Avertissement à la version électronique

La publication du livre « *La Reconstruction phylogénétique. Concepts et Méthodes* » (Masson, 1993), a connu un certain succès, si l'on en juge par le rapide épuisement des stocks. Les mauvaises langues insinueront que les raisons de ce succès tiennent davantage à la prudente politique éditoriale de l'éditeur qui n'a risqué qu'un tirage fort parcimonieux, plutôt qu'à une véritable popularité du livre lui-même. Pourtant il semble bien que plusieurs générations d'étudiants aient su mettre à profit les performances des photocopieuses pour pallier une pénurie vite manifeste…

Depuis la date de parution de ce livre, les méthodes phylogénétiques ont connu un développement considérable. Il s'agit d'une discipline qui n'est plus seulement réservée aux chercheurs avertis. Elle s'adresse également aux étudiants dès les premiers cycles et s'illustre même, avec plus ou moins de bonheur, dans les manuels scolaires des classes terminales. Parallèlement à cet élargissement rapide du public, les revues scientifiques internationales ou nationales, comme d'ailleurs les journaux de vulgarisation, ne craignent plus de publier des arbres phylogénétiques. C'est même devenu une nécessité méthodologique imposée par l'engouement récent pour les sciences de l'évolution. On assiste également à une vaste diversification des applications, qui quittent parfois le domaine traditionnel de la biologie pour s'aventurer vers ceux, peut-être plus incertains, de la linguistique, de l'éthologie ou de la science des textes. Enfin, de multiples ouvrages, pédagogiques ou savants (voir liste ci-dessous ; nous ne conseillons pas tous ces ouvrages au même titre, au lecteur d'en tirer le meilleur parti), ont été publiés ces 10 dernières années, au point qu'il n'existe que l'embarras du choix pour celui qui voudrait étancher sa soif de nouveautés dans le domaine des analyses phylogénétiques.

Compte tenu de cette floraison bibliographique, il nous a quand même semblé utile de remettre à disposition un livre qui conserve encore, aux dires de certains, quelques vertus pédagogiques, malgré ses dix ans d'âge, même si cette qualité s'applique plutôt aux *single malts* qu'aux manuels de phylogénétique. Bien qu'il soit maintenant incomplet ou dépassé sur plusieurs points, il garde cependant l'avantage d'être le seul ouvrage écrit en français à ce jour…

Cette version électronique de « *La Reconstruction phylogénétique. Concepts et Méthodes* » reste, sur le fond, la copie du livre paru en 1993. Il ne s'agit pas d'une nouvelle édition mise à jour. Seules quelques coquilles ont été corrigées. Il est donc de notre devoir d'avertir le lecteur d'aujourd'hui qu'il ne trouvera rien sur les développements les plus récents, qu'il s'agisse d'approches probabilistes (méthodes baysiennes) ou de cladistique structurale (analyse-à-trois-taxons). Cette version « en ligne » a bénéficié du travail d'édition de Yann Bertrand et Régis Debruyne et de l'accueil de la Société Française de Systématique sur son site Web. Qu'ils soient tous chaleureusement remerciés.

Pierre Darlu et Pascal Tassy
Avril 2004

1994. Scotland R.W., Siebert D.J., Williams D.M. *Models in phylogeny reconstruction*. The Systematics Association Publication, special volume n°52 Clarendon Press, London.

1996. Harvey P.H., Leigh Brown A.J., Maynard Smith J., Nee S. *New uses for new phylogenies*. Oxford University Press.

1996. Swofford D.L., Olsen G.J., Waddell P.J., Hillis D.M. Phylogenetic inference. In: *Molecular systematics*, Hillis D.M., Moritz C., Mable B.K. Eds. Sinaueur Associates Inc. Publishers.

1998. Page R.D.M., Holmes E.C. *Molecular evolution. A phylogenetic approach*. Blackwell Science.

1998. Kitching I.J., Forey P.L., Humphries C.J., Williams D.M. *Cladistics, Second Edition; The Theory and Practice of Parsimony Analysis*. The Systematics Association Publication N°11, Oxford Science Publication, Oxford University Press.

2000. Nei M., Kumar S. *Molecular evolution and phylogenetics*. Oxford University Press.

2001 Hall B.G. *Phylogenetic trees made easy. A How-to Manual for Molecular Biologists*. Sinauer Associates, Inc.

2003. Semple C., Steel M. *Phylogenetics*. Oxford Lecture Series in Mathematics and its Applications n°24 Oxford University Press.

2004. Felsenstein J. *Inferring phylogenies*. Sinauer Associates Inc. Publishers.

LA RECONSTRUCTION PHYLOGÉNÉTIQUE

Concepts et méthodes

P. DARLU

P. TASSY

ISBN : 2-225-84229-9

ISSN : 0754-4405

Peut-on se mettre d'accord sur une ressemblance ? Et, d'abord, qu'est-ce que la ressemblance ?

...

La ressemblance se dépare de l'inessentiel. Elle est l'essentiel réintroduit dans le circuit des formes, des idées, des métaphores et des alliances – essentiel conservé des rapports entre objets et parentés d'objets.

Edmond Jabès
Le Livre des Ressemblances, 1976.

TABLE DES MATIÈRES

(see contents page X)

CONTENTS

AVANT-PROPOS

La construction phylogénétique est l'une des disciplines en plein essor des sciences de l'évolution. Depuis une vingtaine d'années elle s'est constituée comme branche autonome. Assurément, la construction d'arbres généalogiques est aussi vieille que la recherche évolutionniste elle-même. Mais, longtemps intuitive, elle est aujourd'hui formalisée et repose le plus souvent sur une base mathématique, ou, à tout le moins, algorithmique.

Ce livre est né de notre expérience en matière de recherche d'abord, expérience de phylogénéticien pour chacun de nous, plutôt de généticien pour l'un et de paléontologue pour l'autre ; mais aussi et surtout, en matière d'enseignement, tant dans le cadre universitaire que dans le cadre de stages de formation permanente du CNRS. Il est vite apparu qu'il restait à écrire un ouvrage pédagogique en langue française accessible aux étudiants, aussi bien qu'aux systématiciens, biologistes, généticiens et paléontologues qui s'intéressent, plus généralement, aux questions d'évolution.

Nombreux sont les ouvrages traitant d'un aspect précis de la recherche phylogénétique : manuels de systématique évolutionniste, manuels de taxinomie numérique, manuels de cladistique comme le récent *The Compleat Cladist* (Wiley *et al.*, 1991). Plus rares sont les manuels de synthèse qui apportent une initiation aux différentes approches phylogénétiques actuellement pratiquées. L'ambition du présent volume est d'offrir un panorama de ces méthodes, ainsi qu'une discussion de leurs particularités, de leurs performances et de leurs limites. Le lecteur trouvera non seulement une introduction aux méthodes cladistique et phénétique de construction d'arbres, qui sont parmi les plus répandues, mais aussi une introduction aux approches probabilistes, moins fréquemment utilisées et discutées. En somme, la fonction de ce livre est d'initier le lecteur aux pratiques phylogénétiques par le concret - par l'exemple -, mais aussi en l'informant sur le

fond, c'est-à-dire sur les principes et modèles qui sous-tendent chaque méthode. Ainsi devraient être plus claires les raisons de choisir telle ou telle méthode, plus argumentées en tout cas que le choix de telle d'entre elles parce que « c'est ce qui se fait dans la maison ». Cet ouvrage n'est pas une introduction aux différents logiciels de construction d'arbres existant sur le marché. Néanmoins l'usage de l'outil informatique a fortement transformé la pratique quotidienne du phylogénéticien. Nous avons donc cherché à intégrer au cours des développements conceptuels touchant à chaque méthode, les procédures spécifiquement informatiques.

Traditionnellement, la construction d'arbres phylogénétiques ressortit au vaste domaine de la systématique, de la taxinomie. Toutefois nous n'avons pas cherché à approfondir la question des liens entre classification et phylogénie. Les livres sur la question sont nombreux (voir par exemple le tout récent ouvrage de Panchen, 1992) et figurent en tout état de cause dans la bibliographie en fin de volume. C'est véritablement la spécificité de la construction de l'arbre, c'est-à-dire l'établissement des liens phylogénétiques, qui fait le sujet du présent livre.

Qu'est-ce qui différencie la construction d'arbres phylogénétiques de celle de simples graphes arborescents ? En quoi se distinguent les différentes approches et méthodes ? Qu'est-ce qui distingue une analyse de distance d'une analyse cladistique, ou bien une analyse de parcimonie d'une analyse probabiliste ? Pourquoi donnent-elles éventuellement des résultats divergents à partir des mêmes données de départ ? Pourquoi utiliser telle méthode plutôt qu'une autre ? Quel est le rapport des constructions phylogénétiques à la théorie de l'évolution en général et aux modèles sur les processus évolutifs en particulier ? Quelle est l'originalité des méthodes probabilistes ? Nous espérons avec ce livre apporter quelques réponses à ces questions (et à bien d'autres), ainsi que des éléments de réflexion à propos des nombreux problèmes qui persistent aujourd'hui dans le champ de la recherche phylogénétique, tant dans le domaine dit traditionnel des recherches morphologiques qu'en biologie moléculaire.

Ce livre s'adresse aux étudiants de 2e et 3e cycles en biologie évolutive, paléontologie incluse, en espérant leur faciliter la tâche dans le dépouillement et la compréhension de la littérature phylogénétique aussi abondante et diversifiée que parfois hermétique. Il s'adresse aussi aux chercheurs non phylogénéticiens qui souhaitent avoir en main un ouvrage sur la question, tant aujourd'hui il est difficile d'apporter des informations évolutives indépendamment de toute construction d'arbre.

De nombreux points abordés dans ce livre sont actuellement discutés voire controversés. Certains sont des sujets de recherches en cours. Nous nous sommes efforcés de ne pas les esquiver, en restant toutefois très succinct. C'est le cas, par exemple, de nouveaux indices et tests relatifs à l'estimation de la robustesse des arbres qui ont été publié durant l'année 1992 et dont l'efficacité fait l'objet de travaux actuels. C'est le cas également des constructions d'arbres réticulés incluant des échanges géniques. Le plus souvent, nous n'avons donc développé que les méthodes les plus répandues.

Cet ouvrage est divisé en huit chapitres. Les trois premiers forment une manière d'introduction historique et conceptuelle des thèmes de la recherche phylogénétique : concepts de généalogie, d'arbre, de ressemblance, de caractère, d'homologie, de taxon. Les trois suivants (IV,V et VI) appartiennent, en gros, à la sphère du cladisme : la méthode cladistique proprement dite, les procédures de parcimonie et la méthode de compatibilité ; approches cladistique et de compatibilité peuvent en effet être regroupées en ce sens qu'il s'agit d'analyses de caractères. Le chapitre IV présente la méthode cladistique à la fois dans ses principes et dans sa pratique, sous un angle plutôt naturaliste. Le chapitre V est lui aussi en grande partie consacré à l'approche cladistique - mais sous un angle plus « informatique », en ce sens qu'y sont développées les procédures en usage dans les logiciels dits de parcimonie. Nous espérons que les redondances dans les chapitres IV et V seront plus comprises comme des correspondances que comme des répétitions. Le chapitre V contient en outre d'autres procédures de parcimonie non cladistiques au sens strict. Le chapitre VI est consacré à la méthode de compatibilité. Le chapitre VII détaille les méthodes phénétiques que l'on rassemble souvent sous le vocable de « taxinomie numérique ». Quoique ces approches se ressemblent dans la mesure où elles sont toutes des analyses de distance, elles se distinguent les unes des autres notamment par leurs différents présupposés relatifs aux processus évolutifs, c'est-à-dire par les critères de conversion d'arbres de distances en arbres phylogénétiques. Le chapitre VIII est consacré aux méthodes probabilistes, méthodes encore marginales dans le domaine de la recherche phylogénétique mais dont on peut attendre une expansion prochaine.

Nous tenons à remercier Hervé Le Guyader sans qui nous n'aurions jamais écrit ce livre. Merci également à Josué Feingold pour le soutien qu'il nous a toujours apporté au sein de l'Unité INSERM U155 et à Michèle Aosaka, Véronique Barriel, Jean Pierre Bocquet-Appel, Daniel Goujet, Hervé Philippe, Dominique Visset pour leurs conseils, suggestions, critiques et aides diverses.

CHAPITRE **I**

DE LA GÉNÉALOGIE À LA PHYLOGÉNIE

1. De Lamarck à Haeckel

Le terme « phylogénie » fut inventé par Ernst Haeckel en 1866 pour définir l'enchaînement des espèces animales et végétales au cours du temps. Jusqu'alors le concept était exprimé par le terme « généalogie ». Ce n'est que dans la dernière édition de l'*0rigine des espèces* (1872) que Charles Darwin introduisit le mot *phylogeny* avec la définition suivante : les lignes généalogiques de tous les êtres organisés. Le mot est resté. Nous définirons la phylogénie comme « le cours historique de la descendance des êtres organisés ».

Haeckel lui-même avait défini la phylogénie comme l'histoire du développement paléontologique des organismes par analogie avec l'ontogénie ou histoire du développement individuel. Les termes « développement » et « évolution » sont tous deux issus de l'embryologie. Pour qualifier les transformations organiques situées dans le temps géologique, le mot « évolution » supplanta progressivement à la fin du XIXe siècle celui de « développement » Haeckel fut l'un des artisans de ce succès qui se fit au détriment de « transformisme », terme synonyme d' « évolutionnisme » et qui reste le plus souvent associé à l'œuvre de J.-B. Monet de Lamarck, quoique ce dernier ne l'utilisât jamais.

Lamarck, en même temps qu'il conçut les bases de la théorie de l'évolution, publia dans sa *Philosophie zoologique* (1809) un schéma de filiation des animaux qui empruntait sa forme à l'image classique de la généalogie qui se lit de haut en bas (figure I-1). Ce schéma est présenté comme la « distribution générale » des animaux. Ce concept lamarckien exceptionnellement fécond s'oppose explicitement à la classification et se veut une construction qui doit exprimer « l'ordre représentant le plus possible celui même de la nature, c'est-à-dire l'ordre qu'elle a suivi dans la production des animaux et qu'elle a éminemment caractérisé par les rapports qu'elle a mis entre les uns et les autres ». Les « rapports » qu'évoque Lamarck sont les « parentés entre les corps vivants ».

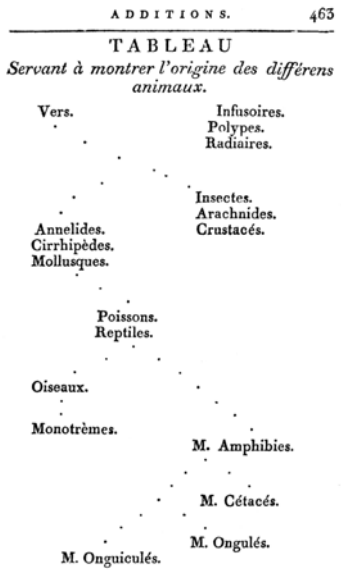


FIGURE I.1. *La filiation des animaux selon Lamarck (1809, vol.2, p.463).*

Lamarck écrit aussi que la « distribution » doit former une « série et non une ramification réticulaire »; cette série devant être « une véritable échelle relativement aux grandes masses, c'est-à-dire les grandes subdivisions du monde animal, les espèces (formant) souvent autour des masses dont elles font partie des ramifications latérales dont les extrémités offrent des points véritablement isolés ». Le concept de phylogénie est donc ébauché chez Lamarck. Contre la légende qui fait de Lamarck un incompris qui fut ignoré de son temps, il convient de souligner la force de ces ébauches lamarckiennes. En outre, c'est un naturaliste lamarckien, Frédéric Gérard, tardivement reconnu, qui conçut en 1845, soit quinze ans avant *l'Origine des espèces,* l'expression « théorie de l'évolution des êtres organisés » dans son sens moderne, preuve, s'il en est, de la fécondité de l'œuvre de Lamarck.

Néanmoins c'est dans l'*Origine des espèces* de Darwin (1859) qu'on trouve à la fois l'idée de phylogénie comme cours historique unique suivi par l'évolution et l'image de l'arbre phylogénétique. Renversant la lecture de la généalogie, la phylogénie est représentée sous la forme d'un arbre – avec un tronc, des branches, des rameaux – qui se lit de bas en haut. La seule illustration incluse dans l'*Origine des espèces* est une image de filiation entre espèces hypothétiques situées dans un contexte stratigraphique : le temps géologique se lit de bas en haut (figure.I-2). Les premières phylogénies jamais publiées se lisent de cette façon, qu'il s'agisse de la phylogénie du monde vivant par Haeckel (1866) où l'échelle du temps n'est qu'implicite (figure I.3) ou de celle par Albert Gaudry (1866) intégrant divers mammifères actuels et fossiles et où l'échelle du temps est explicite (figure I.4).

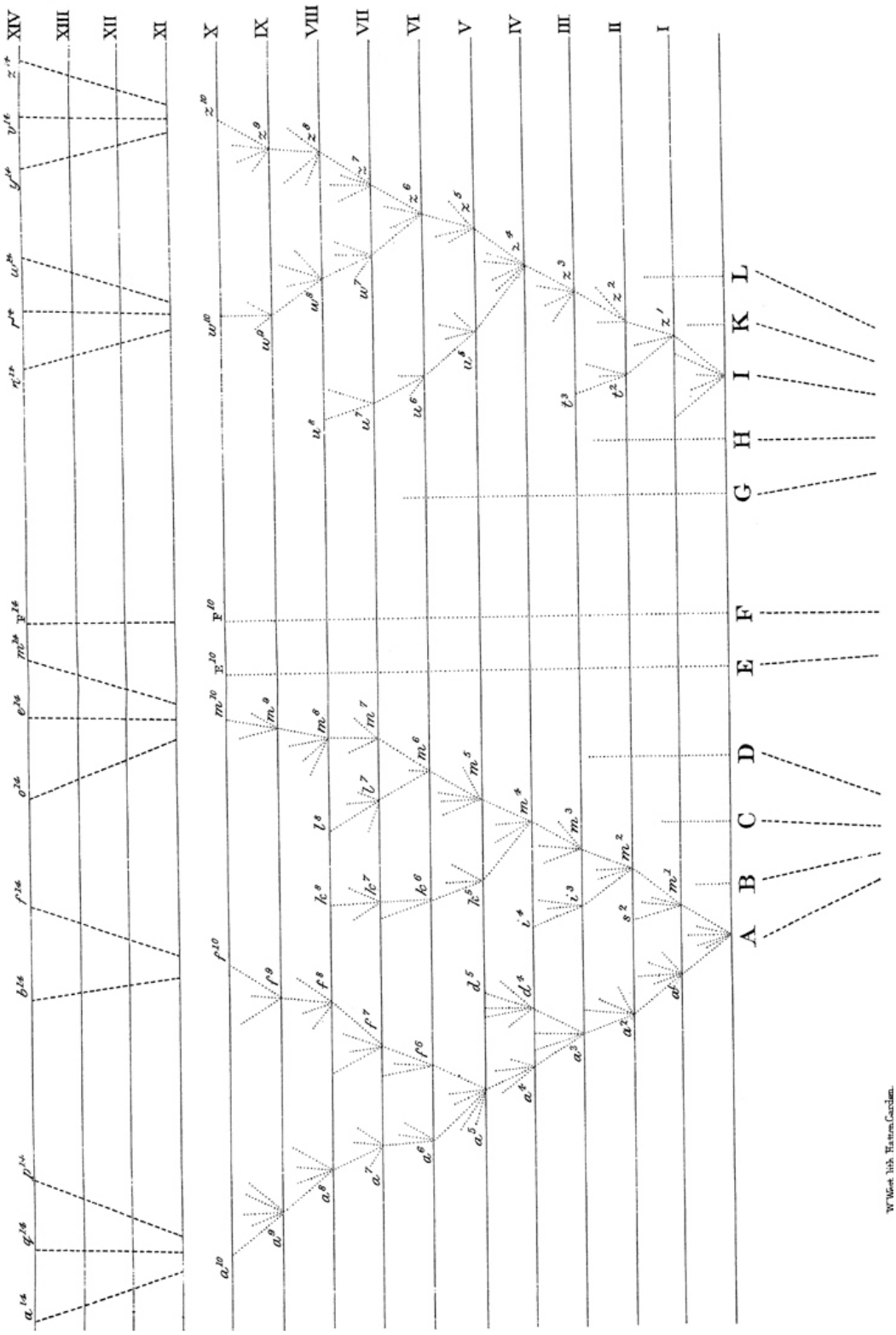


FIGURE I.2. Le modèle de la descendance avec modification selon Darwin (1859, tableau face p.117). A-L : espèces ; I-XIV : étages géologiques.

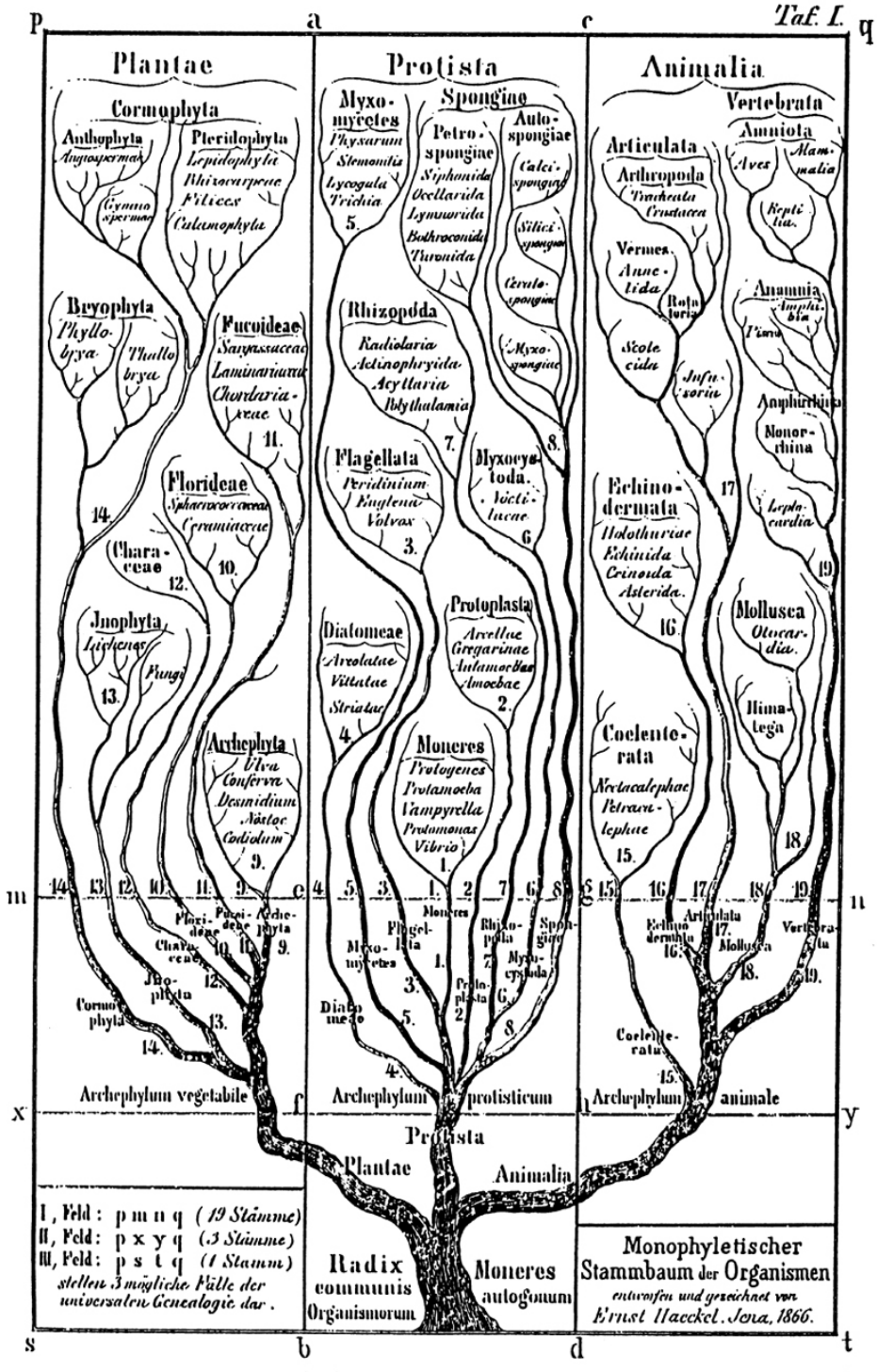


FIGURE I.3. *L'arbre phylogénétique des êtres vivants selon Haeckel (1866, vol.2, pl.1).*

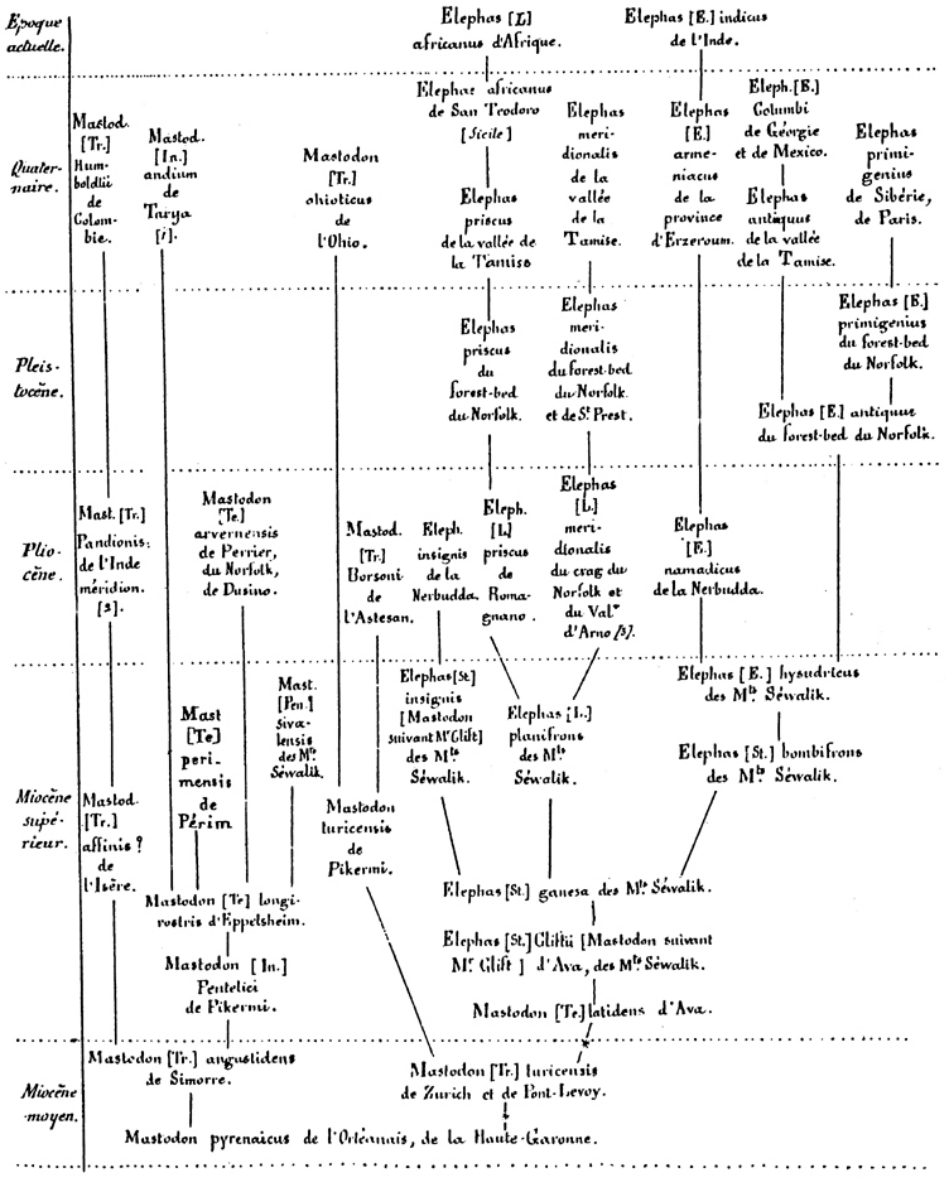


FIGURE I.4. *L'évolution des proboscidiens selon Gaudry (1866).*

Toutefois, donner à différents taxons leur dimension stratigraphique en les situant dans l'échelle des temps géologiques ne revient pas à construire un arbre phylogénétique. Agassiz publia par exemple, dès 1844, une « généalogie de la classe des poissons » située dans un cadre stratigraphique où les positions respectives des différents groupes reconnus peuvent évoquer un arbre. Mais aucune connexion entre ces groupes n'est indiquée. La raison en est que, selon Agassiz, cette « généalogie » n'implique pas de lien évolutif : Agassiz était fixiste.

2. L'arbre phylogénétique

L'arbre phylogénétique est une construction-clé dans l'histoire de la biologie et de la géologie. Son succès opérationnel ne s'est jamais démenti. Imaginons que nous retracions l'histoire de trois espèces choisies arbitrairement (actuelles ou fossiles). En remontant le temps, nous pouvons espérer relier les deux espèces qui dérivent de la même espèce ancestrale ; en remontant plus loin encore, nous rencontrons l'espèce ancestrale des trois espèces. Le dessin en deux dimensions de ces relations de parenté est un arbre composé d'une succession de branchements.

Au reste, on n'a jamais trouvé meilleure façon d'illustrer, en deux dimensions, les relations de parenté entre espèces ou groupes d'espèces, en fonction du temps (verticalement) et de la diversité taxinomique (horizontalement). Dans le cas de la diversité, de façon conventionnelle et le plus souvent symbolique, la dimension horizontale permet aussi une représentation du degré de divergence morphologique de deux branches à partir du point de branchement, ce que Darwin nomme la « somme des modifications ».

