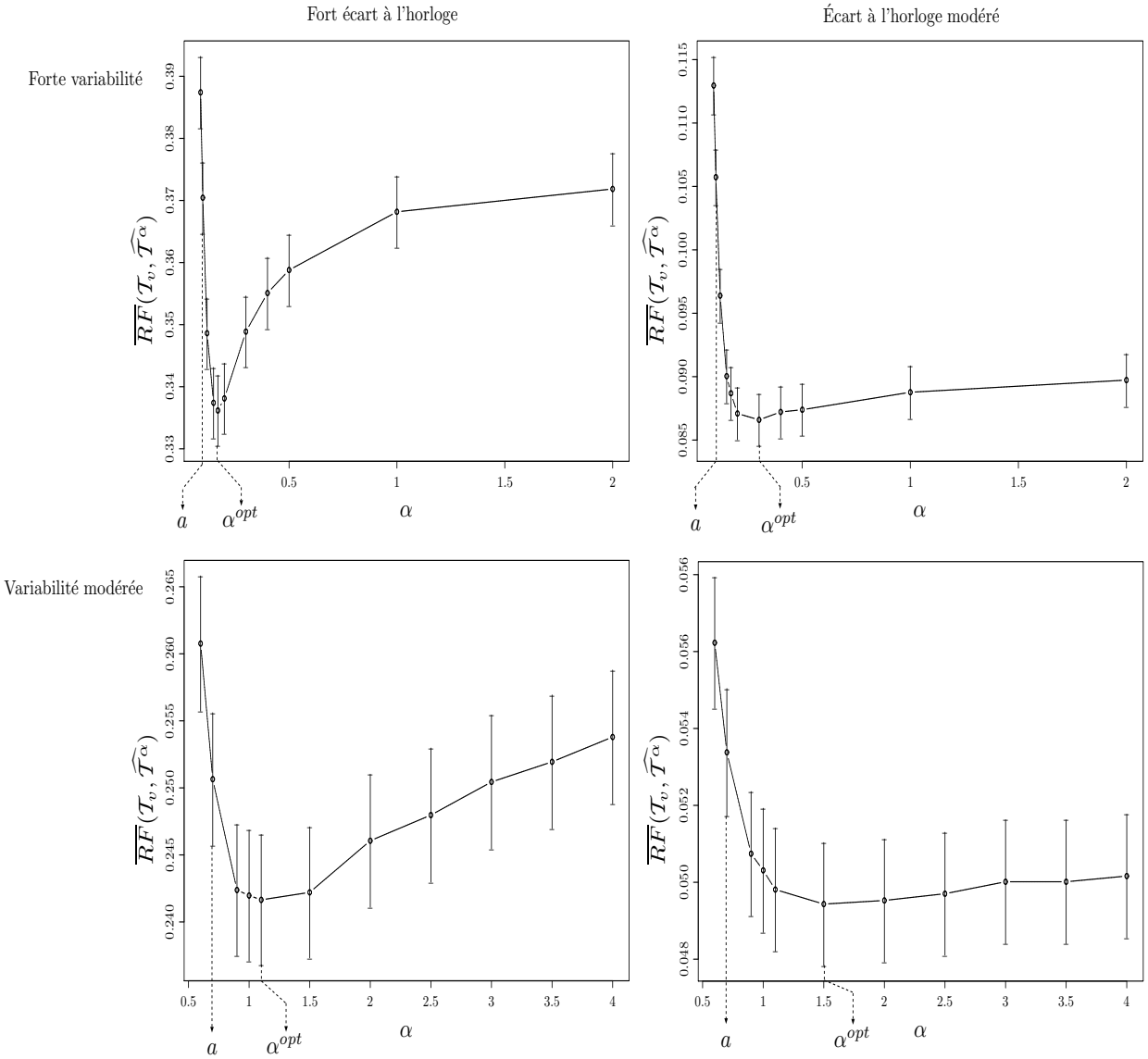


FIG. 4.1 – **Évolution de $\overline{RF}(\mathcal{T}_v, \widehat{\mathcal{T}^\alpha})$ en fonction de α.** Cette figure présente l'évolution de la distance topologique moyenne entre l'arbre vrai et l'arbre inféré par BIONJ à partir de distances évolutives corrigées par α.

présente les fréquences de transitions et transversions observées. Pour une valeur de $\mathbf{p}_{xy}$ fixée, f est une fonction monotone décroissante de a (voir équation 2.17). Lorsque l'hypothèse de l'horloge moléculaire est vérifiée et que les séquences comparées sont infiniment longues, les distances entre séquences sont dites **ultramétriques** et respectent l'inégalité $d_{xy} \leq \max(d_{xz}, d_{yz})$, pour tout x, y et z (Barthélemy et Guénoche, 1991, p.92). Or, cette inégalité reste valide quelle que soit la fonction monotone décroissante de a. On a donc $f(\mathbf{p}_{xy}, \alpha) \leq \max(f(\mathbf{p}_{xz}, \alpha), f(\mathbf{p}_{yz}, \alpha))$ pour tout x, y et z et toute valeur de α. Ceci montre que, lorsque l'horloge moléculaire est respectée et pour des séquences suffisamment longues, la topologie de l'arbre peut être inférée correctement à partir de valeurs de α différentes de celles de a.

Cependant, un tel argument n'est pas complet. Il explique pourquoi une différence entre α et a n'est

pas nécessairement une source d'erreurs dans l'estimation de la topologie, mais il ne justifie pas le fait que α^{opt} soit systématiquement supérieur à a. Ce dernier point est plus difficile à démontrer formellement. Néanmoins, les expressions analytiques des variances des distances estimées montrent que celles-ci sont des fonctions décroissantes exponentiellement de a (voir Tamura et Nei, 1993, par exemple). Par conséquent, les distances estimées à partir de a ont des variances bien supérieures aux distances estimées à partir de $\alpha^{opt} > a$. Ces dernières favorisent donc une diminution des erreurs dans l'estimation de la topologie.

Le décalage observé entre a et α^{opt} indique la nécessité de mettre au point de nouvelles procédures d'estimation du paramètre de forme de la loi gamma. Contrairement au maximum de vraisemblance, les méthodes de distances se prêtent mal à l'évaluation et la maximisation de critères approximant la fiabilité des topologies inférées par l'ajustement des valeurs des paramètres libres du modèle de substitution. Par exemple, faire varier la valeur du ratio transition/transversion se traduit par des variations de la vraisemblance et donc de la confiance accordée à la phylogénie. Transposer cette démarche dans le cadre des méthodes de distances n'est pas direct. Pour cela, il est nécessaire de définir un critère (équivalent à la vraisemblance) permettant de comparer équitablement la fiabilité approximée de phylogénies obtenues à partir de différentes valeurs du paramètre de substitution étudié. Dans ce qui suit, nous décrivons ici un critère permettant d'approximer α^{opt}.

4.3 Approximation de α^{opt}

L'idée est ici de mesurer l'ajustement entre les distances estimées pour une valeur de α, et la topologie de l'arbre déduite de ces distances. Soit $\mathcal{Q}(\Delta^\alpha, \mathcal{T}^\alpha)$, le critère mesurant l'ajustement entre Δ^α, la matrice des distances estimées en utilisant α comme valeur du paramètre de forme de la loi gamma, et $\mathcal{T}^\alpha$, la topologie de la phylogénie estimée à partir de Δ^α. La valeur **efficace** de α, notée α^*, est définie de la façon suivante :

$$\alpha^* = \underset{\alpha \in \mathbb{R}^+}{\operatorname{argmin}} \left(\mathcal{Q}(\Delta^\alpha, \mathcal{T}^\alpha)\right)$$

L'ajustement optimal, au sens du critère proposé, des distances estimées à une distance d'arbre correspond donc à la valeur de α minimisant $\mathcal{Q}$. α^* est censé approximer la valeur optimale de α.

Nous présentons tout d'abord le calcul de la valeur de $\mathcal{Q}$ sur des arbres à quatre UEs, puis sur des phylogénies présentant un nombre d'UEs quelconque. L'efficacité moyenne de ce critère pour l'approximation de α^{opt} est ensuite testée à partir de simulations similaires aux précédentes.

4.3.1 Définition du critère

Soient quatre UEs a, b, c et d et les six distances estimées Δ^α_{ab}, Δ^α_{ac}, Δ^α_{ad}, Δ^α_{bc}, Δ^α_{bd} et Δ^α_{cd} telles que $(\Delta^\alpha_{ab} + \Delta^\alpha_{cd}) \leq (\Delta^\alpha_{ac} + \Delta^\alpha_{bd}) \leq (\Delta^\alpha_{ad} + \Delta^\alpha_{bc})$. Pour une notation plus concise, on écrit $S \leq M \leq L$. Étant donnée cette inégalité, la topologie de l'arbre estimé par la plupart des méthodes de distances est

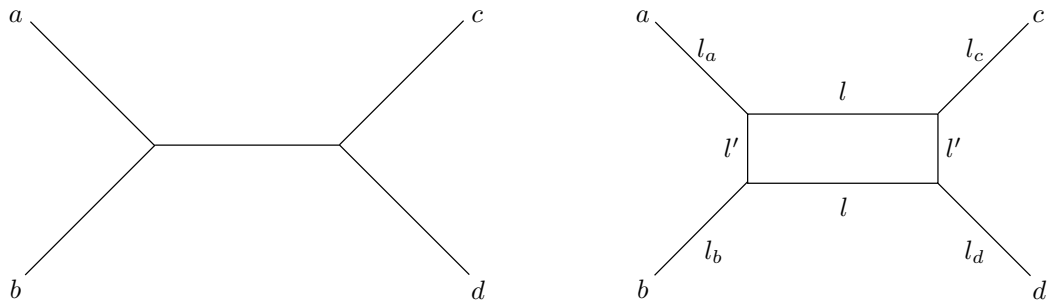


FIG. 4.2 – **Représentation arborée (a) et exacte (b) de distances estimées** L'arbre présenté à gauche est une simple phylogénie non enracinée à quatre UEs. Le graphe situé à droite autorise une représentation exacte des six distances entre les UEs a, b, c et d

$\{a, b\}/\{c, d\}$ (Figure 4.2a)

En pratique, l'ajustement de ces six distances estimées à une distance d'arbre est rarement parfait. Il est néanmoins possible de fournir une représentation géométrique exacte de ces six distances. Le graphe de la Figure 4.2b est un support adéquat pour ceci (Bandelt et Dress, 1992a). Chacune des distances entre UEs s'exprime comme une somme des longueurs d'arêtes dans ce graphe. Ainsi, les six distances estimées sont des combinaisons linéaires des six longueurs d'arêtes et ce système d'équations peut être résolu de manière à obtenir les expressions des longueurs d'arêtes en fonction des distances estimées. En particulier, on a $l = (L - S)/2$ et $l' = (L - M)/2$. Lorsque des distances estimées s'ajustent parfaitement à l'arbre $\{a, b\}/\{c, d\}$, $L = M$ et $l' = 0$. Cette situation est très improbable en pratique car les distances estimées ne correspondent généralement pas exactement à des distances d'arbres. Si l' est proche de l, il est difficile de choisir entre les topologies $\{a, b\}/\{c, d\}$ et $\{a, c\}/\{b, d\}$. En revanche, si l' est petit comparé à l le support pour la topologie $\{a, b\}/\{c, d\}$ est nettement supérieur à celui accordé à $\{a, c\}/\{b, d\}$ ou $\{a, d\}/\{b, c\}$.

La définition formelle de $\mathcal{Q}$ est :

$$\mathcal{Q} = L - M = 2l' \tag{4.1}$$

Ainsi, pour l'exemple de la Figure 4.2, $\mathcal{Q}$ est une mesure de la fiabilité de la branche séparant a et b de c et d. Ce critère évalue aussi l'ajustement des distances estimées à une distance d'arbre : plus sa valeur est grande, plus les distances entre séquences diffèrent de distances d'arbre.

Lorsque le nombre d'UEs dépasse quatre, a, b, c et d sont les racines de quatre sous-arbres A, B, C et D, présentant respectivement $|A|$, $|B|$, $|C|$ et $|D|$ UEs. La distance entre A et B est alors donnée par :

$$\overline{\Delta^{\alpha}_{AB}} = \frac{\sum_{a,b} \Delta^{\alpha}_{ab}}{|A||B|}$$

où Δ^{α}_{ab} est la distance entre les séquences correspondant aux UEs a, appartenant au sous-arbre A, et b, appartenant au sous-arbre B. S, M et L correspondent alors à $(\overline{\Delta^{\alpha}_{AB}} + \overline{\Delta^{\alpha}_{CD}})$, $(\overline{\Delta^{\alpha}_{AC}} + \overline{\Delta^{\alpha}_{BD}})$ et $(\overline{\Delta^{\alpha}_{AD}} + \overline{\Delta^{\alpha}_{BC}})$ respectivement. Lorsque la topologie estimée est $\{A, B\}/\{C, D\}$, S est la plus petite des

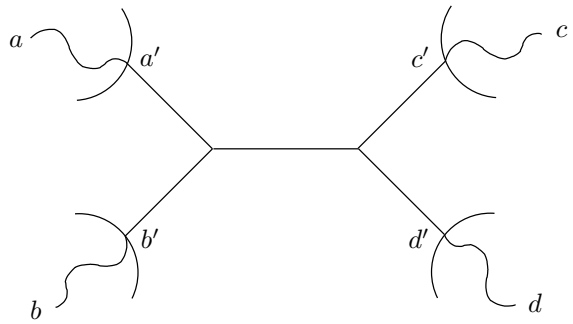


FIG. 4.3 – **Arbre modèle.** a, b, c et d désignent des noeuds externes, tandis que a', b', c' et d' correspondent à des noeuds internes, racines des sous-arbres A, B, C et D, présentant respectivement $|A|$, $|B|$, $|C|$ et $|D|$ UEs.

trois sommes dans la majorité des cas ($\simeq$99% des cas lorsque la phylogénie est estimée par BIONJ à partir de données simulées).

Les valeurs de $\mathcal{Q}$ sont calculées pour chaque branche interne de la phylogénie estimée et le score associé à un arbre est la moyenne de celles-ci. Lorsque les distances estimées sont des distances d'arbres, on démontre facilement que $\mathcal{Q}$ s'annule. On a, en effet :

$$\mathcal{Q} = \frac{1}{|A||D|}\sum_{a,d}\Delta^{\alpha}_{ad} + \frac{1}{|B||C|}\sum_{b,c}\Delta^{\alpha}_{bc} - \frac{1}{|A||C|}\sum_{a,c}\Delta^{\alpha}_{ac} - \frac{1}{|B||D|}\sum_{b,d}\Delta^{\alpha}_{bd}$$

or, chaque distance estimée peut être exprimée de manière à faire apparaître des distances entre noeuds dans l'arbre, noté T ici. Par exemple, on écrit $\Delta^{\alpha}_{ad} = \Delta^{T}_{a'd'} + \Delta^{T}_{aa'} + \Delta^{T}_{dd'}$, où $\Delta^{T}_{a'd'}$, $\Delta^{T}_{aa'}$ et $\Delta^{T}_{dd'}$ sont les longueurs des chemins entre les noeuds a' et d', a et a' et d' et d, respectivement (Figure 4.3). On obtient ainsi :

$$\begin{aligned}
&= \frac{1}{|A||D|}\sum_{a,d}(\Delta^{T}_{a'd'} + \Delta^{T}_{a'a} + \Delta^{T}_{d'd}) + \frac{1}{|B||C|}\sum_{b,c}(\Delta^{T}_{b'c'} + \Delta^{T}_{b'b} + \Delta^{T}_{c'c}) \\
&- \frac{1}{|A||C|}\sum_{a,c}(\Delta^{T}_{a'c'} + \Delta^{T}_{a'a} + \Delta^{T}_{c'c}) - \frac{1}{|B||D|}\sum_{b,d}(\Delta^{T}_{b'd'} + \Delta^{T}_{b'b} + \Delta^{T}_{d'd}) \\
&= \Delta^{T}_{a'd'} + \Delta^{T}_{b'c'} - \Delta^{T}_{a'c'} - \Delta^{T}_{b'd'} \\
&= 0
\end{aligned}$$

Par conséquent, pour des séquences de tailles infinies et un modèle de substitutions exact, et donc $\alpha = a$, les distances estimées sont les vraies distances, c'est à dire des distances d'arbres. Ainsi $\mathcal{Q} = 0$ pour l'arbre vrai, et $\alpha^* = a$, indépendamment de l'écart à l'horloge moléculaire.

L'algorithme 3 décrit la procédure pour le calcul de α^*. La fonction d'estimation des distances (ligne 1) parcours l'alignement de séquences et dénombre les différents patterns de substitutions observés pour chaque couple de séquences. Par exemple, pour la comparaison de deux séquences d'ADN, et pour le modèle K2P, les trois patterns à dénombrer correspondent aux transitions, transversions et identités observées. Les distances évolutives sont des fonctions des fréquences de ces patterns et de la valeur de α.

Ainsi, une fois le dénombrement des patterns effectués, l'estimation des distances nécessite d'appliquer une simple formule analytique dont les paramètres sont les fréquences des transitions et transversions observées ainsi que α (ligne 3). Il n'est alors pas nécessaire de parcourir à nouveau les séquences pour estimer les distances lorsque α varie. Les phylogénies sont estimées par BIONJ (Gascuel, 1997a) puis les distances évolutives sont normées en divisant chaque élément de la matrice de distances par la somme de ses éléments (ligne 2). Cette étape est nécessaire car, pour des valeurs fixées de fréquences de différences observées, les distances évolutives sont décroissantes en fonction de α. Normer les distances permet de s'affranchir de cet effet et comparer ainsi de manière plus équitable plusieurs matrices obtenues à partir de différentes valeurs de α. Diverses stratégies de normalisation, incluant notamment la variance des distances estimées, ont été testées, sans aboutir à une d'amélioration significative des résultats. La valeur du critère est ensuite calculée à partir de la méthode décrite ci-dessus. Enfin, le pas de modification des valeurs de α est ajusté en fonction de la valeur courante de ce dernier paramètre. Lorsque $\alpha < 1.0$, les distances estimées varient fortement en fonction de α et le pas de modification est alors petit. Pour $\alpha > 1.0$, les distances évolutives varient plus modérément et le pas peut alors être de taille supérieure sans risquer de «manquer» un minimum (local ou global) de la fonction.

Algorithme 3: Estimation de α^*.

```
   Calcul de la valeur efficace α*
   début
   |  α_min = 0.06 ; α_max = 5000 ; pas = 0.02 ; Q_min = 1E+10 ;
   |  α ← α_min ;
1  |  Calculer Δ^α;
   |  tant que α < α_max faire
   |  |  Inférer T^α par BIONJ à partir de Δ^α;
2  |  |  Normer Δ^α;
   |  |  Calculer Q(Δ^α, T^α) en appliquant l'expression 4.1;
   |  |  si Q(Δ^α, T^α) < Q_min alors
   |  |  |  Q_min ← Q(Δ^α, T^α);
   |  |  |  α* ← α;
   |  |  fin
   |  |  Mettre à jour le pas de modification de α;
   |  |  α ← α+pas;
3  |  |  Mettre à jour les distances à partir de α;
   |  fin
   fin
   retourner α*;
```

4.3.2 Efficacité moyenne de $\mathcal{Q}$ pour l'approximation de α^{opt}

La Figure 4.4 montre l'évolution des valeurs de $\overline{\mathcal{Q}}(\Delta^\alpha, \mathcal{T}^\alpha)$ dans des conditions expérimentales identiques à celles de la Figure 4.1. Les jeux de données analysés ici sont les mêmes que précédemment. Sur ces courbes figurent les valeurs de a, α^{opt} et α^* et le décalage entre α^{opt} et α^* mesure l'efficacité de $\mathcal{Q}$ pour approximer, en moyenne, α^{opt}.

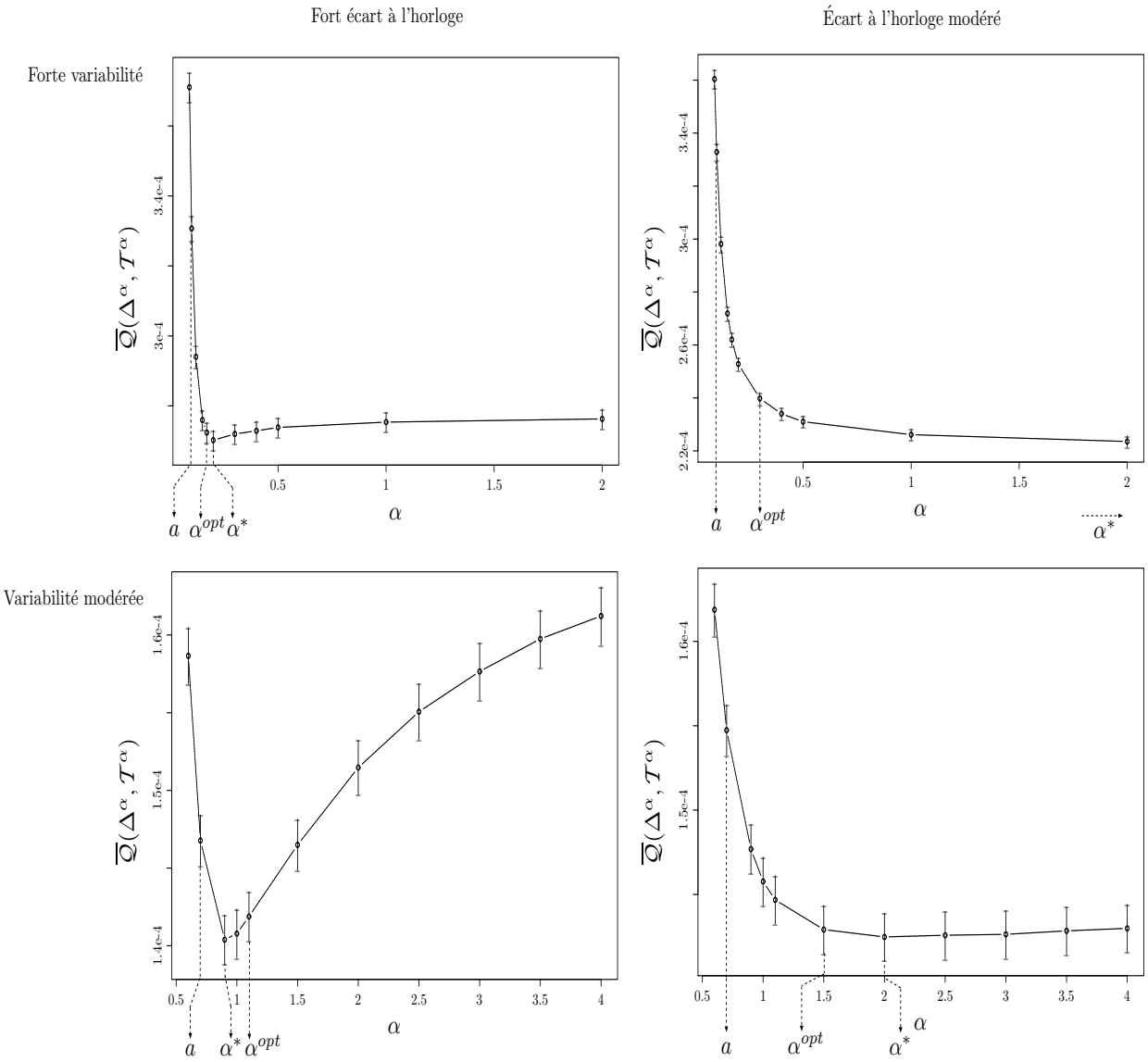


FIG. 4.4 – **Évolution de $\overline{\mathcal{Q}}(\Delta^\alpha, \mathcal{T}^\alpha)$ en fonction de α.** Ces courbes présentent les performances moyennes de $\mathcal{Q}$ pour l'approximation de α^{opt}. α^* est la valeur du paramètre de forme de la loi gamma maximisant l'ajustement entre les distances évolutives et la topologie de l'arbre déduite de celles-ci (voir texte).

L'allure de chaque courbe est similaire à celle du graphique correspondant de la Figure 4.1. Ceci montre que le critère d'arboricité $\mathcal{Q}$ est, en moyenne, un outil satisfaisant pour approximer la fiabilité des topologies. D'ailleurs, dans la majorité des conditions expérimentales testées ici, la valeur de α^* est très proche de celle de α^{opt}. Lorsque l'écart à l'horloge est modéré (courbes de droites) et la variabilité des vitesses entre sites est intense ($a = 0.1$, courbes du haut), le décalage entre valeurs optimales et efficaces de α est important. Cependant, cet écart se traduit pas une différence entre $\overline{RF}(\mathcal{T}_v, \mathcal{T}^{\alpha^{opt}})$ et $\overline{RF}(\mathcal{T}_v, \mathcal{T}^{\alpha^*})$ assez faible et surtout bien moins importante que l'écart entre $\overline{RF}(\mathcal{T}_v, \widehat{\mathcal{T}^{\alpha^{opt}}})$ et $\overline{RF}(\mathcal{T}_v, \widehat{\mathcal{T}^a})$ (voir Figure 4.1 et Tableau 1, Annexe A).

L'aplanissement des courbes au niveau des zones comprises entre α^{opt} et $+\infty$ montre que la probabilité pour que α^* soit supérieur à α^{opt} est plus élevée que la probabilité pour que α^* soit inférieur à α^{opt}. Ramener cette observation aux courbes de la Figure 4.1 montre clairement pourquoi α^* est plus performant que a pour l'estimation de topologies d'arbres. En effet, dans la plupart des cas, surestimer a conduit à des erreurs dans la topologie nettement moins fréquentes que lorsque a est sous-estimé.

Signalons enfin qu'il existe d'autres critères pour mesurer l'arboricité (par exemple Eigen et Winkler-Oswatitsch, 1981 ; Vach, 1992 ; Guénoche et Garreta, 2001). Ceux-ci sont, pour la plupart, basés sur différentes combinaisons des valeurs de S, M, et L. Les résultats obtenus sont semblables et le critère $\mathcal{Q}$, tel qu'il est défini dans l'expression 4.1, est légèrement plus efficace que les autres.

4.3.3 Appliquer $\mathcal{Q}$ à l'inférence phylogénétique

Les courbes précédentes (Figure 4.4) montrent que le critère proposé est efficace, en moyenne, pour approximer α^{opt}. Cependant, celles-ci ne fournissent aucune indication sur l'efficacité de cette approximation sur chaque jeu de données. En effet, nous pouvons imaginer un premier jeu pour lequel $\alpha^{opt} = 0.1$ et $\alpha^* = 0.9$, et pour un second jeu, $\alpha^{opt} = 0.9$ et $\alpha^* = 0.1$. Dans les deux cas, l'approximation de α^{opt} par α^* est médiocre tandis qu'en moyenne $\alpha^* = \alpha^{opt} = 0.5$.

Ainsi, de nouvelles simulations ont été réalisées afin de tester réellement l'efficacité de α^{opt} pour l'inférence de topologies d'arbres à partir de distances évolutives. Des phylogénies et les jeux de séquences correspondants ont été générés suivant une méthode proche de celle utilisée pour les simulations précédentes (voir Annexe A). Pour chaque jeu de données, une première phylogénie est inférée par BIONJ à partir de distances corrigées par α^* et sa topologie est comparée à celle de l'arbre vrai (RF_{α^*}). Une seconde phylogénie est inférée en appliquant BIONJ à partir de distances corrigées par a, la valeur du paramètre de forme de la loi gamma utilisée pour générer les séquences. La topologie de l'arbre obtenu est aussi comparée à celle de l'arbre vrai (RF_a). La comparaison des valeurs de RF_{α^*} à celles de RF_a mesure la diminution d'erreur relative offerte par l'utilisation de α^* comparé à a.

Pour des séquences de 300-pb, les topologies reconstruites à partir de α^* sont plus fiables que celles obtenues à partir de a, et ceci quel que soit l'écart à l'horloge moléculaire dans l'arbre vrai ou la variabilité réelle des vitesses d'évolution entre sites. Les résultats les plus probants sont obtenus lorsque cette variabilité est forte et l'horloge moléculaire est respectée. Dans de telles conditions, la réduction d'erreur est proche de 30% et correspond à une amélioration importante de la fiabilité des topologies inférées.

Pour des séquences de 1,000-pb, la diminution de l'erreur relative est globalement moins importante, même si, dans certaines conditions, celle-ci atteint encore 30%. Ceci s'explique simplement par la disparition progressive du bruit stochastique affectant les distances entre séquences lorsque leurs tailles augmente. Pour des séquences de tailles infinies, la différence entre α^* et a est nulle et les deux topologies d'arbres

sont identiques à celle de l'arbre vrai.

Les topologies inférées par BIONJ à partir de distances évolutives corrigées par α^* ont aussi été comparées à celles obtenues par DNAML (Felsenstein, 1993), un programme d'estimation de phylogénies au sens du maximum de vraisemblance. La vraie valeur du paramètre de forme de la loi gamma est utilisée pour les calculs de vraisemblance. Lorsque l'hétérogénéité des vitesses d'évolution entre sites est forte, les performances obtenues par BIONJ+α^* sont supérieures à celles de DNAML+a. Lorsque l'écart à l'horloge moléculaire dans l'arbre vrai est respectée, la réduction d'erreur dépasse 25%. En revanche, lorsque la variabilité des vitesses entre sites est moins importante, les topologies inférées par DNAML+a sont plus fiables que celles obtenues par BIONJ+α^* et la réduction d'erreur maximale observée approche 15%.

En pratique, la valeur de a est inconnue et α est utilisée pour le calcul des distances évolutives. Du point de vue de l'estimation de paramètres au sens classique, l'objectif est de minimiser l'écart entre a et α ; a correspondant à la valeur «optimale» pour l'estimateur. De manière surprenante, les résultats obtenus (Figure 4.1) indiquent que cet optimum n'est pas la valeur la plus appropriée pour l'inférence phylogénétique à partir de distances, et la méthode d'estimation proposée ici autorise la construction de topologies plus fiables que celles qui pourraient être obtenues si l'on disposait de la vraie valeur du paramètre. Par conséquent, baser la procédure d'estimation de a sur un critère (ici l'arboricité des distances estimées) approximant la fiabilité des phylogénies inférées est une approche efficace pour l'inférence de topologies à partir de distances évolutives.

4.4 Conclusions

L'estimation de paramètres de nuisance est, par définition, indispensable à l'inférence de topologies d'arbres. Les méthodes basées sur la vraisemblance (le maximum de vraisemblance et l'inférence bayesienne) approximent la fiabilité d'une topologie (par sa vraisemblance ou sa probabilité *a posteriori*) en fonction des valeurs des longueurs de branches et des paramètres libres du modèle de substitution. En revanche, les méthodes de distances ne fournissent pas «naturellement» un cadre méthodologique intégrant l'estimation des paramètres de substitution et la construction d'une topologie. Nous proposons ici une solution à ce problème pour l'estimation du paramètre de forme de la loi gamma, modélisant la variabilité des vitesses d'évolution entre sites. La méthode proposée repose sur une mesure de l'arboricité des distances estimées pour différentes valeurs de ce paramètre. Les simulations réalisées démontrent l'efficacité de cette mesure : les topologies des arbres reconstruits à partir de la valeur estimée du paramètre sont plus fiables que celles reconstruites à partir de la valeur sous-jacente aux séquences analysées.

De manière générale, les valeurs estimées par notre approche conduisent à sous-estimer, de manière non-linéaire, les distances évolutives. Ce phénomène est d'autant plus important que l'écart à l'horloge

moléculaire est faible. Quelques arguments théoriques permettent d'expliquer les résultats obtenus. Cependant, aucune démonstration formelle n'a pu être proposée jusqu'ici. Sur ce point, l'approche développée par Rzhetsky et Sitnikova (1996) mériterait d'être considérée attentivement. Ces auteurs exhibent des expressions analytiques des probabilités d'estimer correctement la topologie d'un arbre à quatre UEs, à partir de distances estimées sous le vrai modèle et sous un modèle faux, fournissant des estimations biaisées. Si cette approche peut être généralisée à d'autres modèles et un nombre de séquences supérieur à quatre, elle permettrait d'expliquer de manière formelle les résultats issus de nos simulations.

De manière générale, le problème abordé ici s'inscrit dans le cadre du choix du modèle de substitution le plus adapté à l'estimation de topologies d'arbres fiables. Pour les méthodes de distances, de nombreuses simulations (Saitou et Nei, 1987; Sourdis et Krimbas, 1987; Tajima et Takezaki, 1994) indiquent que sélectionner un modèle conduisant à des variances des estimateurs de distances inférieures aux variances déduites du vrai modèle, est plutôt bénéfique lorsque l'écart à l'horloge moléculaire est faible ou nul (ce qui est conforme aux prédictions de Steel et Penny, 2000). Les travaux de Yang (1997b) aboutissent à des conclusions similaires lorsque les phylogénies sont inférées par maximum de vraisemblance : un modèle ignorant la variabilité des vitesses entre sites, pourtant présente au sein des séquences générées, autorise l'estimation de topologies d'arbres plus fiables que celles déduites du modèle correct. Ce résultat concerne des arbres à quatre UEs respectant l'horloge moléculaire. Cependant, il est probable que cette tendance s'observe également pour des arbres plus réalistes, présentant un nombre d'UEs supérieur à quatre et ne respectant pas exactement l'horloge moléculaire. En effet, une analyse partielle de nos jeux de données simulées à vingt séquences montre que la valeur efficace du paramètre de forme de la loi gamma, telle qu'elle est estimée par notre approche, est plus appropriée que la vraie valeur pour l'inférence de topologies d'arbres par maximum de vraisemblance (résultats non présentés). Ces observations montrent qu'une solution efficace pour le choix du modèle pourrait être indépendante de la méthode utilisée pour l'inférence phylogénétique (maximum de vraisemblance ou distances).

En 1996, Rzhetsky et Sitnikova affirmaient que «[...] dans un futur proche, le choix arbitraire d'un modèle pour chaque cas particulier sera remplacé par des algorithmes mathématiques rigoureux implémentés au sein de logiciels d'utilisation aisée». Même si nos travaux vont dans ce sens, aucun de ces algorithmes génériques n'a encore vu le jour à notre connaissance. La question du choix d'un modèle de substitution adapté à l'inférence de topologies d'arbres reste un problème ouvert.

Chapitre 5

Une nouvelle approche pour l'amélioration itérative de la vraisemblance

Le principe du maximum de vraisemblance constitue un cadre théorique bien connu, et son application à la phylogénie moléculaire est une avancée importante dans le domaine. Malheureusement, cette approche est coûteuse en temps de calculs et la plupart des algorithmes classiques (DNAML (Felsenstein, 1993), fastDNAml (Olsen et al., 1994), PAML (Yang, 1997c), MOLPHY (Adachi et Hasegawa, 1996)) sont limités à l'analyse de jeux de données de tailles réduites ($<$ 100 UEs par arbre). Nous proposons ici une nouvelle méthode d'inférence d'arbres de vraisemblances maximales, basée sur un algorithme d'améliorations itératives de la phylogénie. L'idée est ici d'appliquer simultanément plusieurs modifications locales de la topologie, tout en ajustant l'ensemble des longueurs de branches. Des simulations et l'analyse de jeux de données réels démontre l'efficacité de cette approche tant en termes de fiabilité des topologies estimées que de temps de calcul et de capacité à maximiser la vraisemblance.

Dans un premier temps, nous donnons une description des grandes lignes de l'algorithme puis nous expliquons comment l'exploration de l'espace des topologies d'arbres est ici couplée à une procédure rapide de calcul de la vraisemblance et d'ajustement des longueurs de branches. L'algorithme est ensuite décrit dans sa globalité. Enfin, la fiabilité des topologies inférées, les durées d'exécution et les vraisemblances des phylogénies inférées par cette approche sont comparées aux performances obtenues à partir des méthodes actuelles. Ces travaux seront publiés dans la revue «Systematic Biology».

5.1 Une première description de l'algorithme

La plupart des méthodes d'améliorations itératives actuelles distinguent clairement les perturbations topologiques de l'ajustement des longueurs de branches. Nous l'avons vu au chapitre 3 : l'approche standard consiste à modifier la topologie puis évaluer la phylogénie correspondant après avoir optimisé les longueurs de branches et, éventuellement, les paramètres libres du modèles de substitution. De plus,

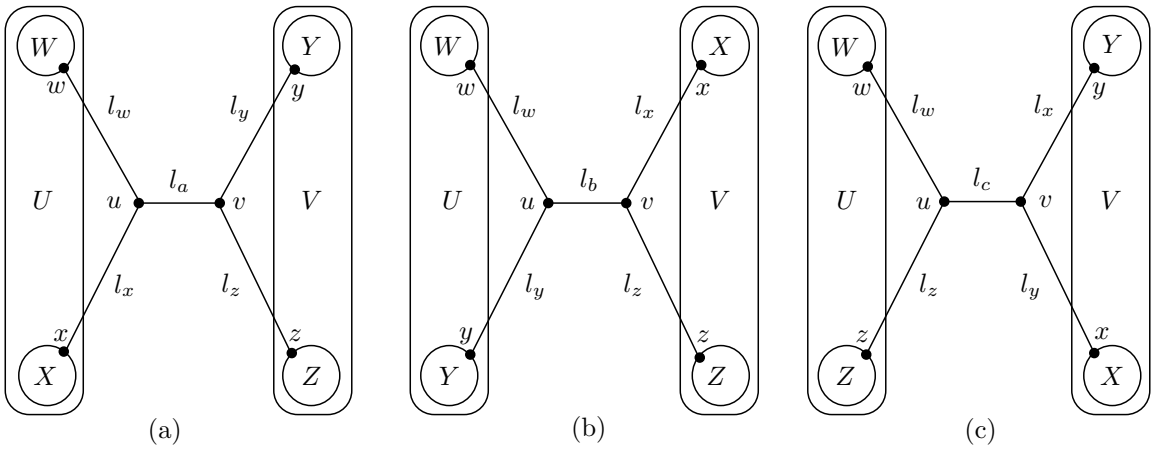


FIG. 5.1 – **Arbre modèle.** Ces trois arbres correspondent aux configurations topologiques qu'il est possible d'atteindre à partir d'un NNI sur la branche joignant u et v. l_a, l_b et l_c sont les trois longueurs maximisant la vraisemblance de la phylogénie lorsque seule cette branche est ajustée (voir texte). W, X, Y et Z sont trois sous-arbres dont w, x, y et z sont les racines respectives.

pour ces algorithmes, les perturbations de la topologie s'effectuent séquentiellement. Ainsi, NNIs, SPRs ou TBRs sont appliqués à partir de chaque branche interne de l'arbre successivement.

La méthode décrite dans ce chapitre diffère des approches classiques sur ces deux points. En effet, longueurs de branches et topologies sont ici optimisées simultanément. De même, les perturbations topologiques interviennent parallèlement autour de différentes branches internes. Les trois étapes de l'algorithme sont les suivantes : (1) pour chaque branche interne, calculer la vraisemblance de la phylogénie dans les trois configurations topologiques obtenues par NNI («Nearest Neighbor Interchange», voir chapitre 3). Dans ces trois cas, seule la longueur de la branche interne est ajustée. (2) Appliquer simultanément la majorité des modifications de topologies et de longueurs de branches conduisant à une augmentation de la vraisemblance. (3) Répéter les étapes (1) et (2) jusqu'à l'obtention d'une phylogénie de topologie et longueurs de branches stables.

Le mouvement élémentaire sur lequel repose cette approche est le NNI. Nous décrivons dans la partie suivante les avantages offerts par ce mouvement du point de vue de la rapidité du calcul de la vraisemblance d'une phylogénie dans différentes configurations topologiques.

5.2 NNIs et mise à jour rapide de la vraisemblance

La Figure 5.1 présente les trois topologies explorées par un NNI appliqué à une branche interne quelconque. Nous supposons ici que la configuration topologique originale est donnée en (a) et (b) et (c) sont donc deux topologies alternatives. De plus, nous considérons que toutes les vraisemblances conditionnelles de l'arbre courant sont connues. Dans la configuration originale, la vraisemblance du sous-arbre U conditionnellement à l'état h à sa racine u se déduit des vraisemblances conditionnelles des sous-arbres W et

X :

$$L_s(u=h) = \left[\sum_{g\in\mathcal{A}} L_s(w=g)P_{hg}(l_w)\right] \times \left[\sum_{g\in\mathcal{A}} L_s(v=g)P_{hg}(l_v)\right]$$

où $L_s(w=g)$ et $L_s(v=g)$ sont les vraisemblances des sous-arbres W et V au site s, conditionnellement à l'état g aux noeuds racines w et v. La vraisemblance conditionnelle du sous-arbre V est obtenue de manière symétrique à partir de celles de Y et Z. Lorsque les vraisemblances conditionnelles de U et V sont connues ainsi que la longueur l_a, la vraisemblance de la phylogénie au site s est calculée en appliquant l'expression 2.12. Le nombre d'opérations à effectuer pour obtenir la valeur de la vraisemblance sur l'ensemble du jeu de données et alors de l'ordre de N, le nombre de sites de la séquence.

Dans la configuration (b), les vraisemblances conditionnelles des sous-arbres U et V se déduisent de celles de W et Y, et X et Z respectivement. Le calcul de la vraisemblance de la phylogénie dans cette seconde configuration nécessite, là encore, un nombre d'opérations de l'ordre de N au lieu de $N \times n$ pour l'algorithme pruning qui requiert un parcours en profondeur de la phylogénie à n UEs (algorithme 1). Ceci s'applique aussi pour le calcul de la vraisemblance de la phylogénie dans la configuration topologique (c). Bien qu'il soit encore possible d'éviter de parcourir toute la phylogénie, soulignons ici que des mouvements de type SPRs et TBRs sont bien plus coûteux en termes de nombre de vecteurs de vraisemblances conditionnelles de sous-arbres à mettre à jour. De plus l'application de mouvements de types NNIs s'accommode d'un nombre restreint d'ajustements des longueurs de branches pour la comparaison des vraisemblances de phylogénies dans différentes configurations topologiques.

Ainsi, pour chaque NNI, seules les valeurs de l_a, l_b et l_c sont optimisées (méthode de Brent, voir chapitre 3). Ces ajustements sont initiés à partir de l, la longueur originale de la branche et les vraisemblances des phylogénies dans ces trois configurations topologiques sont notées L_a, L_b et L_c. Chacune de ces trois étapes d'optimisation fait intervenir les calculs de vraisemblances décrits ci-dessus et sont donc peu coûteux en temps de calculs.

Lorsque L_b est supérieure à L_a et L_c, la configuration topologique (b) est plus vraisemblable que (a) et (c). Dans une telle situation, plus la différence entre les valeurs de L_b et L_a est importante, plus le choix du remplacement de (a) par (b) est fiable. Ainsi, pour chaque branche interne est calculé un score S défini comme suit :

$$S = \max_{\{b,c\}}(L_b - L_a, L_c - L_a)$$

Il est important de noter que L_a, L_b et L_c ne sont pas les valeurs optimales de la vraisemblance de la phylogénie dans les trois configurations topologiques correspondantes. Ceci serait le cas si toutes les longueurs de branches étaient ajustées pour chacune des trois alternatives. Or, seule la longueur de la branche autour de laquelle s'effectue le réarangements, est optimisée. Les longueurs des autres branches sont considérées comme fixes. Nous verrons, dans la partie suivante, que ce choix se justifie lorsque l'on considère l'algorithme dans sa globalité.

5.3 Application simultanée de plusieurs NNIs

Comme nous l'avons souligné au chapitre 3, les valeurs optimales des longueurs de branches sont interdépendantes. Il en est de même pour les configurations topologiques autour de chaque branche interne. Pour cette raison, NNIs et ajustements des longueurs de branches sont généralement appliqués branche par branche. Cependant, cette approche est coûteuse en temps de calcul car elle nécessite la mise à jour de nombreux vecteurs de vraisemblances conditionnelles.

Notre approche consiste à calculer indépendamment les longueurs optimales de l'ensemble des branches et les scores de chacune des branches internes, puis appliquer simultanément la majorité des modifications locales de la topologie et les ajustements de longueurs de branches indiquées. Il est peu probable que plusieurs perturbations locales aboutissent à une topologie globalement optimale. Néanmoins, la stratégie mise en place permet sans doute d'éviter des maxima locaux de la fonction de vraisemblance. Nous revenons sur ce point dans la conclusion du chapitre.

Dans un premier temps, les scores de l'ensemble des branches internes sont classés par ordre décroissant. Lorsqu'aucun score n'est positif, la topologie de l'arbre reste intacte et seule les longueurs de branches sont mises à jour selon la procédure décrite plus bas. Dans le cas contraire, une proportion λ des branches de scores positifs sont soumises aux mouvements topologiques conduisant aux configurations locales de vraisemblances maximales. Lorsque les scores de deux branches adjacentes sont positifs, seule la branche dont le score est le plus fort est soumise au changement de topologie. En effet, deux branches internes adjacentes ont un sous-arbre en commun et appliquer simultanément les NNIs sur ces deux branches n'a pas de sens.

Cette modification de la topologie s'accompagne de l'ajustement des longueurs des branches externes et des branches internes autour desquelles la configuration topologique n'est pas modifiée. Chacune des longueurs originales, notée l, est substituée par $l + \lambda(l_a - l)$. Ainsi, ajustements de longueurs de branches et modifications topologiques sont tous deux régis par le même facteur λ.

Cette procédure de perturbations simultanées ne garantit pas l'augmentation de la vraisemblance d'une étape à la suivante. Lorsque la vraisemblance décroît, la valeur de λ est divisée par deux et la phylogénie originale est modifiée en conséquence. La décroissance de λ se poursuit tant que la vraisemblance diminue. Lorsqu'au moins un score est positif, seule la configuration locale autour de la branche de score maximum est modifiée en dernier recours, assurant ainsi une augmentation de la vraisemblance. Un raisonnement similaire s'applique lorsqu'aucun score n'est positif. Cette stratégie permet donc d'éliminer les «conflits» entre les changements locaux de topologies et de longueurs de branches et garantit la convergence de l'algorithme. La même approche est utilisée par Felsentein et Churchill (1996) pour s'assurer de la croissance de la vraisemblance lors de la procédure d'ajustement de longueurs de branches par la méthode Newton-Raphson (voir chapitre 3). En pratique, la valeur originale de λ est égale à 0.75

et le phénomène de décroissance de la vraisemblance est rare.