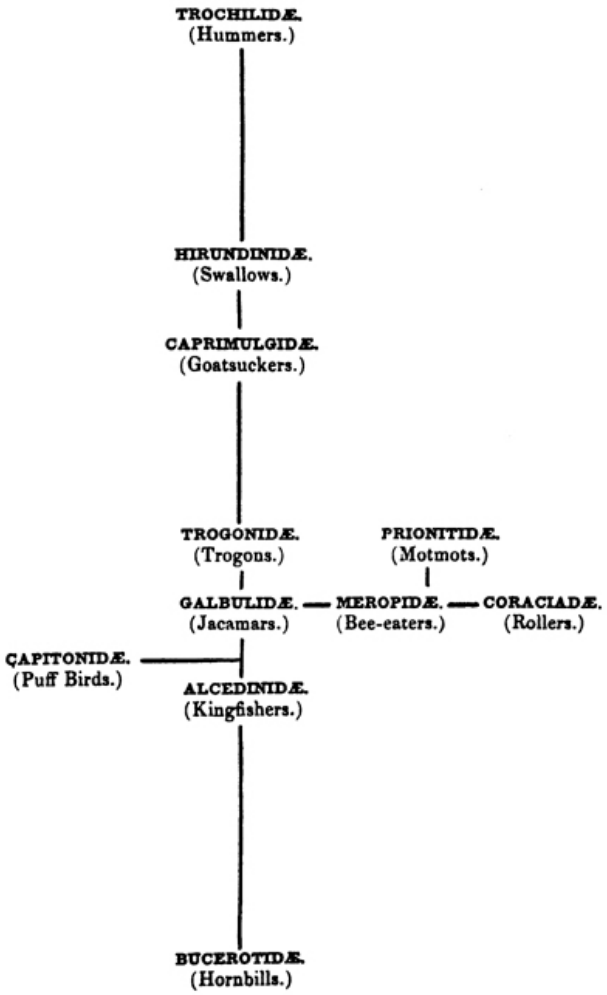


FIGURE I.5. *Classification des oiseaux fissirostres selon Wallace (1856).*

L'arbre phylogénétique rompt – suivant en cela les anticipations lamarckiennes – avec la tradition des représentations en « réseaux » héritées du XVIIIe siècle qui influençaient encore fortement les naturalistes évolutionnistes. Les réseaux, comme les arbres, visent à relier linéairement des groupes selon leurs degrés d'affinités. Cependant la lecture d'un réseau n'impose nullement un point de départ obligé : il n'y a pas de souche à la base du schéma. Nelson et Platnick (1981) ont déjà fait remarquer que A.-G. Wallace – l'autre « père » de l'évolutionnisme – fournit en 1856 un schéma de ce type à propos des affinités des oiseaux de l'ordre des fissirostres. Il nous manque dans un tel schéma qui s'apparente à un arbre non enraciné, l'indication d'un itinéraire : le cours historique (figure I.5). Avec l'arbre phylogénétique, l'itinéraire nous est donné par le point de départ obligé : la racine de l'arbre. Le temps généalogique étant irréversible, l'histoire de la descendance est unique. Le problème de reconstruction d'un fait historique n'admettant qu'une seule solution, l'un des buts de la science évolutionniste du XXe siècle fut – et reste – l'élaboration et le perfectionnement des méthodes de reconstruction phylogénétique.

3. Quelques définitions majeures

Le vocabulaire propre aux phylogénéticiens présente quelques particularités, mais aussi quelques ambiguïtés dans la mesure où il résulte d'un mélange de termes issus de la théorie des graphes, et dont l'usage n'est pas toujours correct, et d'autres termes, généralement plus imagés, hérités de la tradition évolutionniste remontant au XIXe siècle.

Quelques éclaircissements sont donc nécessaires. Partant de définitions de la théorie des graphes, nous expliquerons les termes consacrés maintenant par l'usage, en levant toute ambiguïté.

Les relations entre divers objets (populations, espèces, taxons, unités évolutives) peuvent être représentées selon un diagramme assez général tel que celui de la Figure I.6a. On y distingue des *sommets* et des *liens*, appelés aussi arcs ou arêtes, reliant ces sommets.

3.1. Les sommets

On distingue les sommets internes ou *nœuds* (figure I.6 : N_5, N_6, N_7, N_8) et les sommets externes ou *feuilles* (T_0, T_1, T_2, T_3, T_4). Dans le contexte phylogénétique, ces derniers, pour lesquels on dispose de données observées, sont les *extrémités* ou *taxons terminaux*, ou *unités évolutives* (UE). Les nœuds constituent généralement des taxons ou des UE hypothétiques (UEH) dans la mesure où leur existence n'est pas fondée sur l'observation de caractères mais résulte seulement du processus de reconstruction lui-même.

3.2. Les liens

La relation entre deux sommets constitue un lien appelé souvent *segment* ou *branche*. Deux nœuds internes sont reliés par un *lien interne* (l_{65}, l_{68}, l_{76}, l_{78} et l_{85}) tandis qu'un nœud et une feuille sont reliés par un *lien externe*, désigné généralement comme une branche terminale ou périphérique (l_{50}, l_{16}, l_{27}, l_{37}, l_{48}).

A chacun des liens, on peut associer une *mesure*, comme par exemple une distance, génétique ou autre, une durée, une quantité d'évolution ou un nombre de mutations. On peut également y associer un *poids* qui peut être une vitesse d'évolution, un effectif, un taux de mutation, un coût quelconque etc. Un poids nul le long d'un lien entre deux UE revient à supprimer ce lien, donc à fusionner deux UE.

Par ailleurs un lien peut être *orienté*, c'est-à-dire avoir une mesure et/ou un poids différent selon que la relation est parcourue dans un sens ou dans un autre. Un poids nul le long d'un lien orienté signifie que l'orientation est univoque, d'un nœud vers un autre et non en sens inverse. C'est le sens classique en phylogénétique. On dira que le premier nœud est l'*ancêtre* par rapport au second défini comme *descendant.*

Une autre caractéristique concerne *le nombre de liens attachés à un sommet* : ce nombre peut varier d'un sommet à un autre. Dans le cas de la figure I.6a (liens non orientés), il est égal à 1 pour un sommet externe (T_0 à T_4), à 3 pour N_5, à 4 pour tous les autres. Dans la figure I.6b (liens orientés univoques), il faut distinguer le nombre des liens pointant vers le sommet de ceux qui en partent.

La figure I-6 montre deux types de relations entre UE : dans la première (a) aucun sens d'évolution n'est explicite puisque les liens sont non orientés. En revanche, dans la seconde (b) les liens sont orientés depuis T_0 (UEH ancestrale) jusqu'aux taxons terminaux (T_1 à T_4). Dans les deux cas, la longueur des liens peut être représentée proportionnellement à la mesure reliant les taxons entre eux, ou, au contraire, n'avoir aucune signification autre que de décrire les relations entre les nœuds.

3.3. Réseaux et arbres

Il est nécessaire de distinguer d'abord entre réseaux et arbres, ces derniers pouvant être non enracinés ou enracinés.

3.3.1. Le réseau

Il s'agit d'un graphe connexe (il existe au moins un chemin entre chaque paire de sommets) et cyclique (c'est-à-dire assimilable à une chaîne dont les extrémités coïncident).

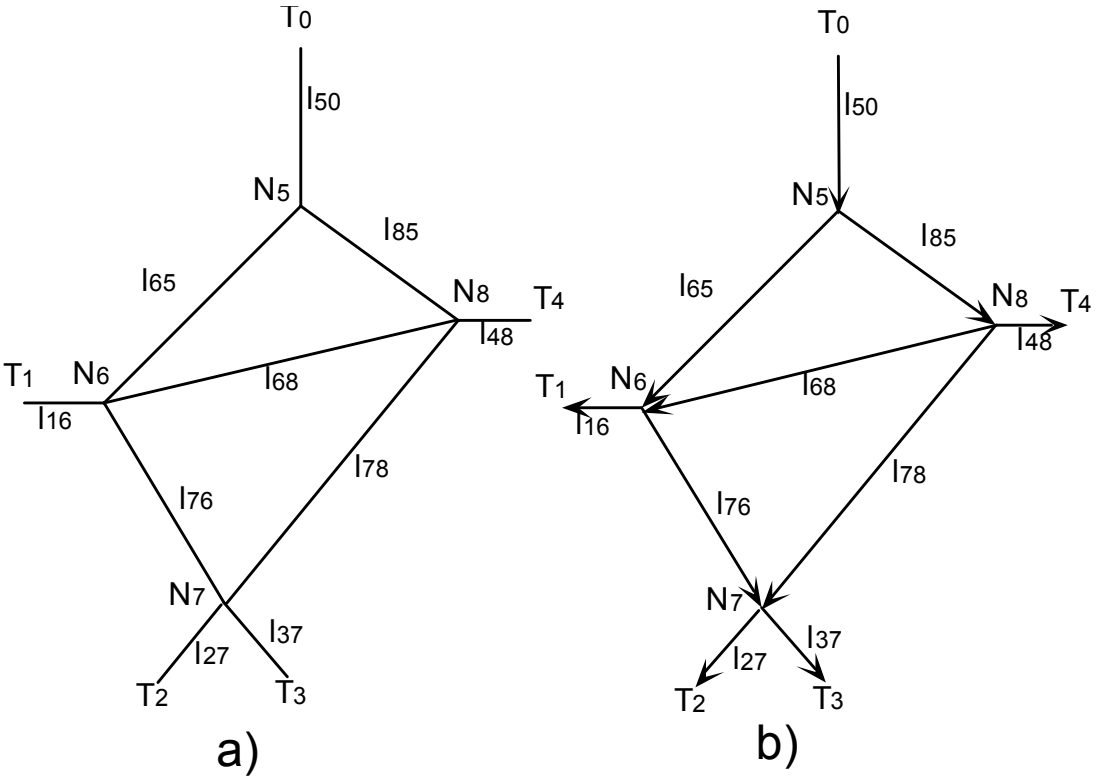


FIGURE I.6. *Schéma des liens non orientés (a) ou orientés (b) entre 5 sommets terminaux ou branches (T_0 à T_4) et 4 nœuds (N_5 à N_8)* (a) *ou orientés* (b).

3.3.2. L'arbre

1) L'*arbre non enraciné* (*unrooted tree*) est un graphe connexe non cyclique. Il n'existe donc pas de boucle, c'est-à-dire qu'un seul et unique chemin permet de passer d'un sommet à un autre. Il n'y a qu'une façon de joindre entre elles deux UE. De ce fait, les sommets terminaux (les feuilles) ne sont reliés aux nœuds internes que par un seul lien non orienté (branche). Cela revient en fait à attribuer un poids nul aux branches qui permettraient un cycle. Par exemple, sur la figure I.6, le lien l_{68} entre les nœuds N_6 et N_8 et le lien l_{78} entre les nœuds N7 et N8 sont supprimés dans la figure I.7. Le terme de réseau (*network*) est souvent considéré comme synonyme d'arbre non enraciné, bien que cet usage ne correspondent pas au vocabulaire de la théorie des graphes (Barthélémy et Guénoche, 1988). Dans la mesure du possible, on évitera d'utiliser ce terme, sauf cas particuliers comme celui où il existe des mélanges entre UE à différents niveaux de l'arbre.

2) L*'arbre enraciné* comporte une contrainte supplémentaire par rapport à l'arbre non enraciné dans la mesure où on est amené à définir une *origine* appelée souvent *racine* ou *ancêtre.* Les *liens* sont alors *orientés* de manière univoque de telle façon qu'un seul lien se dirige sur un sommet tandis que deux liens (dichotomie ou bifurcation) ou plusieurs liens (polytomie ou multifurcation) peuvent en partir. C'est le cas de la figure I.7b, déduite de la figure I.6b en supprimant les liens l_{68} et l_{78}, et où l'ancêtre est T_0.

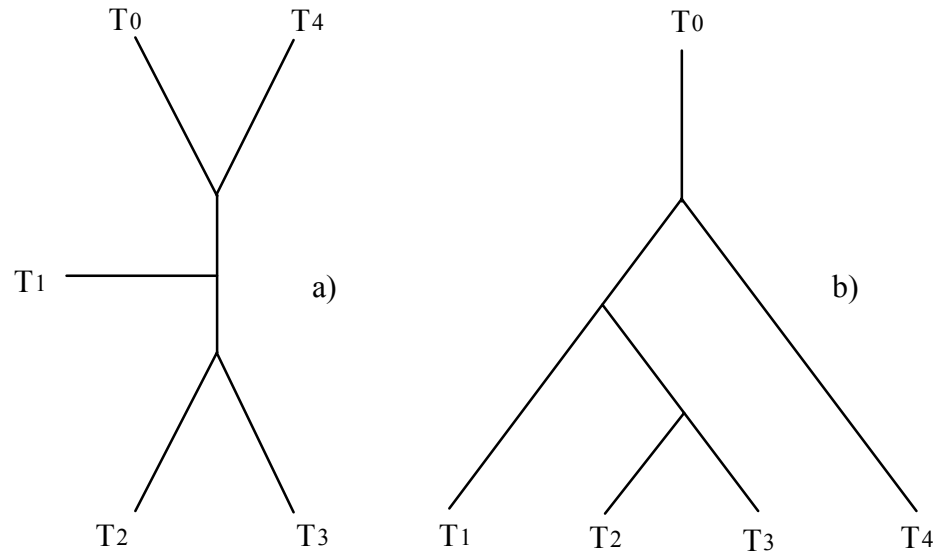


FIGURE I.7. *Schémas des liens non orientés (a) ou orientés (b) quand les liens l_{68} et l_{78} de la figure I.6a et I.6b ont un poids nul. L'arbre (a) est non enraciné tandis que l'arbre (b) possède une racine T_0*

Dans le vocabulaire de la systématique moderne, un usage regrettable met parfois en synonymie les mots *arbre* et *phylogénie*. Or, un arbre n'est pas nécessairement enraciné alors qu'une phylogénie est une histoire et a donc un point de départ, la racine, et une seule orientation de lecture. L'expression d'*arbre phylogénétique* n'est donc pas redondante, tandis que l'expression « phylogénie non enracinée » est à proscrire catégoriquement. Par la suite nous réserverons le plus souvent le terme d'*arbre* pour qualifier un arbre *enraciné*, parlant d'*arbre non enraciné* ou, plus rarement, de *réseau* pour les autres types d'arbres.

3.4. Variétés d'arbres

a — *Le dendrogramme* est un arbre exprimant les liens entre taxons sous la forme d'une succession de branchements. Il ne désigne rien d'autre qu'un arbre dont les éléments terminaux sont les taxons ou UE observés. Ce terme est assez large pour ne rien exprimer quant à la procédure utilisée pour son obtention.

b — *Le cladogramme* est un dendrogramme exprimant les relations phylogénétiques entre taxons et construit à partir de l'analyse cladistique où les points de branchements (les nœuds) sont définis par des synapomorphies. Ce mot a été créé la même année par Mayr (1965) et par Camin et Sokal (1965) avec des sens un peu différents.

c — *Le phénogramme* est un dendrogramme produit par la taxinomie numérique où les relations entre taxons expriment les degrés de similitude globale, défini simultanément par Mayr (1965) et par Camin et Sokal (1965).

d — *Le phylogramme* est un dendrogramme exprimant les branchements cladistiques *et* le degré de divergence adaptative subséquente aux branchements (Mayr, 1969).

4. Combien d'arbres ?

Il n'existe qu'un seul arbre reliant différentes unités évolutives passées ou actuelles : c'est l'arbre évolutif, celui qui raconte l'histoire de la descendance. L'ambition de la reconstruction phylogénétique est de distinguer cet arbre « vrai » parmi l'ensemble des arbres que l'on peut théoriquement reconstruire à partir des différentes UE observées. Il est donc important de connaître le nombre *T* de tous ces arbres possibles. Les méthodes décrites dans ce livre conduisent à effectuer un choix, généralement limité à un seul arbre, parmi tous les arbres possibles.

Dans le cas simple où le nombre *n* d'UE est égal à 4 (A, B, C, D), la figure I.8 montre que 4 arbres non enracinés sont possibles dont trois (T_x, T_y, T_z) possèdent deux nœuds internes et le quatrième (T_w) un seul nœud interne. Ces quatre solutions sont obtenues par l'insertion de la quatrième UE (ici D) sur l'une des trois branches reliant les trois autres UE (A, B et C) ou sur le nœud joignant ces trois branches.

Pour passer d'un arbre non enraciné à un arbre enraciné, il suffit de placer l'ancêtre sur l'une quelconque des branches. On supposera que la position de l'origine est distincte des nœuds et des UE. Pour chacun des arbres T_x, T_y et T_z, il y aura donc 5 localisations possibles pour cette origine, soit 15 arbres différents au total. Avec T_w, 4 positions sont possibles pour cette origine. En conclusion, 19 arbres différents reliant entre elles 4 UE seulement peuvent donc être construits.

De la même façon, si, au lieu de vouloir placer sur un arbre non enraciné de 4 UE une origine ou un ancêtre, on souhaitait agglomérer une cinquième UE, 15 positions seraient également possibles à partir de T_x, T_y et T_z (représentées sur la figure I.8) et 4 à partir de T_w.

Le cas de figure de 4 arbres construits à partir de 4 UE sera privilégié par plusieurs méthodes de reconstruction comme la méthode des invariants (Chapitre V.5).

Le calcul du nombre total d'arbres non enracinés présentant 3 segments par nœuds internes repose sur le raisonnement récurrent suivant (Edwards et Cavalli-Sforza, 1964 ; Cavalli-Sforza et Edwards, 1967 ; Felsenstein, 1978a) : un arbre non enraciné de ce type ayant *n* UE possède :

$n_i = n - 2$ nœuds internes,

$s_i = n - 3$ segments internes,

$s_n = n$ segments externes.

Il est possible de rajouter une UE supplémentaire sur chacun des segments, internes ou externes, donc en $s_i + s_n = 2n - 3$ endroits possibles. Si le nombre des réseaux différents pour $n - 1$ UE est T_{n-1}, ce nombre sera, pour *n* UE, égal à :

$$T_n = T_{n-1} * (2(n-1) - 3)$$

Finalement on peut donc écrire :

$$T_n = \prod_{k=3}^{n} (2k - 5)$$

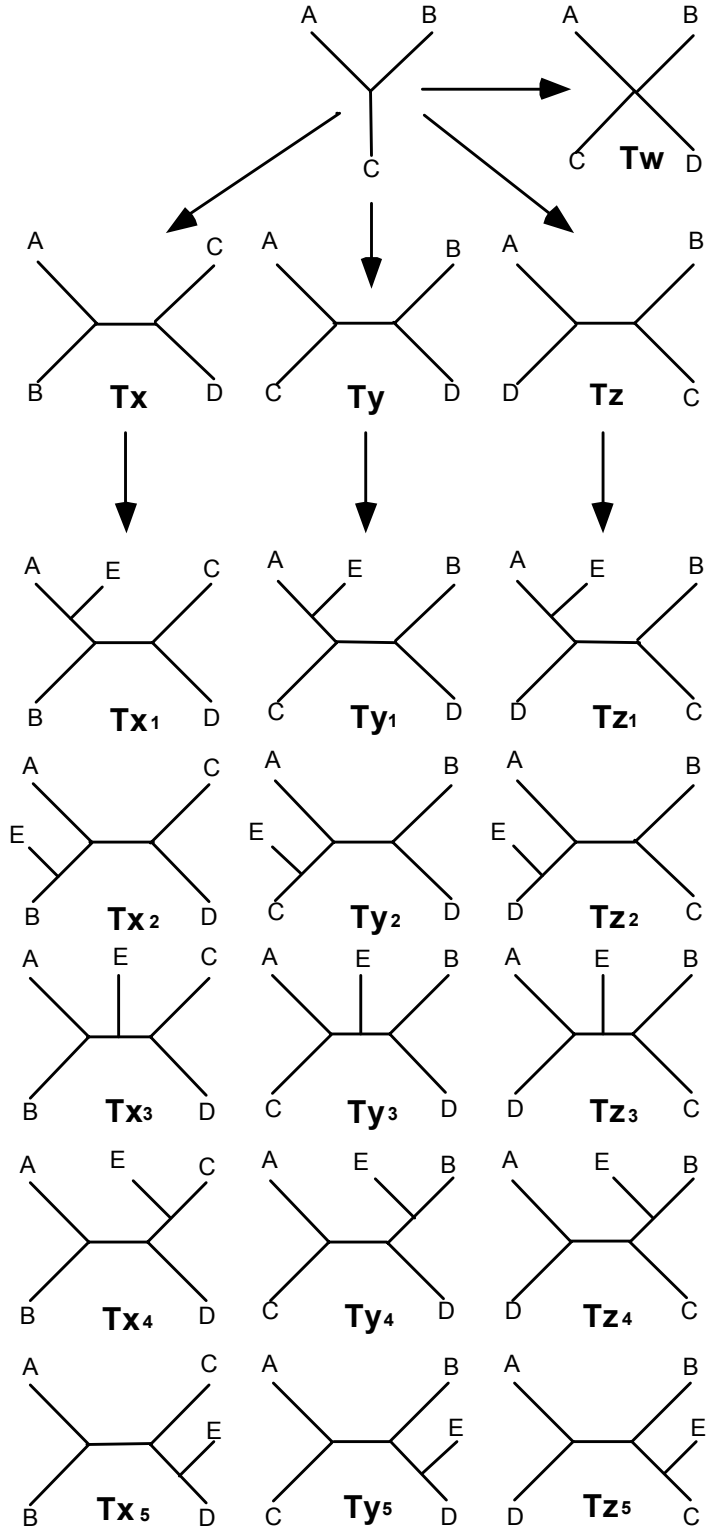


FIGURE I.8. *Représentation des 15 arbres différents non enracinés que l'on peut obtenir à partir de 5 UE : A, B, C, D, E. Lorsque A est considéré comme l'ancêtre, ces 15 arbres deviennent les 15 arbres différents enracinés que l'on peut obtenir à partir de 4 UE différentes (B, C, D et E).*

Le nombre total d'arbres enracinés dichotomiques différents ayant *n* UE terminaux s'obtiendrait, par un raisonnement analogue :

$$T'_n = \prod_{k=2}^{n} (2k - 3)$$

Le nombre d'arbres devient très vite élevé lorsque *n*, le nombre d'UE, augmente. Ainsi, avec $n = 10$ et $n = 20$, les nombres d'arbres dichotomiques sont respectivement :

$$T'_{10} = 34\,459\,425 \text{ et } T'_{20} = 8\,200\,794\,532\,637\,891\,559\,375 > 8*10^{21}$$

On comprend que la recherche exhaustive de l'arbre le plus parcimonieux ou le plus vraisemblable, par énumération de tous les arbres possibles, ne soit réalisable que lorsque *n* ne dépasse pas la dizaine, mais devienne quasiment impossible dès lors que *n* la dépasse, même avec les ordinateurs les plus puissants. Ce résultat implique que soit mise en application une stratégie de recherche de l'arbre le plus parcimonieux ou le plus vraisemblable qui n'impose pas cette recherche exhaustive. Une telle stratégie doit être efficace en ce sens qu'elle ne doit laisser que peu ou pas de chance de ne pas trouver l'arbre recherché. Plusieurs algorithmes seront décrits dans le chapitre V qui permettent de réaliser ce travail.

CHAPITRE **II**

LA PROBLÉMATIQUE PHYLOGÉNÉTIQUE

La construction phylogénétique est une entreprise vénérable qui plonge ses racines dans les oeuvres des grands évolutionnistes du XIXe siècle, Lamarck, Darwin, Haeckel. Mais ses bases conceptuelles, la règle du jeu de construction, n'ont que rarement été abordées sur le fond.

La construction phylogénétique s'appuie sur le concept de base de « la descendance avec modification ». Quels sont les caractères observés chez deux ou plusieurs espèces qui indiquent une proche parenté ? Ce sont ceux hérités de leur ancêtre commun. Le postulat de base est que la ressemblance est intelligible en termes d'ascendance commune. Le problème général est donc celui de l'inférence sur l'ancêtre et sur ses caractères à partir de l'observation des caractères des taxons terminaux (figure II.1).

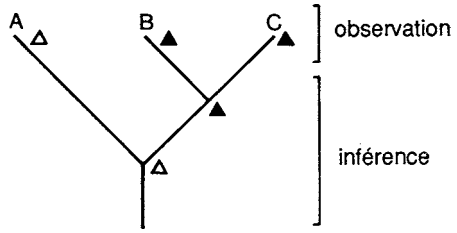


FIGURE II.1. *Le domaine de l'observation et de l'inférence dans la construction phylogénétique (A, B, C : taxons ; triangles : caractères).*

Jusqu'à l'œuvre de taxinomie fondamentale de l'entomologiste Willi Hennig (1950), la construction phylogénétique obéissait au principe du « triple parallélisme » (*threefold parallelism*), une expression conçue par le zoologiste Louis Agassiz (1859) et reprise par Haeckel pour exprimer que l'anatomie comparée, l'ontogénie et la paléontologie fournissent les sources de la reconstruction phylogénétique.

1. Le « triple parallélisme »

Agassiz, l'un des grands anatomistes du XIXe siècle, n'était pas évolutionniste. Ce fut même l'un des opposants les plus résolus au darwinisme. Il ne conçut le principe du « triple parallélisme » que dans la perspective de la classification biologique. Haeckel, darwinien militant, reprit à son compte ce principe dans une perspective évolutionniste. Ce transfert éclaire les liens profonds entre classification et phylogénie. Selon Agassiz, partisan des créations répétées, établir les affinités entre les êtres organisés et les situer dans leur contexte stratigraphique, c'est faire oeuvre de classificateur. Selon Haeckel et les évolutionnistes modernes, une telle pratique relève de la reconstruction phylogénétique. On ne s'étonnera donc pas que les différentes approches méthodologiques concernant la construction des arbres phylogénétiques se sont d'abord opposées au travers des écoles de systématique. Mais quoique les systématiciens s'affrontèrent, et s'affrontent toujours, sur le statut des classifications biologiques et sur la nature des informations qu'elles transmettent, l'accord s'est fait sur nombre de principes d'analyse phylogénétique. Il reste que les débats sur les classifications ne sont pas non plus sans effet sur les pratiques phylogénétiques. Nous y reviendrons à la fin de ce chapitre.

L'anatomie comparée, l'ontogénie et la paléontologie offrirent donc selon Haeckel la triple source des informations phylogénétiques.

1.1. L'anatomie comparée

L'anatomie comparée fournit d'abord des informations collectées sur l'actuel et s'applique naturellement à la morphologie. La recherche des homologies – similitudes liées à la descendance – est la base même de l'anatomie comparée. Le critère primordial d'identification des traits homologues est l'identité de position ou principe des connexions d'Etienne Geoffroy Saint-Hilaire repris par Richard Owen : un organe est homologue chez deux espèces ou plus si, sous quelque forme ou fonction que ce soit, il a les mêmes connexions avec d'autres organes. Dans le bras, l'os impair mince et allongé de l'homme, qui s'articule avec l'omoplate et l'avant-bras, est homologue à celui de la taupe, plus large que haut et dont l'orientation est différente (le coude regarde vers le haut) mais dont les connexions sont identiques : c'est l'humérus. Richard Owen qui, le premier, usa du terme « homologie » (1845) – Geoffroy Saint-Hilaire parlait pour le même concept d'« analogie » – fut un systématicien non darwinien qui ne se préoccupait que de classification et non de phylogénie. La perspective évolutionniste explique l'homologie par la descendance. Les espèces partageant des traits homologues sont apparentées ; le terme « apparenté » n'a pas qu'une signification classificatoire : il a un sens généalogique.

Si l'on se replace dans le contexte de l'anatomie comparée du XIXe siècle, il apparaît donc que le critère de reconnaissance de l'homologie est indépendant de la définition du concept et de son explication évolutionniste. L'identification de l'homologie par l'anatomie comparée repose sur le principe structural des

connexions, et conçoit la ressemblance dans le cadre d'un rapport de position. Il reste que le critère de ressemblance a toujours posé nombre de problèmes aux classificateurs. Différents traits jugés homologues chez différentes espèces indiquaient éventuellement des regroupements différents. Par exemple, de nombreuses espèces de mammifères possèdent un humérus pourvu d'une crête supra-condylienne saillante. D'un point de vue structural cette crête est homologue chez ces espèces : elle est toujours située à l'extrémité distale de l'os, sur la face latérale et au-dessus du condyle articulaire. Or les mammifères qui montrent cette crête, comme le blaireau, le phoque ou l'éléphant, ne sont pas pour autant étroitement apparentés. Les systématiciens du XVIIIe siècle cherchèrent à résoudre le problème des contradictions induites par la répartition des caractères par le principe de « subordination des caractères », conçu par A.-L. de Jussieu. Les caractères « constants » sont plus importants que les caractères « inconstants ». Ce principe, où perce la notion de congruence, fut quelque peu altéré au XIXe siècle et transformé en un principe de recherche des caractères « fondamentaux », les « bons » caractères, ceux dont on pense *a priori* qu'ils sont plus importants que les autres. Si elle contribua à la conception des grands « plans d'organisation » des êtres vivants, l'application de ce principe ne permit pas de résoudre parfaitement les contradictions. Le principe d'homologie, vu comme un rapport de position, ne le permit pas non plus. On sait aujourd'hui que la source des contradictions est le processus évolutif lui-même : des caractères identiques du point de vue de la ressemblance sont apparus indépendamment chez différentes espèces. En réalité, pour être plus efficace du point de vue de la reconstruction phylogénétique, la notion d'homologie doit être soumise à l'application d'un autre principe, celui de congruence. On reprendra cette question dans le chapitre IV, en lui associant la question de l'homologie moléculaire.

Le cadre général de l'application de l'anatomie comparée à la reconstruction phylogénétique est fourni par la notion de complexité. Ce qui est simple est primitif et ce qui est complexe est évolué. Il s'agit bien d'un cadre général et non d'une loi qui ne souffrirait pas d'exception. En effet, la perte d'un organe peut aboutir à un état plus simple que l'état initial. Une espèce évoluée peut sembler, à certains égards, plus simple que son ancêtre.

1.2. L'ontogénie

L'ontogénie fournit une source directe d'observation des transformations. Au cours du développement individuel, des caractères juvéniles peuvent se transformer et changer de forme (et même de fonction) chez l'adulte. L'embryologie fournit une information empirique particulièrement féconde : aux premiers stades du développement les embryons appartenant à divers grands groupes taxinomiques se ressemblent plus entre eux qu'ils ne ressemblent à leurs formes adultes. Autrement dit, au cours du développement, ce qui apparaît d'abord est général, ce qui se transforme subséquemment est particulier (précédence ontogénique). Le pionnier de l'embryologie comparée, K.E. von Baer, énonça quelques lois embryologiques dont celle de la précédence. ontogénique. Von Baer, toutefois, n'était pas darwinien. C'est Haeckel qui interpréta la « loi » énoncée par von Baer sous l'expression devenue fameuse,

quoique vite controversée : « l'ontogénie récapitule la phylogénie ». Autrement dit, ce qui apparaît d'abord dans l'ontogénie est primitif. De la sorte, les données recueillies par l'observation de l'ontogénie peuvent être interprétées à des fins de reconstruction phylogénétique.

La « loi de récapitulation » de Haeckel, dite encore « loi biogénétique fondamentale », ne souffrait, pour son auteur, aucune exception. Or les exceptions, sur lesquelles insistèrent notamment au XXe siècle Sir Gavin de Beer (1930, 1958) et Stephen Jay Gould (1977), discréditèrent d'une certaine manière, la loi biogénétique, d'autant que celle-ci avait été souvent prise au pied de la lettre : au cours du développement, l'embryon tout entier, devrait passer par les stades adultes de ses ancêtres fossiles, ce qui n'est évidemment pas le cas. Nous verrons plus loin que la loi biogénétique redéfinie au niveau des caractères, et non des organismes tout entiers, garde tout son intérêt phylogénétique.

1.3. La paléontologie

La paléontologie fournit des informations directement issues du temps géologique. Les êtres fossilisés nous donnent des éléments de ce qui a véritablement eu lieu au cours du temps, et sont des fragments de la vie passée. L'interprétation phylogénétique des fossiles est subordonnée au principe de l'anatomie comparée : la mise en évidence des homologies. Mais, depuis Gaudry, la position stratigraphique des fossiles est tenue pour fournir une information primordiale vis-à-vis de la reconstruction phylogénétique. L'ancienneté représente un élément pour identifier ce qui est primitif. Si, à l'époque de la publication de la *Philosophie zoologique* ou de l'*Origine des espèces*, les archives paléontologiques n'étaient guère importantes, il n'en est plus de même aujourd'hui. La réalité des fossiles ne pouvant être mise en doute, le critère paléontologique est devenu au cours du XXe siècle le principal critère phylogénétique pour maints biologistes.