5.4 L'algorithme

L'algorithme 4 décrit globalement la méthode proposée. Celui-ci procède à partir de la matrice de séquences homologues, D, et d'un modèle de substitution M (ligne 1). Une première estimation de la phylogénie, notée T_0, est obtenue par BIONJ à partir des distances évolutives estimées par maximum de vraisemblance (ligne 2). Le logarithme de la vraisemblance de T_0, noté $lnL(T_0|D)$, est calculé en utilisant les algorithmes 1 et 2 (ligne 3) et les vecteurs de vraisemblances conditionnelles de chaque sous-arbre pour chaque site sont stockés en mémoire. Les valeurs des paramètres libres du modèle de substitution sont ensuite ajustés par la méthode de la section d'or (ligne 4). Les modifications locales et simultanées de la topologie et des longueurs de branches sont ensuite poursuivies jusqu'à la convergence du logarithme de la vraisemblance. Lorsqu'une série de ces perturbations entraîne une diminution de vraisemblance, la phylogénie courante est contrainte à se rapprocher de l'arbre à l'étape précédente (ligne 6) suivant la méthode décrite plus haut. Enfin, les valeurs des paramètres libres de substitution sont ajustés périodiquement (ligne 7). La valeur de cette période est fixée à quatre.

Le programme PHYML, implémentant cet algorithme, est disponible à l'adresse `http://www.lirmm.fr/~guindon/phyml.html` pour différentes plateformes (Linux, Mac OS X, Win32 et SunOS). Il permet d'analyser des alignements de séquences nucléotidiques sous différents modèles de substitutions (TN93, HKY85, F84, F81, K2P, JC69), et utilise une loi gamma discrétisée pour modéliser la variabilité des vitesses entre sites. Nous travaillons actuellement sur l'implémentation de modèles de substitutions appliqués aux acides aminés.

5.5 Simulations

Des jeux de données ont été générés aléatoirement afin d'évaluer la fiabilité des topologies d'arbres estimés par notre approche et comparer celle-ci aux méthodes d'inférence standard. 5,000 phylogénies comprenant 40 UEs et les 5,000 jeux de séquences homologues correspondantes ont été générées aléatoirement (voir Annexe B)

Les logiciels utilisés dans ces simulations sont : NJ (Saitou et Nei, 1987), Weighbor 1.2 (Bruno et al., 2000), fastDNAml (Olsen et al., 1994), PAUP* 4.0 beta (Swofford, 1999), NJML (Ota et Li, 2001), DNAPARS 3.5 (Felsenstein, 1993) et MrBayes 2.01 (Huelsenbeck et Ronquist, 2001). NJ et Weighbor infèrent un arbre à partir de distances évolutives. Ces dernières sont estimées par DNADIST (Felsenstein, 1993). fastDNAml et PAUP* visent à estimer des phylogénies de vraisemblances maximales à partir de l'insertion d'UEs combinée ici à des perturbations topologiques locales, de type NNIs. NJML est une approche hybride couplant la reconstruction d'un arbre par NJ à des réarangements topologiques locaux

Algorithme 4: Estimation d'une phylogénie de vraisemblance maximale

```
1  Données : D, M
   début
       période = 4;
       précision = 1E − 03;
       λ = 0.75 ;
2      Calculer Δ^M;
       Inférer T_0 à partir de Δ^M par BIONJ;
3      Calculer lnL(T_0|D) et l'ensemble des vraisemblances conditionnelles (algorithmes 1 et 2);
4      Ajuster les paramètres libres du modèle à partir de T_0 et M;
       T'_c ← T_0;
5      répéter
           T_c ← T'_c;
           Calculer L_a, L_b, L_c, l_a, l_b, l_c et S pour chaque branche interne de T_c;
           Classer les branches internes dans l'ordre décroissant des valeurs de S;
           Calculer les valeurs de l_a pour les branches externes;
           λ' ← λ;
           répéter
               T'_c ← T_c;
               Mettre à jour l'ensemble des longueurs de branches de T'_c;
               Réaliser une proportion λ' des modifications topologiques sur T'_c;
               Calculer lnL(T'_c|D);
6              λ' ← λ/2;
           jusqu'à lnL(T'_c|D) > lnL(T_c|D);
           période ← période - 1;
           si période = 0 alors
               période = 4;
7              Ajuster les paramètres libres du modèle à partir de T'_c;
       jusqu'à lnL(T'_c|D) − lnL(T_c|D) < précision;
       retourner T'
   fin
```

guidés par le principe du maximum de vraisemblance. DNAPARS implémente l'algorithme standard pour l'approche du maximum de parcimonie. Enfin, MrBayes permet d'estimer les probabilités *a posteriori* de topologies d'arbres à partir d'une approche de type Monte Carlo (voir chapitre 3). Le logiciel MetaPIGA (Lemmon et Milinkovitch, 2002), implémentant l'algorithme génétique décrit au chapitre 3, n'a pas pu être testé sur ces données car la version actuelle du programme n'autorise pas l'analyse automatique de plusieurs jeux de séquences consécutivement. Enfin, seuls les 1,000 premiers jeux de données ont été analysés par PAUP* et MrBayes afin de limiter les temps de calculs à deux semaines.

La Figure 5.2 présente l'évolution des distances moyennes entre les topologies inférées, $\widehat{\mathcal{T}}$, et celles des arbres vrais, $\mathcal{T}_v$, en fonction de la divergence maximale entre séquences. Cette divergence correspond simplement à la fréquence maximale de différences observées entre paires de séquences. Les résultats obtenus corroborent les observations issues de précédentes simulations (Huelsenbeck et Hillis, 1993; Kuhner et Felsenstein, 1994; Huelsenbeck, 1995; Rosenberg et Kumar, 2001; Ranwez et Gascuel, 2002). Ainsi, deux tendances apparaissent clairement. (1) Lorsque les divergences entre séquences sont trop faibles, les séquences n'offrent pas suffisamment d'informations pour estimer correctement les branches internes

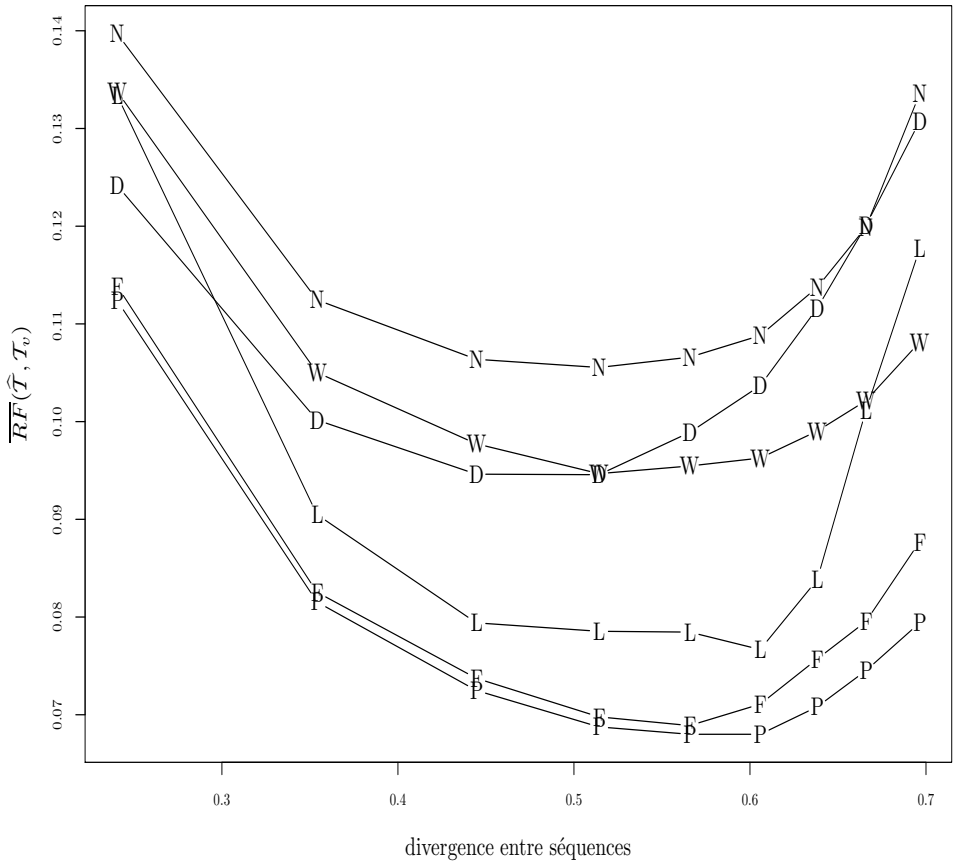


FIG. 5.2 – **Fiabilité des topologies estimées par différentes approches.** Ces courbes présentent l'évolution de la distance topologique moyenne entre l'arbre vrai et les arbres estimés par diverses méthodes. P : PHYML, F : fastDNAml, L : NJML, W : Weighbor, D : DNAPARS, N : NJ.

courtes. (2) Pour des divergences importantes, le phénomène de saturation brouille le signal phylogénétique et inférer des topologies sans commettre d'erreurs est, là-encore, une tâche difficile.

La comparaison des différentes méthodes met en évidence la fiabilité supérieure, en moyenne, des topologies estimées suivant le principe du maximum de vraisemblance. Notons cependant que l'efficacité de NJML n'atteint pas celle de fastDNAml ou PHYML, notamment lorsque la divergence entre séquences est importante. Les performances de ces deux dernières approches sont relativement proches. PHYML tend néanmoins à proposer les topologies les plus fiables pour de fortes divergences entre séquences. L'analyse des 1,000 premiers jeux de données confirme l'efficacité de notre approche : les distances topologiques moyennes obtenues à partir de PAUP*, fastDNAml, MrBayes et PHYML sont égales à 0.086, 0.086, 0.081 et 0.081 respectivement. De plus, après optimisation des longueurs de branches, les vraisemblances des phylogénies estimées par PHYML sont supérieures à celles des arbres vrais dans $\simeq$95% des cas. Ceci indique que l'algorithme proposé permet d'inférer des phylogénies très vraisemblables et une exploration

plus intense de l'espace des topologies ne pourrait pas aboutir à une amélioration importante de la fiabilité des topologies inférées par le principe du maximum de vraisemblance.

5.6 Temps de calculs et optimisation de la vraisemblance

Les temps de calculs des méthodes comparées ci-dessus, auxquelles s'ajoute MetaPIGA (Lemmon et Milinkovitch, 2002), ont ensuite été mesurés à partir de 30 jeux de données simulées présentant 40 et 100 UEs. Les arbres et les séquences ont été générés suivant une procédure identique à celle utilisée précédemment. 218 séquences homologues de génomes procaryotes, issues de la petite sous-unité ribosomique et mesurant 4,182-pb ainsi que 500 séquences du gène *rbc*L de chloroplastes de plantes, mesurant 1,428-pb, ont également été analysées.

Pour ces deux derniers jeux de données, fastDNAml et PAUP* ont été stoppés après dix jours de calculs sur un ordinateur standard. Aussi, MrBayes a été volontairement suspendu après 1,000,000 générations sans avoir atteint l'état stationnaire, tandis que les limitations d'espace mémoire inhérentes à NJML ne permettent pas d'estimer des phylogénies pour des jeux de cette taille. Notons enfin que MetaPIGA et MrBayes procèdent à partir d'arbres aléatoires. Cependant, l'utilisation d'arbres plus réalistes pour points de départs (inférés par une méthode de distance par exemple) auraient sans doute permis d'accélérer les calculs.

Le Tableau 5.1 présente les résultats obtenus. PHYML est nettement plus rapide que les autres programmes basés sur la vraisemblance. Cette différence est particulièrement marquée pour les arbres à 100 UEs : l'estimation par PHYML est obtenue au bout de 12 secondes en moyenne tandis que plus de 25 minutes sont nécessaires à fastDNAml. Les durées d'exécution de PHYML sont en fait plus proches, en termes d'ordre de grandeur, de celles des méthodes de distances que de celles fondées sur la vraisemblance. Pour le jeu de données réelles à 500 UEs, PHYML est même plus rapide que Weighbor, une méthode pourtant basée sur l'analyse de distances évolutives. Ces performances s'expliquent certainement pas le nombre restreint d'étapes de perturbation de la topologie (chiffres en gras dans le tableau) pour aboutir à une estimation convergente.

Les vraisemblances des phylogénies estimées par ces différentes approches ont ensuite été comparées. Pour ce faire, les longueurs de branches des arbres estimés à partir des jeux de données simulés ont été ajustées par une même procédure d'optimisation (Newton-Raphson). Les méthodes ont été classées pour chaque jeu de données en fonction des vraisemblances des phylogénies correspondantes et les rangs moyens ont été calculés à partir des 30 classements obtenus. Pour les arbres à 40 UEs, les valeurs de ces rangs moyens sont : 3.5, 3.4, 3.1, 2.0, 1.9 et 1.7 pour MetaPIGA, NJML, MrBayes, PAUP*, PHYML et fastDNAml, respectivement. Pour 100 UEs : 4.8, 4.3, 3.9, 2.5, 2.1 et 2.0 pour MetaPIGA, NJML, MrBayes, PAUP*, fastDNAml et PHYML, respectivement. Ces résultats montrent donc que la réduction des temps

méthode	Simulations		Données réelles	
	40 UEs 500-pb	100 UEs 500-pb	218 UEs 4,182-pb	500 UEs 1,428-pb
DNADIST + NJ/BIONJ	0.3s	2.3s	50s	2min19s
DNADIST + Weighbor	1.5s	22s	4min52s	58min40s
DNAPARS	0.5s	6s	4min4s	13min12s
PAUP*	3min21s	1h4min	*	*
MrBayes	2min6s	32min37s	*	*
fastDNAml	1min13s	26min31s	*	*
NJML	15s	6min4s	*	*
MetaPIGA	21s	3min27	4h45min	9h4min
PHYML	2.7s (**6.4**)	12s (**8.3**)	8min13s (**15**)	11min59s (**13**)

TAB. 5.1 – **Temps de calculs.** Ceux-ci ont été mesurés sur un PC 1.8 Ghz (1 Go RAM) sous Linux. Les chiffres en gras désignent les nombres moyens d'étapes d'améliorations itératives de la vraisemblance par PHYML

de calcul observée pour PHYML ne s'accompagne pas d'une diminution d'efficacité pour l'optimisation de la vraisemblance. Des conclusions analogues sont obtenues à partir de l'analyse des deux jeux de données réelles.

5.7 Conclusions

L'algorithme présenté dans ce chapitre est fondé sur des méthodes de calcul de la vraisemblance et des mouvements topologiques locaux bien connus. L'originalité se situe ici dans la manière de combiner ces éléments. Les réarangements par NNIs sont avantageux du point de vue de la rapidité de la mise à jour de la vraisemblance, mais n'offrent pas un moyen très efficace pour explorer l'espace des topologies lorsqu'ils sont appliqués sur les branches internes successivement. L'idée est donc de réaliser ces mouvements simultanément, de manière à permettre le passage d'une topologie à une autre très distincte, en une seule étape, tout en modifiant les longueurs de branches.

Appliquer ce type de perturbations revient à accepter des configurations locales non optimales, et se rapproche donc, en ce sens, du principe fondamental des méthodes d'optimisation stochastique. En effet, il est possible que, pour une branche interne donnée, une configuration alternative soit optimale lorsque le reste de l'arbre est intact mais ne le soit plus lorsque les configurations topologiques sont modifiées autour d'autres branches internes. Modifier malgré tout la topologie autour de la première branche revient donc à accepter une configuration non optimale *a posteriori*. Cette dernière est alors corrigée à l'étape suivante de l'algorithme. Une telle stratégie permet donc vraisemblablement d'éviter les minima locaux rencontrés lors de l'application successive de plusieurs NNIs. De plus, réaliser simultanément plusieurs perturbations locales de la topologie consiste à appliquer des pas de modifications de la phylogénie de tailles supérieurs à ceux effectués dans le cadre d'une approche successive. Cet argument explique en grande partie la

réduction des temps de calculs comparé aux méthodes d'améliorations itératives classiques.

Conclusions et perspectives

Les travaux présentés dans ce mémoire s'inscrivent dans le cadre de l'approche statistique en phylogénie moléculaire. Comme nous l'avons vu au cours des trois premiers chapitres, cette discipline se situe à la frontière entre la modélisation statistique et l'algorithmique. Ainsi, les modèles markoviens décrivant l'évolution des séquences constituent autant d'hypothèses plus ou moins sophistiquées sur les mécanismes de substitutions entre bases nucléiques ou acides aminés. Conformément à l'objectif de la statistique inférentielle, cette approche ne prétend pas décider de la véracité de tel ou tel modèle, mais permet plutôt de rejeter ceux correspondant à des hypothèses erronées. Les algorithmes de construction d'arbres interviennent en aval de l'établissement des modèles de substitutions. Le principe général de ces méthodes repose sur l'optimisation d'un critère statistique (les moindres carrés, la vraisemblance, etc.) à travers l'ajustement de la topologie et des longueurs de branches de l'arbre, ainsi que certains paramètres du modèle de substitution.

L'établissement de modèles de substitution réalistes et d'algorithmes de construction d'arbres efficaces vise principalement à améliorer la fiabilité des topologies des phylogénies estimées. Pour ce faire, il est nécessaire de (1) paramétrer de manière adéquate le modèle de substitution, (2) choisir un critère pertinent pour l'évaluation des phylogénies, et enfin, (3) disposer d'algorithmes efficaces pour l'exploration de l'espace des topologies. Nos travaux ont porté sur le premier et le troisième point. Ainsi, la première partie de notre contribution concerne l'estimation de la valeur d'un paramètre libre du modèle de substitution, modélisant la variabilité des vitesses entre sites des séquences homologues alignées. La seconde partie porte sur une méthode d'exploration de l'espace des topologies fondée sur un algorithme d'amélioration itérative de la vraisemblance d'une phylogénie.

La modélisation de la variabilité des vitesses entre sites à partir d'une loi gamma a permi de confirmer l'importance de ce phénomène au sein des séquences nucléiques. Cependant, aucune étude n'a été entreprise afin de caractériser précisément le lien entre l'estimation du paramètre de forme de cette distribution, mesurant l'intensité de l'hétérogénéité des vitesses, et la fiabilité des topologies d'arbres inférés. Ainsi, notre approche, basée sur des simulations, décrit les relations entre la probabilité de recouvrir la topologie correcte d'un arbre phylogénétique à partir de distances évolutives, la valeur du paramètre de forme de la loi gamma, et l'écart à l'horloge moléculaire. Les résultats obtenus montrent que plus les vitesses d'évolution sont homogènes entre lignées (hypothèse de l'horloge moléculaire), plus la valeur optimale du paramètre est supérieure à la valeur sous-jacente au jeu de données analysé. Ainsi, lorsque l'écart à l'horloge moléculaire est faible ou modéré, les modèles ignorant la variabilité de vitesses entre sites, pourtant présente au sein des données, autorisent des inférences de topologies plus fiables que celles

obtenues à partir du modèle correct. Ceci s'explique en partie par des arguments théoriques simples, démontrant la robustesse des méthodes de distances vis à vis de la valeur du paramètre de forme de la loi gamma. À partir de ce constat, nous proposons une méthode d'estimation permettant d'approximer la valeur optimale du paramètre. Les résultats obtenus sont satisfaisants puisque l'utilisation de cette valeur pour calculer les distances évolutives conduit à une précision topologique généralement supérieure à celle obtenue à partir de la vraie valeur (inconnue) du paramètre.

Les perspectives liées à ce travail sont de plusieurs ordres. Le plus important à nos yeux serait d'analyser le lien entre les valeurs du paramètre de la loi gamma minimisant la probabilité d'erreur dans l'estimation de la topologie, et le support des données accordé aux différentes branches internes de l'arbre. En effet, il est attendu que l'application de valeurs du paramètre conduisant à sous-estimer les distances évolutives, et donc leurs variances, aboutisse à une sur-estimation de la fiabilité de branches erronées. Caractériser précisément ce phénomène semble donc indispensable. Il serait aussi souhaitable de définir précisément les contraintes engendrées par l'écart à l'horloge moléculaire sur l'estimation des paramètres du modèle. Dans le cadre des méthodes de distances, lorsque l'horloge moléculaire est respectée, des estimateurs biaisés, de variances inférieures aux estimateurs non-biaisés, sont préférables pour l'inférence topologique. L'idée serait donc de définir un «rayon de sécurité» autour des distances estimées en fonction de l'écart de ces dernières à des distances ultramétriques. En 1997, Atteson a défini un tel rayon dans le cadre de l'inférence d'arbres par NJ et sans faire référence aux contraintes liées à l'écart à l'horloge. Il pourrait être intéressant d'entreprendre des travaux similaires en intégrant ce dernier paramètre. Un autre développement possible serait d'analyser les effets de la sous-estimation de paramètres du modèle de substitution sur l'inférence d'arbres au sens du maximum de vraisemblance. Des résultats préliminaires montrent des tendances similaires à celles observées à partir de méthodes de distances. Si cela est confirmé, il se peut alors que les développements théoriques évoqués ci-dessus doivent s'appliquer dans un cadre plus général que celui de l'inférence d'arbre à partir de distances évolutives. Enfin, il serait intéressant d'appliquer notre approche à l'estimation d'autres paramètres libres du modèle de substitution, tels que le ratio transition/transversion par exemple. Cependant, ce dernier paramètre a généralement un impact moins important sur la phylogénie que celui mesurant l'hétérogénéité des vitesses d'évolution entre sites. Il est donc peu probable que les améliorations des inférences de topologies soient aussi significatives que celles observées ici.

L'estimation de topologies d'arbres suivant le principe du maximum de vraisemblance est parmi les plus performantes en termes de fiabilité. Malheureusement, cette approche est pénalisée par des temps de calculs rédhibitoires, n'autorisant généralement pas l'analyse de jeux de données de plus d'une centaine de séquences. Nous proposons ici un nouvel algorithme d'améliorations itératives de la vraisemblance offrant des performances satisfaisantes à plusieurs niveaux. En effet, les résultats obtenus montrent que (1) la fiabilité des topologies inférées est similaire, voir supérieure, à celle des arbres obtenus à partir des méthodes actuelles les plus efficaces de ce point de vue, (2) les vraisemblances des phylogénies inférées par

notre approche sont au moins similaires à celles issues des approches classiques, (3) les temps de calculs sont de l'ordre de grandeur de ceux des méthodes de distances.

Les perspectives sont, là-encore, nombreuses. Ainsi, les valeurs de bootstrap non paramétrique calculées dans le cadre du maximum de vraisemblance sont souvent issues d'approximations (voir Hasegawa et Kishino, 1994, par exemple). Bien que ces dernières soient relativement précises, appliquer le principe original est clairement plus satisfaisant d'un point de vue théorique. Cependant, les temps de calculs prohibitifs limitent fortement l'application de cette approche en pratique. Or, pour des jeux de données de tailles modérées (<50 UEs), notre méthode autorise le calcul des fréquences de bootstrap pour des durées d'exécution raisonnables. Une autre issue à ce sujet serait de comparer les probabilités de bootstraps estimées par l'approche exacte aux probabilités *a posteriori* de clades. Cette analyse permettrait peut-être d'éclairer d'un jour nouveau la polémique récente concernant la surestimation de la fiabilité des branches mesurée par une approche bayesienne, comparé aux résultats obtenus à partir des fréquences de bootstrap approximées (Suzuki et al., 2002; Douady et al., 2003). Dans la version actuelle de notre algorithme, la valeur du facteur λ, déterminant le nombre de modifications topologiques locales à effectuer, est fixée *a priori*. Ajuster ce paramètre lors du processus de construction mériterait d'être considéré attentivement. Par exemple, dans les premières étapes de la construction, il pourrait être avantageux d'accepter, avec une certaine probabilité, certaines perturbations topologiques autour de branches n'indiquant pas qu'une des deux configurations alternatives présente une vraisemblance supérieure à l'originale. Dans cette situation $\lambda > 1$. Dans les dernières étapes de l'estimation, seuls les changements sur les branches internes indiquant une configuration alternative de vraisemblance supérieure sont acceptés ($\lambda < 1$). λ joue ainsi un rôle analogue à celui de la température dans un algorithme de recuit simulé. Un première implémentation de cette approche offre des résultats très encourageant du point de vue de l'optimisation de la vraisemblance.

Cependant, la perspective la plus intéressante dans l'immédiat est sans nul doute d'ordre biologique. Elle concerne en effet l'application du principe du maximum de vraisemblance à l'analyse de jeux de données classiques, tels que la phylogénie des plantes à fleurs ou l'estimation d'une phylogénie «universelle», pour lesquels sont disponibles un très grand nombre de séquences orthologues. Bien qu'ajouter des séquences ne soit pas nécessairement un gage d'accroissement de la fiabilité des topologies estimées, il est souhaitable de pouvoir construire de telles phylogénies, simplement pour les comparer à celles déjà proposées par d'autres approches, souvent moins fiables. Outre l'augmentation du volume des données analysables, l'utilisation de modèles pertinents mais coûteux en termes de nombre de calculs à effectuer (les modèles de substitutions entre codons, par exemple) est envisageable pour l'analyse de jeux de données de taille standard. Appliquer de tels modèles est souhaitable non seulement pour la fiabilité à accorder aux histoires évolutives estimées, mais aussi pour définir de plus en plus précisément les contraintes évolutives agissant à différentes échelles sur les séquences génétiques.

Notations

UE : unité évolutive
n : nombre d'UEs
D : données
$\mathcal{A}$: alphabet des états pris par les caractères étudiés

N : nombre de sites
N_p : nombre de sites distincts (patterns)

$\mathcal{T}_v$: topologie de la vraie phylogénie
T_v : vraie phylogénie
$\mathcal{T}$: topologie d'une phylogénie quelconque
T : phylogénie quelconque
$\widehat{\mathcal{T}}$: topologie de la phylogénie estimée
$\widehat{T}$: phylogénie estimée
l : longueur d'une branche quelconque
l_{uv} : longueur de la branches joignant les noeuds u et v

Δ : matrice des distances estimées
Δ_{ab} : distance estimée entre les UEs a et b
$\Delta^{\widehat{T}}$: matrice des distances dans la phylogénie estimée
$\Delta_{ab}^{\widehat{T}}$: distance entre les UEs a et b dans l'arbre estimé
d : matrice des vraies distances
d_{ab} : distance vraie séparant a et b

$P_{xy}(t)$: probabilité d'observer l'état y à l'instant t sachant que x était observé à l'instant 0
L : vraisemblance de la phylogénie
L_s : vraisemblance de la phylogénie au site s
lnL : logarithme de la vraisemblance de la phylogénie
lnL_s : logarithme de la vraisembblance de la phylogénie au site s

a : valeur du paramètre de forme de la loi gamma
α : valeur estimée du paramètre de forme
α^{opt} : valeur optimale du paramètre de forme (cf. chapitre 4)
α^* : valeur efficace du paramètre de forme (cf. chapitre 4)

RF : distance topologique de Robinson et Foulds (1979)

π_X : fréquence stationaire de l'état X
f_X : fréquence observée de l'état X
$\mathbf{\Pi}$: vecteur des fréquences stationaires
$\widehat{\mathbf{\Pi}}$: vecteur des fréquences observées
$\mathbf{R}$: matrice des taux de substitutions instantanés

- Annexe A -

Efficient Biased Estimation of Evolutionary Distances When Substitution Rates Vary Across Sites

Stéphane Guindon and Olivier Gascuel

Mol. Biol. Evol. 19(4) : 534-543. 2002

Efficient Biased Estimation of Evolutionary Distances When Substitution Rates Vary Across Sites

Stéphane Guindon and Olivier Gascuel
LIRMM, UMR 9928 Université Montpellier II/CNRS

This paper deals with phylogenetic inference when the variability of substitution rates across sites (VRAS) is modeled by a gamma distribution. We show that underestimating VRAS, which results in underestimates for the evolutionary distances between sequences, usually improves the topological accuracy of phylogenetic tree inference by distance-based methods, especially when the molecular clock holds. We propose a method to estimate the gamma shape parameter value which is most suited for tree topology inference, given the sequences at hand. This method is based on the pairwise evolutionary distances between sequences and allows one to reconstruct the phylogeny of a high number of taxa (>1,000). Simulation results show that the topological accuracy is highly improved when using the gamma shape parameter value given by our method, compared with the true (unknown) value which was used to generate the data. Furthermore, when VRAS is high, the topological accuracy of our distance-based method is better than that of a maximum likelihood approach. Finally, a data set of Maoricicada species sequences is analyzed, which confirms the advantage of our method.

Introduction

Most of the phylogenetic inference methods use an explicit model of sequence evolution. Such a model includes parameters whose values must be estimated. Among these parameters, the variability of substitution rates across sites (VRAS) has been widely studied in the past and remains an important subject in the phylogenetic tree inference domain. Indeed, VRAS is widespread among biological sequences. For example, Sullivan, Holsinger, and Simon (1995) and Yang and Kumar (1996) provided evidence that VRAS occurs in rodent 12S RNA and the D-loop sequences in mitochondrial genomes of many different vertebrates. Rzhetsky, Kumar, and Nei (1995) also built a specific model to describe VRAS among 16S-like ribosomal RNAs. Furthermore, VRAS has a strong effect on tree inference. Yang (1993) and Yang, Goldman, and Friday (1994) have demonstrated a significant improvement of the maximum likelihood (ML) approach (Felsenstein 1981) when the model of sequence evolution incorporates VRAS. The distance-based methods also suffer from this phenomenon. Tateno, Takezaki, and Nei (1994), using simulations with 4-taxon trees, demonstrated a poor robustness of the neighbor-joining method (Saitou and Nei 1987) when VRAS occurs but is not taken into account.

The gamma distribution is most commonly used for modeling rate variation across sites. The shape of this distribution is related to a parameter denoted as a in the text that follows. When a is less than 1, the density function is exponential-like and VRAS is high. Higher values of a (say >2) represent weak variations of substitution rates across sites. When a tends to infinity, all sites evolve at the same rate.

Key words: phylogenetic reconstruction, varying rates of substitution, distance methods, maximum likelihood, computer simulations, Maoricicada.

Address for correspondence and reprints: Olivier Gascuel, LIRMM, UMR 9928 Université Montpellier II/CNRS, 161, Rue Ada, 34392 Montpellier Cedex 5, France. E-mail: gascuel@lirmm.fr.

Mol. Biol. Evol. 19(4):534–543. 2002
 ISSN: 0737-4038

Distances between sequences can be analytically expressed for certain models of sequence evolution, depending on the gamma shape parameter. For the Kimura two-parameter model (K80) (Kimura 1980), the evolutionary distance between two sequences is given by (Jin and Nei 1990):

$$d = \frac{a}{2}\left[(1 - 2P - Q)^{-1/a} + \frac{1}{2}(1 - 2Q)^{-1/a} - \frac{3}{2}\right], \quad (1)$$

where P and Q are the probabilities to observe a transition and a transversion, respectively. An estimate of d is obtained by replacing P and Q by the frequencies of observed transitions and transversions and a by an estimate denoted as α. The expression given previously shows that, for any fixed values of P and Q, d is a decreasing function of a. Hence, when α overestimates a ($\alpha > a$), the evolutionary distance is underestimated.

Both likelihood and parsimony methods have been used to estimate the value of a. Yang (1993) extended the method of Felsenstein (1981) and included VRAS in the ML framework. The estimation of a is usually performed given a specific tree topology. However, when the correct topology is unknown, it is possible to alternate the estimation of a and the tree topology reconstruction, given the value of a. The procedure is stopped when the tree topology does not change between two steps. Unfortunately, this approach involves intensive computation and is only feasible for small data sets (say 30–40 taxa).

The estimation of a in the maximum parsimony framework also relies on a given tree topology, which is supposed to be correct. The computational burden is clearly less than that of ML. Unfortunately, the values of α obtained with this method are not reliable. Indeed, as the number of substitutions between taxa is underestimated, VRAS is underestimated too, and the value of a is overestimated.

The present paper is organized into two parts. The first deals with the best value of α for tree inference using distances. The best or optimal value of α is the value which minimizes the difference between the inferred tree topology and the true topology. Using sim-

ulations we show that evolutionary distances estimated from the true value of the gamma shape parameter are not optimal; underestimated distances provide a better topological accuracy and outperform usual unbiased distances.

In the second part of the paper, we present a method to estimate the optimal value of α. This approach is based on distance algorithms and allows one to deal with numerous taxa (say >1,000). We use simulations and real data to test the accuracy of the method. The results are presented, and finally, we discuss our approach and directions for future research.

The True Value of *a* is not Optimal

In this section we focus on the topological distance between the true tree and the inferred tree and how this depends on the value of α. We first describe our simulations and the results thereafter.

Simulations

A true phylogeny, denoted as *T*, was first generated using the stochastic speciation process described by Kuhner and Felsenstein (1994). The number of taxa was set to 20 and the branch length expectation to 0.03 mutations per site. Using this generating process makes *T* ultrametric (or molecular clock–like). This hypothesis does not hold in most biological data sets, so we created a deviation from the molecular clock. Every branch length of *T* was multiplied by a gamma distributed factor. The mean of the gamma distribution used was equal to 1.0 and the shape parameter, denoted as η, was set to 0.5 or 2.0. The ratio between the mutation rate in the fastest evolving lineage and the rate in the slowest evolving lineage was equal to 3.6 and 2.0, respectively. Therefore, $\eta = 0.5$ corresponds to a strong departure from the molecular clock, and $\eta = 2.0$ to a mild departure. The mean distance between two taxa in such phylogenies is not related to η and is approximately equal to 0.2.

For each *T* thus obtained, a unique set of 1,000-bp sequences was produced, given the pattern of speciation events and branch lengths described by the tree. The K80 model was used, with site to site rate variation following a gamma distribution. The sequences were generated using Seq-Gen (Rambaut and Grassly 1997), with a transition-transversion ratio (TS/TV) of 2.0 and equal base frequencies. Two values for *a* have been tested: 0.1 and 0.7. These values correspond to the first and the third quartiles of the distribution of a series of ML estimates of *a*, which were obtained from the analysis of 16 data sets by Yang (1996). Therefore, 0.1 represents a rather high VRAS, whereas 0.7 corresponds to a medium-low VRAS.

For each sequence set so obtained, several matrices (δ_{ij}^{α}) were computed, depending on the α value used to correct the distances. The values of α flanked the true value *a*. For $a = 0.1$, the values of α lay between 0.09 and 2.0, whereas for $a = 0.7$ the values of α lay between 0.6 and 4.0.

For each distance matrix (δ_{ij}^{α}), a phylogeny, denoted as $\mathcal{T}^{\alpha}$, was inferred using BIONJ (Gascuel 1997). Simulations have been done with other tree building methods, but the results were similar to those presented in this paper. The topology of $\mathcal{T}^{\alpha}$ was then compared with that of the true tree *T* using a topological distance equivalent to that of Robinson and Foulds (1979). It is defined by the proportion of internal branches (or bipartitions) that are found in one tree and not in the other one. This distance varies between 0.0 (both topologies are identical) and 1.0 (they do not share any internal branch). The Robinson and Foulds distance between *T* and $\mathcal{T}^{\alpha}$ is denoted as $RF(T, \mathcal{T}^{\alpha})$ in the text that follows.

We then defined the optimal value of α as the value that minimizes the mean of $RF(T, \mathcal{T}^{\alpha})$, denoted as $\overline{RF}(T, \mathcal{T}^{\alpha})$, given the experimental condition at hand (corresponding here to the values of η and *a*). This optimal value is denoted as α^{opt} and is formally defined as:

$$\alpha^{opt} = \underset{\alpha \in \mathbb{R}^{+}}{\operatorname{argmin}} \left(\overline{RF}(T, \mathcal{T}^{\alpha})\right). \quad (2)$$

Therefore, α^{opt} corresponds to the value that ensures the lowest average topological distance between the true tree *T* and the inferred tree $\mathcal{T}^{\alpha}$, given the conditions at hand.

Results

Figure 1 shows the mean topological distance between the true tree and the inferred tree $(\overline{RF}(T, \mathcal{T}^{\alpha}))$ as a function of the value of α. When the deviation from the molecular clock is strong ($\eta = 0.5$), α^{opt} is close to *a* but remains systematically higher. The difference between α^{opt} and *a* increases when the molecular clock is better satisfied ($\eta = 2.0$). When the molecular clock holds (results not shown), $\overline{RF}(T, \mathcal{T}^{\alpha})$ is a monotonic decreasing function of α, and α^{opt} tends to infinity. In this case, the best topological accuracy is obtained using noncorrected distances, even if VRAS occurs in sequences.

Therefore, underestimated distances outperform unbiased distances when the molecular clock holds. Steel and Penny (2000) showed that, in this case, the correct topology is induced by any monotonic increasing function of the true distances, in particular the Hamming distance between infinite length sequences. Hence, when the noise affecting distance estimates is sufficiently low, the true topology can be retrieved with a high probability even if the distances are not corrected or underestimated. However, this property does not hold when the true distances are not ultrametric, i.e., when the molecular clock is not satisfied.

Such a demonstration explains why correct tree topologies can be retrieved with biased distances. However, it does not explain why, when the molecular clock holds, underestimated distances provide a better topological accuracy than unbiased distances. A widespread idea is that this phenomenon is caused by a decrease in the variance of the distance estimates (Saitou and Nei 1987; Sourdis and Nei 1988; Zharkikh and Li 1993; Schöniger and von Haesler 1993; Tajima and Takezaki 1994; Takahashi and Nei 2000). Because overestimating

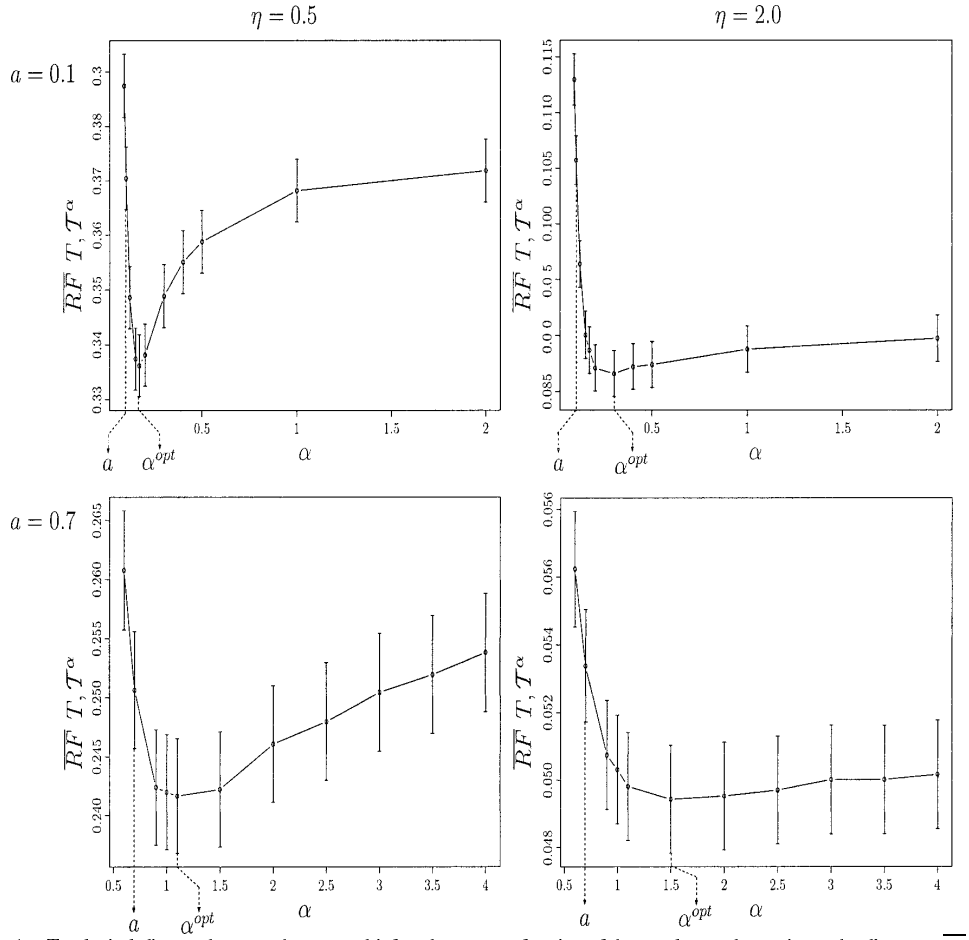


FIG. 1.—Topological distance between the true and inferred trees as a function of the α value used to estimate the distances. $\overline{RF}(T, \mathcal{T}^\alpha)$, average Robinson and Foulds distances between T and $\mathcal{T}^\alpha$; a, true value of the gamma distribution parameter, which is used to generate the data; η, parameter measuring the deviation from molecular clock; α^{opt}, value of α which minimizes $\overline{RF}(T, \mathcal{T}^\alpha)$. Each value of $\overline{RF}(T, \mathcal{T}^\alpha)$ is found by averaging over 1,000 simulated 20-taxon data sets.

a leads to underestimating distances, hence, to a decrease in the variances of the estimates, this explanation could hold there. However, this point remains to be formally demonstrated.

Another interesting point is the comparison between curves for $a = 0.1$ and $a = 0.7$. The region surrounding α^{opt} is indeed much flatter for $a = 0.7$ than for $a = 0.1$. This phenomenon is caused by a shape property of the gamma distribution. When a is small (e.g., near 0.1), the variation of α around a induces a strong variation of distance estimates, and perturbations of tree topologies follow. When a is higher (e.g., =0.7), the variation of α around a produces a small variation of distance estimates, and tree topologies remain more stable. In this case, a large range of values of α around α^{opt} give the same topology as the one obtained with α^{opt}.

In conclusion, the optimal value of the gamma distribution parameter is always higher than the real value of this parameter, and this deviation is the largest when the molecular clock holds.

Approximating α^{opt}

As the topology of T is unknown and represents what is searched for, the value of α^{opt} cannot be estimated from equation (2). We propose in this section a criterion, denoted as Q, to approximate α^{opt}. Q measures the reliability of the inferred tree. The approximation of α^{opt} is denoted as α^* and corresponds to the most reliable tree in the sense of Q. The formal definition of α^* is analogous to equation (2), that is,

$$\alpha^* = \underset{\alpha \in \mathbb{R}^+}{\operatorname{argmin}} \left(Q((\delta_{ij}^\alpha), \mathcal{T}^\alpha)\right). \quad (3)$$

We first describe the computation of Q with four taxa and then for a higher number of taxa. The average ac-

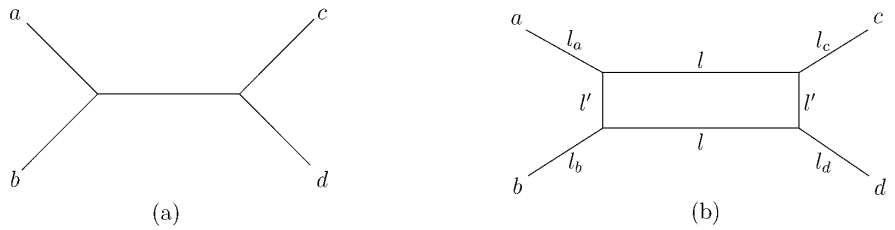


FIG. 2.—Exact representation of estimated distances. *a,* inferred tree. *b,* exact graphical representation of the six estimated distances δ_{ab}, δ_{cd}, δ_{ac}, δ_{bd}, δ_{ad}, and δ_{bc}; we have $L = \delta_{ad} + \delta_{bc} = (l_a + l + l' + l_d) + (l_b + l' + l + l_c)$, $M = \delta_{ac} + \delta_{bd} = (l_a + l + l_c) + (l_b + l + l_d)$, and $Q = L - M = 2l'$.

curacy of Q is found using simulations. Finally, we present a new tree inference method based on this criterion.

Definition of the Criterion with 4-Taxon Trees

Take four taxa denoted as *a, b, c,* and *d,* and the six distances δ_{ab}, δ_{ac}, δ_{ad}, δ_{bc}, δ_{bd}, and δ_{cd}. Assume: $(\delta_{ab} + \delta_{cd}) < (\delta_{ac} + \delta_{bd}) \leq (\delta_{ad} + \delta_{bc})$. The three terms of this inequality are denoted as S (Small), M (Median) and L (Large), respectively. Given this inequality, most of the distance-based methods (in particular BIONJ that is used here) infer the same unrooted topology, denoted as $\{a, b\}/\{c, d\}$ and shown in figure 2*a*. In this case, S can also be defined as the sum of the distances between the two external pairs (external pairs are made of two taxa separated by a single node).

Because of random noise, the fit of the distance estimates to a tree distance is almost always imperfect. However, the graph (Bandelt and Dress 1992) of figure 2*b* provides an exact representation of the six distance estimates. In this graph, the distance between two taxa is equal to the length of the path that separates them, e.g., $\delta_{ad} = l_a + l + l' + l_d$. The set of equations, in which each estimated distance is expressed as a sum of edge lengths, has six degrees of freedom corresponding to l, l', l_a, l_b, l_c, and l_d. Hence, one can express the edge lengths as linear combinations of the distances. In particular, $l = (L - S)/2$ and $l' = (L - M)/2$.

When the fit of the distance estimates to the tree $\{a, b\}/\{c, d\}$ is perfect, $l' = 0$ and the graph of figure 2 becomes a tree. In this case, the 4-point condition (Zaretskii 1965; Buneman 1971) holds, and $L = M$. As explained previously, this situation is not encountered in most real data sets and the edge l' has a positive length.

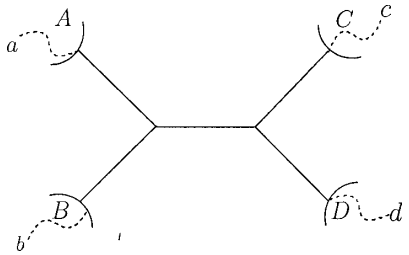


FIG. 3.—Subtrees associated with an internal edge. Each internal edge is associated with four subtrees denoted *A, B, C,* and *D; a, b, c,* and *d* are taxa belonging to these subtrees.