De ce point de vue, la phylogénie des groupes organiques sans archives fossiles, ou pour lesquels l'enregistrement fossile est très lacunaire, était considérée comme largement spéculative. Le rôle prééminent ainsi accordé aux fossiles était tout entier dû à la dimension chronologique, mais non à la manière de traiter l'information qu'ils transmettent. En réalité, les caractères morphologiques des fossiles étaient analysés ni plus ni moins comme ceux des formes actuelles : selon les critères de l'anatomie comparée. Simplement, grâce aux fossiles, les homologies pouvaient, d'une certaine manière, être « datées ». A l'inverse, l'irruption au cours des années soixante de nouveaux caractères qui ne se fossilisent pas (les caractères biochimiques), allait rapidement se traduire par des phylogénies sans fossiles. Celles-ci, par contre-coup, relativisèrent l'importance accordée à la paléontologie. On en vint à minimiser l'information apportée par les fossiles : ceux-ci, souvent fragmentaires, n'apporteraient au fond que de faibles renseignements. Une fois encore, c'était plus la nature des caractères utilisés que la façon d'interpréter l'information transmise qui dicta une telle attitude.

Aucune de ces trois sources d'inférence phylogénétique n'est à l'abri de l'erreur d'interprétation. Des traits tenus pour homologues peuvent n'être que des analogies sans rapport avec la filiation : des similitudes dues aux phénomènes de

convergence ou de parallélisme, c'est-à-dire des traits observés chez telle et telle espèce mais en fait apparus indépendamment. Convergence et parallélisme sont deux phénomènes identiques. On parle conventionnellement de convergence lorsque ces « fausses » similitudes sont rencontrées chez des espèces éloignées, et de parallélisme lorsqu'elles sont rencontrées chez des espèces proches. Nous considérons les deux concepts comme identiques du point de vue de l'analyse des caractères. Dans le cours de l'ontogénie, les phénomènes d'hétérochronie – c'est-à-dire la variation du tempo de développement – peuvent brouiller la reconstruction des états ancestraux. L'arrêt du développement de tel ou tel organe peut aboutir à la persistance chez l'adulte d'un caractère non point ancestral mais tout simplement non transformé.

La paléontologie rencontre les mêmes difficultés que l'anatomie comparée puisque l'analyse des fossiles n'est en quelque sorte que cette dernière appliquée aux formes disparues. Les phénomènes de convergence et de parallélisme sont fréquemment rencontrés.

2. Le concept de ressemblance

Les trois écoles de systématique que l'on reconnaît généralement à des fins pédagogiques – systématique évolutionniste, systématique phénétique et systématique cladistique – peuvent être aisément caractérisées en fonction de leur relation au concept de ressemblance (ou de similitude).

Le cadre conceptuel des approches vues au paragraphe précédent est celui de la similitude. En gros, plus les êtres se ressemblent plus leurs parentés sont étroites. Plus les caractères complexes se ressemblent, plus importantes sont les chances d'être en présence de caractères homologues. Enfin on peut penser *a priori* que des formes globalement similaires qui se succèdent stratigraphiquement sont liées par des relations d'ancêtre à descendant. Or la similitude n'est pas donnée, elle n'existe pas en soi : elle est interprétée. Les critères d'interprétation sont la base des différentes méthodes d'analyse phylogénétique : les inférences généalogiques, sous quelque forme que ce soit, sont fondées sur un traitement de la ressemblance. Rien n'est plus controversé que l'interprétation de la ressemblance à des fins phylogénétiques.

Du point de vue de l'analyse des caractères, le concept de similitude peut être divisé en *homologie* et *homoplasie*. L'homologie est une similitude héritée d'un ancêtre commun, tandis que l'homoplasie est une similitude qui n'est pas héritée d'un ancêtre commun (Simpson, 1961 : 78). Le terme homoplasie a été conçu par Lankester (1870) et signifiait l'apparition indépendante de caractères similaires chez des espèces proches. Aujourd'hui on subdivise l'homoplasie en *convergence* et *réversion*. La convergence est l'apparition indépendante chez deux espèces (ou plus) d'un même caractère. La réversion est l'apparition d'un caractère ayant l'apparence de la morphologie ancestrale. La figure II.2 résume les différentes catégories de similitude. Soit l'arbre (A (BC)). Sur la figure II.2A le caractère x' présent chez B et C est hérité de l'ancêtre de (BC) : homologie. Sur la figure II.2B le caractère x' présent chez A et chez C n'est pas hérité d'un ancêtre commun (homoplasie) : il est apparu deux fois : convergence. Sur la figure II.2C, le

caractère présent chez A et chez C n'est pas hérité d'un ancêtre commun (homoplasie) : il est ancestral chez A et secondairement transformé chez C : réversion. Ainsi, quoique non distinguables, les caractères x' chez A et x' chez C ne sont pas homologues sur la figure II.2B ; de même que le caractère x chez A et x chez C sur la figure II.2C.

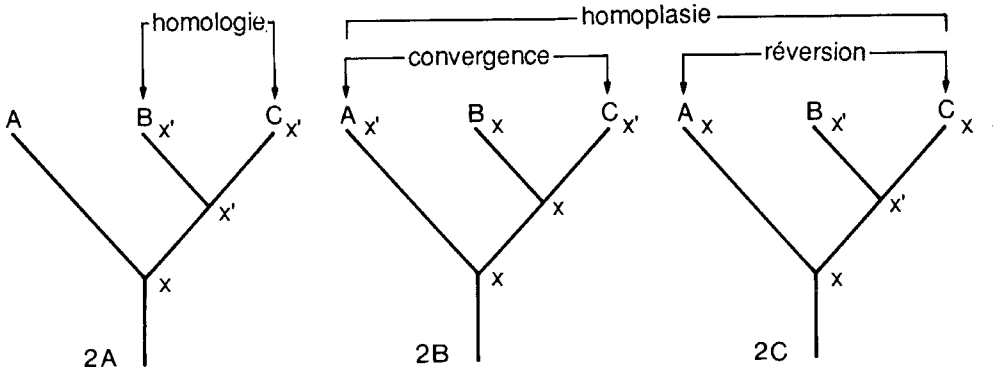


FIGURE II.2. *Les différentes catégories de ressemblance. 2A: ressemblance due à l'homologie ; 2B : ressemblance due à la convergence ; 2C : ressemblance due à la réversion. A, B, C : taxons ; x → x' et x → x' → x : évolution du caractère.*

Pour les systématiciens phénéticiens, adeptes de la « taxinomie numérique » (Sokal et Sneath, 1963 ; Sneath et Sokal, 1973), les combinaisons de taxons ne peuvent être scientifiquement fondées que sur la base de la similitude globale exprimée par des calculs de matrices de distances et d'indices de similitude. Simplement résumée, l'approche phénétique se fonde sur l'analyse, sous forme quantitative, du plus grand nombre de caractères chez les espèces étudiées, homologies et homoplasies mêlées : ce qui se ressemble s'assemble.

Pour les systématiciens évolutionnistes (Simpson, 1961 ; Mayr, 1969), la similitude globale seule ne peut fournir la base de la reconstruction phylogénétique en raison des « fausses » similitudes que sont les homoplasies, c'est-à-dire les convergences et réversions. Dans cette perspective, seule la similitude liée aux traits homologues permet la construction phylogénétique.

Pour les systématiciens partisans de la systématique phylogénétique – encore dénommée cladisme – (Hennig, 1950, 1966 ; Eldredge et Cracraft, 1980 ; Wiley, 1981 ; Nelson et Platnick, 1981 ; Schoch, 1986 ; Matile *et al.*, 1987 ; d'Udekem-Gevers, 1990), le concept même d'homologie doit être raffiné et précisé pour que l'on puisse construire des arbres phylogénétiques précis. Il convient d'identifier les états primitifs (plésiomorphe) et dérivé (apomorphe) des caractères homologues. Pour Hennig, seul le partage par différentes espèces de caractères dérivés est signe de parenté étroite. L'argumentation de Hennig est résumée par la figure II.3. Les espèces B et C partagent un caractère dérivé z' hérité d'un ancêtre commun qui leur est propre : elles sont donc étroitement apparentées. En revanche, le fait que les espèces A et B partagent deux caractères primitifs x et y n'implique pas qu'elles soient étroitement apparentées, même si les caractères x et y sont des homologies chez A et chez B, en ce sens qu'ils ne sont pas des caractères soumis

au phénomène d'homoplasie. Les relations de parenté entre taxons mises en évidence par le partage de traits dérivés sont représentées par un schéma appelé cladogramme (telle la figure II.3). Avec cet exemple on voit l'éclatement définitif du concept de similitude avec l'introduction des notions d'apomorphie et de plésiomorphie et, corrélativement, celle du cladogramme.

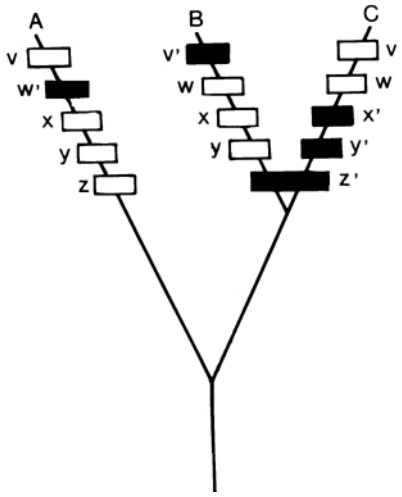


FIGURE II.3. *Schéma d'argumentation phylogénétique selon Hennig. A, B, C : espèces ; v ➞ v', w ➞ w', x ➞ x' , y ➞ y' , z ➞ z' : transformations des caractères (barre blanche : état plésiomorphe ; barre noire : état apomorphe).*

L'interprétation de la similitude, selon qu'elle est globale ou découpée en plésiomorphie, apomorphie et homoplasie, fournit la source des approches contradictoires de la phylogénie que l'on rencontre actuellement. Après des polémiques animées, au cours des années soixante-dix, systématiciens évolutionnistes et cladistes sont aujourd'hui d'accord sur de nombreux aspects touchant aux bases méthodologiques de la construction phylogénétique (Mayr, 1986). Ce n'est que sur le lien entre phylogénie et classification que persistent des désaccords fondamentaux, point qui échappe au sujet de ce livre et que nous n'aborderons pas ici. On trouvera dans Tassy (1986) une illustration des débats sur la nature des classifications biologiques et dans Tassy (1991) une histoire des constructions phylogénétiques. En revanche, à propos de reconstruction phylogénétique, les points de vue phénétique et cladistique se fondent sur des bases opposées.

Les approches concurrentes de la construction phylogénétique en usage actuellement ressortissent donc essentiellement aux deux conceptions du traitement de la similitude. Ou bien l'on admet, malgré la démonstration de Hennig (figure II.3) que la phylogénie peut être construite à partir de la similitude globale : c'est l'approche phénétique. Ou bien l'on admet que seule l'analyse des caractères à des fins de partition en plésiomorphie, apomorphie et homoplasie, permet de construire la phylogénie: c'est l'approche cladistique.

Les approches probabilistes (chapitre VIII), s'éloignent sensiblement des méthodologies phénétique et cladistique. Ces méthodes nécessitent l'adoption préalable et explicite d' un modèle d'évolution des caractères. Une fois ce modèle posé, elles permettent de comparer différents arbres et de choisir le « meilleur » c'est-à-dire le plus vraisemblable au sens statistique du terme.

L'ancien « triple parallélisme » se trouve inféodé à la problématique phylogénétique générale : toutes les analyses des caractères (quels qu'ils soient) portés par tous les taxons (quels qu'ils soient : fossiles ou actuels , espèces ou groupes d'espèces) reposent sur l'identification des états de transformations : plésiomorphe ➞ apomorphe. Il n'y a pas véritablement d'un côté les phylogénies paléontologiques ou, plus généralement, morphologiques, et, de l'autre côté, les phylogénies moléculaires ; mais plutôt des phylogénies fondées sur les différents traitements de la similitude.

Aujourd'hui, les reconstructions phylogénétiques sont grandement facilitées par l'informatique. Les méthodes informatisées de l'analyse phylogénétique sont naturellement issues des travaux des écoles cladistique et phénétique. Parallèlement à la diffusion de l'œuvre de Hennig au cours des années soixante, sont apparus des traitements informatiques de la similitude qui se situent dans la sphère des idées cladistiques : c'est ce que l'on appelle les méthodes de parcimonie. Par ailleurs, les techniques de groupements morphologiques sur la base de la similitude globale ont été adaptées à des fins phylogénétiques au moyen de diverses procédures mathématiques et d'hypothèses sur les processus évolutifs.

Cet ouvrage est donc subdivisé en trois parties principales. L'une consacrée à l'approche cladistique et aux procédures de parcimonie (chapitres IV et V), avec une place à part faite aux analyses dites de compatibilité (chapitre VI). Une deuxième est consacrée aux approches phénétiques (chapitre VII). Une troisième est consacrée à une méthode originale qui n'est issue ni des écoles cladistique ou évolutionniste ni phénétique : c'est une approche probabiliste (chapitre VIII). En préalable à la présentation des différentes méthodes de constructions d'arbres phylogénétiques, le chapitre suivant vise à éclairer succinctement certains concepts de base de la systématique en général, déjà entrevus : les taxons et les caractères.

CHAPITRE **III**

LES OBJETS DE LA PHYLOGÉNÉTIQUE : CARACTÈRES ET TAXONS

1. Les caractères

On appelle caractère tout attribut observable d'un organisme. En tant que tel, le caractère permet de faire des comparaisons entre organismes. En systématique, la notion d'observation du caractère est indissociable de celle de sa représentation. Le moyen par lequel l'observation du caractère devient représentation est le codage, sous quelque forme que ce soit.

D'un point de vue pratique, les expressions « caractère » et « état de caractère » seront parfois considérés comme synonymes. Si la couleur des yeux est un caractère, les yeux bleus sont un état de ce caractère. Dans une comparaison entre organismes qui sont pourvus ou dépourvus de yeux, le caractère discriminant sera la présence ou l'absence de yeux. Dans une comparaison entre organismes pourvus de yeux, le caractère discriminant pourra être, éventuellement, la couleur des yeux. D'un point de vue phylogénétique, pour exprimer une série de transformations d'un caractère ayant deux états *a* et *b*, on pourra dire indifféremment que l'état *a* du caractère se transforme en l'état *b*, ou bien que le caractère *a* se transforme en caractère *b*.

Prenons l'exemple de la morphologie du radius (figure III.1). Un observateur distingue deux morphologies de l'extrémité distale : l'apophyse styloïde est massive ou gracile. Il pourra ainsi comparer des taxons pourvus de l'un ou l'autre des deux types d'apophyse. Afin d'exprimer ces morphologies, il parlera indifféremment du caractère « apophyse styloïde massive » présent ou absent (dans le second cas, l'apophyse est gracile), ou bien de l'état « massif » ou « gracile » du caractère « forme de l'apophyse styloïde ».

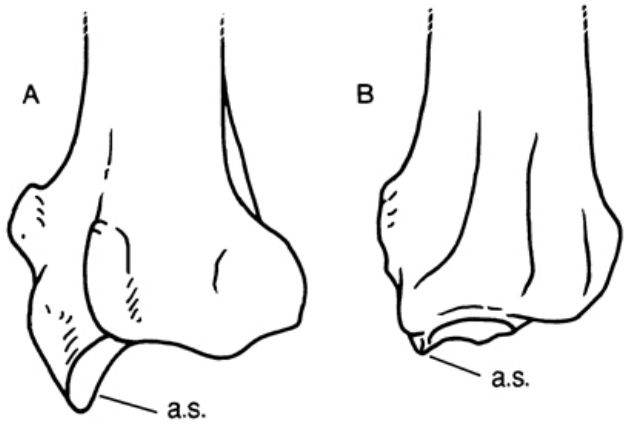


FIGURE III.1. *Caractère et états de caractère: vue antérieure de l'extrémité distale du radius (A : carnivore (*Panthera leo*); B : ongulé (*Oryx dammah*); a.s. : apophyse styloïde. L'apophyse styloïde se présente sous deux états : apophyse styloïde massive (A) ou gracile (B).*

De la même façon, la présence ou l'absence de tel nucléotide dans un site donné de la séquence d'un gène sera décrite généralement comme étant un état de caractère ou, plus rarement, comme un caractère. La présence chez un organisme d'une adénine (A) sur un site S est le caractère de cet organisme ; on préférera cependant dire que l'adénine est l'un des quatre états possibles du caractère « site S », puisque l'adénine est l'une des quatre bases constituant les nucléotides. Un site peut également correspondre à une insertion (I) ou une délétion (D). Dans ce cas il s'agit d'un caractère sous deux états (figure III.2).

```
Homme      ACDGCGACAACAGCGCTDAGIATTACCACTACAIAAAADAATTDTCTICACAGGDTAAAGGCTTADCCGDCGGIG
Chimpanzé  GCDGCGACAACAGCGCTDAGIDTTACCACTACAIAAAADAATTDTCTICACAGGDTCGAGGCTDDATCGDTGGIG
Gorille    ATDGACACAACAGCGTTTAAIATTACCGCTACAIADAAIAADTITCTITACAGADTCGATGCTTAACCGDTGTIG
Orang      GCIAAGGTGGTGCTACCTGADDCCGTTATAGTGDGDDGIGGDGIGTDDTTTGAGICAADTCGADDDTTTACCTDG
Macaque    GTIAAGGTGGTGCTATCDGDDDCCGTTGTAGTGDGDDGIGGDGIGTDDTTTGAAICGADTCGATTATTTACCTDA
Atele      GTIAACGTGTDDDDDDDDDDDDCCGTTGTAGTGDGDDGIGGDAIGTDDTCTGGAIGAAGTCGATTATTCACDDDA
```

FIGURE III.2. *Caractères et états de caractère. Séquences de nucléotides composées de 75 sites alignés chez six espèces. Chaque site correspond à un caractère. Le site 1 présente deux états : A (Adénine) ou G (Guanine). Le site 3 présente deux états correspondant à une insertion I et une délétion D (Sites informatifs de la ψ-η Globine ; d'après Barriel et Darlu, 1990).*

On distingue différents types de caractères :

— les caractères intrinsèques aux espèces étudiées et les caractères extrinsèques ;

— les caractères discrets (discontinus) ou continus.

Les caractères intrinsèques sont les caractères que l'on observe sur les organismes eux-mêmes. Ils sont de nature morphologique, chromosomique, biochimique, physiologique, éthologique (chant des oiseaux par exemple). Les caractères extrinsèques sont les caractères définissant le contexte écologique, géographique et géologique dans lequel se situent ces taxons. Ces caractères ne sont pas utilisés dans la construction phylogénétique.

Pour des raisons qui tiennent à l'histoire des sciences naturelles, la construction phylogénétique se fonde d'abord sur la recherche des homologies pour des caractères morphologiques discrets.

Les caractères morphologiques ont trait à la forme au sens le plus large : ce sont les caractères phénotypiques. L'anatomie externe et interne des êtres vivants a fourni de tout temps la source du savoir morphologique : squelette externe des arthropodes, coquille, forme des feuilles, forme des organes reproducteurs, constitution du bois, squelette interne des vertébrés, systèmes nerveux et vasculaire, système digestif et caractères myologiques, système reproducteur et placentation etc. Tous ces caractères peuvent être transcrits sous la forme de caractères discontinus, discrets, subdivisés en deux états ou plus (états multiples). Ils peuvent parfois être exprimés sous forme quantitative, c'est-à-dire métrique. Sous ce dernier aspect, l'information morphologique est le plus souvent traitée par des méthodes multivariées (analyses factorielles) plutôt que par des méthodes phylogénétiques.

L'apparition, ces trente dernières années, de caractères biochimiques a réactivé la problématique phylogénétique tout en s'intégrant parfaitement dans le cadre général de l'analyse de la similitude. Les caractères biochimiques se traitent différemment selon qu'il s'agit de caractères discrets ou continus :

— les séquences d'acides aminés dans les protéines ou les séquences de nucléotides dans l'ADN ou l'ARN fournissent des données (comparaison des sites) qui sont analysées comme des caractères discontinus, souvent par les mêmes méthodes que celles conçues pour des caractères discrets morphologiques ;

— les données de réactions immunologiques, d'hybridation d'ADN ou de fréquences alléliques s'expriment sous forme quantitative : réaction immunitaire plus ou moins forte d'une espèce par rapport à un antisérum, taux d'hybridation plus ou moins élevé entre les brins d'ADN appartenant à deux espèces différentes, fréquences variables de différents allèles selon les espèces. Ce type de caractère ne peut être analysé, dans une perspective phylogénétique, que par des méthodes quantitatives.

Le polymorphisme allélique mis en évidence par l'électrophorèse est un exemple de caractère qui peut être traité comme un caractère discret ou comme un caractère continu. En effet, si l'on ne considère que la présence ou l'absence de telle forme allélique dans une espèce, il s'agit d'un caractère discret (voir chapitre IV, paragraphe 6.2). En revanche si l'on prend également en considération la fréquence de ces formes alléliques dans les différentes espèces, il s'agit alors d'un caractère continu. Cette deuxième façon contient une information plus riche, même s'il s'avère difficile de la prendre en compte dans une analyse phylogénétique.

2. Les taxons

Un taxon est un groupe d'organismes reconnu en tant qu'unité formelle à chacun des niveaux de la classification (Simpson, 1961). *Elephas maximus*, l'éléphant d'Asie, est un taxon de rang spécifique ; *Elephas* est un taxon de rang générique ; les Elephantidae sont un taxon de rang familial ; les Proboscidea sont un taxon de rang ordinal, etc.

Le mot « taxonomie », sous cette orthographe, est selon son inventeur (Candolle, 1813) construit à partir de *taxis* (arrangement) et de *nomos* (loi) et signifie « la théorie des classifications ». L'orthographe fut corrigée en « taxinomie » par Littré. Le mot « taxon » ne fut inventé et introduit dans la nomenclature botanique que 137 ans plus tard (Lam, 1950) ; son premier usage dans la littérature zoologique est plus tardif encore (Mayr *et al.*, 1953).

Deux concepts ont été plus récemment dérivés de celui de taxon. Il s'agit de l'UTO : Unité Taxinomique Opérationnelle (*OTU, Operational Taxonomic Unit* de Sokal et Sneath, 1963) et de l'UTH : Unité Taxinomique Hypothétique (*HTU : Hypothetical Taxonomic Unit* de Farris, 1970). Selon leurs concepteurs, l'UTO est l'unité pragmatique soumise à l'investigation, tandis que l'UTH est l'ancêtre hypothétique d'un nombre donné d'UTOs, reconstruit en même temps qu'est reconstruit l'arbre phylogénétique.

En dernier lieu, l'UE : Unité Evolutive (EU : *Evolutionary Unit* de Meacham, 1984) est l'organisme étudié dans une analyse phylogénétique : il correspond donc à l'UTO des phénéticiens.

Par extension et simplification, on considérera dans ce livre les « taxons terminaux » comme les éléments de base de l'analyse phylogénétique, qu'ils soient reconnus formellement, c'est-à-dire classifiés, ou bien qu'ils soient des UTOs ou UEs. Ils correspondent aux « taxons liminaux » de Dupuis (1988), c'est-à-dire les « feuilles » de l'arbre au sens de la théorie des graphes (Barthélemy et Guénoche, 1988 ; d'Udekem-Gevers, 1990).

L'espèce est généralement considérée comme un taxon à part : l'espèce seule aurait un statut biologique objectif (paragraphe 2.1). Mais les analyses phylogénétiques n'ont pas toutes des espèces comme objets directs d'étude. Ces derniers peuvent être des taxons de rang infra-spécifique : des sous-espèces ou des populations ; ou bien de rang supra-spécifique : des genres, des familles, des ordres, etc. Le lien logique qui rassemble ces différents objets d'étude est double : il réside d'une part dans la dimension taxinomique des objets, et, d'autre part, dans leur dimension phylogénétique.

Dimension taxinomique : les taxons sur lesquels s'applique l'enquête phylogénétique, c'est-à-dire ceux dont on essaie de mettre en évidence les relations de parenté, sont les taxons terminaux. Ils doivent avoir une identité. Ils doivent être reconnaissables, ne serait-ce que par un seul attribut qui leur soit propre.

Dimension phylogénétique : les taxons terminaux doivent représenter une section non arbitraire de l'arbre phylogénétique. Mais si les taxons sont des groupements d'organismes de rang générique ou familial (ou au-delà), ils ont eux-

mêmes une histoire : ils sont composés d'espèces dont les relations ont la forme d'un arbre. Ces taxons terminaux doivent être - en principe - des groupes naturels ou groupes monophylétiques, ou encore « monophylons ».

2.1. L'espèce et les taxons infra-spécifiques

Les mécanismes évolutifs opèrent au niveau des organismes classés dans la catégorie de base de la hiérarchie linnéenne : l'espèce.

Il en ressort *a priori* que l'espèce devrait donc être l'unité de base de l'arbre phylogénétique. Une espèce conçue dans une perspective biologique est un pool génétique fermé, en cela dissocié des autres espèces définies semblablement : entre individus appartenant à des espèces différentes existe une barrière empêchant l'interfécondité. Le lien objectif invoqué pour justifier le statut privilégié accordé à l'espèce réside dans le critère d'interfécondité. La spéciation est la production d'une nouvelle espèce (ou plus) à partir d'une espèce ancestrale (espèce mère ou espèce souche) ; autrement dit, l'apparition de nouveaux pools génétiques à partir d'un pool génétique ancestral. De la sorte, la phylogénie est, strictement parlant, l'histoire des spéciations depuis la première forme vivante, il y a près de quatre milliards d'années jusqu'à la diversité biologique actuelle (2 millions d'espèces vivantes recensées, avec probablement 6 à 12 millions de plus qui restent à décrire, selon les estimations actuelles).

Parce qu'il met en évidence les relations de parenté entre espèces nées de spéciations successives, l'arbre phylogénétique est une hiérarchie. Cette hiérarchie est le fruit de l'histoire. Mais au niveau des espèces et des populations, c'est-à-dire au niveau où opèrent les mécanismes de la spéciation, la phylogénie reste largement inconnue. En effet, on est loin de connaître les relations de parenté entre les deux millions d'espèces vivantes connues. Pour bien des groupes on n'a même aucune idée des relations phylogénétiques entre espèces. Les constructions phylogénétiques sont des représentations partielles, plus ou moins importantes ou exhaustives, de l'« arbre de la vie ». Ce sont des hypothèses émises sur les relations de parentés entre espèces ou groupes d'espèces choisis pour diverses raisons, à partir d'un ensemble de données observées et interprétées : les caractères.

Si du point de vue biologique les individus rangés dans la catégorie espèce ont un statut particulier (dû à l'interfécondité), il n'en est pas tout à fait de même d'un point de vue taxinomique.

L'espèce elle-même est en effet un agrégat de populations. Les populations, de tailles fort variables selon les espèces, correspondent, si elles sont bien délimitées géographiquement aux sous-espèces des nomenclatures zoologiques. En général, les populations ne sont pas génétiquement isolées les unes des autres. C'est pourquoi les relations entre populations sont le plus souvent de nature réticulaire et pas seulement hiérarchique. La théorie évolutionniste fait de la population le lieu des mécanismes évolutifs, et, par là-même, l'unité de l'évolution. Dans la mesure où la population, et non l'espèce, est cette unité, l'espèce elle-même devient un taxon comme un autre : un regroupement d'unités qui lui sont subordonnées. L'examen des relations de parenté entre populations ou taxons infra-spécifiques de quelque statut que ce soit, aboutit le plus souvent à la

construction d'un réseau si les unités analysées ne sont pas isolées les unes des autres. Les relations intra-spécifiques de ce type sont nommées tokogéniques par Hennig (1966) : les relations entre les taxons terminaux ne sont alors pas nécessairement hiérarchiques. Si le concept de phylogénie se conçoit comme l'histoire des ruptures dans les pools génétiques et de l'apparition de pools génétiques nouveaux, isolés les uns des autres, l'histoire des parentés entre les constituants de ces pools (entre ces populations) sera plutôt une tokogénie qu'une phylogénie. Néanmoins l'usage a prévalu qui maintient le terme phylogénie pour des reconstructions d'arbres dont les taxons terminaux sont des populations.

Le système phylogénétique a pour but d'émettre des hypothèses de parenté entre unités vues comme des ensembles reconnaissables, isolés des autres unités définies semblablement, et qui constituent donc des ensembles fermés. Ces unités peuvent être des espèces, mais en tout état de cause, elles ne seront pas forcément identiques aux unités de l'évolution. En effet, l'unité sur laquelle s'expriment les mécanismes évolutifs est la population (réduite éventuellement à quelques individus), non nécessairement isolée des autres populations appartenant à la même espèce. Ces unités phylogénétiques, les plus petits ensembles fermés, ne sont donc pas nécessairement des espèces au sens biologique du terme. Mais comme la population n'a pas de statut taxinomique, les phénéticiens et les cladistes ont tenté de résoudre le dilemme de deux façons différentes.

L'une, l'approche des phénéticiens, est de nier à l'espèce biologique tout statut privilégié. La conception de l'UTO des phénéticiens est liée à ce problème. L'objet des investigations systématiques, le taxon terminal, est un objet de convention, « opérationnel » et les groupements de ces objets sont opérés sur la base de la similitude globale exprimée au moyen d'indices mathématiques. Cette approche n'est sujette à aucune contradiction logique dans un système de pure ressemblance globale. Mais elle se heurte à deux points de vue. Point de vue phylogénétique : la similitude globale n'est pas nécessairement un indicateur de la proche parenté. Point de vue biologique : par exemple, chez les espèces à fort dimorphisme sexuel la simple mesure de la similitude globale n'associe pas toujours mâle et femelle dans une même UTO.

L'autre approche s'inscrit dans une perspective phylogénétique. Elle vise à définir l'espèce dans un contexte historique. Tel est le point de vue de Hennig (1966) selon lequel une espèce n'est qu'un élément de l'arbre phylogénétique situé entre deux spéciations. Autrement dit l'espèce en tant que « bio-espèce temporelle » (***t****ime bio-species* de Bonde, 1981), est un pool génétique borné historiquement par deux ruptures : celle qui lui a donné naissance et celle où il se subdivise à son tour. Cette définition résout le dilemme entre espèce biologique (achronique) et espèce chronologique, dans la mesure où chacune des deux définitions d'espèces répond à celle de pool génétique fermé. Mais alors un autre paradoxe surgit de cet accord. L'objectivité du concept d'espèce biologique réside - en principe - dans le comportement des membres de l'espèce vis-à-vis de la reproduction, et non dans la nomenclature du systématicien. En revanche, l'espèce chronologique ainsi définie n'obéit qu'à une hypothèse formulée par le systématicien : celle de la succession des spéciations, c'est-à-dire l'hypothèse phylogénétique. Rares sont les biologistes qui ont accepté la définition hennigienne de l'espèce. Mais la question de la dimension chronologique de l'espèce demeure, et elle seule est pertinente dans le système phylogénétique.