If l' is small compared with l, the support for the topology $\{a, b\}/\{c, d\}$ is higher than that for $\{a, c\}/\{b, d\}$. If l and l' are close, one cannot clearly choose between $\{a, b\}/\{c, d\}$ and $\{a, c\}/\{b, d\}$. Note that this uncertainly is not necessarily translated into a small internal branch length in the inferred tree (at least, when using least squares branch length estimates). If $l \simeq l' \simeq 0$, the internal edge of the inferred tree is close to zero, and the data support a star tree.

The Q criterion is then:

$$Q = 2l' = L - M. \quad (4)$$

Hence, Q assesses the reliability of the inferred internal edge. This criterion also measures the fit of the distance estimates to a tree distance: the larger the value of Q, the more the distance estimates differ from a tree distance.

Definition of the Criterion with *n*-Taxon Trees

Let n, the number of taxa, be larger than four. Each of the $n - 3$ internal branches of the inferred tree defines four subtrees, denoted as *A, B, C,* and *D* (fig. 3). Let $\overline{\delta_{AB}}$ be the mean of the estimated distances between subtree *A* and subtree *B,* i.e.,

$$\overline{\delta_{AB}} = \frac{\sum_{a \in A} \sum_{b \in B} \delta_{ab}}{n_A \cdot n_B},$$

where n_A and n_B are the numbers of taxa in subtrees *A* and *B,* respectively (fig. 3). Let $\overline{\delta_{AC}}$, $\overline{\delta_{AD}}$, $\overline{\delta_{BC}}$, $\overline{\delta_{BD}}$, and $\overline{\delta_{CD}}$ be defined in the same way. S, M, and L now correspond to $(\overline{\delta_{AB}} + \overline{\delta_{CD}})$, $(\overline{\delta_{AC}} + \overline{\delta_{BD}})$, and $(\overline{\delta_{BC}} + \overline{\delta_{AD}})$, respectively. S is then defined by both external pairs. S is also the smallest of the three sums in most practical cases (99% of cases with the data sets used in the previous section, when inferring the trees with BIONJ).

The value of the criterion for the focused edge is then obtained using equation (4). The value of the criterion for the whole tree is equal to its mean value for every internal branch. However, as the criterion only makes sense when branches have positive length, the negative or null branches are not taken into account. Q is then a global measure of internal branch reliability. The value of Q is null when the distances are tree-like. Therefore, assuming that the evolutionary model used to estimate the distances is satisfied, α^* converges to the true value α of the gamma shape parameter when the sequence length increases.

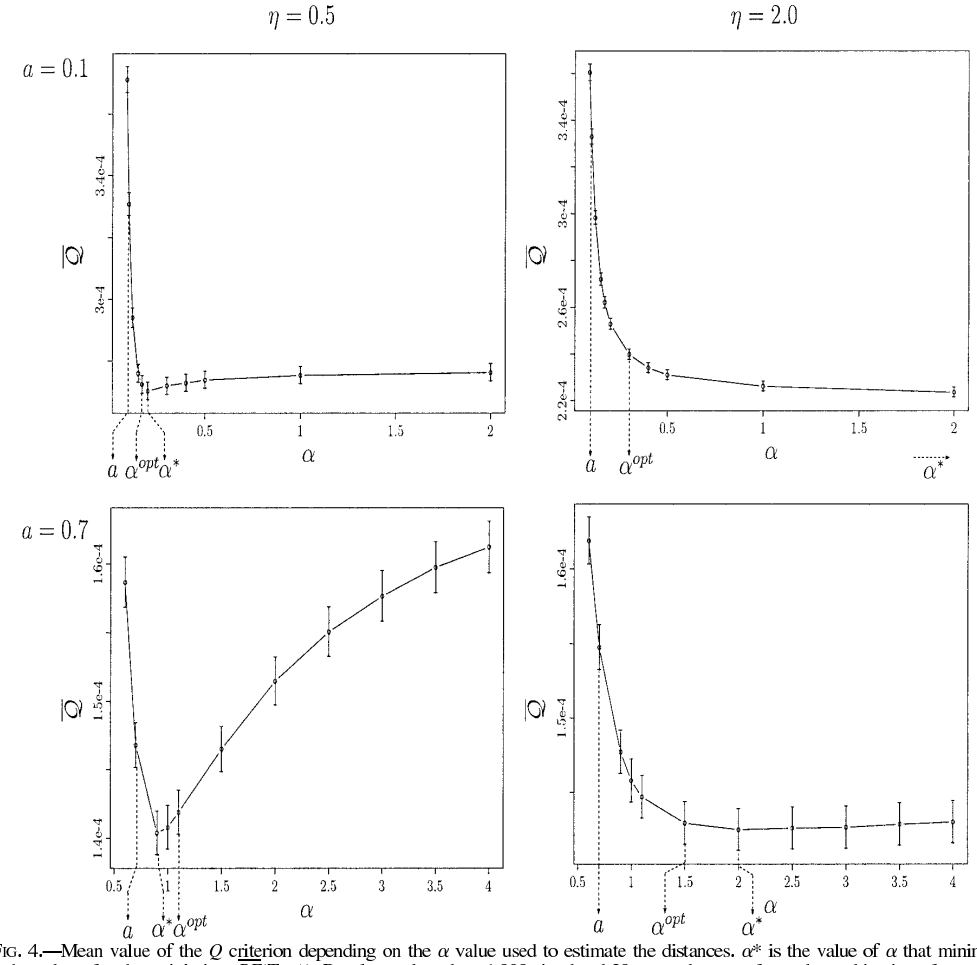


FIG. 4.—Mean value of the Q criterion depending on the α value used to estimate the distances. α^* is the value of α that minimizes $\overline{Q}$. α^{opt} is the value of α that minimizes $\overline{RF}(T, \mathcal{T}^\alpha)$. Results are based on 1,000 simulated 20-taxon data sets for each combination of η and a.

The time complexity of the computation of Q for one branch is equal to $O(n^2)$ in the worst case ($n_A = n_B = n_C = n_D = n/4$). The worst case complexity for n taxa is then $O(n^3)$, but in practice it is often lower. This worst-case time complexity is equal to that of NJ-like tree building algorithms, so the Q criterion can be used with large data sets. For example, with $n = 500$, the computing time to build a tree using BIONJ is equal to 11.43 s, whereas the time to compute Q is equal to 2.21 s (PentiumIII, 750 MHz).

Mean Performance of Q in Approximating α^{opt} Using Simulations

The performance of Q is shown in figure 4. The curves are obtained in the same manner as the ones in figure 1; but instead of the Robinson and Foulds distance, the ordinate reports now the value of the Q criterion. This value is averaged over 1,000 data sets for each experimental condition, and α^* is obtained by considering the mean values of Q and not a single value as used in equation (4). Therefore, figure 4 provides a view on the mean accuracy of Q in approximating α^{opt}.

The curves of figures 4 and 1 are similar, and $\overline{Q}$ appears to be relatively accurate in approximating α^{opt}. However, when $a = 0.1$ and $\eta = 2.0$, which corresponds to a strong VRAS and a moderate deviation from the molecular clock, the curve of figure 4 is a monotonic decreasing function and α^* tends to infinity, whereas $\alpha^{opt} \simeq 0.3$. In spite of this difference, $RF(T, \mathcal{T}^{\alpha^*}) = 0.1893$ is close to $\overline{RF}(T, \mathcal{T}^{\alpha^{opt}}) = 0.1951$, whereas the accuracy obtained with a is much lower: $\overline{RF}(T, \mathcal{T}^a) = 0.2296$. Indeed, table 1 indicates that $\overline{RF}(T, \mathcal{T}^{\alpha^*})$ is always inferior to $\overline{RF}(T, \mathcal{T}^a)$ and very close to $\overline{RF}(T, \mathcal{T}^{\alpha^{opt}})$. Therefore, the performance of α^* in reconstructing T is similar to that of α^{opt}, even when α^* is remote from α^{opt}. However, results in table 1 have to be interpreted carefully as the values of α^* are obtained from 1,000 data sets which are generated under the same evolutionary conditions, whereas parameter estimation in

Table 1.
Topological Accuracy of the Inferred Tree when Distances are Corrected with a, α^{opt} or α^*

Γ param.		$\eta = 0.5$	$\eta = 2.0$
$a = 0.1$	$\overline{RF}(T, \mathcal{T}^a)$	0.3704	0.2296
	$\overline{RF}(T, \mathcal{T}^{\alpha^{opt}})$	0.3362	0.1893
	$\overline{RF}(T, \mathcal{T}^{\alpha^*})$	0.3381	0.1951
$a = 0.7$	$\overline{RF}(T, \mathcal{T}^a)$	0.2506	0.1134
	$\overline{RF}(T, \mathcal{T}^{\alpha^{opt}})$	0.2416	0.1122
	$\overline{RF}(T, \mathcal{T}^{\alpha^*})$	0.2424	0.1127

NOTE.—η, parameter measuring the deviation from molecular clock; a, true value of the gamma distribution parameter, which is used to generate the data; α^{opt}, optimal value of α; α^*, our approximation of α^{opt}; $\overline{RF}(T, \mathcal{T}^a)$, $\overline{RF}(T, \mathcal{T}^{\alpha^{opt}})$, and $\overline{RF}(T, \mathcal{T}^{\alpha^*})$, average topological accuracies that are obtained with a, α^{opt}, and α^*, respectively. Results are based on 1,000 simulated 20-taxon data sets for each combination of η and a.

the frame of phylogenetic inference is done from a single data set. A better view of the performance of Q is given in the results section.

It must be underlined that numerous other criteria have been tested in this study (e.g., Eigen and Winkler-Oswatitsch 1981; Vach 1992; Guénoche and Garreta 2001), but none of these performed as well as Q.

Using Q for Phylogenetic Inference

Given a set of homologous sequences, several (δ_{ij}^{α}) distance matrices are computed. The α values are obtained from a predefined sample with size r. In this study, the r values of α ranged from 0.1 to 5,000. Between 0.1 and 3.0, the step was equal to 0.02, between 3.0 and 10, to 0.1, whereas the remaining α values were 10, 50, 100, 500, 1,000 and 5,000. These increasing steps are explained by the necessity to concentrate on the area where a small variation of α likely involves some perturbations in the inferred topology. The calculation of the different (δ_{ij}^{α}) matrices is very fast. Indeed, the transition, transversion, and identity frequencies are computed only once, which requires $O(n^2 l)$ computing time where l is the sequence length. The (δ_{ij}^{α}) distances matrices are obtained by correcting these three frequencies with the corresponding α values using equation (1) in the case of K80 model; the computational burden for the r matrices is then equal to $O(n^2 r)$. The $\mathcal{T}^{\alpha}$ phylogenies are inferred from the (δ_{ij}^{α}) distance matrices using BIONJ (Gascuel 1997). The values of Q for the various values of α are then computed using both the (δ_{ij}^{α})'s and $\mathcal{T}^{\alpha}$'s. Finally, we select the tree $\mathcal{T}^{\alpha^*}$ that minimizes $Q((\delta_{ij}^{\alpha}), \mathcal{T}^{\alpha})$ among the r inferred trees. The whole time complexity is equal to $O(n^2 l + n^2 r + n^3 r)$, where the three terms correspond to: (1) counting the observed mutations, (2) computing the distance matrices, and (3) inferring the trees and computing Q. Practical computing times are given in the next section, and a PHYLIP compatible program, called GAMMA, is available from http://www.lirmm.fr/~w3ifa/MAAS/.

Results

We first compare the performance which is obtained using our approximation, α^*, to the performance that would be obtained if the true value of a was known. Then, we compare the topological accuracy of our method with the one of ML (Felsenstein 1981; Yang 1993). Finally, we illustrate our approach using sequences from Maoricicada species (Buckley, Simon, and Chambers 2001).

$\mathcal{T}^{\alpha^*}$ versus $\mathcal{T}^a$

We performed simulations in a way similar to that described previously. Three deviations from the molecular clock were used: $\eta = 0.5$ and $\eta = 2.0$, as previously, while the molecular clock (*MC*) held in the third case. The evolution of the sequences along the trees was simulated using three values of the a gamma shape parameter: 0.1, 0.7, and 2.0. The sequences were 300 or 1,000 bp long, and each data set contained 20 taxa. For each of these data sets, two trees were inferred. The first was built with BIONJ from the (δ_{ij}^{a}) matrix, where a was the value used to generate the sequences. The second was built with BIONJ by using the $(\delta_{ij}^{\alpha^*})$ matrix, where α^* was the value computed by our method. Both inferred topologies were compared with the true topology T. We then obtained the two topological distances $RF(T, \mathcal{T}^a)$ and $RF(T, \mathcal{T}^{\alpha^*})$ and computed the average (denoted as $\overline{RF}$) of these distances over 4,000 data sets with a and η being fixed. For each of the experimental conditions, we also computed the relative error decrease induced by the use of α^* instead of the (unknown) true value a. This corresponds to the ratio $[\overline{RF}(T, \mathcal{T}^{\alpha^*}) - \overline{RF}(T, \mathcal{T}^a)]/\overline{RF}(T, \mathcal{T}^a)$, which is negative when α^* performs better than a. Finally, a sign test was used to check the statistical significance of our findings.

The results are displayed in table 2. With 300-bp sequences, the three topologies inferred using α^* present less errors than those inferred using a, whatever the values of a and η. The best results occur when VRAS is strong ($a = 0.1$) and when the molecular clock holds. In this case, the relative decrease in topological error is close to 30%, which is highly significant and corresponds to much better inferred topologies.

For 1,000-bp sequences, the results are similar. However, the relative decrease in topological error is lower than before for seven of the nine experimental conditions. When $a = 2.0$ and $\eta = 2.0$, the topological accuracy is better using a than α^*, but the difference is not statistically significant. On the other hand, we still obtain an error decrease of about 30% in some cases (e.g., *MC* and $a = 0.1$). For longer sequences, the performances of α^* and a should become close, simply because (δ_{ij}^{a}) tends to be tree-like; therefore, α^* becomes close to a.

In conclusion, our method is remarkably accurate because its results are better than those that would be obtained if the real value of the gamma shape parameter was known. Its relative topological accuracy increases when VRAS is strong and when the deviation from the molecular clock is slight or null.

Table 2.
Topological Accuracy of *a* Versus α*

	Length		$\overline{RF}$ (*RED*)		
			η = 0.5	η = 2.0	MC
a = 0.1 ...	300 bp	[BIONJ + *a*]	0.499	0.389	0.332
		[BIONJ + α*]	0.438 (−12.2%)*	0.298 (−23.2%)*	0.238 (−28.3%)*
	1,000 bp	[BIONJ + *a*]	0.367	0.226	0.165
		[BIONJ + α*]	0.351 (−4.4%)*	0.190 (−15.9%)*	0.116 (−29.8%)*
a = 0.7 ...	300 bp	[BIONJ + *a*]	0.367	0.223	0.165
		[BIONJ + α*]	0.352 (−4.1%)*	0.204 (−8.4%)*	0.146 (−11.5%)*
	1,000 bp	[BIONJ + *a*]	0.257	0.116	0.075
		[BIONJ + α*]	0.254 (−0.9%)*	0.114 (−1.8%)*	0.066 (−11.4%)*
a = 2.0 ...	300 bp	[BIONJ + *a*]	0.334	0.190	0.134
		[BIONJ + α*]	0.328 (−2.0%)*	0.184 (−2.8%)*	0.130 (−3.0%)*
	1,000 bp	[BIONJ + *a*]	0.227	0.096	0.061
		[BIONJ + α*]	0.226 (−0.5%)*	0.097 (+0.9%)	0.058 (−4.0%)*

NOTE.—η, parameter measuring the deviation from molecular clock; MC, the molecular clock holds in the true tree; *a*, true value of the gamma shape parameter; [BIONJ + *a*], the trees are inferred with BIONJ from distances corrected with *a*; [BIONJ + α*], the trees are inferred with BIONJ from distances corrected with α*; $\overline{RF}$, mean values of $RF(T, \mathcal{T}^a)$ and $RF(T, \mathcal{T}^{\alpha^*})$ computed from 4,000 20-taxon data sets; *RED*, relative error decrease between $\overline{RF}(T, \mathcal{T}^a)$ and $\overline{RF}(T, \mathcal{T}^{\alpha^*})$; the statistical significance of each value is checked with the sign test: * → $P \leq 0.05$.

$\mathcal{T}^{\alpha^*}$ versus $\mathcal{T}^{ML}$

The results of our approach are now compared with that of ML. We used DNAML from the PHYLIP package (Felsenstein 1989) to build the ML trees. VRAS was modeled by a four category discretized gamma distribution using the true value *a* of the gamma shape parameter. In the same way, the TS/TV ratio was set to its real value, i.e., 2.0. Under such conditions, ML likely performs better than if *a* and TS/TV were unknown and had to be estimated from the sequences.

The values of η and *a* were identical to the previous ones, the sequences were 300 bp long and each data set contained 20 taxa. We computed the mean Robinson and Foulds distance between the true tree and the ML tree, $\overline{RF}(T, \mathcal{T}^{ML})$, and the relative deviation $[\overline{RF}(T, \mathcal{T}^{\alpha^*}) - \overline{RF}(T, \mathcal{T}^{ML})]/\overline{RF}(T, \mathcal{T}^{ML})$ assessed the difference of performance between our method and ML.

The results are displayed in table 3. When VRAS is strong (*a* = 0.1), the tree topology inference is better using BIONJ with α* than ML with *a*. For example, when the molecular clock holds, the relative decrease in topological error is about 26% with our method. When *a* = 0.7 and *a* = 2.0, this property does not hold anymore. For example, ML trees are better than ours by about 12%–15%, when *a* = 2.0, which corresponds to a low VRAS. However, it must be underlined that ML trees are likely less accurate in real cases where *a* and the TS/TV ratio are unknown.

As phylogenetic inference methods are sensitive to the number of taxa analyzed, we have done supplementary simulations with 10-taxon trees. The results are similar to those obtained with 20-taxon trees, that is, the tree topology inference is better using BIONJ with α* than ML with *a* when VRAS is strong (*a* = 0.1), irrespective of the deviation from the molecular clock. The mean relative error decrease averaged over the three values of η is then close to 18%, in favor of our method (17% with 20-taxon trees). On the other hand, the topological accuracy is better with ML than with our method for *a* = 0.7 and *a* = 2 with a mean relative error decrease close to 14% in favor of ML trees (8% with 20-taxon trees).

Simulations with more than 20-taxon trees have not been carried out as it takes more than 1 week to run the

Table 3.
Comparison of our Approach and Maximum Likelihood

		$\overline{RF}$ (*RED*)		
		η = 0.5	η = 2.0	MC
a = 0.1 ...	[ML + *a*]	0.471	0.365	0.325
	[BIONJ + α*]	0.433 (−7.92%)*	0.303 (−16.78%)*	0.238 (−26.74%)*
a = 0.7 ...	[ML + *a*]	0.317	0.184	0.139
	[BIONJ + α*]	0.344 (+8.54%)*	0.192 (+4.38%)	0.141 (+1.84%)
a = 2.....	[ML + *a*]	0.311	0.171	0.118
	[BIONJ + α*]	0.337 (+8.45%)*	0.197 (+14.76%)*	0.132 (+11.63%)*

NOTE.—[ML + *a*], trees are inferred by maximum likelihood using the true value *a* of the gamma shape parameter; [BIONJ + α*], trees are inferred with BIONJ from distances corrected with α*; $\overline{RF}$, averages of $RF(T, \mathcal{T}^{ML})$ and $RF(T, \mathcal{T}^{\alpha^*})$; *RED*, relative error decrease between $RF(T, \mathcal{T}^{ML})$ and $RF(T, \mathcal{T}^{\alpha^*})$; the statistical significance of these values are checked with sign tests (* → $P < 0.05$). Results are based on 300 simulated 20-taxon data sets for each combination of η and *a*.

Table 4.
Computing Times Required by our Method and by DNAML

	[BIONJ + α*]	[ML + *a*]
n = 100	38.6 s	>3 days
n = 50	5.7 s	≃6 h
n = 20	0.8 s	≃15 min

NOTE.—*n*, number of taxa. The given values represent the time needed to infer one phylogeny with *n* taxa. These experiments have been performed with a PentiumIII, 750 Mhz computer.

tests with 20-taxon trees. Most of this computational time amount is caused by the building of ML trees. We have done supplementary simulations to compare more precisely the computational time required by both methods. We measured the time needed on a PentiumIII, 750 MHz computer by both methods to infer 20-, 50-, or 100-taxon trees from data sets being generated as described previously. Results are given in table 4. Our method is clearly more efficient than ML. For example, with 50 taxa, our method requires ≃6 s, whereas ML requires ≃6 h. This clearly precludes to bootstrap the data in the case of ML, whereas this task is easily achieved when using our method. Moreover, 3 days of computation are needed by ML with 100-taxon trees, which make its use rather unrealistic, whereas our method only requires ≃40 s.

We did not use fastDNAml (Olsen et al. 1994) (which is faster than DNAML) because it does not have the ability to handle the gamma distribution. However, refined implementations of DNAML, for example based on ideas from fastDNAml, would significantly reduce the computing times given here (although remaining much slower than our distance-based method).

Application to Maoricicadas Sequences

To illustrate our approach, we analyzed 25 orthologous sequences of the Maoricicada species (Buckley, Simon, and Chambers 2001). These sequences are 1,520 bp long and contain two mitochondrial regions which have been concatenated. The first is the COI gene, the second is the region from the $tRNA^{Asp}$, A8 and A6 genes. This data set was previously collected and analyzed by Buckley, Simon, and Chambers (2001) and Buckley et al. (2001). These authors used and compared different models of substitution and rate heterogeneity. All the variants of the Jukes and Cantor (1969), Kimura (1980), and Hasegawa, Kishino, and Yano (1985) models were rejected against the variants of the general-time reversible (GTR) model (Yang 1994). The rate heterogeneity model with best fit was obtained when partitioning the characters into first, second, and third codon positions and all $tRNA^{Asp}$ sites and then estimating the gamma shape parameter separately for each of the four categories (Γ_4 model). The ML estimate of α was equal to 0.168 when considering all sites together. Hence, Maoricicada sequences seem to follow a more sophisticated pattern of evolution than simple models, such as Jukes and Cantor's or Kimura's, and VRAS is relatively strong in these sequences. Moreover, the ML tree that is inferred presents a moderate deviation from molecular clock (figure 6 in Buckley et al. 2001).

Our method was used in the same way as previously described (i.e., K80 model and $0.1 < \alpha < 5{,}000$). We obtained for α^* a value of 5,000 ($\simeq\infty$) which implies that the fit of the estimated distances to a tree distance is optimal when VRAS is not taken into account. The phylogeny inferred with BIONJ, given the ($\delta_{ij}^{\alpha^*}$) matrix is shown in figure 5. The topology of this tree is similar to the one inferred with ML using the GTR + Γ_4 model, but three differences appear. The first difference concerns the position of the two *M. cassiope* species. These two sequences and both of *M. tenuis* constitute a monophyletic clade in the tree of Buckley et al. (2001). However, this clade is not well supported by the data, so the position of *M. cassiope* in our tree is also a plausible one (T. Buckley, personal communication). In the same manner, the position of *M. phaeoptera* differs in the two trees, but neither of these two positions is well supported. The third difference is more interesting and concerns the monophyly of the three *M. campbelli* sequences. This monophyly is retrieved in our tree but not by the tree of Buckley et al. (2001), despite it being very likely for several biological reasons (T. Buckley, personal communication). Note that this monophyletic clade is not recovered by BIONJ when using K80-distances and $\alpha = 0.168$, the ML value of *a*. Even if the bootstrap proportion corresponding to this clade is not very high (0.478, against 0.384 for Buckley et al.'s clade), it is worth noting that this biologically likely fact is retrieved, despite an apparently low amount of information in the data.