Le statut particulier des populations allopatriques (distribuées dans des aires disjointes) a mené certains cladistes à une autre révision énergique du statut de l'espèce : toute population reconnaissable serait une espèce. Selon Nelson et Platnick (1981), l'espèce est le plus petit ensemble identifiable d'organismes qui se reproduisent entre eux ayant un ensemble unique de caractères. D'après cette définition la « sous-espèce » est comprise comme une espèce. De la sorte, on maintiendrait aux unités évolutives (les populations) le statut formel d'espèce. A vrai dire, peu importe le statut formel des taxons terminaux. Ce qui importe est que les caractères distinctifs soient fixés dans les taxons terminaux, qu'il s'agisse d'espèces ou de sous-espèces. Même si les caractères sont polymorphes une telle exigence est acceptable, dès lors qu'une explication évolutive de ce polymorphisme est possible. Cela revient à dire qu'une hiérarchie (une histoire) peut émerger de l'étude des variations du polymorphisme.

2.2. Taxons supra-spécifiques

Ces taxons terminaux sont - en principe - des groupes naturels ou groupes monophylétiques.

Ces groupes renferment la totalité de la descendance à partir d'une espèce ancestrale. C'est à cette seule condition qu'ils possèdent une dimension phylogénétique qui est, au sens strict, chronologique : une date d'origine et une date de différenciation, voire, éventuellement, une date d'extinction ; cela, pour tous les membres du taxon terminal et *a contrario* pour aucun des taxons situés hors du groupe en question. La figure III.3 montre les relations de parenté entre trois taxons terminaux A, B et C, avec B et C étroitement apparentés. Ces taxons sont des taxons de rang supra-spécifique dont la figure III.3B montre la composition et l'histoire. Par exemple, le taxon C est composé des espèces 5 à 10. C'est un groupe monophylétique qui rassemble tous les descendants de l'espèce ancestrale c, celle à partir de laquelle s'est différencié le taxon C. Sa date d'origine est aussi celle du taxon B, proche parent de C. L'ancêtre b du taxon (BC) est nécessairement plus ancien que l'ancêtre c du taxon C inclus dans (BC). La dimension chronologique des taxons implique que la structure de l'arbre phylogénétique est hiérarchique. La hiérarchie – l'emboîtement des taxons – est tributaire du degré d'ancienneté des espèces ancestrales. Dès lors que les taxons

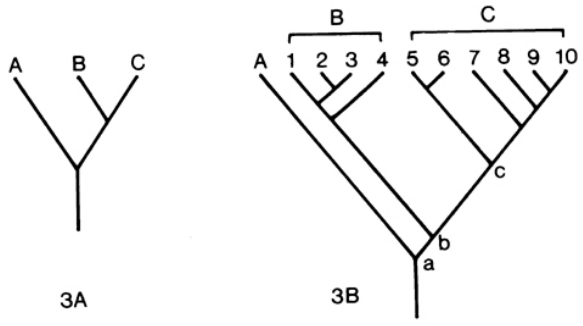


FIGURE III.3. *3A : relations de parenté entre trois taxons terminaux de rang supra-spécifique, A, B et C. 3B : composition et histoire des taxons terminaux B et C détaillées au niveau de l'espèce ; a, b, c : ancêtres.*

de rang supra-spécifique sont des groupes supposés naturels, les phylogénies représentées au niveau supra-spécifique sont comparables aux phylogénies représentées au niveau spécifique, à la seule différence qu'aucun taxon de rang supérieur à l'espèce ne peut être tenu pour un ancêtre naturel.

Dans le cadre phylogénétique, les caractères sont considérés comme les nouveautés évolutives. En tant que tels, dans la mesure où ce sont eux qui sont transmis de génération en génération, les caractères sont les unités de base de l'évolution. Ainsi est gommée la question de l'influence des catégories de classification sur l'analyse phylogénétique.

Or les approches contradictoires de la construction phylogénétique s'opposent toutes vis-à-vis du statut des caractères et de leur traitement. Selon l'approche phénétique, la totalité des caractères (phénotypiques ou génotypiques) permet seule des inférences phylogénétiques. Dans la pratique, cette « totalité » se limite au plus grand nombre de traits que l'on peut observer et que l'on analyse ensuite par le biais de distances. Dans cette perspective, la similitude globale dûment quantifiée serait le préalable à toute enquête phylogénétique (Sneath et Sokal, 1973).

Selon l'approche phylogénétique (évolutionniste aussi bien que cladistique), c'est au contraire l'examen de caractères libres d'évoluer indépendamment les uns des autres qui permet de dégager les traits ayant un sens phylogénétique : les homologies, les seuls traits à être signes de parenté.

Bien entendu, toutes les reconstructions phylogénétiques ne sont que des hypothèses : hypothèses sur le statut de groupe naturel des taxons et sur les relations de parenté. En tant qu'hypothèse scientifique, une construction phylogénétique obéit à des règles. Chaque règle s'inscrit dans l'une ou l'autre des approches méthodologiques concurrentes dont l'explication est le but de ce livre. L'hypothèse scientifique doit être testable et heuristique, c'est-à-dire pouvoir être soumise à réfutation et aider à la découverte, permettre des prédictions. Il reste que la notion de test pour des constructions historiques est particulièrement délicate. Rien ne ressemble plus à une « bonne » phylogénie, qu'une phylogénie réfutée, c'est-à-dire « mauvaise ». Ce point ne sera abordé que succinctement dans les chapitres suivants ; il mériterait néanmoins de vastes développements.

CHAPITRE **IV**

LA MÉTHODE CLADISTIQUE

1. Qu'est-ce que l'analyse cladistique ?

Les principes de l'analyse cladistique ont été élaborés par l'entomologiste Willi Hennig, quoique certains concepts et certaines méthodes formalisés par Hennig peuvent être parallèlement rencontrés dans la littérature chez des contemporains (Wagner, 1961), voire de lointains prédécesseurs tel Mitchell (1901) redécouvert par Nelson et Platnick (1981). Les principes de base du cladisme sont énoncés dans les ouvrages de taxinomie fondamentale de Hennig (1950, 1966) ainsi que dans sa synthèse sur la phylogénie des insectes (Hennig, 1969, 1981). Il existe par ailleurs de nombreux manuels de systématique qui se réclament des principes du cladisme (Eldredge et Cracraft, 1980 ; Wiley, 1981 ; Nelson et Platnick, 1981; Ax, 1984 ; Schoch, 1986).

Hennig n'a pas utilisé dans ses différents ouvrages les termes « cladisme », « analyse cladistique », « cladogramme », ou tout simplement « clade », tous dérivés de la racine grecque *klados* (branche). Le cladisme (ou la cladistique) y est dénommée « systématique phylogénétique », le cladogramme est un « schéma d'argumentation phylogénétique », le clade est un « groupe monophylétique ». Le succès du cladisme et des termes associés est dû aux auteurs anglo-américains.

Dans le système cladistique, la phylogénie est reconstruite à l'aide d'une analyse de caractères qui vise à identifier les états plésiomorphe (= primitif) et apomorphe (= dérivé). Les parentés entre les taxons étudiés sont identifiées sur la base des seuls états apomorphes partagés par tel et tel taxon, ce que l'on appelle les synapomorphies. Les synapomorphies sont imputées à un héritage à partir d'une espèce ancestrale propre aux taxons qui les possèdent. Les groupes ainsi construits sont monophylétiques.

1.1. Apomorphie, plésiomorphie et groupes naturels

Le principe de base de l'analyse cladistique est donc la mise en évidence des séries de transformation des caractères de l'état plésiomorphe vers l'état apomorphe, c'est-à-dire de type a ➞ a'. De telles séries sont appelées « morphoclines » (Maslin, 1952).

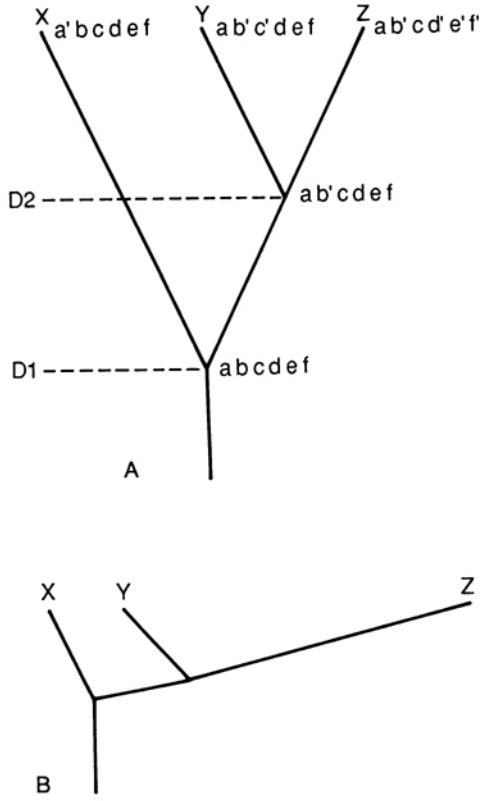


FIGURE IV.1. *A : Relations phylogénétiques entre trois taxons terminaux X, Y et Z. Transformation des caractères : a ➞ a', b ➞ b', c ➞ c', d ➞ d', e ➞ e', f ➞ f'. D1: date d'origine du taxon (Y,Z) ; D2 : date de différenciation du taxon (Y,Z). B : même schéma où les longueurs des branches correspondent au degré de divergence morphologique.*

La figure IV.1 montre les relations de parentés entre trois taxons X, Y et Z, construites à partir de l'analyse de 6 caractères, ou plus exactement, de 6 séries de transformation de caractères. Les taxons Y et Z sont étroitement apparentés parce qu'ils partagent le même caractère transformé b' : la même apomorphie. Ils ont en commun une espèce ancestrale qui n'est pas en même temps l'espèce ancestrale de X. L'hypothèse fondamentale de l'analyse cladistique est que le même caractère dérivé b' observé chez Y et chez Z est hérité de l'espèce ancestrale de (Y,Z).

Les séries de transformations sont elles-mêmes des hypothèses qui sont émises à partir de critères explorés dans le paragraphe 4. Du point de vue de la similitude globale, on peut juger que d'après la figure IV.1B, les taxons X et Y se ressemblent plus que chacun d'eux ne ressemble à Z (X et Y partage trois caractères non transformés, les plésiomorphies d, e et f, alors que Y et Z ne partagent que le caractère apomorphe b'). De ce cas de figure, on conclue que les partages de caractères non transformés, plésiomorphes, n'indiquent pas une étroite parenté phylogénétique. On nomme symplésiomorphie le partage d'un caractère plésiomorphe par deux ou plusieurs taxons. Le groupe (Y,Z) est dit *monophylétique*, tandis que le groupe (X,Y) fondé sur des symplésiomorphies, est

dit *paraphylétique*. Un groupe paraphylétique ne renferme pas tous les descendants d'une espèce ancestrale et, par conséquent, n'a pas d'histoire propre : ce n'est pas un groupe naturel. Sur la figure IV.1 la date d'origine D1 du groupe paraphylétique (X,Y) est aussi celle du taxon monophylétique (Y,Z). On ne peut invoquer de différenciation pour le groupe paraphylétique (X,Y) puisque l'émergence de Y est aussi celle de Z. La date de différenciation D2 est celle du taxon monophylétique (Y,Z), c'est aussi la date d'origine du taxon Y et du taxon Z.

Les groupes monophylétiques de la figure IV.1 sont identifiés par la présence d'au moins une apomorphie ; ici a' pour X, b' pour (Y,Z), c' pour Y et d', e', f' pour Z. La figure IV.1 est appelée cladogramme. Les deux taxons étroitement apparentés Y et Z sont appelés *groupes frères* (ou *espèces sœurs* si les taxons terminaux sont des espèces). Par ailleurs, X est le groupe frère de (Y,Z). Selon la convention cladistique acceptée par tous les auteurs, la topologie du cladogramme peut s'écrire linéairement (X(Y,Z)) ou bien ((Y,Z)X).

Les notions d'apomorphie et de plésiomorphie sont des notions relatives. L'expression «le caractère apomorphe du taxon T...» ne doit pas laisser croire qu'un caractère est en soi apomorphe ou plésiomorphe. La présence de cinq doigts à la main et au pied est une synapomorphie des tétrapodes mais c'est une symplésiomorphie pour les groupes inclus dans les tétrapodes, comme l'homme (primates) et le lézard (squamates). Sur la figure IV-1, le caractère b' est apomorphe pour le groupe (Y,Z) : synapomorphie de (Y,Z). Il est plésiomorphe à l'intérieur du groupe (Y,Z) c'est une plésiomorphie pour Y et pour Z.

Les caractères présents dans un seul taxon terminal – qui ne permettent pas d'émettre des hypothèses de parenté entre groupes – sont des autapomorphies (par exemple, le caractère d' pour Z, le caractère a' pour X et le caractère c' pour Y). Si un taxon terminal est constitué de groupes d'espèces, les autapomorphies de ce taxon sont les synapomorphies des espèces qu'il regroupe. Par exemple, les caractères d', e' et f' sont les synapomorphies des espèces incluses dans le taxon Z.

La figure IV.1 montre que les caractères observés chez chacun des taxons terminaux ne se présentent pas tous au même niveau évolutif. Chez le taxon Y le caractère b est évolué (apomorphe) tandis que les caractères d, e et f sont primitifs (plésiomorphes). Cette constatation est triviale : des caractères se transforment, d'autres pas. Le nombre de doigts à la main de l'homme est primitif, tandis que le cortex de son cerveau est dérivé. Une telle association de caractères dont les niveaux évolutifs sont différents est dénommée « hétérobathmie des caractères » par Hennig. Le processus évolutif responsable de ce type de distribution des états de transformation – vitesses d'évolution différentes des caractères – fut appelé « évolution en mosaïque » par de Beer (1954), à l'occasion de l'étude de l'*Archaeopteryx*, le célèbre oiseau jurassique.

A la suite de Hennig et par pure convention, il est fréquent de présenter sur un schéma de relations de parenté les états apomorphes sous la forme d'une barre noire et les états plésiomorphes sous la forme d'une barre blanche (figure IV.2).

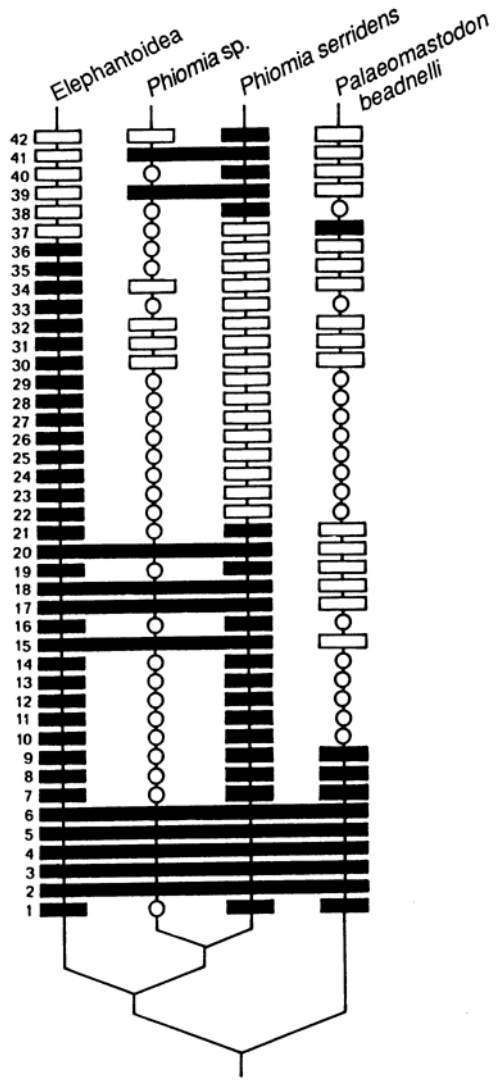


FIGURE IV.2. *Relations de parenté chez les proboscidiens éléphantiformes. 1–42: caractères ; barre blanche : état plésiomorphe ; barre noire : état apomorphe ; cercle : caractère manquant (d'après Tassy, 1982).*

Dans cet exemple, le problème traité est celui de la monophylie d'un groupe de proboscidiens (Elephantoidea) et des parentés de certaines espèces réputées primitives (*Phiomia serridens* et *Palaeomastodon beadnelli*). La distribution des caractères dérivés montre que le taxon Elephantoidea possède un grand nombre d'apomorphies qui lui sont propres (autapomorphies) de telle sorte qu'est admise l'hypothèse de monophylie du groupe. D'autre part, *Phiomia serridens* se rapproche des Elephantoidea en raison du partage d'au moins six caractères apomorphes restés plésiomorphes chez *Palaeomastodon beadnelli*. Cet exemple paléontologique inclut des espèces fossiles (*Phiomia* sp. et *Palaeomastodon beadnelli*) chez lesquelles de nombreux caractères n'ont pu être observés en raison de la fossilisation (restes fragmentaires). Les lacunes d'observation sont

nombreuses (cercles blancs de la figure IV.2). Ces lacunes ne sont pas un obstacle à la reconstruction phylogénétique. Par exemple, si les hypothèses de parenté sont justes, on peut prévoir – par congruence de caractères et en l'absence d'homoplasie – que les traits 7 à 14 se présenteront lors de découvertes à venir sous leur état apomorphe chez *Phiomia* sp.. Sur ce cladogramme, l'échelle du temps est implicite : la position relative des ancêtres au niveau des dichotomies indique la séquence chronologique mais les datations ne sont pas inscrites.

1.2 Images cladistiques

La distribution des caractères plésiomorphes et apomorphes sur un cladogramme obéit à plusieurs symboliques toutes équivalentes. La figure IV.3 résume quatre façons de représenter quatre séries de transformations (w ➞ w', x ➞ x', y ➞ y' et z ➞ z') chez trois taxons A, B et C, sachant que l'état dérivé v' est partagé par les trois taxons.

Cette symbolique n'est pas sans connotation sur la narration évolutive. Il y a plusieurs façons d'exprimer la même information phylogénétique. Les figures IV.3.A-B et IV.3.C-D diffèrent par la manière de placer les états dérivés. Sur la figure IV-3 A-B, ils sont situés sur les branches ; sur la figure IV.3.C-D, ils sont situés sur les nœuds en bout de ces branches (nœuds internes : ancêtres ; nœuds externes : taxons terminaux). D'un point de vue narratif, à partir de la figure IV.3.A-B, on dira plutôt que l'état dérivé w' est apparu sur la lignée menant au taxon (BC) et que l'état dérivé z' est apparu sur la lignée menant au taxon C. Au contraire, à partir de la figure IV.3.C-D, on dira plutôt que l'état dérivé w' est présent chez le plus récent ancêtre du taxon (B,C) – ou bien est propre au taxon (BC) – et que l'état dérivé z' est présent chez le taxon C. Il y a là deux manières d'exprimer la même information phylogénétique, la première manière évoque plutôt le processus, la seconde – plus structurale – évoque le résultat. Par ailleurs, les figures IV.3 A-D présentent les différents états primitifs et dérivés, tandis que sur les figures IV.3.B-C sont notés les seuls états dérivés.

1.3. Cladogramme et arbre phylogénétique

Il est d'usage d'opposer en systématique évolutive le cladogramme et l'arbre phylogénétique. Le premier montre la distribution des caractères et les parentés entre les groupes étudiés qui en sont déduites. Il représente la phylogénie sous la forme d'une succession de dichotomies (appelées aussi nœuds du cladogramme), chacune correspondant à un ancêtre construit à partir des synapomorphies de ses descendants. Aucun des taxons terminaux étudiés n'est, *a priori*, tenu comme un ancêtre au sens strict.

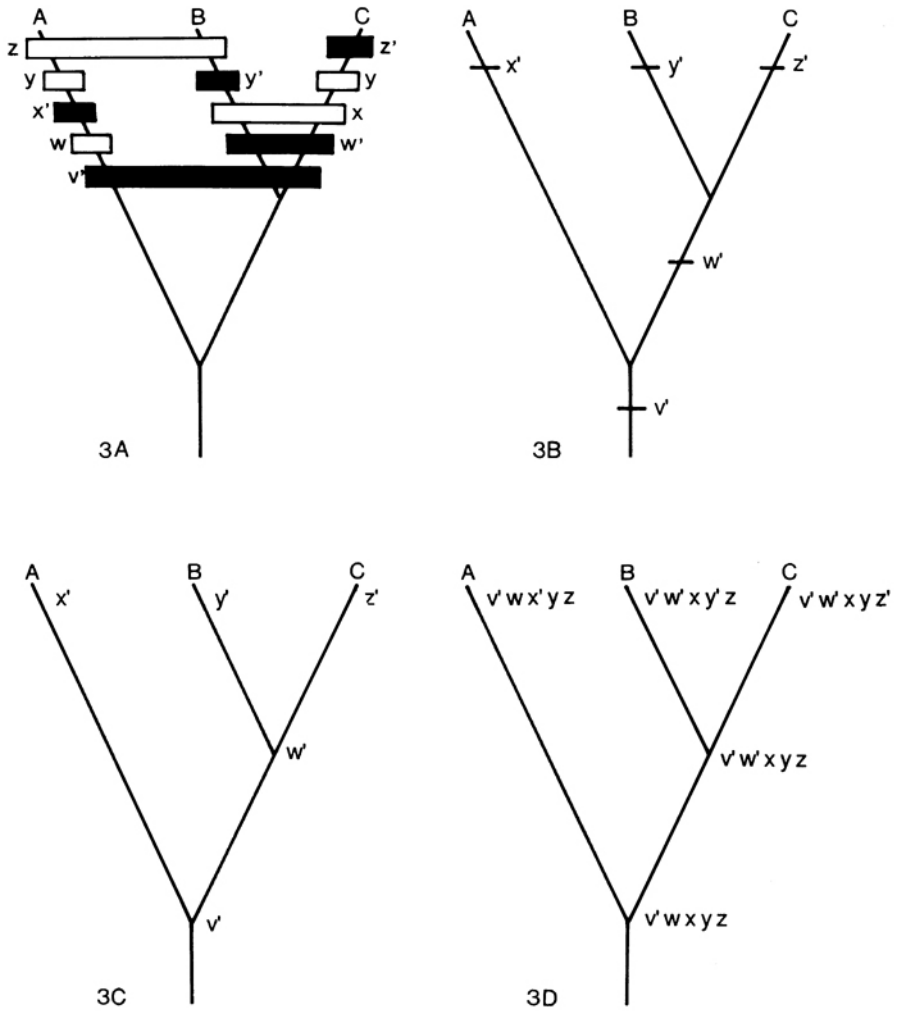


FIGURE IV.3 . *Images cladistiques : quatre façons de représenter quatre séries de transformations de caractères et les parentés de trois taxons A, B et C. L'état dérivé v' est présent chez les trois taxons.*

L'arbre phylogénétique apporte les mêmes informations que le cladogramme mais, en outre, on peut y adjoindre l'échelle du temps : les taxons terminaux sont inscrits dans l'échelle géologique. La divergence morphologique (le nombre d'autapomorphies) peut être symbolisée par le plus ou moins grand éloignement des deux groupes frères à partir du point de branchement (représentant l'espèce ancestrale), c'est-à-dire par des longueurs de branches inégales (figure IV.1B).

Enfin, un taxon terminal de rang spécifique peut se révéler être dépourvu d'autapomorphies et avoir les traits d'une espèce ancestrale hypothétique (tel nœud du cladogramme). Dans un cladogramme, où tous les taxons sont terminaux, ce taxon ancestral apparaîtra comme apparenté de façon égale à ses deux descendants, c'est-à-dire sous la forme d'une trifurcation. Prenons l'exemple de la figure IV.4. Sur l'arbre phylogénétique 4A, X est l'espèce ancestrale des espèces A et B. L'arbre 4B montre la position que prend l'ancêtre lorsqu'il est pris

comme un taxon terminal. Cette trifurcation résume trois arbres dichotomiques possibles (X parent de A, X parent de B, X parent de A et B) sachant qu'est nulle la longueur de la branche menant à X sur les arbres 4C et 4D, ainsi que la longueur menant de la branche commune à (A,B) jusqu'à X sur l'arbre 4E. Le cladogramme et l'arbre phylogénétique ne sont donc pas des constructions contradictoires mais complémentaires.

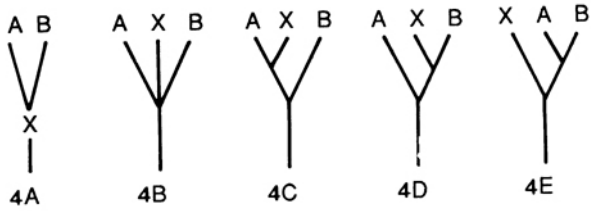


FIGURE IV.4. *Représentation cladistique d'un ancêtre analysé comme un taxon terminal (4A : arbre phylogénétique ; 4B, 4C, 4D, 4E : cladogrammes compatibles avec l'arbre 4A).*

1.4. Ancêtres

L'ancêtre en systématique cladistique est nécessairement de rang spécifique ou infra-spécifique. Les « groupes ancestraux », de rang supraspécifique, ne sont pas des groupes naturels. Ce sont des groupes paraphylétiques dont les membres sont proches parents de divers autres groupes. Dans les cladogrammes, les ancêtres sont des ancêtres hypothétiques, appelés souvent « morphotype ancestral hypothétique », situés aux nœuds du schéma. Ils sont construits à partir de l'analyse des taxons terminaux: leurs attributs sont déduits de ceux de leurs descendants.

Le statut de l'espèce ancestrale reste un des points les plus controversés de la systématique même si cette question ne joue pas un rôle fondamental dans les constructions phylogénétiques. En effet un ancêtre au sens strict ne peut avoir comme caractères dérivés que ceux partagés par ses descendants, c'est-à-dire des symplésiomorphies à l'intérieur du groupe formé par l'espèce ancestrale et ses descendants. Si un taxon tenu pour un ancêtre se révèle posséder un caractère dérivé propre (une autapomorphie), c'est un taxon qui a divergé à partir de l'ancêtre. L'ancêtre ne peut être que paraphylétique et par conséquent non identifiable selon les critères cladistiques. Comme on vient de l'envisager dans le paragraphe précédent (figure IV.4), il reste qu'un groupe fossile, de rang spécifique, dépourvu d'autapomorphie apparaît dans un cladogramme en tant que taxon terminal identique à l'ancêtre qu'il partage avec son espèce sœur dans un cladogramme. Si son âge est compatible avec une position ancestrale, il est équivalent à l'ancêtre de l'arbre phylogénétique. On n'est toutefois jamais assuré que ses membres sont étroitement apparentés puisque seule la synapomorphie permet d'émettre une telle hypothèse.

2. Homologie et orthologie

Il ressort des pages qui précèdent que l'analyse cladistique est une méthode de reconstruction de la phylogénie qui se fonde sur la reconnaissance des homologies à leur niveau de synapomorphie. Comme on l'a vu plus haut, le concept d'homologie est né de l'anatomie comparée et a été appliqué depuis bientôt deux siècles aux analyses morphologiques. Le concept d'homologie s'applique-t-il de façon identique à tous les types de caractères ?

2.1. Définition et critères de l'homologie

Le savoir zoologique et botanique qui s'est progressivement accumulé depuis que les naturalistes observent le monde vivant et fossile, a intégré le concept central d'homologie à partir duquel furent émises les hypothèses historiques d'arrangement des êtres vivants. Mais il convient de rappeler, une fois encore, la distinction essentielle entre la définition de l'homologie et le critère de reconnaissance de l'homologie. En morphologie, le principe des connexions permet de reconnaître ce qui est effectivement semblable. La déduction évolutionniste d'une telle observation est que la similitude en question est due au phénomène de descendance. Mais on n'observe pas une homologie, on pose une hypothèse d'homologie à partir d'une observation. L'homologie est une hypothèse : une hypothèse sur l'ascendance. Comme l'exprime de façon concise Walter Fitch (in Lewin, 1987) : « il est important de faire la distinction entre l'observation et la conclusion ».

La définition de l'homologie est simple : est homologue ce qui est hérité d'une ascendance commune. Autrement dit, un trait partagé par différentes espèces est homologue parce qu'il est hérité d'un ancêtre commun propre à ces espèces. Or l'ascendance commune est identifiée grâce à la mise en évidence d'une homologie. Comment identifier l'homologie sans connaître *a priori* la phylogénie ?

Le critère de reconnaissance de l'homologie est triple (Patterson, 1982, 1987) : critères de ressemblance, de non-coexistence, et de congruence.

Le *critère de ressemblance,* vu au chapitre II, est lié au principe des connexions.

Le *critère de non-coexistence* permet de distinguer l'homologie vraie de l'homologie dite sérielle : deux caractères généalogiquement homologues ne peuvent coexister dans un même organisme. L'exemple plaisant donné par Patterson est celui des anges : la théorie selon laquelle le bras humain et l'aile des oiseaux sont homologues au sens généalogique, sera réfutée lorsqu'on découvrira des anges munis à la fois de bras et d'ailes. En revanche, le bras et la jambe de l'homme sont une homologie sérielle : ils sont construits selon le même patron ; ils coexistent dans un même organisme et l'un ne descend pas de l'autre, et inversement.

Le *critère de congruence* permet de superposer les arbres construits à partir de différents caractères : les caractères homologues sont congruents. Ils permettent

de construire les mêmes arbres phylogénétiques. Ce point est développé dans le paragraphe IV.3.

La confusion largement entretenue entre le critère de reconnaissance de l'homologie et la déduction phylogénétique qui en est tirée est à la source de nombreux débats sur la signification de l'homologie en dehors du terrain de l'anatomie comparée, si fréquenté qu'il ne recèle plus, en apparence, de controverses.

Quelle est la nature de l'homologie en matière de biologie moléculaire ?

Il y a quelques années, une dizaine de biologistes des molécules réveillèrent le vieux débat sur l'homologie (Reeck *et al.*, 1987), cette fois à propos de l'utilisation du mot et du concept dans les travaux de séquençage de protéines ou d'acides nucléiques. Le débat est exemplaire car il permet d'aborder un point central propre aux analyses moléculaires. Reeck *et al*. s'insurgèrent sur l'utilisation fréquente en biologie moléculaire du terme «homologue» dans le sens de «similaire» : si deux séquences se ressemblent, il conviendrait de parler de «similitude entre séquences» et non d'«homologie ». L'homologie ne devrait s'appliquer qu'aux inférences phylogénétiques.

Fitch (1970) a proposé de réserver le terme homologie aux traits morphologiques et a créé le terme « orthologie» pour qualifier le concept de similitude moléculaire due à la descendance. Orthologie s'oppose à «paralogie» (autre terme créé par Fitch) qui est la similitude due à la duplication de gènes, indépendamment de toute spéciation. Dans ce cas, la similitude moléculaire est acquise indépendamment d'un ancêtre commun.

Le fond de la question réside néanmoins dans la phase analytique de l'observation qui transfère l'observation de la similitude dans le domaine de son interprétation phylogénétique. Ce transfert fait, d'une part, une homologie ou une orthologie, ou d'autre part, une homoplasie. En quoi les données morphologiques et moléculaires se distinguent-elles ?

L'anatomie comparée (paléontologie incluse) est une discipline ancienne et nombre d'hypothèses d'homologies sont aujourd'hui considérées comme des faits d'observation car rien n'est venu les infirmer. On reviendra sur la nature de ce « rien » dans la quatrième partie de ce chapitre. Patterson (1987), en illustrant la solidité de ce savoir morphologique, écrit : « si l'on compare les crânes, ou les membres, des vertébrés (...) depuis les requins jusqu'aux mammifères, nous pouvons être sûrs que nous avons à faire à de vraies homologies (un crâne est un crâne et non une éventuelle duplication (...) ». L'affirmation de Patterson signifie d'abord qu'un crâne de requin ou de mammifère, est construit selon le même schéma. A l'analyse des structures adultes s'ajoute celle de l'ontogénie. Même si un crâne de requin ne ressemble guère à un crâne humain, on est assuré que les crânes, tout au moins leurs parties fondamentales, sont homologues. En effet, le crâne est d'abord l'enveloppe des centres nerveux supérieurs (encéphale, capsules otique, optique, nasale). Dans l'embryon, les précurseurs de ces enveloppes, autrement dit la construction du crâne primordial, sont identiques chez un requin et chez un mammifère quelconque. On en déduit que requins et mammifères descendent d'une espèce ancestrale commune : les similitudes crâniennes sont dues au phénomène de descendance. D'après Patterson, les hypothèses

d'homologie moléculaire – les orthologies – seraient d'un autre ordre, plus statistique.

Mais selon Goodman (1989), les hypothèses d'orthologie des séquences nucléiques sont plus solides que ne le laisse penser Patterson. En dehors des cas où, effectivement, il est difficile de distinguer séquences orthologues et paralogues, il n'y aurait que deux sources potentielles d'erreurs dans la construction des arbres moléculaires : 1) quand les mutations successives dans un même site sont fréquentes et 2) quand l'alignement des séquences se révèle problématique.

2.2. Alignement et mutations multiples

Les comparaisons des séquences moléculaires sont dépourvues d'ambiguïté du point de vue des états de caractères étudiés : une leucine est une leucine, une guanine est une guanine quelle que soit la position taxinomique de son propriétaire. Mais la question de l'orthologie se pose au niveau des comparaisons. Prenons l'exemple de la séquence d'un gène codant pour une protéine. La séquence s'exprime sous la forme d'un enchaînement de nucléotides. Les séquences de ce gène chez différentes espèces sont plus ou moins longues : des phénomènes d'insertion et de délétion de nucléotides sont responsables de ces différences. La comparaison des séquences nécessite la réalisation d'un alignement. De cet alignement de séquences naîtront toutes les inférences phylogénétiques, notamment la comparaison des nucléotides, site par site. Par exemple, la présence chez telle et telle espèce d'une adénine ou d'une guanine en

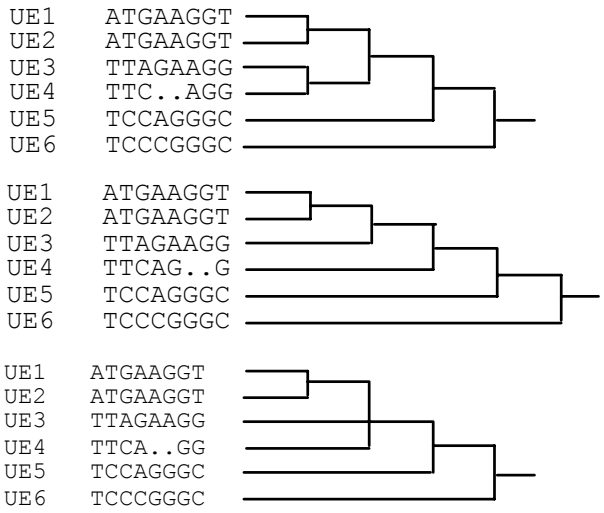


FIGURE IV.5. *Différents alignements de 6 séquences comprenant 8 sites nucléotidiques. Les deux premiers alignements conduisent à des arbres différents de même longueur minimum (11 pas), tandis que le troisième donne ces deux arbres (l'arbre consensus est représenté) ; pour tous ces arbres, I.C. = I.R. = 1 (voir chapitre V.4). La pondération différentielle des transitions et des transversions (poids de 1 et 2 respectivement) ou des indels (poids de 2) ne modifie pas la conclusion.*

un site de la séquence, devient le caractère moléculaire dont il s'agit de tirer la signification phylogénétique. L'innovation évolutive est le remplacement d'une adénine par une guanine – ou l'inverse -, ou une insertion/délétion chez telles espèces. Les hypothèses d'homologie et toutes les constructions phylogénétiques qui en dérivent sont induites par l'alignement des séquences dont la qualité réside dans l'appréciation de leur degré de congruence : on aligne les séquences de telle manière qu'elles se ressemblent le plus, c'est-à-dire que leur superposition implique le minimum de différences site par site. Les méthodes d'alignement des séquences conditionnent donc les hypothèses d'orthologie et, conséquemment, les constructions d'arbres moléculaires. C'est là une difficulté propre aux données moléculaires. La figure IV.5 illustre un cas simple où des alignements différents mais équivalents en terme de coût conduisent à des arbres différents.

Une autre difficulté réside dans la superposition dans un site donné de mutations semblables. Les événements évolutifs affectant la séquence d'un gène reposent, entre autres, sur le remplacement d'une base par une autre. Pour un site donné les états de caractères sont les quatre bases : présence d'une adénine (A), d'une guanine (G), d'une cytosine (C) ou d'une thymine (T) dans le cas de l'ADN (ou bien A, G, C, et uracile U dans le cas de l'ARN). Des mutations successives A ➞ G ➞ A peuvent être responsables de l'observation dans deux séquences d'une même base A à un site donné, alors qu'il n'y aurait pas homologie mais réversion. Les réversions ne sont pas rares en morphologie mais on conviendra qu'avec seulement 4 états de transformations possibles, les erreurs d'interprétation phylogénétique dues aux mutations successives dans la structure des gènes peuvent être fréquentes.

Il reste qu'une fois réalisé l'alignement des séquences, la construction de cladogrammes repose sur les hypothèses de synapomorphies. Qu'il s'agisse de données morphologiques ou moléculaires, quelle que soit la difficulté de construire lesdites hypothèses, l'analyse cladistique s'applique à tout type de caractère discret.

3. Une méthode hypothético-déductive

L'analyse cladistique est une méthode profondément empirique. C'est en quelque sorte une simple mise en conformité des observations, chaque observation ayant *a priori* la même valeur. L'approche cladistique prétend ne faire appel à aucun modèle sur le processus évolutif. Que les caractères évoluent à des vitesses égales ou à des vitesses différentes n'influe pas, en principe, sur l'analyse. Toutefois la pondération différentielle des caractères permet d'exprimer dans un cadre cladistique des hypothèses sur le comportement des caractères. C'est notamment le cas des transformations des nucléotides : ce point sera abordé au paragraphe 6 ; des contraintes peuvent être imposées *a priori* sur les modes d'évolution des caractères dans le cadre d'une analyse de parcimonie (voir paragraphe 5). Mais ces options, qui doivent être justifiées en amont de l'analyse, ne conditionnent pas la mise en pratique de l'analyse cladistique.

En règle générale, l'application de la méthode n'exige aucune autre contrainte que l'hypothèse selon laquelle les caractères évoluent indépendamment les uns des autres. Or cette exigence tient de l'observation empirique. On a observé depuis longtemps que les taxons possèdent à la fois des caractères restés à l'état primitif et des caractères présents à l'état dérivé. L'homme, comme la salamandre et la tortue, a gardé 5 doigts au pied alors que les circonvolutions de son cerveau se sont sensiblement transformées. L'ornithorynque pond encore des oeufs mais il a acquis pour nager dans l'obscurité un système de sonar que n'ont pas acquis la plupart des mammifères vivipares.

Néanmoins, l'observation n'est jamais neutre. Des caractères observés comme des traits distincts peuvent en réalité être liés pour diverses raisons (déterminisme génétique commun, contraintes biomécaniques, etc.). Par exemple des traits ostéologiques et des traits myologiques liés à la réduction du nombre de doigts ne devraient pas être comptés comme des caractères indépendants.

On peut aussi invoquer la théorie évolutionniste pour justifier l'hypothèse d'évolution indépendante des caractères. La spéciation vue comme la production d'un nouveau pool génétique n'implique pas que tout le patrimoine génétique et, conséquemment, phénotypique, soit transformé. Seuls quelques traits divergent et se fixent en raison de la barrière d'interfécondité. Deux espèces filles gardent de leur espèce mère nombre de caractères inchangés.

3.1. Le principe de parcimonie

L'analyse cladistique peut être qualifiée de méthode hypothético-déductive. Hypothèses sur le sens des transformations de caractères et déductions sur les affinités phylogénétiques caractérisent la méthode, qui reste néanmoins, à l'inverse des théories scientifiques comme celles de la physique, une méthode historique. L'application de la méthode nécessite le refus des hypothèses *ad hoc,* ou, tout au moins, leur minimisation, c'est-à-dire l'application du principe de parcimonie.

Qu'est ce qu'une hypothèse *ad hoc* en matière de construction phylogénétique ? C'est l'hypothèse d'une transformation de caractère partagée par 2 ou plusieurs taxons et qui n'est pas due à une ascendance commune. L'hypothèse de base admise par l'analyse cladistique est que le même caractère dérivé observé chez deux taxons (ou plus) est dû à l'héritage à partir d'une espèce ancestrale propre.

Prenons l'exemple de la figure IV.6A. Sachant que le même caractère dérivé 1 est observé chez B et C, le cladogramme (figure IV.6A1) montre B et C en position de groupes frères et compte deux hypothèses de transformations évolutives : une première hypothèse est celle de la transformation du caractère 1 chez l'ancêtre de (B,C) : cette transformation correspond à une hypothèse généalogique. Une seconde hypothèse est celle de la transformation du caractère 2 chez C.

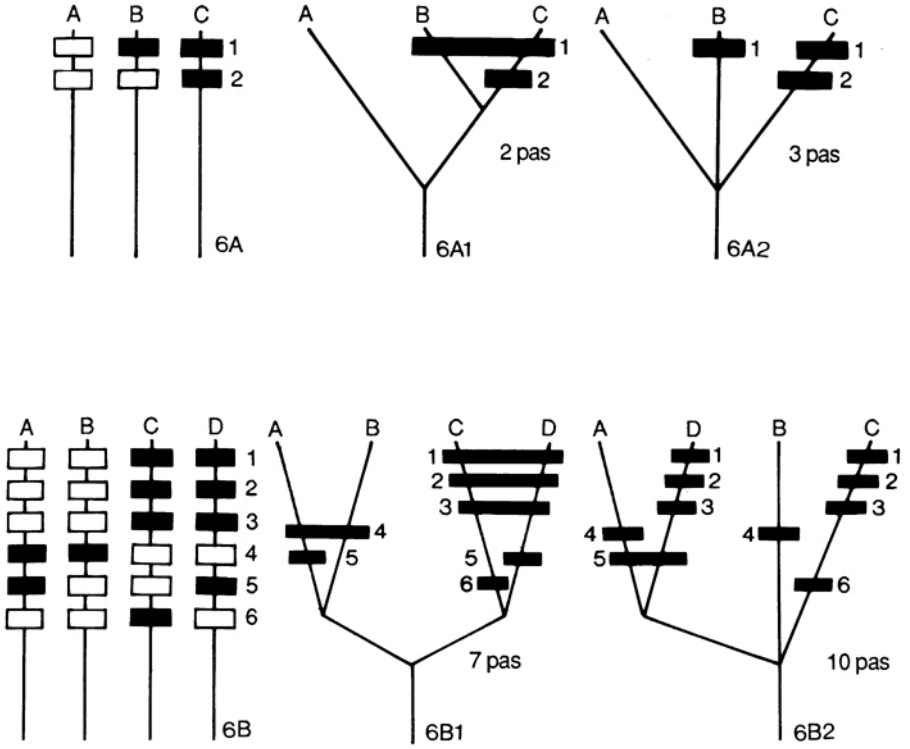


FIGURE IV.6 *Analyse cladistique et principe de parcimonie. A : relations de parenté entre 3 taxons A, B et C ; 1-2 : caractères. B: relations de parenté entre 4 taxons A, B, C et D ; 1-6 : caractères. Barre noire : état apomorphe, barre blanche : état plésiomorphe.*

L'arbre IV.6A2 n'implique pas que les taxons B et C soient proches parents. Il compte trois hypothèses de transformations évolutives : aucune ne supporte une hypothèse de parenté : deux hypothèses de transformation du caractère 1 chez B d'une part et chez C d'autre part, et une hypothèse de transformation du caractère 2 chez C. Cet arbre implique que le caractère dérivé partagé par B et C n'est pas hérité d'un ancêtre commun. Plus long que l'arbre IV.6A1 – 3 pas évolutifs (*steps* au sens de Camin et Sokal, 1965) au lieu de 2 – il contient une hypothèse *ad hoc* parfaitement inutile pour rendre compte de la distribution des caractères.

La minimisation des hypothèses *ad hoc* permet de lever des contradictions dans les distributions de caractères. Sur la figure IV.6B l'absence de congruence entre les regroupements construits à partir des caractères 1, 2, 3 et 4 d'une part, et à partir du caractère 5 d'autre part (figures IV.6B1-B2), implique la nécessité d'invoquer des hypothèses *ad hoc*. Le taxon D ne peut être à la fois proche parent de C avec lequel il partage 3 apomorphies (caractères 1, 2 et 3) et de A avec lequel il ne partage qu'une apomorphie (caractère 5). Si les observations sont justes, il convient donc de minimiser le nombre d'hypothèses *ad hoc*. Cette minimisation consiste à ne compter qu'une seule transformation pour chacun des caractères 1, 2, 3 et 4 (hypothèse généalogique) et deux transformations indépendantes pour le caractère 5 : soit au total 7 pas pour les 6 caractères. En

conséquence D est tenu pour proche parent de C et non de A (figure IV.6B1). L'hypothèse inverse (figure IV.6B2) implique une seule transformation pour le caractère 5 et deux transformations pour chacun des caractères 1, 2, 3 et 4, soit au total 10 pas pour les 6 caractères : c'est une hypothèse moins économique.

Dans le cadre précis d'observations effectuées sur des taxons terminaux, le principe de parcimonie permet, sachant quels sont les états dérivés des caractères, d'évaluer la quantité des caractères dus à l'ascendance (les synapomorphies) et la quantité des caractères qui ne sont pas dus à l'ascendance (les homoplasies : convergences et réversions). Les homoplasies sont les hypothèses *ad hoc* puisque non liées à l'ascendance. Ce sont les « changements évolutifs supplémentaires » de d'Udekem-Gevers (1990).

Le principe de parcimonie a un rôle plus fondamental encore puisque lui seul permet de poser des hypothèses de synapomorphie (figure IV.6, A1). Hennig ne fait pas référence explicite au principe d'économie mais il renvoie à la notion de congruence qui en est un corollaire. La question de savoir si Hennig a préconisé la parcimonie a fait l'objet de controverses (nous y reviendrons à propos des analyses de compatibilité). Mais ce point d'histoire des sciences est tout à fait secondaire.

Sur la nature hypothético-déductive de la méthode cladistique, Hennig (1966 p.21) écrit : « La présence de caractères apomorphes chez différentes espèces fournit toujours un motif pour suspecter une parenté ; leur origine par convergence ne devrait pas être envisagée *a priori* ». Ce point est important, car toute hypothèse de synapomorphie peut, en réalité, être erronée et être du ressort de l'homoplasie. Dans le cas de la figure IV.6B1, le caractère 1 partagé par C et D est apparu une fois. Mais en réalité il est peut-être apparu, indépendamment, une fois chez C et une fois chez D. Le raisonnement par l'absurde peut être suivi pour chacun des autres caractères qui ont servi à la construction de la figure IV.6B1. Dans ce cas, les groupes (A,B) et (C,D) ne sont pas légitimes ; il n'y a pas d'hypothèse sur les parentés des taxons. Le même raisonnement vaut pour le caractère 5 qui a effectivement pu apparaître une fois chez A et une fois chez D (comme le montre la figure IV.6B1). Si chacun des caractères dérivés contenus dans la figure IV.6B est tenu pour être apparu indépendamment chez les taxons qui les portent, aucun schéma de relations de parenté ne peut être construit. Naturellement, le choix dans l'infinité des solutions non parcimonieuses, la construction de n'importe quel groupement ou bien l'absence de toute construction sont étrangers à la démarche cladistique. Le but de l'analyse phylogénétique est la construction d'un schéma relationnel qui ne soit pas arbitraire : un tel schéma doit pouvoir être soumis à réfutation par l'introduction de nouveaux caractères et/ou de nouveaux taxons. L'arbre le plus court est celui qui permet ce type de contrôle.

3.2. La notion de congruence

La démonstration qui précède montre que le principe de congruence n'est autre que le principe de parcimonie ou d'économie d'hypothèses.

La figure IV.6B implique 7 hypothèses de transformations : c'est l'arbre le plus court en nombre de transformations ou pas, compte-tenu de la distribution des caractères. C'est aussi une synthèse des phylogénies construites caractère par

caractère. La congruence entre chacune des images phylogénétiques correspondant aux caractères 1,2,3,4, aboutit à l'arbre ((A,B) (C,D)). En revanche, l'image phylogénétique donnée par le caractère 5, ((A,D),(B,C)), n'est pas congruente avec les trois précédentes. L'arbre le plus court élimine la contradiction entre les différentes images phylogénétiques et explique la distribution du caractère 5 par une homoplasie (convergence).

Une autre façon de résoudre l'absence de congruence ou de lever la contradiction entre distributions de différents caractères est de retourner, à l'issue de l'analyse, à la définition des caractères. Ce « retour » aux caractères est fréquent en morphologie où l'identification du caractère, primordiale, peut toujours être remise en cause. Le caractère 5 (barre noire) jugé comme similaire chez les taxons A et D (figure IV.6B2) est-il réellement similaire ? C'est ce que Hennig (1966) nomme la phase de « contrôle, correction, nouveau contrôle » (*checking, correcting, and rechecking*). Un examen attentif pourrait montrer que la similitude partagée par A et D n'est que superficielle : il ne s'agit pas du même caractère. Les tests de la congruence et de la ressemblance ont alors levé l'hypothèse d'homologie pour l'état apomorphe présumé du caractère 5 chez A et chez D. On peut aussi aboutir au résultat inverse : le nouvel examen ne permet pas de distinguer le caractère 5 de A du caractère 5 de D ; le test de la ressemblance est passé mais non celui de la congruence. Ce dernier est donc le test le plus sévère de l'homologie (Patterson, 1988) ; on admettra alors que le « même » trait est apparu indépendamment chez A et chez D. Le partage du « même » caractère par A et D n'est pas dû à une espèce ancestrale propre à (A,D). Le caractère 5 n'est pas suffisant pour réfuter la distribution des autres caractères.

Le principe de congruence dont se réclame Hennig repose donc lui-même sur le principe de parcimonie ou d'économie d'hypothèses.

4. Les critères d'identification du sens de transformation des caractères

La méthode cladistique de reconstruction phylogénétique, qui est fondée sur l'analyse des caractères, quels qu'ils soient, repose sur l'identification du sens de leurs transformations, ou polarité (plésiomorphe ➞ apomorphe) et, comme on l'a vu, sur le principe d'économie d'hypothèses (parcimonie). Il est d'usage d'invoquer quatre critères d'identification du sens des transformations de caractères de type a ➞ a'. Les deux principaux sont le critère de comparaison extra-groupe et le critère ontogénique. Ces deux critères ressortissent au principe de parcimonie. S'y ajoutent deux critères, jugés accessoires dans la mesure où ils ne s'appliquent pas indépendamment des critères précédents, le critère paléontologique et le critère chorologique. Les données moléculaires ne sont analysées qu'au travers du critère de comparaison extra-groupe.

4.1. Le critère de comparaison extra-groupe

Hennig (1966) ne nomme pas ainsi ce critère mais emploie l'expression « analyse de groupes apparentés » due à Maslin, (1952 : *related groups*). Une telle analyse se situe dans ce que Hennig appelle les « corrélations de morphoclines ». L'expression « comparaison extra-groupe » (*outgroup comparison*) est due à Wiley (1976) et l'usage l'a consacrée.

Le critère se définit comme suit : si un caractère observé dans le groupe étudié est également présent à l'extérieur du groupe (c'est-à-dire dans le ou les extra-groupes), il est plésiomorphe pour le groupe étudié ; s'il n'est présent qu'à l'intérieur du groupe étudié il est apomorphe.

Le critère permet d'identifier le degré d'universalité du caractère examiné (ou de l'état du caractère). On a vu précédemment que les notions d'apomorphie et de plésiomorphie sont des notions relatives. Le critère de comparaison extra-groupe vise à identifier le niveau précis (tel nœud du cladogramme) où le caractère est apomorphe.

Contrairement à ce qui est souvent écrit, la comparaison extra-groupe ne doit pas se restreindre au seul groupe frère du groupe étudié (s'il est connu). Si l'on souhaite retirer toute ambiguïté à l'analyse, elle doit être appliquée à plusieurs groupes extérieurs au groupe étudié (voir notamment Farris, 1982 ; Maddison *et al.* 1984).

4.1.1. Combien d'extra-groupes ?

Les figures IV.7 à IV.11 montrent l'application du critère de comparaison extra-groupe. Le problème est celui d'identifier les parentés dans un groupe de 3 taxons (A, B et C) et, accessoirement, de tester la monophylie du groupe formé par A, B et C. Dans les exemples illustrés par les figures IV.7 à IV.10, il est posé *a priori* que l'extra-groupe choisi (ou les extra-groupes) n'est pas étroitement apparenté à l'un quelconque des membres du groupe étudié.

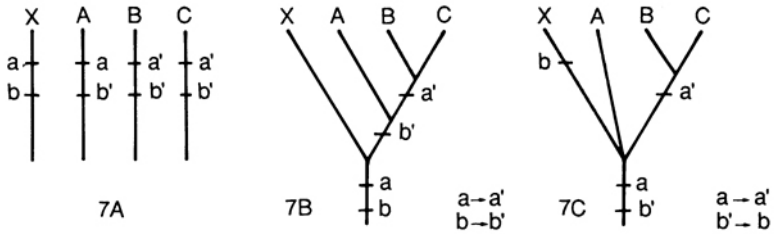


FIGURE IV.7. *Application du critère de comparaison extra-groupe. A, B, C : taxons analysés ; X : extra-groupe ; a-a', b-b' : caractères.*

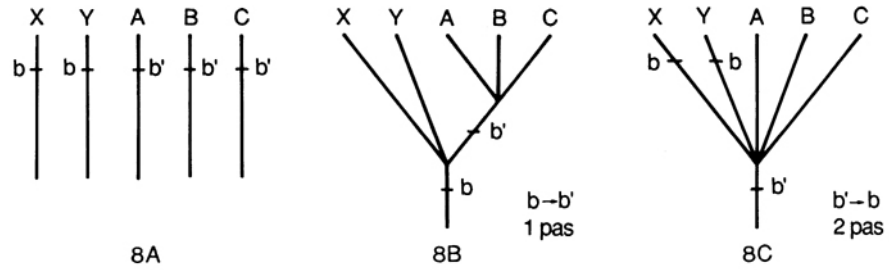


FIGURE IV.8. *Application du critère de comparaison extra-groupe. A, B, C : taxons analysés ; X, Y : extra-groupes ; b-b' : caractères.*

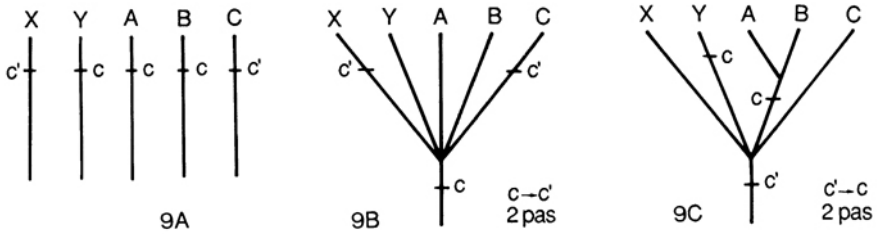


FIGURE IV.9. *Application du critère de comparaison extra-groupe. A, B, C : taxons analysés ; X, Y : extra-groupes ; c-c' : caractères.*

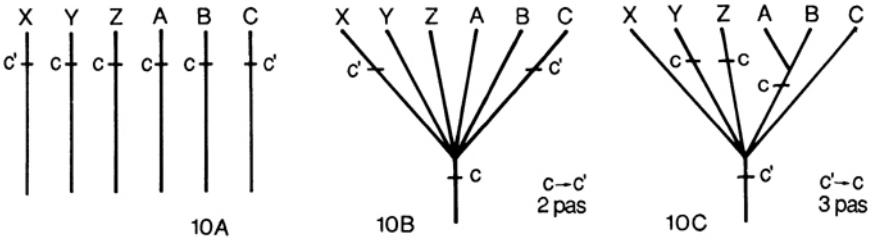


FIGURE IV.10. *Application du critère de comparaison extra-groupe. A, B, C : taxons analysés ; X, Y, Z : extra-groupes ; c-c' : caractères.*

La figure IV.7 montre l'utilisation d'un seul extra-groupe (X). Deux caractères sont analysés, se présentant chacun sous deux états : a-a' et b-b'. L'état a est présent à la fois dans le groupe étudié (chez A) et à l'extérieur du groupe étudié (chez X) : il est plésiomorphe pour le groupe étudié. *A contrario*, l'état a' qui n'est présent qu'à l'intérieur du groupe étudié (chez B et chez C) est apomorphe. La transformation a ➞ a' représente la synapomorphie de (B,C).

En revanche, l'interprétation des états b et b' n'est pas dénuée d'ambiguïté. L'état b' n'est présent que dans le groupe étudié. La comparaison ne porte donc que sur deux groupes : l'ensemble formé par A, B et C d'une part, et d'autre part l'extra-groupe X. L'état ancestral des deux groupes peut être b (la transformation est b ➞ b'), auquel cas le groupe (A,B,C) est monophylétique (figure IV.7B) Il peut tout aussi bien être b', auquel cas la transformation b' ➞ b caractérise l'extra-

groupe X : le groupe (A,B,C) n'apparaît pas comme un groupe monophylétique (figure IV.7C).

Cet exemple montre la nécessité d'introduire plusieurs extra-groupes dans l'analyse, si l'on souhaite tester la monophylie du groupe étudié. Ces extra-groupes ne doivent pas être étroitement apparentés : ils ne doivent pas former un groupe monophylétique, car, dans ce cas, il n'y aurait en fait qu'un seul extra-groupe et nous serions ramenés au cas de figure précédent.

La figure IV.8 montre l'utilisation de deux extra-groupes (X et Y). L'état b', qui est absent chez X et Y mais présent chez tous les membres du groupe étudié, est donc dérivé pour ce groupe. La transformation est b ➞ b'. Cette hypothèse est fondée sur le principe de parcimonie : elle ne « coûte » qu'un pas ; la transformation inverse b' ➞ b coûterait deux pas (chez X et chez Y, qui, rappelons-le, ne sont pas étroitement apparentés) (Figure IV.8C).

Les caractères peuvent être affectés par l'homoplasie, aussi bien chez les membres du groupe étudié que chez les extra-groupes.

Prenons l'exemple de la figure IV.9. Les états c et c', présents chacun dans le groupe étudié, sont également présents chez les extra-groupes : c chez Y et c' chez X. Le critère de comparaison extra-groupe ne permet pas d'identifier la polarité du caractère. La transformation c ➞ c' « coûte » deux pas et ne permet pas de construction dichotomique (figure IV.9B). La transformation inverse c' ➞ c « coûte » également deux pas et résout partiellement l'arbre (figure IV.9C).

L'introduction d'un troisième extra-groupe (fig. IV.10A) lève l'ambiguïté introduite par l'homoplasie et permet d'opter pour l'une ou l'autre des hypothèses de transformation. L'état c est présent chez Z. La transformation c ➞ c' ne « coûte » que deux pas (mais ne permet pas de résoudre le problème phylogénétique (figure IV.10B). La transformation inverse c' ➞ c résout partiellement le problème mais « coûte » un pas de plus : trois apparitions indépendantes de l'état c (figure IV.10C) : cette hypothèse est rejetée. On admettra donc que la transformation est c ➞ c' et que l'état c' est apparu indépendamment chez X et chez C (figure IV.10B).

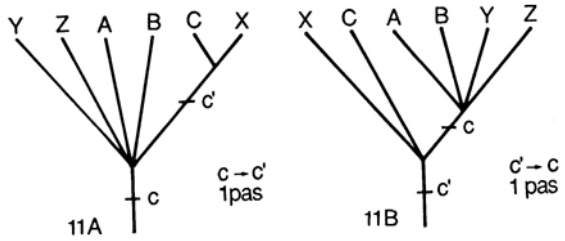


FIGURE IV.11. *Application du critère de comparaison extra-groupe. A, B, C : taxons analysés ; X, Y, Z : extra-groupes ayant la liberté de s'insérer entre A, B et C ; c-c' : caractères.*

Ces exemples montrent que l'application du critère de comparaison extra-groupe ne permet de polariser les caractères sans ambiguïté que par l'application du principe de parcimonie.

En revanche, la qualité d'extra-groupe accordée à tel ou tel taxon dépend d'un choix purement empirique ou bien d'une hypothèse phylogénétique préexistante. Ce choix peut être erroné, comme peut être erronée l'hypothèse préalable de la monophylie du groupe étudié. Prenons l'exemple des figures IV.10 et IV.11. Contrairement au postulat qui nous a guidé jusqu'à présent, admettons qu'un extra-groupe (ou plusieurs) puisse être inclus dans le groupe étudié. Cela revient à dire que le choix des taxons en tant qu'extra-groupe est erroné : tel taxon habituellement considéré comme extérieur au groupe étudié (sur des bases de classification traditionnelle ou d'analyse de similitude globale) est en réalité apparenté à l'un des membres du groupe étudié. La situation de la figure IV.10A peut alors être illustrée plus simplement encore que ne l'indique la figure IV.10B. Les figures IV.11A et IV.11B montrent que les deux transformations c ➞ c' et c' ➞ c sont possibles (un pas seulement à chaque fois) avec des arbres évidemment différents. Par cet exemple, on voit que l'application du principe de parcimonie conduit à rejeter le statut d'extra-groupe choisi *a priori* pour certains taxons.

4.1.2. Le choix des extra-groupes et les limites d'application du critère

Les limites d'application du critère de comparaison extra-groupe sont celles de l'observation. L'hiatus morphologique entre les taxons du groupe étudié et les extra-groupes est parfois tel que la polarisation de nombreux caractères n'est pas concluante. Si, par exemple, la morphologie du 5e métatarsien est le caractère analysé et si le 5e doigt manque chez l'extra-groupe, le critère ne sera pas opérationnel. Les extra-groupes doivent donc posséder des caractères pertinents : si l'on veut résoudre un problème de parentés phylogénétiques des primates on évitera de choisir comme extra-groupes des salamandres ou des oiseaux ! Morphologie et molécules n'échappent pas à cette exigence. La paléontologie le montre, l'extinction est responsable des hiatus existant entre des groupes tant actuels que fossiles. Certains groupes fossiles paraissent ainsi « totalement » isolés et il est alors difficile d'identifier des caractères morphologiques permettant des comparaisons. En outre, sachant que de nombreux taxons fossiles ne sont représentés dans nos archives paléontologiques que par certains types de caractères (denture notamment) les éléments de comparaison sont réduits d'autant.