In summary, the Maoricicada tree inferred using our method is close to the ML tree but also proposes an original and biologically relevant group of taxa. It should be noted that the sequences present a strong VRAS and a low deviation from *MC,* which likely explains our good results (see the previous comparison between our method and ML).

Conclusions

This paper contains two main parts. In the first part, we show that the best value of α (α^{opt}) for tree inference from evolutionary distances is not equal to its true value (*a*). The lower the deviation from the molecular clock, the larger α^{opt} is relative to *a* and the more the optimal distances underestimate the true distances. This finding corroborates the observations from many authors (e.g., Saitou and Nei 1987; Sourdis and Krimbas 1987; Tajima and Takezaki 1994), established under many different experimental conditions without VRAS, where uncorrected/corrected for multiple substitutions distances were compared.

Given these observations, we propose a method to approximate the optimal value of α. We use a criterion that measures the reliability of the inferred tree, and our approximation (α^*) corresponds to the value which optimizes this criterion. Simulation results demonstrate the topological accuracy of our method because performance is better using α^* than using the (unknown) true

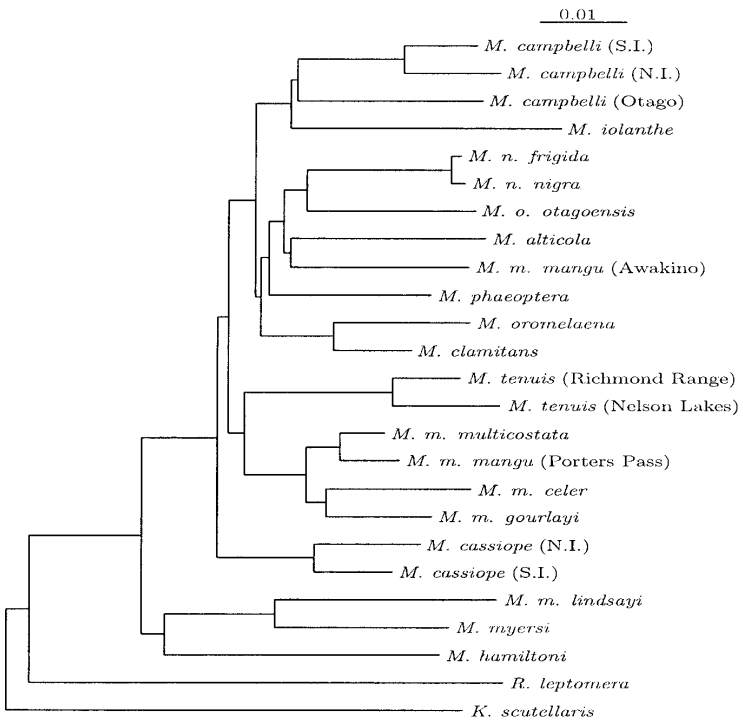


FIG. 5.—Phylogeny of Maoricicada species. This tree has been built with BIONJ from distances estimated with the K80 model and our $\alpha^*(\simeq\infty)$value.

value a. In numerous realistic experimental conditions, we obtain a relative decrease in topological error of about 30%. The comparison with the ML approach leads to unexpected results. Indeed, when VRAS is strong, our method seems to be more efficient than ML. This result is of importance because the always increasing amount of biological data confirms that VRAS is widespread and often very strong, notably in the first and second codon positions (Buckley et al. 2001). Moreover, our analysis of the Maoricicada sequences shows that correcting the distances by α^* yields a plausible topology with biologically likely clades which are not retrieved by ML and more sophisticated models.

As pointed out before, different authors have already described the improvement of topology inference induced by underestimating evolutionary distances when the molecular clock holds. However, no fully convincing explanation of this phenomenon has been given so far. A line of approach could be to extend some of the ideas presented by Rzhetsky and Sitnikova (1996).

In this study we compared various criteria to estimate α^*, and we selected the criterion that best performed in simulations. However, other criteria and other tree building algorithms could be combined to achieve better performance. Moreover, the approach presented here could likely be used to estimate other parameters involved in sequence evolution models.

Acknowledgments

Thanks to Thomas Buckley for providing Maoricicada data and for his comments on our findings, and to Nicolas Galtier, Olivier Elemento, and Andy McKenzie for their suggestions for improvement of the paper.

LITERATURE CITED

BANDELT, H.-J., and A. DRESS. 1992. Split decomposition: a new and useful approach to phylogenetic analysis of distance data. Mol. Biol. Evol. **1**:242–252.

BUCKLEY, T. R., C. SIMON, and G. K. CHAMBERS. 2001. Exploring among-site rate variation models in a maximum likelihood framework using empirical data: effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. Syst. Biol. **50**:67–86.

BUCKLEY, T. R., C. SIMON, H. SHIMODAIRA, and G. K. CHAMBERS. 2001. Evaluating hypotheses on the origin and evolution of the New Zealand Alpine Cicadas (Maoricicada) using multiple-comparison tests of tree topology. Mol. Biol. Evol. **18**:223–234.

BUNEMAN, P. 1971. The recovery of trees from measures of dissimilarity. Pp. 387–395 *in* F. R. HODSON, D. G. KENDALL, and P. TAUTA, eds. Mathematics in archeological and historical sciences. University Press, Edinburgh.

EIGEN, M., and R. WINKLER-OSWATITSCH. 1981. Transfer-RNA: the early adaptor. Die Naturwissenschaften **68**:217–228.

FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17**:368–376.

———. 1989. PHYLIP (phylogeny inference package). Version 3.2. Cladistics **5**:164–166.

GASCUEL, O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. **14**:685–695.

GUÉNOCHE, A., and H. GARRETA. 2000. Can we have confidence in a tree representation? Pp. 45–56 *in* O. GASCUEL and M.-F. SAGOT, eds. Computational biology, LNCS 2066. Springer, Berlin.

HASEGAWA, M., H. KISHINO, and T. YANO. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial-DNA. J. Mol. Evol. **22**:160–174.

JIN, L., and M. NEI. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol. Biol. Evol. **7**:82–102.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–132 *in* H. N. MUNRO, ed. Mammalian protein metabolism, Vol. III, Chap. 24. Academic Press, New York.

KIMURA, M. 1980. A simple method for estimating evolutionary rates base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

KUHNER, M., and J. FELSENSTEIN. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. **11**:459–468.

OLSEN, G. J., H. MATSUDA, R. HAGSTROM, and R. OVERBEEK. 1994. fastDNAml: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Comput. Appl. Biosci. **10**:41–48.

RAMBAUT, A., and N. GRASSLY. 1997. Seq-Gen: an application for the Monte-Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. **13**:235–238.

ROBINSON, D. F., and L. R. FOULDS. 1981. Comparison of phylogenetic trees. Math. Biosci. **53**:131–147.

RZHETSKY, A., S. KUMAR, and M. NEI. 1995. Four-cluster analysis: a simple method to test phylogenetic hypotheses. Mol. Biol. Evol. **12**:163–167.

RZHETSKY, A., and T. SITNIKOVA. 1996. When is it safe to use an oversimplified substitution model in tree-making? Mol. Biol. Evol. **13**:1255–1265.

SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

SCHÖNIGER, M., and A. VON HAESLER. 1993. A simple method to improve the reliability of tree reconstruction. Mol. Biol. Evol. **10**:471–483.

SOURDIS, J., and C. KRIMBAS. 1987. Accuracy of phylogenetic trees estimated from DNA sequence data. Mol. Biol. Evol. **4**:159–166.

SOURDIS, J., and M. NEI. 1988. Relative efficiencies of the maximum parsimony and distance-based methods in obtaining the correct phylogenetic tree. Mol. Biol. Evol. **5**: 298–311.

STEEL, M., and D. PENNY. 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol. Biol. Evol. **17**:839–850.

SULLIVAN, J., K. E. HOLSINGER, and C. SIMON. 1995. Among-site variation and phylogenetic analysis of 12s rRNA in sigmontine rodents. Mol. Biol. Evol. **12**:988–1001.

TAJIMA, F., and N. TAKEZAKI. 1994. Estimation of evolutionary distance for reconstructing molecular phylogenetic trees. Mol. Biol. Evol. **11**:278–286.

TAKAHASHI, K., and M. NEI. 2000. Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. Mol. Biol. Evol. **17**:1251–1258.

TATENO, Y., N. TAKEZAKI, and M. NEI. 1994. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. Mol. Biol. Evol. **11**:261–277.

VACH, V. 1992. The Jukes-Cantor transformation and additivity of estimated genetic distances. Pp. 141–150 *in* M. SHADER, eds. Analysing and modeling data and knowledge. Springer, Berlin.

YANG, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. **10**:1396–1401.

———. 1994. Estimating the pattern of nucleotide substitution. J. Mol. Evol. **10**:105–111.

———. 1996. Among-site rate variation and its impact on phylogenetic analyses. TREE **11**:367–372.

YANG, Z., N. GOLDMAN, and A. FRIDAY. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol. Biol. Evol. **11**:316–324.

YANG, Z., and S. KUMAR. 1996. Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. Mol. Biol. Evol. **13**: 650–659.

ZARETSKII, K. 1965. Construction d'un arbre sur la base d'un ensemble de distances entre ses feuilles. USpekHi Math. Nauk. **20**:90–92 [in Russian].

ZHARKIKH, A., and W.-H. LI. 1993. Inconsistency of the maximum parsimony method: the case of five taxa with a molecular clock. Syst. Biol. **42**:113–125.

DAN GRAUR, reviewing editor

Accepted December 6, 2001

- Annexe B -

A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood

Stéphane Guindon and Olivier Gascuel

Syst. Biol. accepté

January 10, 2003

Running head: FAST MAXIMUM LIKELIHOOD TREE INFERENCE

A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood

Stéphane Guindon and Olivier Gascuel[1]

LIRMM, CNRS

161, Rue Ada. 34392-Montpellier cedex 5-France

{guindon,gascuel}@lirmm.fr

Équipe Méthodes et Algorithmes pour la Bioinformatique

http://www.lirmm.fr/w3ifa/MAAS

tel: (33)+0 4 67 41 85 47.

fax: (33)+0 4 67 41 85 00.

Keywords: phylogeny, maximum-likelihood, algorithm, computer simulations, RDPII project, *rbc*L

[1]Corresponding author

Abstract

The increase in the number of large data sets, combined with current complex probabilistic sequence evolution models, necessitates fast and reliable phylogeny reconstruction methods. We describe a new approach, based on the maximum likelihood principle, which clearly satisfies these requirements. The core of this method is a simple hill climbing algorithm that adjusts tree topology and branch lengths simultaneously. This algorithm starts from an initial tree built by a fast distance-based method, and refines this tree so as to improve its likelihood at each step. Due to simultaneous adjustment of the topology and all branch lengths, only a few steps are sufficient to reach an optimum. Using extensive and realistic computer simulations, we show that the topological accuracy of this new method is at least as high as that of the existing maximum likelihood programs, and much higher than the performance of distance-based and parsimony approaches. Moreover, the computing time dramatically decreases in comparison with other maximum likelihood packages, while the likelihood maximization ability tends to be higher. For example, only 12 minutes are required on a standard computer to deal with a data set consisting of 500 *rbc*L sequences with 1,428 bp from plant plastids, thus reaching a speed analogous to that of some popular distance-based and parsimony algorithms. This new method is implemented in the PHYML program that is freely available on our web page.

INTRODUCTION

The size of homologous sequence data sets has dramatically increased in recent years and many of them now involve one or several hundreds of taxa. Moreover, current probabilistic sequence evolution models (Swofford et al., 1996; Page and Holmes, 1998), notably including rate variation among sites (Uzzell and Corbin, 1971; Jin and Nei, 1990; Yang, 1996), require an increasing number of calculations. Therefore, the speed of phylogeny reconstruction methods is becoming a significant criterion, as important as their accuracy (Lemmon and Milinkovitch, 2002).

The maximum likelihood (ML) approach, first introduced in the field by Felsenstein (1981), is especially accurate for building molecular phylogenies. Numerous computer studies (Huelsenbeck and Hillis, 1993; Kuhner and Felsenstein, 1994; Huelsenbeck, 1995; Rosenberg and Kumar, 2001; Ranwez and Gascuel, 2002) have shown that it can recover the correct tree from simulated data sets to a greater extent than other methods. Another important advantage of the ML approach is the ability to compare different trees and evolutionary models within a statistical framework (Whelan et al., 2001). However, like all optimality criterion-based phylogenetic reconstruction approaches, ML is hampered by computational difficulties, which makes it impossible to obtain the optimal tree with certainty even from moderate data sets (Swofford et al., 1996). So all practical methods rely on heuristics which search for near-optimal trees in reasonable computing time. Moreover, the problem is especially difficult with ML, because the tree likelihood not only depends on the tree topology but also on numerical parameters, including branch lengths, and even computing the optimal values of these parameters on a single tree is computationally hard (Chor et al., 2000).

The usual heuristic method, notably implemented in the popular PHYLIP (Felsenstein, 1993) and PAUP* (Swofford, 1999) packages, is based on hill climbing. An initial tree is constructed by stepwise addition and then subjected to a topological rearrangement or "branch swapping"; the branch lengths of the resulting tree are optimized and the tree likelihood is computed; when this new tree is more likely than the current tree, it becomes the new current tree, otherwise a new branch swapping is tested; the procedure stops when no branch swapping improves the current best tree. Despite significant decreases in computing times, notably in

fastDNAml (Olsen et al., 1994), this heuristic becomes impracticable with several hundreds of taxa. This is mainly due to the two-level strategy, which separates branch lengths and tree topology optimization. Indeed, most calculations are done to optimize the branch lengths and then to evaluate the likelihood of trees, which finally are rejected.

New methods have thus been proposed. Strimmer and von Haesler (1996), and others, assemble four taxon (quartet) trees inferred by ML, in order to reconstruct a complete tree. However, the results of this approach have not been very satisfactory to date (Ranwez and Gascuel, 2001). Ota and Li (2000,2001) describe NJML, an algorithm in the continuation of Adachi and Hasegawa's (1996). NJML first build a tree by the fast distance-based Neighbor Joining (NJ) algorithm (Saitou and Nei, 1987); the unreliable branches of this initial tree are then detected using the bootstrap procedure (Felsenstein, 1985), and resolved by computing the branch lengths and the likelihood of alternative resolutions. In this way, NJML avoids spending time on the well-supported parts the tree. Ranwez and Gascuel (2002) propose another combination of distance-based and ML approaches, with the computation speed of the former and topological accuracy midway between both. This method relies on triplets of taxa, and then share a divide-and-conquer strategy with the quartet approach.

Stochastic optimization is another attractive strategy. Bayesian methods (Rannala and Yang, 1996; Mau, 1996; Li, 1996; Simon and Larget, 2000; Huelsenbeck and Ronquist, 2001) usually rely on Markov Chain Monte Carlo (MCMC) algorithms. However, the aim is mainly to evaluate the posterior probability of alternative trees or hypotheses, rather than to speed up the computing time. Salter and Pearl (2001) describe a simulated annealing algorithm which is substantially faster than DNAML (Felsenstein, 1993) and PAUP*. This algorithm simultaneously perturbs branch lengths and tree topology, and then accelerates the computations in comparison with standard hill climbing. Finally, the genetic algorithms recently proposed by Lewis (1998) and Lemon and Milinkovitch (2002) are specially efficient. These algorithms randomly navigate in the tree space, perturbing a population of trees by modifying their branch lengths and topology, combining these trees to obtain better trees, and selecting the best trees until an optimum is reached. In this way, large phylogenies containing hundreds of taxa are obtained in few hours on a standard computer (Lemmon and Milinkovitch, 2002). Moreover,

an efficient parallel implementation of the genetic optimization approach is described by Brauer et al. (2002).

However, the hill climbing principle is usually considered faster than stochastic optimization and sufficient for numerous combinatorial optimization problems (Aarts and Lenstra, 1997), notably when the function to be optimized does not fully reflect the overall reality of the problem at hand. This feature is clearly that of phylogenetic reconstruction as we do not know the real substitution process which occurred. Moreover, despite the ever increasing size of data bases, the length of sequences used to build phylogenies remains limited, so that sampling variations in the likelihood are inevitable, even when the chosen evolutionary model fits the sequences well.

This article presents a new simple hill climbing algorithm which avoids the limits of the previous ones. The tree topology and branch lengths of a unique tree are simultaneously and progressively modified, so that the tree likelihood increases at each step until an optimum is reached. During this process, we also adjust the model parameters, such as the transition/transversion ratio or the gamma shape parameter that accounts for rate variation among sites. This algorithm is implemented in the PHYML package, which is much faster than other existing ML programs, including MetaPIGA (Lemmon and Milinkovitch, 2002). In the following, we first present this algorithm. Then we compare PHYML to other packages using extensive computer simulations and two large data sets comprising 218 and 500 taxa. It is shown that PHYML is at least equivalent to other ML programs, both in terms of topological accuracy and likelihood maximization, while having a speed analogous to some popular distance-based and parsimony methods.

Method and algorithms

The principle is to start from an initial tree, constructed by a fast distance-based algorithm, and to refine it. We first present how every branch can be adjusted, independently of the other branches, so as to maximize tree likelihood. We then show how this process extends to tree swapping, without much more computation. Branch length optimization and tree swapping define possible modifications of the current tree, and we explain how these modifications are

selected and combined. Finally, we present the whole method, including model parameter optimization. For the sake of simplicity, the method is described for nucleotide sequences, but its principle could be extended to proteins.

Branch length optimization

Let e be the branch under consideration and l its current length. U and V are the subtrees at the two extremities of e, with roots u and v (Figure 1(a)). We assume that these subtrees are fixed. We then define the conditional likelihood $L(i = h/U)$ as the probability of observing the data at site i at the tips of U, given that node u has nucleotide h. When U is reduced to taxon u, $L(i = h/U)$ is equal to 1 if site i of taxon u has nucleotide h, and 0 otherwise. $L(i = h/V)$ has the same meaning when V (and v) replaces U (and u). We also assume, as usual, that the sequence substitution model is homogeneous, stationary and time reversible. The a priori probability of nucleotide h is then denoted as π_h , while $P_{hh'}$ is the probability for the nucleotide h to become h' in interval l. With this notation, and assuming that all sites evolve at the same rate, the likelihood of the whole tree may be written as (Felsenstein, 1981):

$$L = \prod_i \sum_{h,h' \in \{A,C,G,T\}} \pi_h L(i = h/U) L(i = h'/V) P_{hh'}(l). \qquad (1)$$

Moreover, this equation is easily adapted to incorporate site to site variation using a discrete rate distribution (Yang, 1994). Equation (1) applies to internal and external branches as well. In the latter case, either U or V is reduced to a single contemporary taxon. When the conditional likelihoods $L(i = h/U)$ and $L(i = h/V)$ are known for every site, computation of Equation (1) is fast and requires $\mathcal{O}(s)$ time, where s denotes the sequence length, i.e. a time proportional to s.

Since U and V are fixed, likelihood L in Equation (1) only depends on l, and we adjust l by maximizing L. This optimization of one parameter function is achieved using Brent's (1973) method. This very simple method does not require function derivatives and in our experiments the computational speed was similar to that of Newton-Raphson's method (Olsen et al., 1994;

Felsenstein and Churchill, 1996; Yang, 2000). After optimization, the likelihood of tree (a) is augmented; the new tree likelihood value is denoted as L_a, while l_a is the new length of e.

Tree swapping

Let e now be an internal edge. e defines four subtrees, as shown in Figure 1. When swapping these subtrees, which corresponds to a "nearest neighbor interchange", the initial configuration (a) is changed into (b) or (c) (Figure 1). Consider configuration (b). Subtree U now contains subtrees W and Y, while V contains subtrees X and Z. However, we assume that W, X, Y, and Z are unchanged from configuration (a), as well as branch lengths l_W, l_X, l_Y, and l_Z. The conditional likelihood of U for any given site i is then equal to (Felsenstein, 1981):

$$L(i = h/U) = \left(\sum_{g \in \{A,C,G,T\}} L(i = g/W) P_{hg}(l_W) \right) \left(\sum_{g \in \{A,C,G,T\}} L(i = g/Y) P_{hg}(l_Y) \right), \qquad (2)$$

and the conditional likelihood of V is obtained by symmetry from X and Z. Once the conditional likelihoods of U and V have been computed, we adjust the length of e by Equation (1), as explained above. We thus obtain the likelihood of configuration (b), denoted as L_b, and l_b which is the corresponding branch length of e. L_c and l_c are defined and obtained in the same way for configuration (c). When the conditional likelihoods of W, X, Y, and Z are known, the computation of Equation (2) for all sites is fast and requires $\mathcal{O}(s)$. So computing L_b and L_c has essentially the same cost as computing L_a.

If L_b is larger than L_a and L_c, then configuration (b) is more likely than the two other configurations. Moreover, the larger the gap between L_b and L_a, the more confident we are in the swap from the current configuration (a) to configuration (b). So when L_b is larger than L_a and L_c, we say that e defines (b) as a possible swap, with score $S = L_b - L_a$. The same holds by symmetry when L_c is larger than L_a and L_b.

Selecting and combining these modifications

Changing l into l_a or performing a possible swap increases the tree likelihood. However, edges are not independent, and when simultaneously modifying two edges with values computed as described above, we cannot be sure that the tree likelihood will increase. So the standard approach is to perform one modification at a time; after each modification, the conditional likelihoods (and even all branch lengths in case of branch swapping) are updated. However, this is a slow solution, since conditional likelihood (and branch length) updating is time consuming.

Our approach involves first independently computing all modifications, i.e. the optimal lengths of all branches and possible swaps around all internal branches, and then simultaneously applying "most" of these modifications to the current tree. This latter step is performed as follows:

- The possible swaps are ranked according to their scores S; when two possible swaps correspond to two adjacent branches, they have one subtree in common and only the best swap is conserved; we then apply a proportion λ of the remaining swaps to the current tree, starting from the higher values of S.

- For external branches and internal branches that do not correspond to a possible swap (or that have not been retained in the previous selection), we change the current branch length l into $l + \lambda(l_a - l)$, i.e. we apply a proportion λ of the change that has been computed using Equation (1).

- Having $\lambda = 1$ would simultaneously apply all possible modifications, while $\lambda = 0$ would leave the current tree unchanged. In fact, λ is analogous to the "safeguard" often used in Newton-Raphson and other numerical optimization methods to ensure convergence (Felsenstein and Churchill, 1996). We start with a high λ value but check that the tree likelihood increases. In the (rare) cases where the likelihood decreases, λ is divided by 2, the tree is modified accordingly, and we again check the likelihood; when it still decreases, λ is divided by 2 again, and the process is repeated until we get a tree with higher likelihood than the current tree. With a low λ value, only the best swap is performed and branch lengths remain identical, which corresponds to a tree that is better than the

current tree, thus ensuring convergence. Moreover, λ is reset to its initial value after each refinement stage. For example, with the 218-taxon ribosomal data set, this backward movement occurs only once during the whole refinement process. PHYML uses 0.75 as initial λ value, but this is not a sensitive parameter. Nearly identical trees (but different run times) are obtained with λ in the $[0.1, 1.0]$ range.