Le critère de comparaison extra-groupe s'applique à tous les caractères discrets, notamment moléculaires. Un site donné dans une séquence de nucléotides peut être assimilé à un caractère. L'état a de la figure IV.7 peut être une guanine (G) et l'état a' une cytosine (C). La polarité est alors G ➞ C : substitution d'une cytosine à une guanine. Mais les séquences des extra-groupes, tant des gènes que des protéines, peuvent être tellement différentes de celles des autres groupes qu'il est parfois difficile de les aligner afin de les comparer aux autres séquences. Dans ces conditions, l'extra-groupe ne permet pas la polarisation. Par exemple, sur la figure IV.7, si l'extra-groupe X possède une adénine là où A, B et C possèdent une guanine ou une cytosine, la polarisation des transformations chez A, B et C est impossible. A l'inverse, les séquences peuvent être tellement semblables (on dit « conservées ») qu'elles ne permettent pas de construire une phylogénie. L'explication évolutionniste de ces deux cas de figure tient à la vitesse d'évolution des séquences étudiées. Le taux de mutation de

l'ADN mitochondrial, par exemple, est tel qu'il n'est pas possible de comparer les séquences pour des taxons ayant divergé depuis longtemps. Le taux de mutation du cytochrome c est si bas, au contraire, que les séquences d'acides aminés à l'intérieur d'un ordre de mammifères comme celui des Primates, sont identiques. En revanche, ce faible taux de mutations permet de comparer les grands embranchements ayant divergé depuis plus d'un milliard d'années (Fitch et Margoliash, 1967). Dans ce cas, l'éloignement des extra-groupes n'entraîne plus de difficultés.

4.1.3. Extra-groupes et parcimonie

Il reste que dans une analyse phylogénétique fondée sur de nombreux caractères, la congruence entre les hypothèses de polarité permet souvent de lever les ambiguïtés entraînées par l'homoplasie, y compris celles affectant les caractères des extra-groupes. Dans le cas de la figure IV.9, si de nombreux caractères, autres que c, supportent un arbre de configuration (A(B,C)), la polarité du caractère c est c ➞ c' (deux pas), la polarité inverse c' ➞ c impliquant trois pas.

Selon ce point de vue, le critère de comparaison extra-groupe est une analyse structurale et non une analyse des processus ayant mis en place les caractères. Le critère permet la construction d'un schéma relationnel (le cladogramme) tiré de la seule interprétation parcimonieuse de la distribution des caractères. Ce dernier aspect du critère de comparaison extra-groupe – une simple application du principe de parcimonie – est rejeté par nombre de biologistes pour lesquels la Nature n'a pas à se plier à une contrainte d'ordre logique.

Malgré ces réticences, plus conceptuelles que pratiques, le critère de comparaison extra-groupe, tel qu'il a été défini ici, est le critère le plus généralement utilisé dans les constructions phylogénétiques fondées sur l'analyse des caractères, qu'il s'agisse de traits morphologiques pris sur l'actuel ou le fossile, ou de traits biochimiques.

4.2. Le critère ontogénique

Le critère ontogénique se situe dans la sphère de la « loi biogénétique fondamentale » ou « loi de la récapitulation » : l'ontogénie récapitule la phylogénie, autrement dit l'embryon récapitule lors de son développement la succession des états ancestraux. Quoique selon Hennig (1966, p.96) un rejet total de la « loi biogénétique fondamentale » est certainement injustifié, il est patent que pour nombre de biologistes, cette « loi » passe pour infirmée, notamment à la suite des travaux de Garstang dans les années vingt, de de Beer à partir des années trente, et plus récemment de Gould (1977). A l'inverse, la « loi biogénétique fondamentale » reformulée récemment par Nelson (1973a,b, 1978 ; Nelson et Platnick, 1981) est présentée comme le critère primordial de la reconstruction phylogénétique. Aussi à la question de savoir si la phylogénie diffère substantiellement de l'ontogénie, le paléontologue britannique Colin Patterson n'hésite pas à répondre par la négative (Patterson, 1983, p.27).

4.2.1 La reformulation de la « loi biogénétique »

La reformulation de la « loi biogénétique » par Nelson (1973b, p.330 ; 1978, p. 327) est la suivante : « étant donné la transformation ontogénique d'un caractère depuis un état plus général vers un état moins général, l'état plus général est primitif et l'état moins général est évolué ».

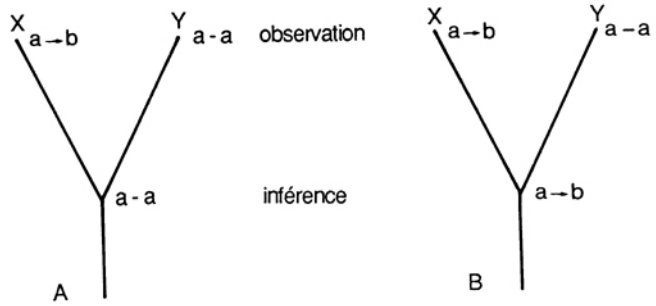


FIGURE IV.12 *Application du critère ontogénique. X, Y : espèces ; a, b : états du caractère ontogénique (a – a : même état à deux phases du développement ; a ➞ b : transformation ontogénique). A : solution parcimonieuse (une transformation) ; B : solution non parcimonieuse (deux transformations : chez l'ancêtre, et suppression chez Y).*

Comparons deux espèces X et Y (figure IV.12A) dont l'une (X) montre la transformation ontogénique a ➞ b d'un caractère et l'autre (Y) ne montre pas de transformation : a – a. L'état a est le plus général parce qu'il est présent à la fois chez X et chez Y; il est plésiomorphe. L'état b présent une fois (chez X) est le moins général : il est apomorphe. Traduits en termes évolutionnistes cette affirmation revient à considérer que, pour ce qui est du caractère étudié, l'espèce Y donne une image de la condition ancestrale. Traduits en termes haeckeliens, en termes de processus, l'affirmation revient à dire que l'ontogénie du caractère – de a vers b – récapitule la phylogénie puisque a est plésiomorphe (ancestral) et b apomorphe : cette situation correspond au processus dit de récapitulation.

Un exemple simple que nous paraphraserons ici est celui donné par Nelson (Nelson, 1978, p. 326 ; Nelson et Platnick, 1981, p. 331) à propos des poissons plats (soles et limandes). Prenons deux espèces X et Y. L'espèce Y (une sardine) possède un oeil de chaque côté du crâne (caractère a), l'espèce X (une sole) possède deux yeux du même côté (caractère b). Lequel des deux caractères est primitif ? Supposons que l'étude de l'ontogénie montre que les embryons des deux espèces ont le caractère a, et que durant le développement de l'espèce X le caractère a se transforme en caractère b. On peut alors répondre à la question : le caractère a est primitif, le caractère b est dérivé.

On n'a pas observé de transformation évolutive. On a observé que le caractère a est plus général que le caractère b : a est présent chez les espèces X et Y et b n'est présent que chez l'espèce X (la sole). La réponse faite en termes évolutionnistes : a est primitif (c'est-à-dire ancestral), est fondée sur le degré de généralité du caractère a et sur l'observation de la transformation ontogénique

a ➞ b. L'inférence phylogénétique implique que, pour les caractères étudiés, l'espèce Y donne une image de l'ancêtre, de telle sorte qu'on ne retient qu'une seule transformation phylogénétique (a ➞ b) chez l'espèce X (cet exemple est celui illustré par la figure IV.12A). C'est pourquoi le critère ontogénique est souvent tenu pour une technique directe d'investigation phylogénétique : la transformation ontogénique est observée, par opposition au critère de comparaison extra-groupe (anatomie comparée et paléontologie incluses) où la transformation est inférée.

L'un des exemples les plus célèbres de récapitulation est celui des poches viscérales (plus communément nommées fentes branchiales, quoique ce terme doit être réservé au stade adulte) que l'on observe chez les embryons de tétrapodes, dont l'homme, comme chez ceux des poissons. Alors que les poches restent ouvertes chez les poissons, elles se ferment chez l'homme et les tétrapodes. La présence d'un pharynx chez l'homme est due à des transformations ontogéniques de la région branchiale ; la trompe d'Eustache y est tout ce qui reste des poches viscérales. On sait par ailleurs que l'homme et le chimpanzé sont deux espèces proches. La descente du larynx qui permet le langage articulé chez le jeune humain, ne se produit pas chez le chimpanzé. La descente (transformation) est une addition terminale.

L'ontogénie de la région branchiale chez les tétrapodes en général et chez l'homme en particulier, récapitule la phylogénie en ce sens que le pharynx (poches fermées) est précédé par des poches ouvertes. Avant d'être en position basse, le larynx humain, jusqu'à l'âge d'un an et demi à deux ans, a la même position que chez le chimpanzé, lequel, de ce point de vue, représente l'état ancestral.

Du point de vue phylogénétique on dira que les poches viscérales – qui persistent chez les poissons adultes (fentes branchiales : branchies) – sont une symplésiomorphie à l'intérieur des chordés (Chordata). La fermeture des poches chez l'adulte est une synapomorphie des tétrapodes. Quant au déplacement du larynx, c'est une autapomorphie d'*Homo sapiens*.

La relecture et la reformulation de la loi de Haeckel par Nelson relèvent du raisonnement formel. Nelson analyse la seule distribution des caractères de façon à en tirer un schéma relationnel ; c'est ce que l'on peut appeler une analyse structurale (par opposition à une analyse des processus), qu'on pourrait même qualifier de structuraliste. En effet, Nelson ne fait en aucune manière référence à un processus de récapitulation qui télescoperait au cours du développement d'un individu tous ses états ancestraux adultes. C'est pourquoi la reformulation de la loi biogénétique rappelle par son style les « lois » du développement de von Baer, dépourvues de perspective évolutionniste. Rappelons que les deux premières lois du développement énoncées par von Baer (1828) étaient formulées ainsi : 1) les caractères généraux d'un grand groupe apparaissent plus tôt dans l'embryon que les caractères spéciaux; 2) Les caractères moins généraux se développent à partir des caractères plus généraux.

La « loi biogénétique » reformulée par Nelson suppose d'abord que l'observation ne nous trompe pas *a priori.* Elle suppose ensuite que le transfert de la transformation ontogénique (observée) dans un contexte de transformation

phylogénétique (inférée) peut se faire simplement en minimisant les événements évolutifs. Le raisonnement ressortit au principe de parcimonie.

Prenons à nouveau l'exemple de la figure IV.12. Si l'absence de transformation est considérée comme un état ancestral, la transformation compte pour une innovation (un pas évolutif sur la branche X) et exprime l'état évolué (figure IV.12A). Ce cas de figure correspond au processus de récapitulation (cas dit de péramorphose). Si, au contraire, la transformation ontogénique est tenue pour illustrer l'état ancestral (figure IV.12B), elle compte pour un pas chez l'ancêtre. L'absence de transformation est alors une perte de transformation, soit un deuxième pas. Cette situation est une solution moins parcimonieuse, qui postule que la transformation ancestrale (un pas) est précédée dans l'histoire par une absence de transformation, ce qui correspond au premier cas de figure. Choisir une hypothèse moins parcimonieuse pour expliquer un même ensemble de données, c'est introduire une hypothèse *ad hoc* : celle-ci ne s'impose qu'en fonction d'autres observations. De la même façon, les cas de convergence vus dans le paragraphe IV.1, n'apparaissent comme tels qu'en posant *a priori* les hypothèses de synapomorphies. Dans la figure IV.6B il est moins parcimonieux d'envisager 2 pas pour la transformation du caractère 5 chez A et chez D qu'un seul pas chez l'ancêtre de (A,D). Mais la prise en compte de l'ensemble des 6 caractères montre que l'hypothèse la plus parcimonieuse (7 pas au total) implique une distribution non parcimonieuse du caractère 5.

L'observation d'une *absence* de transformation n'est interprétée comme le résultat d'une *perte* de transformation (signifiant que l'existence de la transformation a précédé généalogiquement sa disparition) que si d'autres caractères permettent de penser ainsi. En termes de processus, ce cas de figure correspond à la paedomorphose – c'est-à-dire la persistance à l'état adulte de caractères juvéniles par arrêt ou ralentissement du développement somatique (néoténie) ou par accélération du développement germinal (progénèse). On a vu que dans la comparaison extra-groupe, la réalité de l'homoplasie ne réfutait pas celle de la synapomorphie. Dans le cas du critère ontogénique, la paedomorphose réfute-t-elle la « loi biogénétique » reformulée par Nelson et aujourd'hui appelée « règle de Nelson » (Wheeler, 1990) ?

4.2.2 Réfutation et parcimonie

De Beer (1958) et Gould (1977) se sont appuyés sur les processus non récapitulatifs (c'est-à-dire sans addition terminale d'états de transformation), pour minorer fortement, voire nier, l'importance de la « loi biogénétique ». Il n'est pas question de minimiser ici l'intérêt de l'étude des processus et de l'hétérochronie, c'est-à-dire la variation du tempo du développement. Mais les modèles de développement mettant en évidence les différents rôles que joue l'hétérochronie (Gould, 1977 ; Alberch *et al*, 1979) sont fondés sur la reconnaissance préalable de l'état ancestral : l'appréciation des hétérochronies dépend d'hypothèses phylogénétiques préalables. Or la question qui est posée ici est autre : c'est précisément celle de l'apport de l'ontogénie à l'identification de l'état ancestral. La reformulation de la « loi biogénétique » a pour but l'inférence des états primitif et dérivé dans le champ ontogénique.

Une première objection faite à Nelson est qu'il est impossible de tirer des conclusions phylogénétiques à partir d'observations obtenues sur deux espèces seulement (Kluge, 1985, p.22). L'argument est le suivant et s'applique à la figure IV.12 : si une transformation ontogénique est présente chez un ancêtre (figure 12B), seule l'absence de transformation chez un descendant doit compter pour un pas : de la sorte, l'absence de transformation comme la transformation peuvent être tenues à égalité comme deux situations plausibles chez l'ancêtre (cet exemple se rapporte précisément aux processus de paedomorphose). Auquel cas, l'argument ontogénique seul ne peut être utilisé à des fins phylogénétiques. A l'inverse, Nelson compte toute transformation pour un pas ; de la sorte, si nous inférons la présence d'une transformation chez un ancêtre puis sa suppression chez un descendant (figure IV.12B), nous nous situons donc dans une situation non parcimonieuse dont la validité est à justifier à l'aide d'autres observations ontogéniques ou d'autres observations de caractères.

Une seconde objection faite à Nelson porte sur la nature réfutable de son énoncé de « loi biogénétique ». Un même modèle de comparaison d'espèces aux ontogénies postulées, a été proposé par Voorzanger et van der Steen (1982) et par Kluge (1985) afin de démontrer que la parcimonie ne permettait pas de trancher entre des hypothèses contradictoires et que, sur ce plan, la « loi » n'était pas de nature réfutable (figure IV.13).

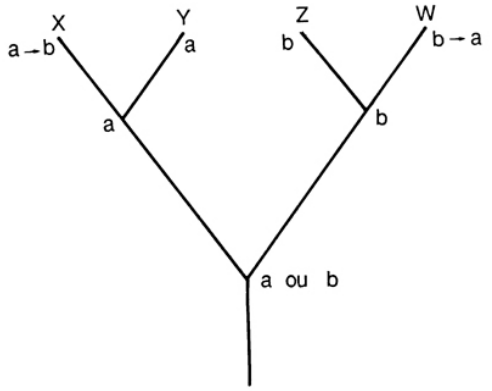


FIGURE IV.13. *Un essai de réfutation de la « règle de Nelson » . W, X, Y, Z : taxons ; a ➞ b et b ➞ a sont deux séries de transformations ontogéniques contradictoires observées chez X et chez W. Chez les espèces Y et Z on n'observe pas de transformation ontogénique.*

Prenons l'exemple de la figure IV.12 et essayons de réfuter l'hypothèse 12A à l'aide de deux autres espèces, Z et W dont nous observons les caractères ontogéniques a et b. On obtient la figure IV.13. La transformation observée chez W est b ➞ a, alors que Z ne présente que l'état b durant son développement. L'application de la « règle de Nelson » aux seules espèces X et Y aboutirait à la conclusion que a est plus primitif que b, tandis que son application aux espèces Z et W aboutirait à dire que b est plus primitif que a, ce qui est contradictoire. Si l'on considère simultanément les espèces X, Y, Z et W, la réfutation n'est pas possible car les deux combinaisons ancestrales contradictoires sont également possibles : l'ancêtre possède aussi bien a que b. Cet exemple implique en effet que

l'ontogénie n'est pas orientée et que les deux transformations contradictoires a ➞ b et b ➞ a coexistent, autrement dit, pour reprendre l'exemple précédent, le pharynx naît des fentes branchiales aussi bien que les fentes branchiales naissent du pharynx. Nelson (1985, p. 36) ne retient pas cet argument et considère au contraire que ce cas n'a jamais été rencontré et que les « relations entre caractères sont universelles ». En réalité, des ontogénies contradictoires ont été signalées chez des arthropodes dont le développement est discontinu, chaque stade ayant sa propre phylogénie (acariens du groupe des Oribatida ; André, 1988 ; Bonde 1984) cite aussi un exemple signalé par Wingstrand chez des mouches, les stratyomidés. Dans une telle situation, l'observation ontogénique n'est phylogénétiquement intelligible que confrontée au critère de comparaison extra-groupe. Que de tels exemples soient rares, pour ne pas dire exceptionnels (André ne cite qu'un exemple tiré de travaux anciens de Grandjean), n'est peut-être dû qu'à la rareté des travaux sur l'ontogénie complexe des arthropodes. Mais ces cas sont remarqués chez des êtres dont l'ontogénie est discontinue avec des remaniements cellulaires complexes. Nelson fonde au contraire son raisonnement sur des ontogénies continues. Il reste qu'il existe une réfutation de l'universalité de la « règle de Nelson » sur la base de l'observation empirique. La « règle de Nelson » doit être comprise comme une stratégie de recherche, non une loi biologique.

4.2.3. la paedomorphose et l'exception à la règle

Les objections à la « règle de Nelson » que l'on retiendra dans ce paragraphe sont celles issues de la reconnaissance des processus non additifs, c'est-à-dire non récapitulatifs (paedomorphose) : ceux-ci réfuteraient aussi bien la « règle de Nelson » que la « loi » de Haeckel.

Ces objections, auxquelles on a déjà répondu en référence à la parcimonie, tirent d'une certaine manière leur source dans l'histoire de l'embryologie et dans l'œuvre même de Haeckel.

L'exemple classique de paedomorphose, ici de néoténie, est celui de l'axolotl. En 1865, le zoologiste français A. Duméril avait découvert le phénomène de néoténie en observant la métamorphose d'un axolotl en salamandre. Chez cet amphibien, la métamorphose ne se fait pas dans la nature : chez l'adulte – c'est-à-dire l'individu capable de se reproduire – persistent des caractères larvaires comme l'existence de branchies. Dans son *Anthropogénie*, Haeckel (1877, p. 392) considère l'axolotl (et d'autres salamandres à branchies) comme un représentant primitif des amphibiens, « au plus bas degré », par opposition aux autres amphibiens dont « les branchies disparaissent chez l'adulte ». On sait que les branchies de l'axolotl adulte, au contraire, ne font pas de l'axolotl un ancêtre mais sont la persistance d'un trait juvénile par arrêt du développement somatique. De Beer (1958), Gould (1977) et bien d'autres, considèrent donc que les phénomènes de paedomorphose réfutent l'argument ontogénique tel qu'il est formulé par Haeckel ou par Nelson. L'axolotl n'est pas un « ancêtre » des amphibiens et la conclusion de Haeckel – qui n'admettait pas d'exception à sa « loi » – était erronée. On ne pourrait alors échapper à la conclusion selon laquelle l'ontogénie n'est pas une source fiable d'information phylogénétique : ce n'est que grâce à l'adjonction d'autres types d'observations que nous pourrions espérer comprendre dans un cadre phylogénétique les transformations ontogéniques.

Mais il n'y a pas de paradoxe à reconnaître à la fois l'erreur de Haeckel et le bien – fondé de la « règle de Nelson ». Nelson se réfère à l'ontogénie des caractères, non celle des organismes tout entiers. Il n'y a pas « d'organismes récapitulatifs » mais récapitulation dans le développement de tel ou tel caractère. Dans le cas de l'axolotl, bien que l'exemple soit devenu un « classique » de la littérature évolutionniste, ce n'est que tout récemment qu'a été menée (Kraus, 1988) une analyse empirique de 41 caractères observés chez les salamandres du genre *Ambystoma*, dont l'axolotl. La conclusion de cette analyse cladistique est sans ambiguïté : les caractères dus à la paedomorphose (34% des caractères) n'ont pas brouillé l'image phylogénétique fondée sur l'application du principe de parcimonie, mais, surtout, n'ont pu précisément être mis en évidence qu'à partir de ce principe, c'est-à-dire, pour ce qui est de l'argument ontogénique, la « règle de Nelson ». Les prétendus caractères primitifs de l'axolotl, ne pèsent pas lourd par comparaison avec les caractères de salamandres terrestres, et, plus précisément, ceux propres au genre *Ambystoma*, un genre qui ne se situe pas à l'origine des amphibiens.

Un contre-exemple est toutefois donné par Mabee (1989) à propos de la phylogénie de poissons osseux perciformes. L'analyse empirique de 63 caractères chez 29 espèces de la famille des Centrarchidae a conduit cet auteur à conclure très nettement que « le critère ontogénique n'est pas un critère phylogénétique valable » (Mabee, 1989, p.415). La raison en est la suivante : les processus d'addition terminale (qui correspondent à la « règle de Nelson ») ne dépasseraient pas 51 % des transformations envisagées, et plus vraisemblablement ne représenteraient que 33 % des cas, un chiffre qui, d'après Mabee, ne permettrait pas de supporter la « règle de Nelson ». Mais l'examen attentif de cet exemple montre qu'en fait plusieurs arbres parcimonieux rendent comptent des données avec des topologies différentes. Autrement dit, le degré de contradiction entre les hypothèses de transformations des caractères est tel que le choix d'un arbre parmi d'autres semble relever de l'arbitraire. L'exemple apparaît plus comme un cas de non-résolution d'un problème phylogénétique que comme une réfutation de la « règle de Nelson ».

Il est difficile d'aborder le critère ontogénique sans faire la part de l'expérimentation, au cœur des analyses de l'ontogénie. Que dit l'expérimentation qui révèle le caractère néoténique de telle ou telle ontogénie ? Que la paedomorphose est un étrange « falsificateur », au sens de Popper, car elle n'est mise en évidence que lorsqu'elle n'existe plus : lorsque la métamorphose – c'est-à-dire la transformation – a eu lieu.

En l'absence d'expérimentation, les caractères néoténiques ne peuvent être révélés comme tels qu'en association avec d'autres caractères, non néoténiques cette fois, qui permettent l'édification de l'hypothèse phylogénétique. Pour mettre en évidence les exceptions à la règle, il faut appliquer la règle. Les processus non additifs ne nous autorisent pas à éliminer l'ontogénie du domaine de l'inférence phylogénétique, sauf à admettre que seules les méthodes dépourvues de risques d'erreur doivent être appliquées, ce qui pose des contraintes extrêmes à l'activité scientifique, voire l'empêche purement et simplement. A ce titre, le critère de comparaison extra-groupe devrait aussi être éliminé : pris isolément un caractère peut nous induire en erreur (par exemple les états b-b' de la figure IV.7 et c-c' de la figure IV.9).

4.3. Les critères paléontologique et chorologique

4.3.1 Le critère paléontologique

Le critère paléontologique appelé encore « critère de la précédence géologique » est souvent cité comme le premier critère de polarisation des caractères (Hennig, 1966 ; Mayr, 1986).

Le critère s'énonce comme suit : si, dans un groupe monophylétique, l'état d'un caractère est présent chez les fossiles anciens et l'autre état est présent chez les fossiles plus récents, le premier est l'état plésiomorphe, le second est l'état apomorphe.

Le critère est opératoire si les parentés entre fossiles ne sont pas trop lointaines. C'est pourquoi Hennig a spécifié que le critère s'applique pour les fossiles appartenant à un groupe monophylétique. Or le problème phylogénétique est bien d'identifier l'étroitesse des liens et la monophylie des groupes. C'est pourquoi ce critère est cité comme critère auxiliaire qui ne peut être appliqué indépendamment des critères principaux (comparaison extra-groupe et ontogénie) si l'on veut éviter la circularité du raisonnement fondé sur l'équation ancien = primitif. On peut citer à ce sujet la boutade de Nelson et Platnick (1981) : appliquer sans discernement ce critère revient à considérer que les blattes qui infestent les caves des immeubles des grandes villes sont plus évoluées que les mammouths qui vivaient il y a 15.000 ans : la comparaison des caractères n'a pas de sens. Or, toute ironie mise de côté, il convient de préciser que blattes et mammouths appartiennent bien à un groupe monophylétique, les Metazoa (sans même remonter aux eucaryotes). Tout est donc question de discernement. Ici le discernement ne se conçoit pas sans le critère de comparaison extra-groupe. Il apparaît ainsi que la position stratigraphique est un caractère extrinsèque à l'organisme, tout comme sa distribution géographique (voir paragraphe suivant). La primauté des caractères intrinsèques en analyse cladistique fait aussi du critère de précédence géologique un critère auxiliaire.

En réalité, dans un groupe monophylétique de faible amplitude (aux caractères moyennement divergents) on peut s'attendre à ce que de nombreux caractères des fossiles anciens soient primitifs par rapport à ceux des fossiles récents ou des formes actuelles. Mais il est tout aussi commun de rencontrer des taxons anciens ayant évolué à leur manière et dont les caractères ont subi des transformations inconnues chez les taxons plus récents.

Ces deux constatations empiriques montrent que le critère, qui ne peut s'appliquer indépendamment de tout autre critère principal, doit être manié avec précaution.

L'exemple des blattes et des mammouths montre que la multiplicité des branchements (la diversification taxinomique avec acquisition de caractères) est responsable de l'ambiguïté introduite par l'application du critère de précédence géologique indépendamment de tout autre critère.

L'application stricte du critère n'est possible qu'au niveau spécifique (ou populationnel), dans le cas d'une lignée phylétique. La lignée phylétique est l'enchaînement au cours du temps d'ancêtres et de descendants sans production de

diversité taxinomique, c'est-à-dire sans branchement. Sans branchement, pas de possibilité de divergence avec acquisition de caractères autapomorphes (sauf, bien entendu, pour l'espèce terminale). Dans ce cas précis, ce qui est ancien est nécessairement primitif par rapport à ce qui est plus récent. Une telle lignée est parfois aussi appelée lignée anagénétique.

Lignées paléontologiques

Le critère de la précédence géologique est souvent appliqué en paléontologie indépendamment de tout autre critère. La superposition stratigraphique est alors tenue comme le révélateur du sens du morphocline : le morphocline devient un chronocline. Ce type d'approche est judicieusement qualifié de « stratophénétique » par Gingerich (1979) puisqu'il repose sur la reconnaissance de la similitude et sur son orientation selon la stratigraphie.

La figure IV.14 montre la phylogénie d'un groupe de mammifères fossiles nord-américains construite à partir d'un caractère : le logarithme de la surface occlusale de la première molaire inférieure. Le choix de ce caractère est dû au fait que, généralement, la taille moyenne des individus d'une espèce est liée à celle de la surface occlusale de la molaire. Les échantillons ont été recueillis dans un seul bassin sédimentaire et sont situés selon l'axe des ordonnées en fonction de leur position stratigraphique. On constate augmentation ou réduction de taille, au cours du temps. En fonction des divergences biométriques, des espèces phylétiques sont reconnues, c'est-à-dire des enchaînements d'espèces sans branchement (par exemple *Pelycodus trigonodus,* de petite taille, et *Pelycodus abditus*, son descendant postulé, de plus grande taille). Des espèces apparues par division (cladogénèse) sont également reconnues, cette fois lorsque deux populations contemporaines sont distinguées sur des bases statistiques (par exemple, *Pelycodus frugivorus* contemporain de *P. jarrovii*, et significativement plus petit). Ce modèle phylogénétique implique que l'enregistrement fossile ne souffre pas d'hiatus, que l'évolution s'est faite sur place : aucun événement phylogénétique n'a impliqué, dans d'autres bassins sédimentaires, les espèces dont on a recueilli les dents. Enfin, l'évolution du caractère est tenue pour une phylogénie d'organismes : la surface occlusale de la première molaire inférieure doit refléter l'évolution des espèces elles-mêmes. Par exemple, si *Pelycodus trigonodus* est considéré comme l'ancêtre de *P. abditus*, on admet que tous les autres caractères de ces deux espèces évoluent de concert. Cet exemple montre donc les limites de la méthode, lorsque celle-ci ne tient compte d'aucun autre critère.

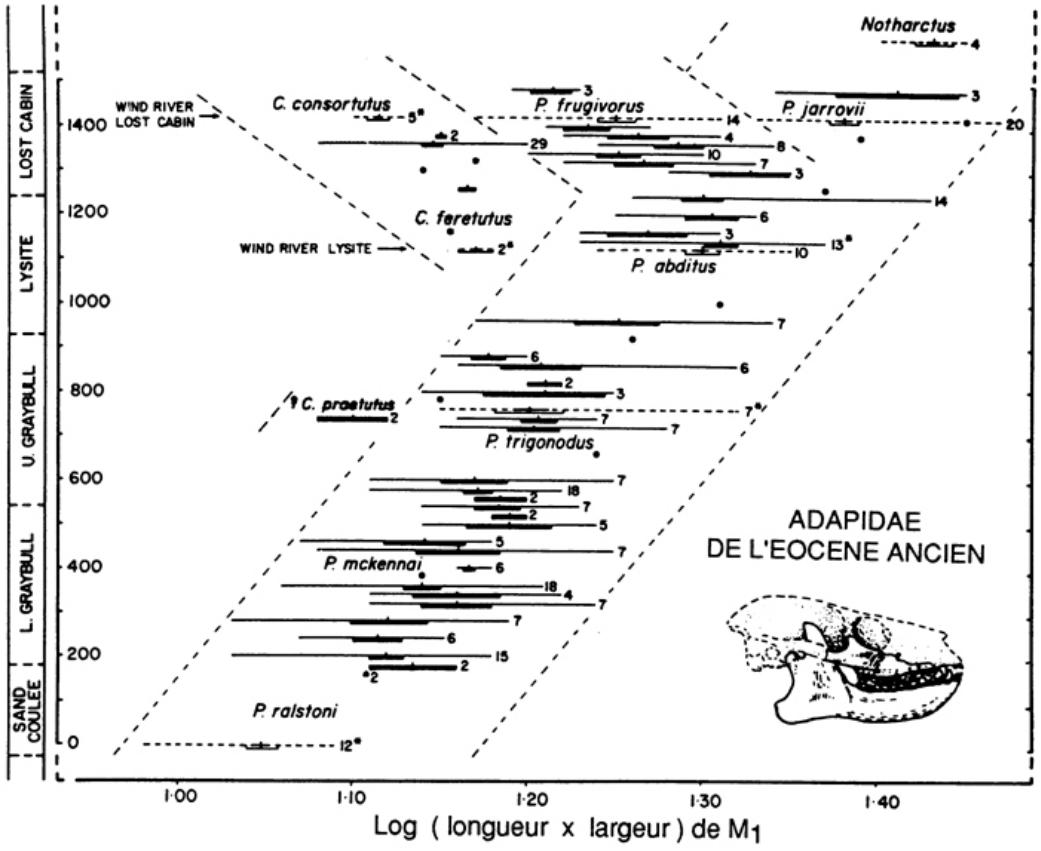


FIGURE IV.14. *Un exemple de stratophénétique. Extension stratigraphique et parentés des primates de la famille des Adapidae (genres* Pelycodus *et* Copelemur*) dans l'Eocène ancien du Bassin de Big Horn en Amérique du Nord. Abscisses : logarithme de la surface de la première molaire inférieure. Ordonnées : section stratigraphique. Lignes horizontales : amplitude de variation de l'échantillon ; barres verticales : moyenne ; sections en gras des lignes horizontales : erreur standard sur la moyenne ; lignes de tirets : parentés postulées. D'après Gingerich (1979).*

Dans la plupart des cas (comme dans la figure IV.14), de telles constructions phylogénétiques expriment l'évolution au cours du temps d'un caractère choisi pour sa pertinence (notamment la facilité de fossilisation et, en conséquence, le nombre élevé d'échantillons), ou bien d'un indice qui résume plusieurs caractères. Dans le premier cas, la phylogénie est en fait une phylogénie de caractères. Dans le second cas, des combinaisons contradictoires de caractères dues au fait que l'ancien n'est pas nécessairement primitif, sont gommées par l'utilisation d'un seul indice. Celui-ci ne fait que résumer grossièrement l'évolution des caractères.