Whole method

We have seen in the previous sections how the possible modifications are computed and how they are combined to refine the current tree. We detail here how these components are incorporated in the whole method, which is described step by step.

(a) A pairwise evolutionary distance matrix is computed from the sequences, by an algorithm analogous to DNADIST (Felsenstein, 1993). This step necessitates comparing all sequence pairs and then requires $\mathcal{O}(n^2 s)$ time, where n is the number of taxa.

(b) An initial tree is built from this matrix, using BIONJ (Gascuel, 1997). Tests with other distance-based methods led to identical results, so the main criterion at this step is computational speed. BIONJ is just as fast as NJ, but slightly more accurate, and requires $\mathcal{O}(n^3)$ time.

(c) The conditional likelihoods $L(i = h/U)$ are computed for all sites and every subtree U, as well as the likelihood of the whole tree, using Equation (2) and (1), respectively. These computations are achieved using an algorithm similar to Adachi and Hasegawa's (1996), which requires $\mathcal{O}(ns)$ time.

(d) The values of the free parameters of the substitution model (i.e., the transition/transversion ratio and the gamma shape parameter measuring the variability of substitution rates among sites) are adjusted so as to increase the likelihood of the starting phylogeny. This is achieved using the Golden Section numerical optimization method (Press et al., 1988). When the model involves several free parameters, they are iteratively optimized one after the other until convergence. The parameter estimates so obtained are dependent from the starting tree. However, this dependency is slight (Yang, 1996). Moreover, the free parameters are periodically re-estimated during the refinement process (every four stages in the current version of PHYML).

(e) The current tree is iteratively refined until convergence, as described in the previous section. Each refinement stage involves: (1) computing the possible modifications of every branch, (2) applying a λ proportion of these modifications to the current tree, (3) checking that the tree likelihood increases and, if necessary, returning to step (2) with a lower λ value. Step (1) requires $\mathcal{O}(s)$ time per branch and then $\mathcal{O}(ns)$ for the whole tree. Step (2) is fast, basically in $\mathcal{O}(n)$. And step (3) performs likelihood computations as described in (c), i.e. requires $\mathcal{O}(ns)$ time. Moreover, after step (3) all conditional likelihoods have been updated, and then a new refinement stage can start.

(f) Tree refinement stops when there are no more possible swaps and when the branch lengths are stable. The current tree is then returned.

The time complexity of model parameter, topology and branch length optimization (step (c) to (f)) is then $\mathcal{O}(pns)$, where p basically represents the number of refinement stages that have been performed. Even when this analysis does not clarify some (bound but significant) parameters, e.g. the number of iterations required by Brent's method to optimize branch lengths, it reveals why our method is so fast. With the 218-taxon data set, p is only equal to 15 and, in practice, p is always smaller than n (see Table 1). This explains why our $\mathcal{O}(pns)$ ML optimization has computing time in the same range as distance methods such as NJ, BIONJ and Weighbor, which require $\mathcal{O}(n^2s + n^3)$ time, including distance estimation.

Results

Computer simulations

We generated 5,000 random phylogenies, each comprising 40 taxa, using the standard speciation process described in Kuhner and Felsenstein (1994). This process makes the trees molecular clock-like, so we created a deviation from this model by multiplying every branch length by $(1+X)$, where X followed an exponential distribution with expectation λ. The λ value represents the extent of deviation and was identical within each tree, but different from tree to tree and equal to $0.2/(0.001 + U)$, where U was uniformly drawn from $[0, 1]$. The smaller U, the larger λ and the larger is the deviation from molecular clock. Finally, the tree length was rescaled

by multiplying every branch length by $(0.4 + 8.6V)/T$, where T is the total tree length and V was identical within each tree but different from tree to tree and uniform in $[0, 1]$. This scaling made the tree length uniformly distributed in the $[0.4, 9.0]$ range.

The phylogenies generated in this way have a broad variety of deviations from molecular clock and evolutionary rates. The branch length mean is equal to 0.06 mutations per site, with the 5%, 25%, 50%, 75% and 95% quantiles about equal to 0.0015, 0.01, 0.03, 0.07 and 0.20, respectively. The ratio between the length of the longest and the shortest lineages measures the deviation from molecular clock, with the perfect molecular clock having a ratio of 1. The mean of this ratio, among the 5,000 phylogenies, is equal to 3.4, with the 5%, 25%, 50%, 75% and 95% quantiles about equal to 1.3, 2.3, 3.2, 4.2 and 6.4, respectively. These values come from an analysis of substitution rates in various organisms (Page and Holmes (1998)) and of numerous recently published phylogenies; they should then cover the features of almost all real data sets, even when the extreme values, notably the highest divergence rates, are likely rare.

Sequences 500 bp in length were generated from these phylogenies using Seq-Gen (Rambaut and Grassly, 1997) under the Kimura two parameter model (Kimura, 1980), with a transition/transversion ratio of 2.0. The 5,000 data sets (phylogenies and sequences) obtained in this way are available on our web page.

Topological accuracy

Using these data sets, we compared PHYML with numerous other packages: NJ, Weighbor 1.2 (Bruno et al., 2000), DNAPARS 3.5 (Felsenstein, 1993), NJML+ (Ota and Li, 2001), fastDNAml (Olsen et al., 1994), PAUP* 4.0 beta (Swofford, 1999) and MrBayes 2.01 (Huelsenbeck and Ronquist, 2001). NJ and Weighbor are distance-based and were combined with DNADIST 3.6 (Felsenstein, 1993), DNAPARS uses the parsimony principle, while the other programs implement ML approaches. We did not test MetaPIGA (Lemmon and Milinkovitch, 2002) because no batch version allowing for multiple data sets was available. Moreover, for computing time reasons, PAUP* and MrBayes were only run on the first 1,000 data sets, and only nearest neighbor interchanges were used in PAUP*. MrBayes was run with a random starting tree, 30,000 generations and a sampling frequency of 10, while the resulting consensus phylogeny was

built from the last 1,500 trees. The options for NJML were: bootstrap threshold = 90% and composite mode for likelihood computation. Other packages were used with default options, supplying the simulation settings (e.g. the sequence length or the transition/transversion ratio) when required.

The topological accuracy of these various methods was measured by the standard Robinson and Foulds (1979) distance between the inferred tree and the true tree. This distance corresponds to the proportion of internal branches that are found in one tree and not in the other one. Its value ranges from 0.0 (both topologies are identical) to 1.0 (they do not share any branch in common). The value of this distance was plotted against the maximum pairwise divergence in the data set under consideration, with the (uncorrected) divergence between two sequences being simply the proportion of sites where both sequences differ. The results are displayed in Figure 2, where the 5,000 original points corresponding to each method are smoothed by averaging over a slipping window of length 1,000.

These results are in good accordance with expectations and with previously published simulations (Huelsenbeck and Hillis, 1993; Kuhner and Felsenstein, 1994; Huelsenbeck, 1995; Rosenberg and Kumar, 2001; Ranwez and Gascuel, 2002). When the divergence rate is low, phylogeny reconstruction is hard as there is not enough information in the data to estimate the short internal edges. On the other side, with a high divergence rate, saturation corrupts the phylogenetic signal and reconstruction is also hard. This explains why all methods perform better with medium divergence rates. The best region for parsimony corresponds to low rates, as expected since it assumes that multiple substitutions are rare, while distance-based methods (that account for multiple substitutions) tend to perform better than parsimony with high substitution rates. The performance of NJML, which combines distance-based and ML approaches, is midway between both. However, the best approach is clearly ML. Both fastDNAml and PHYML outperform all other methods, and PHYML even tends to improve fastDNAml with high substitution rates. Indeed, PHYML and fastDNAml are very close concerning likelihood optimization, except with high substitution rates, where PHYML is slightly better (results not shown). Moreover, for about 95% of the data sets, both packages infer trees with likelihood identical to or higher than the likelihood of the true tree, which indicates that there is little room for accuracy

improvement by further optimizing the tree likelihood. This is confirmed by the average accuracy of the various ML programs on the first 1,000 data sets, i.e. 0.086, 0.086, 0.081 and 0.081, for PAUP*, fastDNAml, MrBayes and PHYML, respectively. So we do not expect to achieve major improvements on these data sets by any ML method, including MetaPIGA.

Computing times and likelihood optimization

We compared the computing time of the various methods with 30 data sets comprising 40 taxa, and 30 data sets with 100 taxa, both being generated as described above (and available on our web page). We also used two large real data sets. The first set contains 218 prokaryotic sequences with 4,182 bp from the small ribosomal subunit, and was downloaded from the RDPII project web page (`http://rdp.cme.msu.edu/download/SSU_rRNA/alignments/`). The second set includes 500 *rbc*L sequences with 1,428 bp from plant plastids, and was obtained from `http://www.cis.upenn.edu/~krice/treezilla/record.nex`.

The computing time was measured on a PC Pentium IV 1.8 GHz (1 Go RAM) running with Linux. Basically, all methods were run as described above. However, fastDNAml and PAUP* were not run on the two larger data sets because even the 218-taxon set required more than two days of computations. Moreover, with these two data sets, we were not able to obtain any result with NJML, seemingly for memory size reasons, while MrBayes was stopped after 1,000,000 generations without having reached stable likelihood values. MetaPIGA was run by hand using four metapopulations (default option). With the 100-taxon data sets, MrBayes was run with 200,000 generations and a consensus tree was built from the last 10,000 trees. These parameters were chosen to converge on stable likelihood values, but we did not explore a large range of settings, first preferring computation speed. In fact, this criterion was used for MrBayes, but also for the other packages. So it is likely that other relevant speed/performance compromises could be found for any of these programs, and our results must therefore not be over-interpreted. The basic aim of this study was to check the good qualities of PHYML, rather than to provide an extensive comparison of existing packages.

The results are displayed in Table 1. PHYML is clearly faster than other ML programs. For example, with 100 taxa it requires 12 seconds while fastDNAml requires about 25 minutes.

With 500 taxa, PHYML requires only about 12 minutes, while MetaPIGA requires more than 9 hours. In fact, the computing time with PHYML is in the same range as that with NJ, Weighbor and DNAPARS.

We also checked that the speed of PHYML is not offset by lower performance in optimizing the tree likelihood. For a fair comparison, the branch lengths of trees inferred by the various ML packages were re-optimized using the same Newton-Raphson procedure and the tree likelihood was recomputed. We then sorted the (ML) methods with respect to their loglikelihood values and computed the mean of their rank by averaging over the 30 data sets analyzed. For the 40-taxon data sets, we found 3.5, 3.4, 3.1, 2.0, 1.9 and 1.7 for MetaPIGA, NJML, MrBayes, PAUP*, PHYML and fastDNAml respectively, with 100 taxa, the result was 4.8, 4.3, 3.9, 2.5, 2.1 and 2.0 for MetaPIGA, NJML, MrBayes, PAUP*, fastDNAml and PHYML respectively. Average loglikelihood confirms this ordering, but with a much lower contrast. For example, MetaPIGA is quite close to PHYML (-14,371 and -14,369, respectively, for the 100-taxon trees).

For the two large real data sets, MetaPIGA and PHYML were run with the HKY model and both programs adjusted the transition/transversion ratio. With the 218-taxon set, the loglikelihood of the MetaPIGA tree is higher than that of PHYML (-156,727/-156,860 in terms of loglikelihood), but the converse holds with the 500-taxon set (-100,631/-100,208). Moreover, when applying PHYML to the MetaPIGA tree for 218 taxa, it performs 6 swaps and improves the loglikelihood by -156,692. Therefore, it appears that PHYML is not only fast but also provides trees with high likelihood.

Conclusion

PHYML is freely available on our web page. The current version implements several models of nucleotide sequence evolution: JC69 (Jukes and Cantor, 1969), F81 (Felsenstein, 1981), K2P (Kimura, 1980), F84 (Felsenstein, 1993), HKY (Hasegawa et al., 1985) and TN93 (Tamura and Nei, 1993), and a discrete gamma distribution (Yang, 1994) can be used to account for variable substitution rates among sites. The parameters of these models can be either user defined or fitted to the data by likelihood maximization. PHYML can also be used to refine a user-supplied

tree. Proteins should be dealt with in the near future.

Regarding its simplicity, the performance of our algorithm is quite surprising. It is not only much faster than the standard approach but also slightly better in terms of topological accuracy and likelihood maximization. In fact, it seems that adjusting all branch lengths and the tree topology together avoids being trapped too early in local optima. The algorithm does not follow the slope corresponding to a unique branch or a unique swap, but moves in a direction that improves the whole tree, and by striding this way avoids getting lost in local irregularities of the likelihood landscape. However, testing more intense topological rearrangements or introducing some randomness in the search are interesting directions for future research.

Acknowledgments

Thanks to Michel C. Milinkovitch for providing us with useful advices about MetaPIGA, and Nicolas Galtier and Marc Robinson-Rechavi for their suggestions for improvement of the paper.

References

Aarts E. and J. K. Lenstra. 1997. Local search in combinatorial optimization. Wiley, Chichester.

Adachi J. and M. Hasegawa. 1996. Molphy version 2.3. programs for molecular phylogenetics based on maximum likelihood. In Ishiguro M., G. Kitagawa, Y. Ogata, H. Takagi, Y. Tamura and T. Tsuchiya., eds., Computer Science Monographs 28. The Institute of Statistical Mathematics, Tokyo.

Brauer M., M. Holder, L. Dries, D. Zwickl, P. Lewis and D. Hillis. 2002. Genetic algorithms and parallel processing in maximum-likelihood phylogeny inference. Mol. Biol. Evol. 19:1717–1726.

Brent R. P. 1973. Algorithm for Minimization without Derivatives. Prentice-Hall.

Bruno W., N. D. Socci and A. L. Halpern. 2000. Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. Mol. Biol. Evol. 17:189–197.

Chor B., M. D. Hendy, B. R. Holland and D. Penny. 2000. Multiple maxima of likelihood in phylogenetic trees: an analytic approach. Mol. Biol. Evol. 17:1529–1541.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 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) 3.6a2. Distributed by the author, Department of Genetics, University of Washington, Seattle.

Felsenstein J. and G. Churchill. 1996. A hidden markov model approach to variation among sites in rate of evolution. Mol. Biol. Evol. 13:93–104.

Gascuel O. 1997. BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. 14:685–695.

HASEGAWA M., H. KISHINO AND T. YANO. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial-DNA. J. Mol. Evol. 22:160–174.

HUELSENBECK J. P. 1995. Performance of phylogenetic methods in simulation. Syst. Biol. 44:17–48.

HUELSENBECK J. P. AND D. HILLIS. 1993. Success of phylogenetic methods in the four-taxon case. Syst. Biol. 42:247–264.

HUELSENBECK J. P. AND F. RONQUIST. 2001. MrBayes: Bayesian inference of phylogeny. Bioinformatics 17:754–755.

JIN L. AND M. NEI. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol. Biol. Evol. 7:82–102.

JUKES T. H. AND C. R. CANTOR. 1969. Evolution of protein molecules. In MUNRO H. N., ed., Mammalian Protein Metabolism vol. III chap. 24:21–132. Academic Press, New York.

KIMURA M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111–120.

KUHNER M. K. AND J. FELSENSTEIN. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11:459–468.

LEMMON A. AND M. C. MILINKOVITCH. 2002. The metapopulation genetic algorithm: an efficient solution for the problem of large phylogeny estimation. Proc. Natl. Acad. Sci. U.S.A. 99:10516–10521.

LEWIS P. 1998. A genetic algorithm for maximum likelihood phylogeny inference using nucleotide sequence data. Mol. Biol. Evol. 15:277–283.

LI S. 1996. Phylogenetic tree construction using Markov chain Monte Carlo. Ph.D. thesis, Ohio State University.

MAU B. 1996. Bayesian phylogenetic inference via Markov Chain Monte Carlo methods. Ph.D. thesis, Wiscontin University.

OLSEN G. J., H. MATSUDA, R. HAGSTROM AND R. OVERBEEK. 1994. fastDNAml: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Comput. Appl. Biosci. 10:41–48.

OTA S. AND W.-H. LI. 2000. NJML: A hybrid algorithm for the neighbor-joining and maximum-likelihood methods. Mol. Biol. Evol. 17:1401–1409.

OTA S. AND W.-H. LI. 2001. NJML+: An extension of the NJML method to handle protein sequence data and computer software implementation. Mol. Biol. Evol. 18:1983–1992.

PAGE R. D. M. AND E. C. HOLMES. 1998. Molecular Evolution: a phylogenetic approach. Blackwell Science Ltd, Osney Mead, Oxford.

PRESS W. H., B. P. FLANNERY, S. A. TEUKOLSKY AND W. T. VETTERLING. 1988. Numerical Recipies in C. Press Syndicate of the University of Cambridge, Cambridge.

RAMBAUT A. AND N. GRASSLY. 1997. Seq-gen: an application for the Monte Carlo Simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235–238.

RANNALA B. AND Z. YANG. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J. Mol. Evol. 43:304–311.

RANWEZ V. AND O. GASCUEL. 2001. Quartet-based phylogenetic inference: improvements and limits. Mol. Biol. Evol. 18:1103–11016.

RANWEZ V. AND O. GASCUEL. 2002. Improvement of distance-based phylogenetic methods by a local maximum likelihood approach using triplets. Mol. Biol. Evol. 19:1952–1963.

ROBINSON D. F. AND L. R. FOULDS. 1979. Comparison of weighted labelled trees. In Lectures Notes in Mathematics vol. 748:119–126. Springer, Berlin.

ROSENBERG M. S. AND S. KUMAR. 2001. Traditional phylogenetic reconstruction methods reconstruct shallow and deep evolutionary relationship equally well. Mol. Biol. Evol. 19:1823–1827.

SAITOU N. AND M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406–425.

SALTER L. A. AND D. K. PEARL. 2001. Stochastic search strategy for estimation of maximum likelihood phylogenetic trees. Syst. Biol. 50:7–17.

SIMON D. AND B. LARGET. 2000. Bayesian analysis in molecular biology and evolution (BAMBE), version 2.03 beta. Department of Mathematics and Computer Science, Duquesne University.

STRIMMER K. AND VON A. HAESELER. 1996. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. Mol. Biol. Evol. 13:964–969.

SWOFFORD D. L. 1999. PAUP*: Phylogenetic Analysis by Parsimony* and other methods. Sinauer, Sunderland, MA.

SWOFFORD D. L., G. J. OLSEN, P. J. WADDEL AND D. M. HILLIS. 1996. Phylogenetic inference. In HILLIS D. M., C. MORITZ AND B. K. MABLE., eds., Molecular Systematics chap. 11. Sinauer, Sunderland, MA.

TAMURA K. AND M. NEI. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10:512–526.

UZZELL T. AND K. W. CORBIN. 1971. Fitting discrete probability distributions to evolutionary events. Science 172:1089–1096.

WHELAN S., P. LIÒ AND N. GOLDMAN. 2001. Molecular phylogenetics: state-of-the art methods for looking into the past. Trends in Genetics 17:262–272.

YANG Z. 1994. Maximum-likelihood estimation of phylogeny from DNA sequences when substitute rates differ over sites: approximate methods. J. Mol. Evol. 39:306–314.

YANG Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. TREE 11:367–372.

YANG Z. 2000. Maximum likelihood estimation on large phylogenies and analysis of adaptative evolution in human influenza virus A. J. Mol. Evol. 51:423–432.

Table 1: Average run times. The computing times were measured on a 1.8 Ghz (1 Go RAM) PC with Linux. The bold number in brackets is the (average) number of refinement stages performed by PHYML.

Table 1

Method	Simulations		Real data	
	40 taxa 500 bp	100 taxa 500 bp	218 taxa 4,182 bp	500 taxa 1,428 bp
DNADIST + NJ/BioNJ	0.3s	2.3s	50s	2min19s
DNADIST + Weighbor	1.5s	22s	4min52s	58min40s
DNAPARS	0.5s	6s	4min4s	13min12s
PAUP*	3min21s	1h4min	*	*
MrBayes	2min6s	32min37s	*	*
fastDNAml	1min13s	26min31s	*	*
NJML	15s	6min4s	*	*
MetaPIGA	21s	3min27	4h45min	9h4min
PHYML	2.7s (**6.4**)	12s (**8.3**)	8min13s (**15**)	11min59s (**13**)

Figure 1: The three alternative topological configurations around an internal branch. W, X, Y and Z are four subtrees. l_w, l_x, l_y, l_z are the lengths of the four branches connected to the roots of W, X, Y and Z, respectively. These lengths are the same in the three topological configurations. U and V are the subtrees on the left and on the right of the internal branch. l_a, l_b, l_c are the internal branch lengths that maximize the likelihoods of the corresponding phylogenies.

Figure 2: Topological accuracies of various tree building methods as a function of the divergence between sequences N: NJ, W: Weighbor, L: NJML , D: DNAPARS, F: fastDNAml, P: PHYML. BIONJ (used to build the starting tree in PHYML) is midway between NJ and Weighbor.

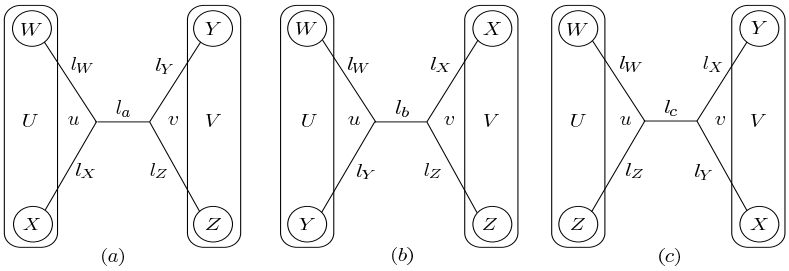

W
Y
l_W
l_Y
U
u
l_a
v
V
l_X
l_Z
X
Z
(a)
W
X
l_W
l_X
U
u
l_b
v
V
l_Y
l_Z
Y
Z
(b)
W
Y
l_W
l_X
U
u
l_c
v
V
l_Z
l_Y
Z
X
(c)

Figure 1

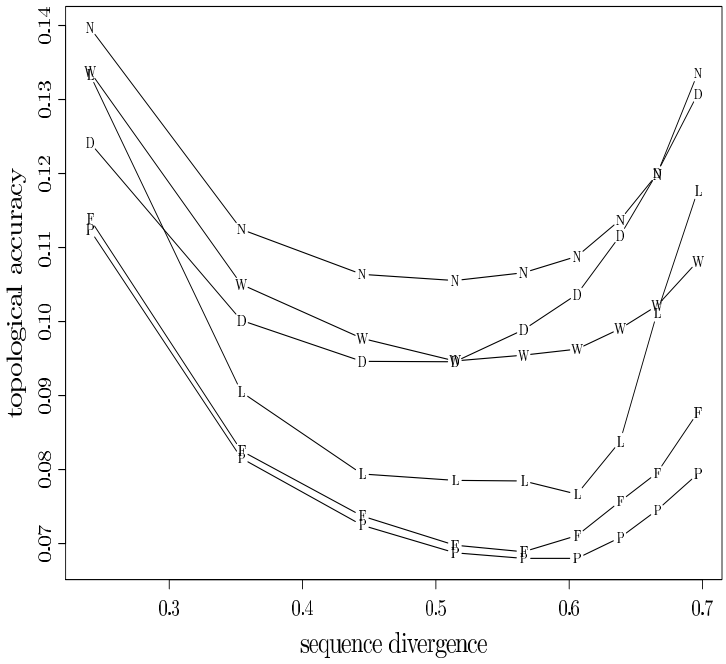

topological accuracy
0.07
0.08
0.09
0.10
0.11
0.12
0.13
0.14
0.3
0.4
0.5
0.6
0.7
sequence divergence

Figure 2

Bibliographie

Aarts, E. et Lenstra, J. K. (1997). *Local search in combinatorial optimization*. Wiley, Chichester.

Adachi, J. et Hasegawa, M. (1996). Molphy version 2.3. programs for molecular phylogenetics based on maximum likelihood. In Ishiguro, M., Kitagawa, G., Ogata, Y., Takagi, H., Tamura, Y., et Tsuchiya, T., editeurs, *Computer Science Monographs*, number 28. The Institute of Statistical Mathematics, Tokyo.

Andrieu, G. (1997). *Estimation par intervalle d'une distance évolutive*. PhD thesis, Univ. Montpellier II.

Atteson, K. (1997). The performance of the NJ method of phylogeny reconstruction. In Mirkin, B., McMorris, F., Roberts, F., et Rhzetsky, A., editeurs, *Mathematical Hierarchies and Biology*, pages 133–148. American Mathematical Society.

Bandelt, H.-J. et Dress, A. (1992a). A canonical decomposition theory for metrics on a finite set. *Advances Math*, 92 :47–105.

Bandelt, H.-J. et Dress, A. (1992b). Split decomposition : A new and useful approach to phylogenetic analysis of distance data. *Mol. Biol. Evol.*, 1 :242–252.

Barry, D. et Hartigan, J. A. (1987). Asynchronous distance between homologous DNA sequences. *Biometrics*, 43 :261–276.

Barthélemy, J. et Guénoche, A. (1991). *Trees and proximities representations*. Wiley.

Barthélemy, J.-P. et Guénoche, A. (1988). *Les arbres et les représentations des proximités*. Méthodes + Programmes. Masson.

Brent, R. (1973). *Algorithms for Minimization without Derivatives*. Prentice-Hall.

Brown, K. S. (1999). Deep green rewrites evolutionary history of plants. *Science*, 285 :990–991.

Bruno, W., Socci, N., et Halpern, A. (2000). Weighted neighbor joining : a likelihood-based approach to distance-based phylogeny reconstruction. *Mol. Biol. Evol.*, 17 :189–197.

Bulmer, M. (1991). Use of the method of generalized least squares in reconstructing phylogenies from sequence data. *Mol. Biol. Evol.*, 8 :868–883.

Dayhoff, M., Schwartz, R., et Orcutt, B. (1978). A model of evolutionary change in proteins. volume, 5, pages 345–352. National Biomedical Research Foundation, Washington, D. C.

D'Erchia, A., Gissi, C., Pesole, G., Saccone, C., et Arnason, U. (1996). The guinea-pig is not a rodent. *Nature*, 381 :597–600.

Desper, R. et Gascuel, O. (2002). Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *Journal of Computational Biology*, 9 :687–705.

Douady, C. J., Delsuc, F., Boucher, Y., Doolittle, W. F., et Douzery, E. J. P. (2003). Comparison of bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Mol. Biol. Evol.*, 20 :248–254.

Edwards, A. W. F. et Cavalli-Sforza, L. L. (1964). Reconstruction of evolutionary trees. *Systematics Association Publication*, 6 :67–76.

Eigen, M. et Winkler-Oswatitsch, R. (1981). Transfer-rna : The early adaptor. *Die Naturwissenschaften*, 68 :217–228.

Etzold, T. et Argos, P. (1993). SRS-an indexing and retrieval tool for flat file data libraries. *CABIOS*, 9 :49–57.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences : a maximum likelihood approach. *J. Mol. Evol.*, 17 :368–376.

Felsenstein, J. (1989). Phylogeny inference package (version 3.2). *Cladistics*, 5 :164–166.

Felsenstein, J. (1993). *PHYLIP (Phylogeny Inference Package version) 3.6a2*. Distributed by the author, Department of Genetics, University of Washington, Seattle.

Felsenstein, J. et Churchill, G. (1996). A hidden Markov model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.*, 13 :93–104.

Fitch, W. M. et Margoliash, E. (1967). Construction of phylogenetic trees. *Science*, 155 :279–284.

Galtier, N. (1997). *L'approche statistique en phylogénie moléculaire : influence des composition en bases variables*. PhD thesis, Université Claude Bernard - Lyon I.

Galtier, N. (2001). Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol. Biol. Evol.*, 18 :866–873.

Galtier, N. et Gouy, M. (1995). Inferring phylogenies from DNA sequences of unequal base compositions. *Proc. Natl. Acad. Sci. U S A.*, 92 :11317–11321.

Galtier, N., Tourasse, N., et Gouy, M. (1999). A nonhyperthermophilic common ancestor to extant life forms. *Science*, 283 :220–221.

Gascuel, O. (1994). A note on Sattah and Tversky's, Saitou and Nei's, and Studier and Keppler's algorithms for inferring phylogenies from evolutionary distances. *Mol. Biol. Evol.*, 11 :961–963.

Gascuel, O. (1997a). BIONJ : an improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.*, 14 :685–695.

Gascuel, O. (1997b). Concerning the NJ algorithm and its unweighted version UNJ. In Mirkin, B., McMorris, F., Roberts, F., et Rzhetsky, A., editeurs, *Mathematical Hierarchies and Biology*, pages 149–170. Am. Math. Society, Providence, USA.

Gascuel, O. (2000). On the optimization principle in phylogenetic analysis and the minimum-evolution criterion. *Mol. Biol. Evol.*, 17 :401–405.

Gascuel, O., Bryant, D., et Denis, F. (2001). Strengths and limitations of the minimum evolution principle. *Syst. Biol.*, 50 :621–627.

Gaut, B. et Lewis, P. (1995). Success of maximum likelihood phylogeny inference in the four-taxon case. *Mol. Biol. Evol.*, 12 :152–162.

Golding, G. B. (1983). Estimates of DNA and protein sequence divergence : an examination of some assumptions. *Mol. Biol. Evol.*, 1 :125–142.

Gouy, M., Milleret, F., Mugnier, C., Jacobzone, M., et Gautier, C. (1984). ACNUC : a nucleic acid sequence data base and analysis system. *Nucleic Acids Res.*, 12(2) :121–127.

Graur, D., Hide, W., et Li, W.-H. (1991). Is the guinea-pig a rodent ? *Nature*, 351 :649–652.

Gu, X., Fu, Y., et Li, W. (1995). Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Mol. Biol. Evol.*, 12 :546–557.

Gu, X. et Zhang, J. (1997). A simple method for estimating the parameter of substitution rate variation among sites. *Mol. Biol. Evol.*, 14 :1106–13.

Guindon, S. et Gascuel, O. (2002). Efficient biased estimation of evolutionary distances when substitution rates vary across sites. *Mol. Biol. Evol.*, 19 :534–543.

Guindon, S. et Gascuel, O. (2003). A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*

Guénoche, A. et Garreta, H. (2000). Can we have confidence in a tree representation ? In *Computational biology, LNCS 2066*, pages 45–46.

Hasegawa, M., Kishino, H., et Yano, T. (1985). Dating of the Human-Ape splitting by a molecular clock of mitochondrial-DNA. *J. Mol. Evol.*, 22 :160–174.

Higgins, D. G., Thompson, J. D., et Gibbson, J. T. (1996). Using CLUSTAL for multiple sequence alignments. In Doolittle, R. F., editeur, *Methods in enzymology*, pages 383–401. Academic Press, San Diego.

Huelsenbeck, J. P. (1995). Performance of phylogenetic methods in simulation. *Syst. Biol.*, 44 :17–48.

Huelsenbeck, J. P. (2002). Testing a covariotide model of DNA substitution. *Mol. Biol. Evol.*, 19 :698–707.

Huelsenbeck, J. P. et Bollback, J. P. (2001). Empirical and hierarchical bayesian estimation of ancestral states. *Syst. Biol.*, 50 :351–366.

Huelsenbeck, J. P. et Hillis, D. (1993). Success of phylogenetic methods in the four-taxon case. *Syst. Biol.*, 42 :247–264.

Huelsenbeck, J. P. et Ronquist, F. (2001). MrBayes : Bayesian inference of phylogeny. *Bioinformatics*, 17 :754–755.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R., et Bollack, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, 294 :2310–2314.

Jin, L. et Nei, M. (1990). Limitations of the evolutionary parsimony method of phylogenetic analysis. *Mol. Biol. Evol.*, 7 :82–102.

Jones, D., Taylor, W., et Thornton, J. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.*, 8 :275–282.

Jukes, T. et Cantor, C. (1969). Evolution of protein molecules. volume, III, chapitre 24, pages 21–132. Academic Press, New York.

Kidd, K. et Sgaramella-Zonta, L. (1971). Phylogenetic analysis : concepts and methods. *Am. J. Human. Genet.*, 23 :235–252.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, 16 :111–120.

Kimura, M. (1981). Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Acad. Sci. USA*, 78 :454–458.

Kirkpatrick, S., Gelatt, C. D., et Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220 :671–680.

Kondrashov, A., Sunyaev, S., et Kondrashov, F. (2002). Dobzhansky-Muller incompatibilities in protein evolution. *Proc. Natl. Acad. Sci. U S A.*, 99 :14878–14883.

Kuhner, M. K. et Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.*, 11 :459–468.

Lanave, C., Preparata, G., Saccone, C., et Serio, G. (1984). A new method for calculating evolutionary substitution rates. *J. Mol. Evol.*, 20 :86–93.

Larget, B. et Simon, D. L. (1999). Markov chain Monte Carlo algorithms for the bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.*, 16 :750–759.

Lemmon, A. et Milinkovitch, M. (2002). The metapopulation genetic algorithm : an efficient solution for the problem of large phylogeny estimation. *Proc. Natl. Acad. Sci. USA.*, 99 :10516–10521.

Lewis, P. (1998). A genetic algorithm for maximum likelihood phylogeny inference using nucleotide sequence data. *Mol. Biol. Evol.*, 15 :277–283.

Li, S. (1996). *Phylogenetic tree construction using Markov chain Monte Carlo.* PhD thesis, Ohio State University.

Lopez, P., Forterre, P., et Philippe, H. (1999). The root of the tree of life in the light of the covarion model. *J. Mol. Evol.*, 49 :496–508.

Maidak, B., Cole, J., Parker, C., Garrity, G., Larsen, N., Li, B., Lilburn, T., McCaughey, M., Olsen, G., Overbeek, R., Pramanik, S., Schmidt, T., Tiedje, J., et Woese, C. (1999). A new version of the RDP (Ribosomal Database Project). *Nucl. Acids. Res*, 27 :171–173.

Maidak, B., Larsen, N., McCaughey, M., Overbeek, R., Olsen, G., Fogel, K., Blandy, J., et Woese, C. (1994). The Ribosomal Database Project. *Nucl. Acids. Res*, 22 :3485–3487.

Maidak, B., Olsen, G., Larsen, N., Overbeek, R., McCaughey, M., et Woese, C. (1996). The Ribosomal Database Project (RDP). *Nucl. Acids. Res*, 24 :82–85.

Maidak, B., Olsen, G., Larsen, N., Overbeek, R., McCaughey, M., et Woese, C. (1997). The RDP (Ribosomal Database Project). *Nucl. Acids. Res*, 25 :109–111.

Maidak, B. L., Cole, J. R., Lilburn, T. G., Parker, C. T., Saxman, P. R., Farris, R. J., Garrity, G. M., Olsen, G. J., Schmidt, T. M., et Tiedje, J. M. (2001). The RDP-ii (Ribosomal Database Project). *Nucl. Acids. Res*, 29 :173–174.

Maidak, B. L., Cole, J. R., Lilburn, T. G., Parker, C. T., Saxman, P. R., Stredwick, J. M., Garrity, G. M., Li, B., Olsen, G. J., Pramanik, S., Schmidt, T. M., et Tiedje, J. M. (2000). The RDP (Ribosomal Database Project) continues. *Nucl. Acids. Res*, 28 :173–174.

Mau, B. (1996). *Bayesian phylogenetic inference via Markov Chain Monte Carlo methods*. PhD thesis, Wiscontin University.

McQuitty, L. L. (1966). Similarity analysis by reciprocal pairs for discrete and continuous data. *Educational and Psychological Measurement*, 26 :825–831.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., et Teller, E. (1953). Equation of state calculations by fast computing machines. *J. of Chem. Phys.*, 21 :1087–1092.

Mitchison, G. et Durbin, R. (1995). Tree-based maximum likelihood substitution matrices and hidden markov model. *J. Mol. Evol.*, 41 :1139–1151.

Murphy, W. J., Eizirik, E., Johnson, W. E., Y.P., Z., Ryder, O., et O'Brien, S. (2001). Molecular phylogenetics and the origins of placental mammals. *Nature*, 409 :614–618.

Muse, S. V. (1995). Evolutionary analyses when nucleotides do not evolve independently. In Nei, M. et Takahata, N., editeurs, *Current Topics on Molecular Evolution*, pages 115–124. University Park, Penn. State Univ.

Needleman, S. G. et Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48 :443–453.

Olsen, G., Matsuda, H., Hagstrom, R., et Overbeek, R. (1994). fastDNAml : a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Comput. Appl. Biosci.*, 10 :41–48.

Ota, S. et Li, W.-H. (2001). NJML+ : an extension of the NJML method to handle protein sequence data and computer software implementation. *Mol. Biol. Evol.*, 18 :1983–1992.

Page, R. et Holmes, E. (1998). *Molecular Evolution : a phylogenetic approach*. Blackwell Science Ltd, Osney Mead, Oxford.

Pauplin, Y. (2000). Direct calculation of a tree length using a distance matrix. *J. Mol. Evol.*, 51 :41–47.

Perrière, G., Duret, L., et Gouy, M. (2000). HOBACGEN : database system for comparative genomics in bacteria. *Genome Res.*, 10 :379–385.

Philippe, H. (2000). Opinion : long branch attraction and protist phylogeny. *Protist*, 151 :307–316.

Philippe, H., Lopez, P., Brinkman, H., Budin, K., Germot, A., Laurent, J., Moreira, D., Muller, M., et Guyader, H. L. (2000). Early branching or fast evolving eukaryotes ? an answer based on slowly evolving positions. *Proc. R. Soc. Lond. B. Biol. Sci.*, 267 :1213–1221.

Press, W., Flannery, B., Teukolsky, S., et Vetterling, W. T. (1988). *Numerical Recipes in C.* Press Syndicate of the University of Cambridge, Cambridge.

Purvis, A. et Quicke, D. L. J. (1997). Building phylogenies : are the big easy ? *TREE*, 12(2) :49–50.

Rannala, B. et Yang, Z. (1996). Probability distribution of molecular evolutionary trees : a new method of phylogenetic inference. *J. Mol. Evol.*, 43 :304–311.

Ranwez, V. (2002). *Méthodes efficaces pour reconstruire des phylogénies suivant le principe du maximum de vraisemblance.* PhD thesis, Université Montpellier II.

Ranwez, V. et Gascuel, O. (2001). Quartet-based phylogenetic inference : improvements and limits. *Mol. Biol. Evol.*, 18 :1103–11016.

Ranwez, V. et Gascuel, O. (2002). Improvement of distance-based phylogenetic methods by a local maximum likelihood approach using triplets. *Mol. Biol. Evol.*, 19 :1952–1963.

Robinson, D. et Foulds, L. (1979). Comparison of weighted labeled trees. volume, 748, pages 119–126. Springer, Berlin.

Rodriguez, F., Olivier, J. L., Marin, A., et Medina, J. R. (1990). The general stochastic model of nucleotide substitution. *J. Theo. Biol.*, 142 :485–501.

Rogers, J. et Swofford, D. (1999). Multiple local maxima for likelihoods of phylogenetic trees : a simulation study. *Mol. Biol. Evol.*, 16 :1079–1085.

Rosenberg, M. et Kumar, S. (2001). Traditional phylogenetic reconstruction methods reconstruct shallow and deep evolutionary relationship equally well. *Mol. Biol. Evol.*, 19 :1823–1827.

Rzhetsky, A., Kumar, S., et Nei, M. (1995). Four-cluster analysis : a simple method to test phylogenetic hypotheses. *Mol. Biol. Evol.*, 12 :163–167.

Rzhetsky, A. et Nei, M. (1993). Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Mol. Biol.Evol.*, 10 :1073–1095.

Rzhetsky, A. et Nei, M. (1994). Unbiaised estimates of the number of nucleotide substitutions when substitution rate varies among different sites. *J. Mol. Evol.*, 38 :295–299.

Rzhetsky, A. et Sitnikova, T. (1996). When is it safe to use an oversimplified substitution model in tree-making ? *Mol. Biol. Evol.*, 13 :1255–1265.

Saitou, N. et Nei, M. (1987). The neighbor-joining method : a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4 :406–425.

Salter, L. et Pearl, D. (2001). Stochastic search strategy for estimation of maximum likelihood phylogenetic trees. *Syst. Biol.*, 50 :7–17.

Sattah, S. et Tversky, A. (1977). Additive similarity trees. *Psychometrika*, 42 :319–345.

Schöniger, M. et von Haesler, A. (1994). A stochastic model for the evolution of autocorrelated DNA sequences. *Mol. Phylogeny Evol.*, 3 :240–247.

Schöniger, M. et von Haesler, A. (1995). Performance of the maximum likelihood, neighbor joining, and maximum parsimony methods when sequence sites are not independent. *Syst. Biol.*, 44 :533–547.

Simon, D. et Larget, B. (2000). *Bayesian analysis in molecular biology and evolution (BAMBE), version 2.03 beta*. Department of Mathematics and Computer Science, Duquesne University.

Sokal, R. R. et Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38 :1409–1438.

Sourdis, J. et Krimbas, C. (1987). Accuracy of phylogenetic trees estimated from DNA sequence data. *Mol. Biol. Evol.*, 4 :159–166.

Steel, M. (1994). The maximum likelihood point for a phylogenetic tree is not unique. *Syst. Biol.*, 43 :560–564.

Steel, M. et Penny, D. (2000). Parsimony, likelihood, and the role of models in molecular phylogenetics. *Mol. Biol. Evol.*, 17 :839–50.

Strimmer, K. et Moulton, V. (2000). Likelihood analysis of phylogenetic networks using directed graphical models. *Mol. Biol. Evol.*, 17 :875–881.

Strimmer, K. et von Haeseler, A. (1996). Quartet puzzling : a quartet maximum-likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.*, 13 :964–969.

Studier, J. et Kepler, K. (1988). A note on the neighbor-joining algorithm of saitou and nei. *Mol. Biol. Evol.*, 5 :729–731.

Sullivan, J., Holsinger, K. E., et Simon, C. (1995). Among-site variation and phylogenetic analysis of 12S rRNA in sigmontine rodents. *Mol. Biol. Evol.*, 12 :988–1001.

Suzuki, Y., Glazko, G. V., et Nei, M. (2002). Overcredibility of molecular phylogenies obtained by bayesian phylogenetics. *Proc. Natl. Acad. Sci. USA*, 99 :16138–16143.

Swofford, D. (1999). *PAUP* : Phylogenetic Analysis by Parsimony* and other methods*. Sinauer, Sunderland, MA.

Swofford, D., Olsen, G., Waddel, P., et Hillis, D. (1996). Phylogenetic inference. In Hillis, D., Moritz, C., et Mable, B., editeurs, *Molecular Systematics*, chapitre 11. Sinauer, Sunderland, MA.

Tajima, F. et Takezaki, N. (1994). Estimation of evolutionary distance for reconstructiong molecular phylogenetic trees. *Mol. Biol. Evol.*, 11 :278–286.

Tamura, K. et Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.*, 10 :512–526.

Tateno, Y., Takezaki, N., et Nei, M. (1994). Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. *Mol. Biol. Evol.*, 11 :261–77.

Tavaré, S. (1986). Some probabilistic and statistical problems on the analysis of DNA sequences. *Lec. Math. Life Sci.*, 17 :57–86.

Thorne, J. L., Kishino, H., et Felsenstein, J. (1991). An evolutionary model for maximum likelihood alignement of DNA sequences. *J. Mol. Evol.*, 33 :114–124.

Thorne, J. L., Kishino, H., et Felsenstein, J. (1992). Inching toward reality : an improved likelihood model of sequence evolution. *J. Mol. Evol.*, 34 :3–16.

Tourasse, N. et Gouy, M. (1997). Evolutionary distances between nucleotide sequences based on the distribution of substitution rates among sites as estimated by parsimony. *Mol. Biol. Evol.*, 14 :287–98.

Uzzell, T. et Corbin, K. (1971). Fitting discrete probability distributions to evolutionary events. *Science*, 172 :1089–1096.

Vach, V. (1992). The jukes-cantor transformation and additivity of estimated genetic distances. In Shader, M., editeur, *Analysing and Modeling Data and Knowledge*, pages 141–150. Springer.

Vach, W. (1989). Least squares approximation of additive trees. In Opitz, O., editeur, *Conceptual and numerical analysis of data*, pages 230–238. Springer Verlag.

Wakeley, J. (1993). Substitution rate variation among sites in hypervariable region 1 of human mitochondrial dna. *J Mol Evol*, 37(6) :613–23.

Wakeley, J. (1994). Substitution rate variation among sites and the estimation of transition bias. *Mol. Biol. Evol.*, pages 436–442.

Whitley, D. (1994). A genetic algorithm tutorial. *Statistics and Computing*, 4 :65–85.

Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.*, 10 :1396–1401.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites : approximate methods. *J. Mol. Evol.*, 39 :306–314.

Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. *TREE*, 11(9) :367–372.

Yang, Z. (1997a). Are big trees indeed easy ? *TREE*, 12(9) :357.

YANG, Z. (1997b). How often do wrong models produce better phylogenies ? *Mol. Biol. Evol.*, 14 :105–108.

YANG, Z. (1997c). PAML : a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, 13 :555–556.

YANG, Z. (2000). Maximum likelihood estimation on large phylogenies and analysis of adaptative evolution in human influenza virus A. *J. Mol. Evol.*, 51 :423–432.

YANG, Z., GOLDMAN, N., et FRIDAY, A. (1994). Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol. Biol. Evol.*, 11 :316–324.

YANG, Z. et KUMAR, S. (1996). Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. *Mol. Biol. Evol.*, 13 :650–659.

YANG, Z. et RANNALA, B. (1997). Bayesian phylogenetic inference using DNA sequences : a markov chain Monte Carlo method. *Mol. Biol. Evol.*, 14 :717–724.

ZHARKIKH, A. (1994). Estimation of evolutionary distance between nucleotide sequences. *J. Mol. Evol.*, 39 :315–329.

ZUCKERKANDL, E. et PAULING, L. (1962). *Horizons in Biochemistry*, chapitre Molecular disease, evolution, and genic heterogeneity, pages 189–225. Elsevier, Amsterdam.