Il reste que le modèle de la lignée phylétique, le seul où s'appliquerait strictement le critère paléontologique implique que l'enregistrement fossile est complet : les organismes anciens appartiennent à la population-mère et les organismes récents appartiennent aux populations-filles.

Lignées et cladogrammes

La lignée phylétique de la figure IV.15A est construite à partir de trois séries de transformations tandis que celle de la figure IV.16A est construite à partir d'une seule série. Elles sont respectivement compatibles avec les cladogrammes IV.15B et 16B. Les cladogrammes représentent la même distribution des caractères et des transformations que les lignées phylétiques, mais toutes les espèces sont tenues pour des taxons terminaux. En revanche, les représentations en lignées phylétiques indiquent que les ancêtres et les descendants sont identifiés. Elles représentent un système « fermé » qui implique que la totalité de l'information est obtenue : il n'y a pas de lacune dans les archives fossiles qui permettrait de renverser la polarité du caractère établie selon le chronocline. Les cladogrammes sont, au contraire, des systèmes « ouverts ». Des populations – non encore découvertes – peuvent s'intégrer dans les cladogrammes sans altérer les transformations de caractères. L'espèce D peut, par exemple, avoir un ancêtre commun avec C et être contemporaine de C dans un autre bassin sédimentaire que celui qui a livré les espèces A, B, C, D des figures IV-15A ou IV-16A.

La figure IV.17 montre un cas extrême où les lacunes de l'enregistrement fossile masquent le processus évolutif de telle façon que l'évolution est l'inverse de celle supposée à partir de la superposition stratigraphique. Le chronocline est a ➞ a' ➞ a"➞ a'". Le morphocline est en fait a'"➞ a"➞ a'➞a. Une telle situation peut être identifiée si d'autres fossiles apparentés à ce groupe sont découverts avec, outre leurs caractères propres, les caractères les rapprochant de l'ensemble A, B, C, D, notamment le caractère a'" ; autrement dit, par l'application du critère de comparaison extra-groupe.

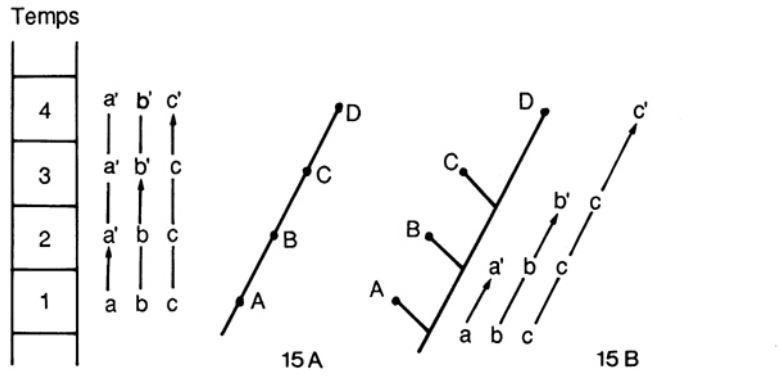


FIGURE IV.15. *Le critère paléontologique. 1-4 : étages géologiques ; A, B, C, D : espèces ; a – a', b – b', c – c' : séries de transformations des caractères . 15A: lignée phylétique ; 15B : cladogramme (arbre dichotomique).*

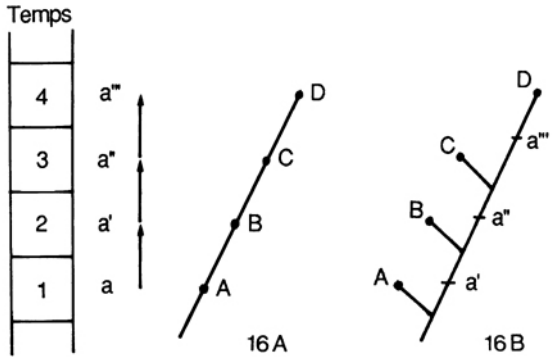


FIGURE IV.16. *Le critère paléontologique. 1-4 : étages géologiques ; A, B, C, D : espèces ; a – a′ – a″ – a‴ : série de transformations du caractère. 16A : lignée phylétique ; 16B : cladogramme (arbre dichotomique).*

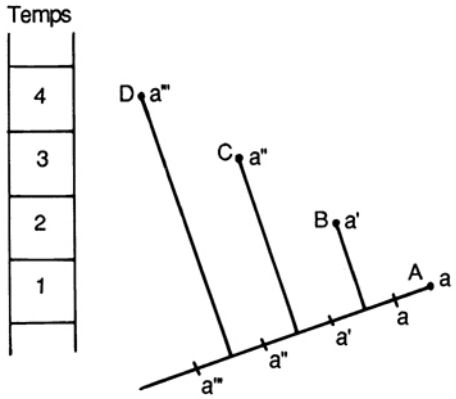


FIGURE IV.17. *Réfutation du critère paléontologique. 1–4 : étages géologiques ; A, B, C, D : espèces ; a‴ – a″ – a′ – a : série de transformations du caractère. La séquence géologique (chronocline) est l'inverse de celle de la série de transformation (morphocline).*

Pour conclure, il convient de rappeler que le critère de la précédence géologique, n'a été discuté que dans le but de polariser les caractères, ce pourquoi il est conçu. L'usage ou le non-usage des données paléontologiques à des fins de construction phylogénétique est un tout autre problème. Ne pas utiliser des données paléontologiques, ou les relativiser en raison des hiatus qui persistent dans les archives fossiles, revient à refuser de l'information, ce qui ne peut être légitimé.

4.3.2. Le critère de progression chorologique

La chorologie est la distribution géographique des êtres vivants.

On admet que lorsqu'une espèce X se subdivise en deux espèces Y et Z, l'état transformé a' apparaît chez l'espèce Z qui s'est le plus éloignée géographiquement de l'espèce initiale (figure IV.18). La conséquence de ce point de vue est que la distribution géographique permet d'établir des hypothèses phylogénétiques.

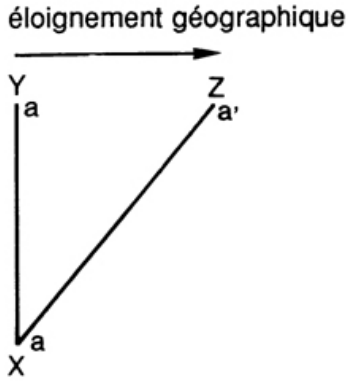


FIGURE IV.18. *Subdivision d'une espèce X en deux espèces Y et Z avec divergence d'un caractère a vers a' chez l'espèce éloignée géographiquement de l'espèce initiale. X et Y sont présents dans la même région géographique.*

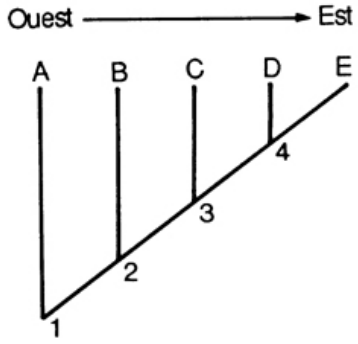


FIGURE IV.19. *Progression chorologique d'Ouest en Est des espèces A à E impliquant une augmentation corrélative du nombre des synapomorphies en passant du nœud 1 au nœud 4.*

Sur la figure IV.19, l'augmentation des synapomorphies depuis l'espèce A jusqu'à l'espèce E est liée au gradient géographique d'Ouest en Est.

La méthode est généralement appliquée lorsque les données morphologiques ne suffisent pas à résoudre un problème de parenté et ce, au niveau spécifique. Cependant, la méthode ne se restreint pas à ce niveau taxinomique et peut s'appliquer à des groupes supra-spécifiques monophylétiques distribués dans des unités géographiques bien circonscrites (Hennig, 1966). Dans ce cas particulier, le rang des taxons n'a pas d'importance.

Le critère de progression chorologique est invoqué le plus souvent comme test des hypothèses fondées sur la morphologie. Il est manifeste, à la lecture de la littérature cladistique, que ce critère est pratiquement abandonné. Son utilisation à des fins de reconstruction phylogénétique suppose en effet une hypothèse de parenté préexistante, à partir de laquelle est posée la localisation de la région ancestrale : c'est bien par conséquent un critère auxiliaire.

Actuellement, les distributions géographiques ne sont pas à la source d'hypothèses cladistiques. Au contraire, l'histoire des distributions géographiques des taxons est entièrement déduite des caractères intrinsèques : tel est le principe de base de la « biogéographie historique » au sens de Nelson et Platnick (1981). On trouvera chez ces auteurs un exposé des méthodes de construction de cladogrammes d'aires à partir des cladogrammes de taxons, cette question échappant au sujet du présent livre.

4.4. Polarisation et construction cladistique

De ce qui précède, il ressort qu'aucun des critères d'orientation des transformations de caractères n'est absolu. Leur application peut toujours entraîner des erreurs sur tel ou tel caractère : toute hypothèse de synapomorphie peut être une erreur, c'est-à-dire relever de l'homoplasie.

La juxtaposition des phylogénies de caractères permet d'évaluer le degré de congruence des données. La phylogénie des taxons ne se conçoit donc qu'au travers des phylogénies de caractères. Plus les caractères, appartenant à différents systèmes biologiques, sont nombreux, plus informatif est le résultat.

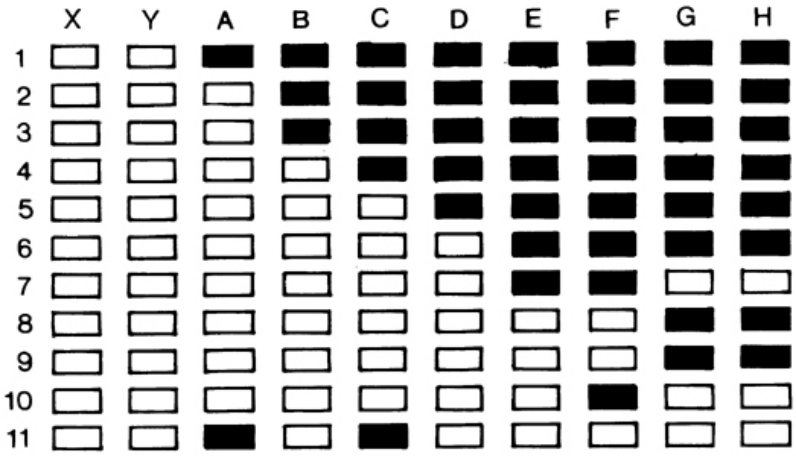
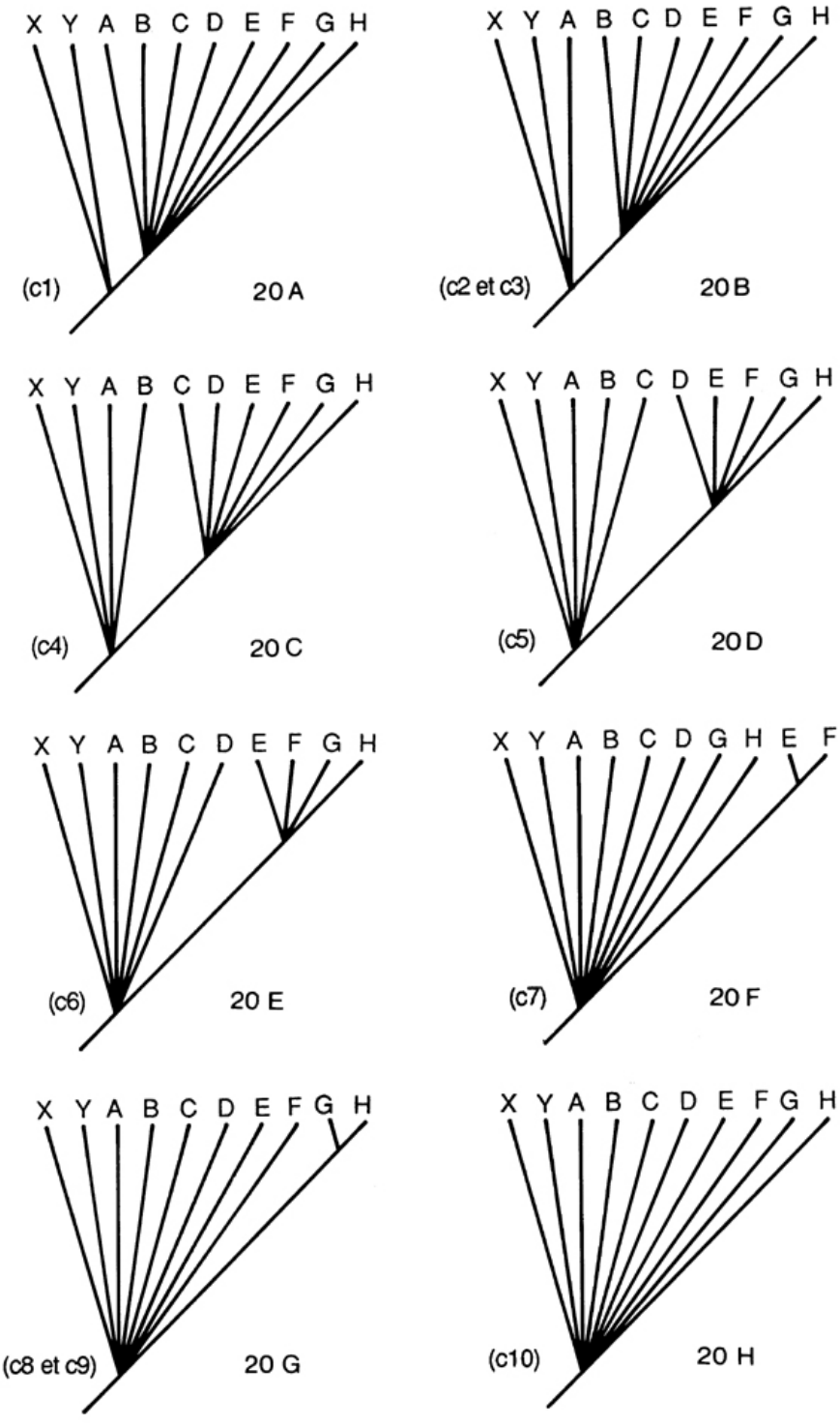


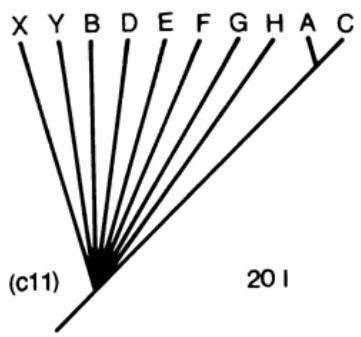
TABLEAU IV.1. *Matrice de 11 caractères pour 8 taxons (A-H : groupe étudié) et 2 extra-groupes (X, Y). Chaque caractère est représenté par deux états. La polarité est donnée par le critère de comparaison extra-groupe : la parcimonie indique que pour chacun des caractères, l'état barre blanche est plésiomorphe et l'état barre noire est apomorphe.*

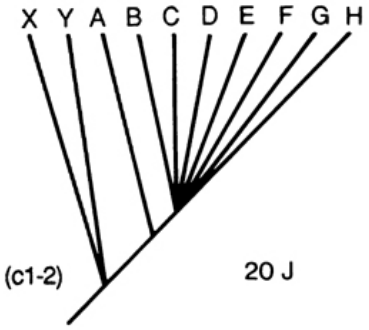
La figure IV.20 résume le fonctionnement de l'analyse cladistique au moyen du critère de comparaison extra-groupe. A partir du tableau IV-1 il est possible de construire la phylogénie de chacun des caractères (figure IV.20A-I). Sachant que X et Y sont les extra-groupes, la phylogénie du caractère 1 permet de regrouper (A,B,C,D,E,F,G,H) (figure IV.20A). Les phylogénies des caractères 2 et 3 sont identiques, elles regroupent (B,C,D,E,F,G,H). La phylogénie du caractère 4 regroupe (C,D,E,F,G,H) etc. Aucun caractère ne donne seul la phylogénie des taxons. Celle-ci est donnée par la mise en congruence des neuf images phylogénétiques, autrement dit par leur addition.

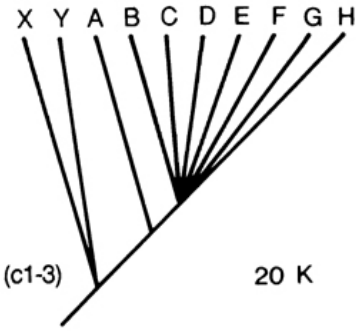
L'addition des phylogénies des caractères 1, 2 et 3 permet de construire l'arbre (A(B,C,D,E,F,G,H)) (figure IV.20K) . L'addition des caractères 1, 2, 3 et 4 permet de construire l'arbre (A(B(C,D,E,F,G,H))) (figure IV.20L) etc. La phylogénie du caractère 11 correspond à l'arbre ((A,C)(B,D,E,F,G,H))) (figure IV.20I). Cet arbre n'est pas congruent avec l'arbre construit à partir des caractères 1 à 10 (figure IV.20Q). Les figures IV.20R et 20S sont les deux compromis possibles. En l'absence de congruence, la parcimonie choisit entre les images phylogénétiques et permet de sélectionner la figure IV.20 R (12 pas) plutôt que la figure IV.20S (14 pas). Les caractères 2, 3 et 4 contredisent le caractère 11 : pour ce caractère, l'hypothèse de synapomorphie de (A,C) est une erreur. L'opération de mise en congruence des phylogénies de caractères peut être extrêmement laborieuse si les taxons sont nombreux et si les contradictions sont nombreuses (forte homoplasie ou « bruit »). Grâce à l'usage d'algorithmes, l'outil informatique permet de résoudre, autant que faire se peut, ces situations délicates.

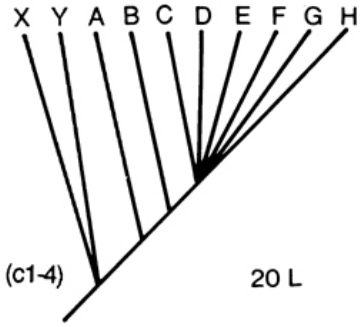
FIGURE IV.20. *(pages suivantes) Analyse cladistique de 8 taxons (A-H), 2 extra-groupes (X,Y) et 11 caractères, à partir du tableau IV.1. A-I : cladogrammes obtenus pour chacun des caractères 1 à 11 (20A : caractères 1 ; 20B : caractères 2 et 3 ; 20C : caractère 4 ; 20D : caractère 5 ; 20E : caractère 6 ; 20G : caractères 8 et 9 ; 20H : caractère 10 ; 20I : caractère 11). 20J à 20S : combinaisons des cladogrammes 20A à 20I (20J : caractères 1 et 2, 20K : caractères 1 à 3, 20L : caractères 1 à 4, 20M : caractères 1 à 5, 20N : caractères 1 à 6, 20O : caractères 1 à 7, 20P : caractères 1 à 9, 20Q : caractères 1à 10, 20R : caractère 1 à 11 (12 pas), 20S : caractère 1 à 11 (14 pas)).*

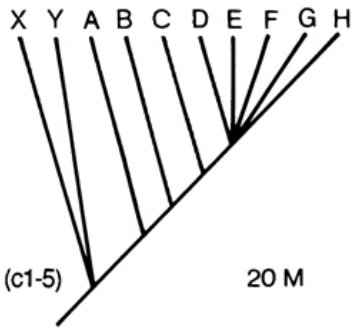

X Y A B C D E F G H
(c1)
20 A
X Y A B C D E F G H
(c2 et c3)
20 B
X Y A B C D E F G H
(c4)
20 C
X Y A B C D E F G H
(c5)
20 D
X Y A B C D E F G H
(c6)
20 E
X Y A B C D G H E F
(c7)
20 F
X Y A B C D E F G H
(c8 et c9)
20 G
X Y A B C D E F G H
(c10)
20 H

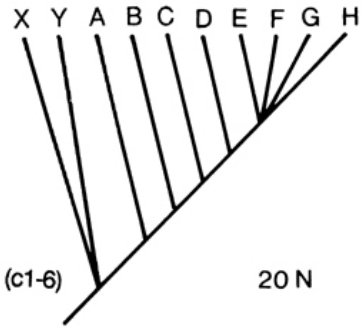

X Y B D E F G H A C
(c11)
20 I

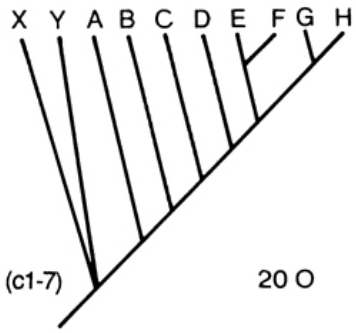

X Y A B C D E F G H
(c1-2)
20 J

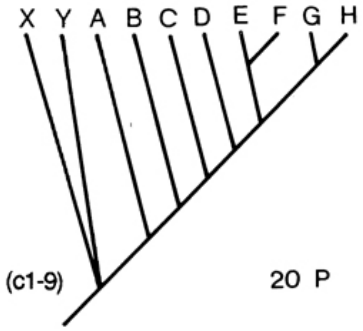

X Y A B C D E F G H
(c1-3)
20 K

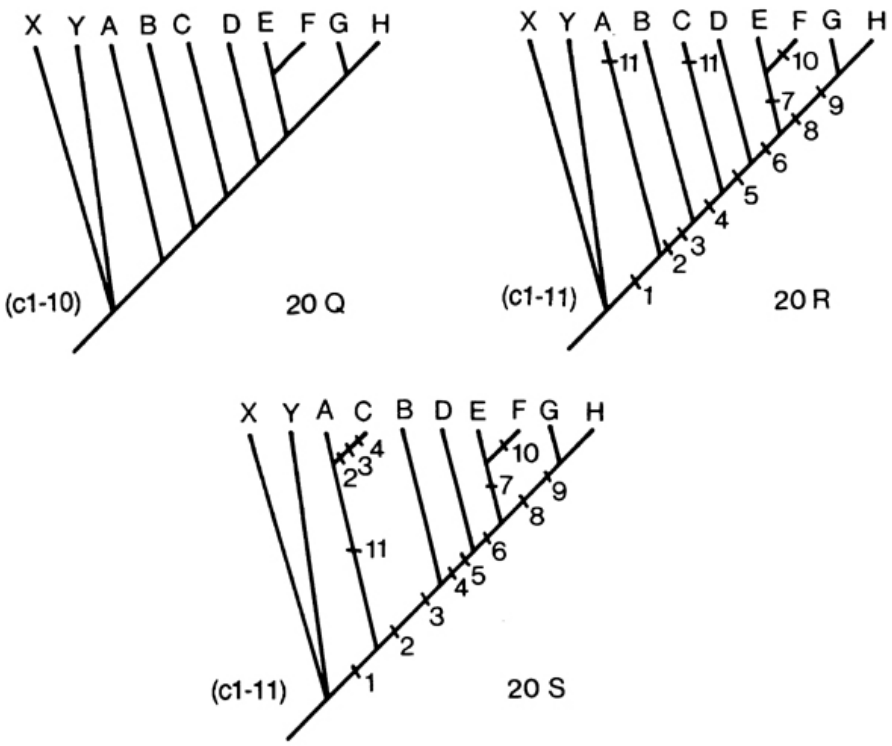

X Y A B C D E F G H
(c1-4)
20 L


X Y A B C D E F G H
(c1-5)
20 M


X Y A B C D E F G H
(c1-6)
20 N


X Y A B C D E F G H
(c1-7)
20 O


X Y A B C D E F G H
(c1-9)
20 P


X Y A B C D E F G H
(c1-10)
20 Q
X Y A B C D E F G H
11
11
10
7
9
8
6
5
4
3
2
1
(c1-11)
20 R
X Y A C B D E F G H
4
3
2
10
7
9
8
11
6
5
4
3
2
1
(c1-11)
20 S

CHAPITRE **V**

LES PROCÉDURES DE PARCIMONIE

1. La recherche de l'arbre le plus court

La recherche du sens de l'évolution des caractères, de l'état primitif vers l'état dérivé revient à exprimer la similitude sous une forme binaire : plésiomorphe / apomorphe, barre blanche / barre noire, 0 – 1 etc. Une telle approche de la ressemblance se prête donc particulièrement à un traitement informatique. Des algorithmes de recherche de l'arbre le plus court, le plus parcimonieux, ont été conçus depuis une vingtaine d'années afin de résoudre les problèmes complexes de construction phylogénétique. Tous sont fondés sur le principe de parcimonie. Certains logiciels y font explicitement référence tel PAUP, dû à D. Swofford, abréviation de *Phylogenetic Analysis Using Parsimony*. Tous reposent sur l'algorithme dit de Wagner conçu par Farris (Kluge et Farris, 1969 ; Farris, 1970). Mais, depuis, de nombreux autres algorithmes plus performants ont été découverts, dont certains sont exacts et donnent avec certitude l'arbre le plus court, tandis que d'autres sont heuristiques.

Indépendamment de la systématique phylogénétique de Hennig, dès 1963, Edwards et Cavalli-Sforza ont invoqué explicitement le principe de parcimonie à propos de génétique des populations : l'estimation la plus plausible d'un arbre évolutif est celle qui fait appel à la quantité minimale d'évolution. Le problème immédiat qui nous occupe ici est celui de préciser et quantifier cette « quantité minimale d'évolution ».

Dans les années soixante surgirent des méthodes se réclamant explicitement ou implicitement du principe de parcimonie tant pour des analyses de distances (voir chapitre VI) (Cavalli-Sforza et Edwards, 1967 ; Fitch et Margoliash, 1967) que pour des analyses cladistiques (Camin et Sokal, 1965 ; Kluge et Farris, 1969). Ce sont ces dernières qui nous intéressent ici. L'usage fréquent dans la littérature d'expressions telles « arbre minimal », « arbre le plus court », « arbre le plus parcimonieux », a trait à des arbres construits selon des méthodes cladistiques aussi bien que phénétiques. Aussi ne sera-t-il question dans ce paragraphe que de parcimonie au sens cladistique, celle qui permet de construire un arbre phylogénétique minimal (comptant le minimum de transformations) par addition

des phylogénies de caractères, selon des algorithmes (paragraphe 1.2) qui réalisent la procédure illustrée par la figure IV.20, quoique pas nécessairement à l'identique dans son déroulement.

1.1. Modèles de parcimonie

On distingue trois types de parcimonie, selon que l'on impose ou non des contraintes sur les transformations de caractères et selon la nature de ces contraintes. Ces dernières influent sur la topologie, la longueur des branches et la longueur totale des arbres.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
A	0	1	0	0	0	0	0	0	1
B	1	0	1	1	1	0	0	0	0
C	1	0	0	1	1	1	1	1	0
D	1	1	1	1	1	1	1	1	1
E	0	0	0	1	1	1	1	1	1
X	0	0	0	0	0	0	0	0	0

TABLEAU V.1. *Matrice de 9 caractères pour 6 taxons (A-E, X étant l'extra-groupe). Chaque caractère est représenté par deux états codés 0 et 1.*

1.1.1. Parcimonie de Wagner

Dans ce modèle (Kluge et Farris, 1969 ; Farris, 1970) convergences et réversions sont acceptées *a priori* (0→1 *et* 1→0).

La figure V.1A illustre une analyse de parcimonie de type Wagner : c'est l'analyse de parcimonie sans contrainte imposée au mode de transformation des caractères. Pour la simplicité de l'exposé on admet que A, B, C, D et E forment ensemble un taxon monophylétique.

Le cladogramme (13 pas) implique que le caractère 1 se transforme deux fois (homoplasie) : une fois chez l'ancêtre de (B,C,D,E) : (0→1) et une fois chez E (réversion 1→0). Les caractères 2, 3 et 9 apparaissent chacun deux fois (convergences). La combinaison (D,E) est due à un caractère homoplasique (caractère 9) qui est à la fois synapomorphie de groupe (D,E) et apomorphie du taxon A.

1.1.2. Parcimonie de Camin-Sokal

Ce modèle (Camin et Sokal, 1965) n'autorise que les convergences (0→1 *ou* 1→0 selon l'état ancestral). Les réversions sont exclues. La connaissance *a priori* de l'état ancestral est donc nécessaire. Une telle analyse à partir du tableau V.1 interprète le caractère 1 différemment : l'état dérivé est apparu indépendamment chez B, chez C et chez D, autrement dit trois fois. L'arbre a la même configuration que dans le cas de l'analyse de Wagner mais compte 14 pas, soit un pas de plus

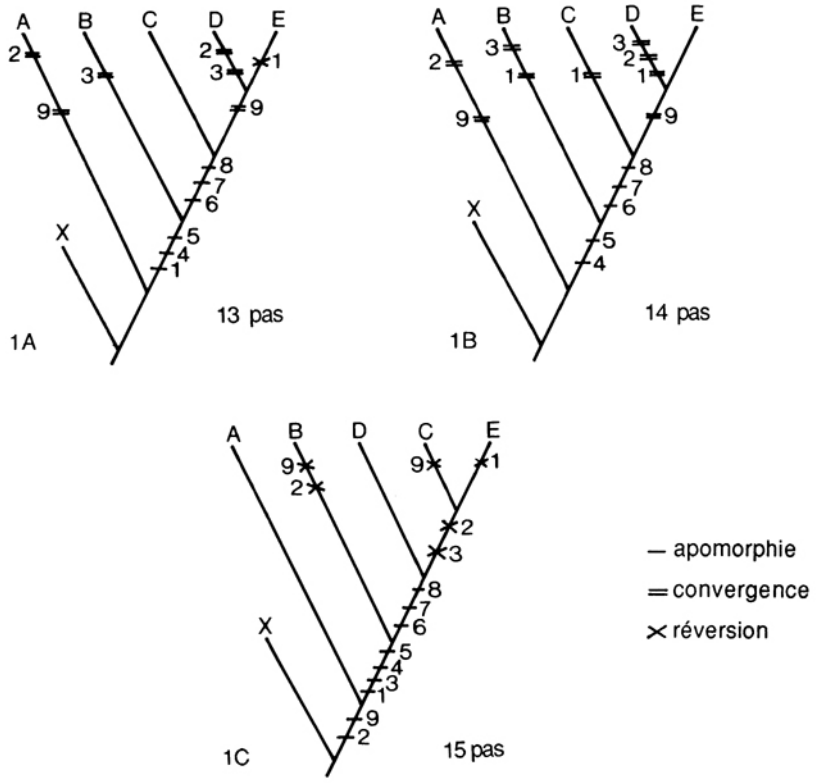


FIGURE V.1. *Cladogrammes construits à partir du tableau V.1. A : selon le modèle de parcimonie dit de Wagner (convergences et réversions admises) : 13 pas. B : selon le modèle de parcimonie dit de Camin-Sokal (réversion non admise) : 14 pas. C : selon le modèle de parcimonie dit de Dollo (convergence non admise) : 15 pas. Dans tous les cas de figure, on admet que le groupe (A,B,C,D,E) est monophylétique.*

(figure V.1B). Mais, compte tenu de la contrainte imposée (réversion interdite), c'est l'arbre le plus court : si l'on contraignait le groupe (B,C,D) à être monophylétique, sur la base du caractère 1, l'arbre compterait 16 pas.

1.1.3. Parcimonie de Dollo

Ce modèle (Le Quesne, 1972 ; Farris 1977a) n'accepte que les réversions et exclut les convergences. C'est l'une des applications du concept de « caractère dérivé unique » de Le Quesne (1972) (l'autre ressortit à l'analyse de compatibilité, voir chapitre VI). Selon Le Quesne, il est plus facile de perdre un caractère (réversion : retour apparent à l'état initial) que d'acquérir en parallèle un même caractère (convergence). L'expression « parcimonie de Dollo » est trompeuse. La « loi de Dollo » (du nom du paléontologue belge Louis Dollo) implique au contraire que le retour à l'état ancestral est impossible.

L'analyse appliquée au tableau V.1 donne un arbre différent des précédents, dans lequel le groupe (C,E) est monophylétique. L'arbre compte 15 pas (figure V.1C). Dans cet arbre, quatre caractères sont homoplasiques et soumis à réversion (1, 2, 3 et 9). Le seul moyen de construire un arbre sans convergence est de situer les caractères 2 et 9 sous leur état dérivé à la racine de l'arbre ; le caractère 2 est réverse chez B et chez (C,E), tandis que le caractère 9 est réverse chez B et chez C. Les caractères 1 et 3 sont dérivés pour (B(D(C,E))), 1 est réverse chez E et 3 est réverse chez (C,E). les réversions 2 et 3 sont les synapomorphies de C et E.

Comme on peut le constater avec cet exemple, ce modèle accepte des apparitions multiples des mêmes réversions (ici les caractères 2 et 9). Autrement dit, des réversions apparaissent par convergence. Mais, à la différence de la parcimonie de Camin-Sokal, ces convergences impliquent toujours un retour à l'état initial.

1.1.4. Parcimonie, longueur de l'arbre et sens de l'évolution

Les contraintes apportées au comportement des caractères (parcimonie de Camin-Sokal ou de Dollo) aboutissent à des arbres plus longs que celui obtenu lorsque les caractères sont autorisés à évoluer « dans tous les sens ». C'est pourquoi le mode de parcimonie dit de Wagner est considéré comme celui reflétant l'application pure et simple du principe de parcimonie. L'introduction d'options qui excluent les convergences ou les réversions sont le fait d'hypothèses *ad hoc* justifiées par d'autres considérations que la simple mise en congruence des phylogénies de caractères.

Dans le cas de la parcimonie de Wagner, la longueur de l'arbre est indépendante de la position de la racine, même si l'orientation des transformations dépend du choix de cette racine, c'est-à-dire du taxon pris comme ancêtre ou extra-groupe. Dans l'exemple de la figure V.1A, on a choisi X comme extra-groupe ; la lecture de l'arbre à partir de n'importe quel autre taxon choisi comme point de départ (par exemple A) ne changera pas le nombre de pas mais changera le sens des transformations. Autrement dit, l'arbre minimal mesure 13 pas quels que soient le ou les taxons terminaux choisis comme point de départ (voir paragraphe 3 de ce chapitre). Au contraire, l'usage des modèles de parcimonie de Camin-Sokal et de Dollo conduit à définir explicitement l'état initial des caractères.

1.2. Algorithmes exacts et heuristiques

A partir de n taxons terminaux, on peut construire $(2n-3)!/(2^{n-2}(n-2)!)$ arbres dichotomiques (Cavalli-Sforza et Edwards, 1967). Ainsi, pour 10 taxons terminaux il existe 34.459.425 arbres (voir paragraphe I.4). La comparaison d'un si grand nombre d'arbres afin de découvrir l'arbre le plus court est une opération extraordinairement laborieuse. En réalité, la recherche de l'arbre minimal est un problème qui rentre dans le cadre de ce que l'on appelle en algorithmique les problèmes NP-complets (NP pour *non deterministic polynomial*, c'est-à-dire « polynomial non déterminé »). Un algorithme est dit polynomial si son exécution demande un nombre minimum d'opérations, borné par une fonction polynomiale

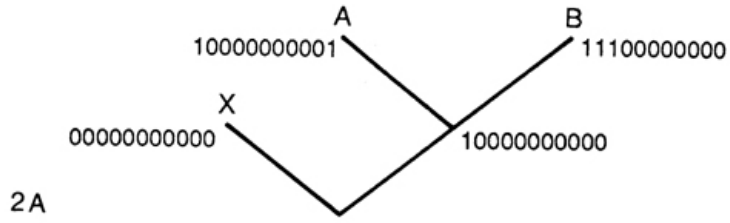
de la taille des données d'entrées. Dans ce cas le problème est traitable. En revanche, pour un problème NP-complet, il n'existe pas d'algorithme polynomial pour le résoudre, mais on ne peut pas démontrer non plus qu'il n'est pas traitable (Barthélemy et Guénoche, 1988 ; d'Udekem-Gevers, 1990).

Si l'on revient une nouvelle fois à la figure IV.20 – qui est pourtant un exemple simple et remarquablement cohérent – on peut saisir facilement ce qui fait de la recherche de l'arbre le plus court un exercice si difficile. Dans cet exemple, la procédure a été découpée en plusieurs étapes, chacune correspondant au cladogramme défini par un caractère. La superposition de tous les cladogrammes de caractères permet de choisir le cladogramme de congruence optimale qui sera le cladogramme des taxons terminaux, c'est-à-dire l'arbre le plus court. Les algorithmes de recherche de l'arbre le plus court agglomèrent les taxons terminaux les uns aux autres de telle façon qu'à chaque insertion d'un taxon supplémentaire, le schéma exprime le nombre de pas minimal nécessaire pour rendre compte de la distribution des caractères (le tableau V.2 et la figure V.2 illustrent le même exemple que la figure IV.20). De la sorte, la longueur minimale finale, dite encore « globale », ne dépend pas que des relations locales entre taxons terminaux, mais de toutes les relations possibles entre les taxons terminaux. C'est pourquoi un nombre élevé de taxons rend très vite l'opération véritablement astronomique.

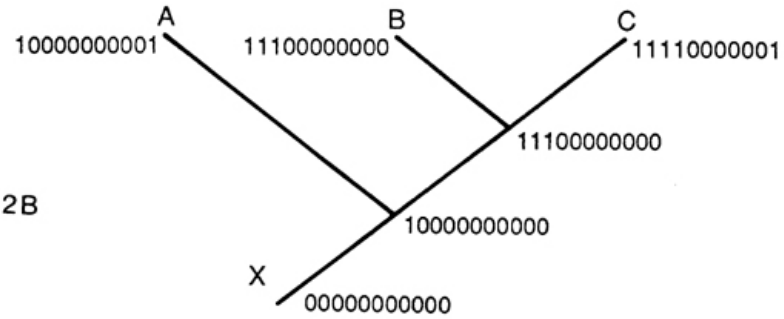
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
A	1	0	0	0	0	0	0	0	0	0	1
B	1	1	1	0	0	0	0	0	0	0	0
C	1	1	1	1	0	0	0	0	0	0	1
D	1	1	1	1	1	0	0	0	0	0	0
E	1	1	1	1	1	1	1	0	0	0	0
F	1	1	1	1	1	1	1	0	0	1	0
G	1	1	1	1	1	1	0	1	1	0	0
H	1	1	1	1	1	1	0	1	1	0	0
X	0	0	0	0	0	0	0	0	0	0	0

TABLEAU V.2. Matrice de 11 caractères pour 9 taxons (A-H et X pris comme extra-groupe). Chaque caractère est représenté par deux états codés 0 et 1

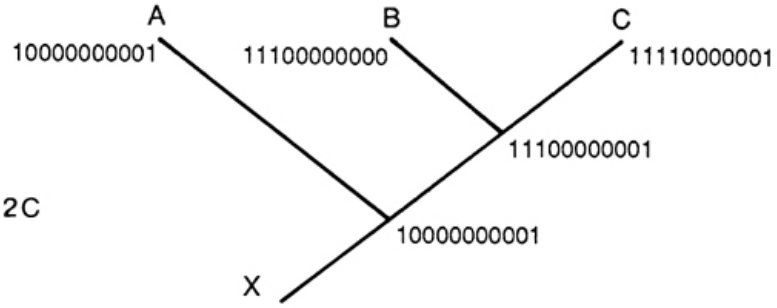
FIGURE V.2. *(Pages suivantes) Procédure de l'analyse de parcimonie construite à partir du tableau V.2. Le modèle de parcimonie est celui dit de Wagner (convergences et réversions admises). Figure 2A: insertion du taxon B sur A en fonction des états de caractères, sachant que X est l'extra-groupe ; figures 2B-M: insertions des taxons C à F. Le cladogramme final (figure 2M) nécessite 12 pas.*


A
B
10000000001
11100000000
X
00000000000
10000000000

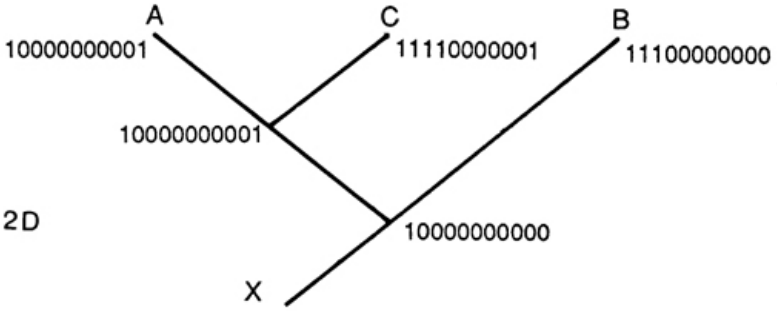
2A


A
B
C
10000000001
11100000000
11110000001
11100000000
10000000000
X
00000000000

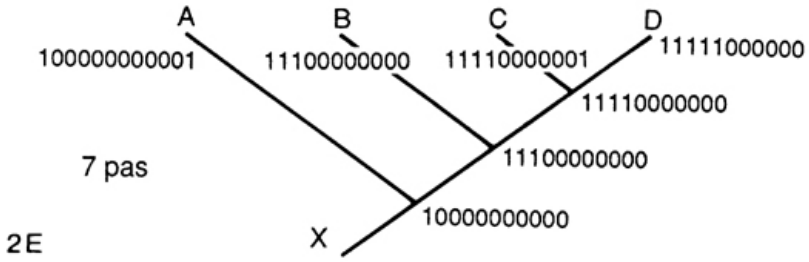
2B


A
B
C
10000000001
11100000000
11110000001
11100000001
10000000001
X

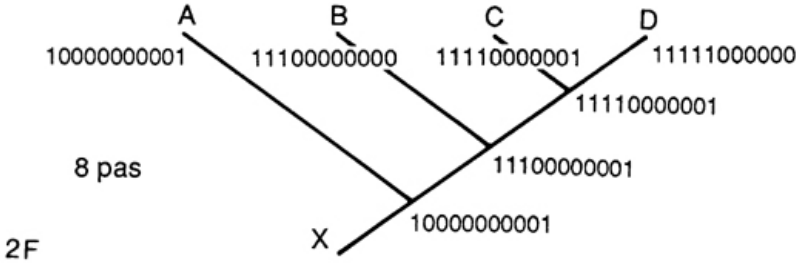
2C


A
C
B
10000000001
11110000001
11100000000
10000000001
10000000000
X

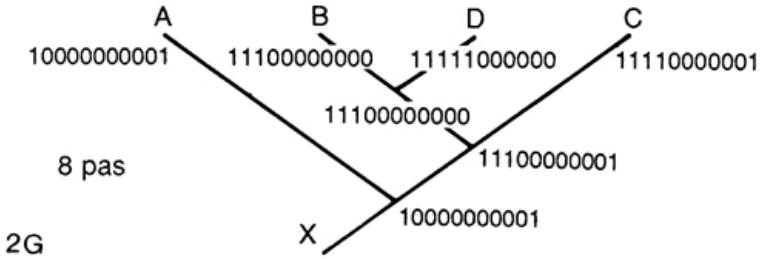
2D


A
B
C
D
100000000001
11100000000
11110000001
11111000000
11110000000
11100000000
7 pas
10000000000
X

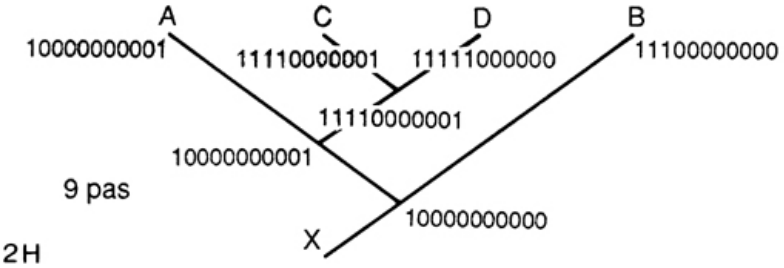
2E


A
B
C
D
10000000001
11100000000
11110000001
11111000000
11110000001
11100000001
8 pas
10000000001
X

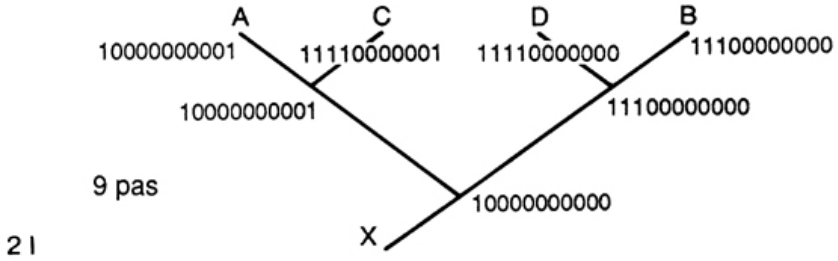
2F


A
B
D
C
10000000001
11100000000
11111000000
11110000001
11100000000
11100000001
8 pas
10000000001
X

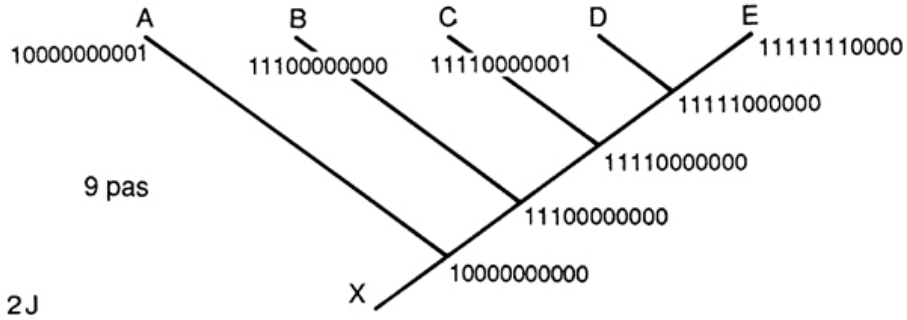
2G


A
C
D
B
10000000001
11110000001
11111000000
11100000000
11110000001
10000000001
9 pas
10000000000
X

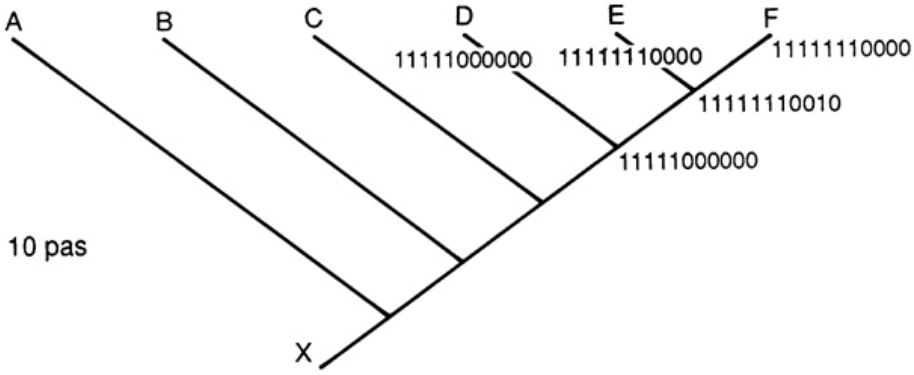
2H


A
C
D
B
10000000001
11110000001
11110000000
11100000000
10000000001
11100000000
9 pas
10000000000
X

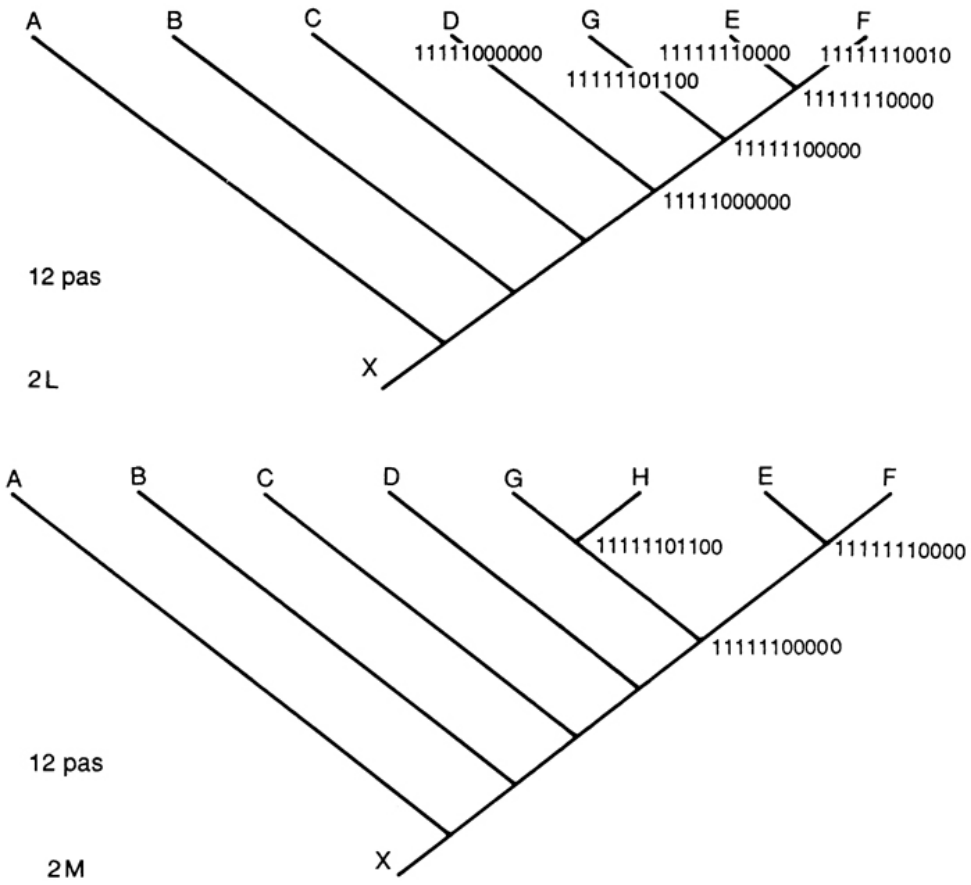
2 I


A
B
C
D
E
10000000001
11100000000
11110000001
11111110000
11111000000
11110000000
9 pas
11100000000
10000000000
X

2J


A
B
C
D
E
F
11111000000
11111110000
11111110000
11111110010
11111000000
10 pas
X

2K



1.2.1 La procédure de parcimonie

La figure V.2 illustre la procédure de l'analyse de parcimonie et correspond à ce que Kluge et Farris (1969) ont appelé la méthode de construction d'arbres dits de Wagner (voir paragraphe suivant). Par rapport à la figure IV.20, on a réduit la comparaison extra-groupe au seul taxon terminal X, afin de simplifier la présentation.

La méthode consiste à agglomérer dans une *première étape* deux taxons terminaux à la racine (l'extra-groupe). Ces deux premiers taxons peuvent être n'importe lesquels parmi ceux étudiés. On obtiendrait un même arbre minimal

final en partant de n'importe quel taxon pris comme extra-groupe. Seule la polarisation des caractères changerait. On a ici choisi les deux premiers figurant dans la matrice : A et B (arbre 2A).

La *deuxième étape* est l'insertion d'un troisième taxon terminal, C (arbre 2B). On voit qu'il est plus économique d'insérer la branche menant à C sur la branche menant à B (arbre 2B et 2C) que sur la branche menant à A (arbre 2D). Une seule topologie parcimonieuse (6 pas) où B et C sont apparentés, est donnée par deux hypothèses différentes de transformations des caractères : selon l'arbre 2B le caractère dérivé 11 (état 1) est une convergence entre A et C ; selon l'arbre 2C le caractère dérivé 11 (état 1) est dérivé pour (A,B,C) et réverse (état 0) chez B. L'arbre 2D, où les taxons A et C sont apparentés, demande 7 pas.

La *troisième étape* est l'insertion d'un quatrième taxon terminal, D. La solution la plus économique (7 pas) est celle où D se branche avec C (arbre 2E), l'ensemble (C,D) se branchant avec B. La solution est obtenue à partir de l'arbre 2B, celle où le caractère 11 est une convergence entre A et C. L'insertion de D a levé l'ambiguïté sur le comportement du caractère 11: la réversion est moins économique (arbre 2F). La proche parenté de (B,D) et de C est également moins économique (arbre 2G), comme sont moins économiques les autres solutions : ((A(C,D))B) et ((A,C)(D,B)) (arbres 2H et 2I) qui comptent toutes un ou deux pas de plus que l'arbre 2E.

Les *4e, 5e, 6e et 7e étapes*, à savoir les insertions des taxons terminaux E, F, G et H, confirment les relations de parenté entre A, B, C et D, illustrées par l'arbre 2E. Comme aucune homoplasie n'affecte les caractères portés par les taxons terminaux E, F, G et H, il n'y a, à chaque étape, qu'une topologie parcimonieuse (arbres 2J-M).

Cet exemple est simple mais il montre la nécessaire mise en mémoire, pas à pas, de toutes les solutions locales possibles, y compris les solutions non parcimonieuses. Si, au cours de l'exécution de l'algorithme, le taxon terminal D avait été inséré, lors de la troisième étape, à partir de l'arbre 2C, le résultat final aurait été erroné en ce sens que l'arbre de longueur minimale n'aurait pas été trouvé (l'arbre 2F compte un pas de plus que l'arbre 2E).

De la même façon, admettons que les taxons E à H portent des caractères tels que l'arbre de longueur minimale ne puisse être obtenu qu'à partir de l'arbre 2D. Si l'algorithme avait éliminé les topologies 2D, 2H et 2I pour des raisons locales (l'insertion de C et D), l'arbre de longueur minimale ne pourrait pas être découvert.

Pour répondre au défi que représente la comparaison simultanée de tous les arbres possibles, des algorithmes exacts et des algorithmes heuristiques ont été élaborés. Les algorithmes exacts garantissent la solution optimale (l'arbre le plus court). Cependant, ils ne sont efficaces (c'est-à-dire qu'ils donnent le résultat en un temps de calcul raisonnable) qu'avec des données restreintes. Les algorithmes heuristiques ne consomment qu'un faible temps de calcul quand les données sont importantes (grand nombre de taxons et de caractères) mais ils ne garantissent pas toujours la découverte de l'arbre minimal.

1.2.2. Méthode, algorithme et arbre de Wagner

Le botaniste W.H. Wagner Jr. conçut, parallèlement à Hennig, une méthode d'analyse des caractères nommée *groundplan divergence analysis* (Wagner, 1961). Cette méthode, ainsi que les algorithmes qu'elle utilise et les arbres qu'elle produit, sont dits «de Wagner» dans la présentation et la formalisation qu'en donnent Kluge et Farris (1969) et Farris (1970). C'est une méthode de construction d'arbre où la distinction entre les états ancestraux et les états dérivés d'un caractère est fondée sur le critère de comparaison extra-groupe. Wagner (1984) signale que sa méthode fut d'abord informatisée par Lellinger en 1965. Ce travail, une thèse non publiée, passa inaperçu, semble-t-il. L'approche de *groundplan divergence* elle-même ne fut relativement bien connue qu'à partir de la publication du travail de Kluge et Farris (1969) qui y faisait explicitement allusion.

L'algorithme de Wagner, tel qu'il a été décrit par Farris et dont on a dit précédemment qu'il était à l'origine de tous les logiciels de parcimonie, n'est utilisé tel quel dans aucun des logiciels actuellement disponibles (Swofford, 1985). Comme les principes de base de cet algorithme restent cependant toujours valables, ils seront brièvement présentés. En revanche, les algorithmes actuellement disponibles, bien que plus performants que l'algorithme de Wagner, ne seront pas décrits, la plupart n'étant pas communiqués par leurs auteurs. On peut trouver néanmoins en français dans d'Udekem-Gevers (1990) une présentation et une analyse de l'algorithme heuristique utilisé par le logiciel MIX de *Phylip (Phylogeny Inference Package* , version 3.1) conçu par J. Felsenstein.

La méthode de Wagner poursuit conjointement un double objectif :

— Connecter les UE entre elles, en construisant un arbre de telle façon que le nombre total de transformations de caractères soit minimal. Cet arbre peut être enraciné (Kluge et Farris, 1969) ou non enraciné (Farris, 1970).

— Etablir les états des caractères aux nœuds de l'arbre. Cette inférence est effectuée en maximisant les synapomorphies et en minimisant les homoplasies. Les nœuds prennent ainsi le statut d'ancêtre, d'« unité taxinomique hypothétique » (ou « unité évolutive hypothétique ») auxquelles sont attribuées des informations de même nature que celles qui définissent les UE.

L'algorithme présenté ici est celui proposé par Farris (1970) pour des arbres non enracinés :

— Soit x_{Ah} l'état du caractère h chez l'UE A. Les états d'un caractère sont mesurés sur une échelle d'intervalle de telle façon que la différence entre deux états soit un nombre entier qui corresponde au nombre de transformations nécessaires pour passer d'un état à l'autre. Cela peut s'appliquer aussi bien à des états binaires (x_{Ah} = 0 ou 1) qu'à des états multiples, additifs ou non (Farris, 1970 ; Fitch, 1971).

— La différence entre deux UE A et B est définie par la distance Manhattan (voir Chapitre VII), établie à partir de K caractères :

$$d_{AB} = \sum_{h=1}^{K} | x_{Ah} - x_{Bh} |$$

Cette différence représente exactement le nombre de pas nécessaires pour relier l'UE A à l'UE B et réciproquement.

La procédure d'agglomération est la suivante (figure V.3) :

1) On choisit de connecter les UE entre lesquelles cette différence est maximale : A et B par exemple.

2) Une autre UE est ensuite connectée sur la branche AB en un nœud Y. Le choix de cette UE se fait sur le critère suivant : il s'agit de l'UE telle que la différence entre cette UE (C par exemple) et le nœud Y soit maximale. Cette différence est estimée par :

$$d_{CY} = \frac{1}{2}(d_{AC} + d_{BC} - d_{AB})$$

3) Les états des caractères de l'unité évolutive hypothétique Y seront définis par la règle suivante. On attribue les états des caractères aux nœuds en leur donnant la valeur médiane des états des caractères des 3 UE qui l'entourent :

$$x_{Yh} = médiane\ (x_{Ah}\ ,\ x_{Bh}\ ,\ x_{Ch}\ )$$

C'est la règle qui s'impose puisque c'est elle qui assure le minimum de transformations entre A, B, C et Y.

4) Puisque les états des caractères sont maintenant connus pour Y, il est possible de poursuivre le processus d'agglomération en choisissant à la fois une nouvelle UE et la branche sur laquelle l'insérer, en s'aidant du même critère que celui défini en 2).

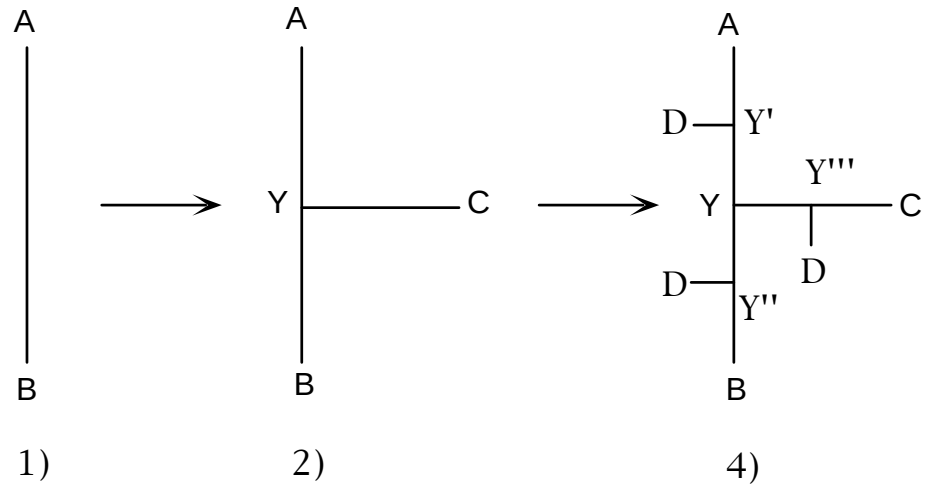


FIGURE V.3. *Insertion d'une nouvelle UE (ici C) et d'un intermédiaire (Y) sur une branche AB déjà constituée. L'étape suivante consiste à insérer une autre UE (D) sur l'une des branches, ici en Y', Y'' ou Y'''.*

Par exemple on retiendra l'UE D et le point d'insertion Y''' si la différence:

$$d_{DY'''} = \frac{1}{2}\left(d_{CD} + d_{YD} - d_{YC}\right)$$

est la différence maximale parmi toutes les différences envisageables ($d_{DY'}$, $d_{DY''}$, $d_{DY'''}$).

Notons que cette différence se calcule également en fonction des différences entre UE (et non en fonction des différences entre nœuds et UE) puisque :

$$d_{YD} = \frac{1}{2}\left(d_{AD} + d_{BD} - d_{AB}\right)$$

$$d_{YC} = \frac{1}{2}\left(d_{AC} + d_{BC} - d_{AB}\right)$$

On a donc :

$$d_{DY'''} = \frac{1}{2}\left[d_{CD} + \frac{1}{2}\left(d_{AD} + d_{BD}\right) - \frac{1}{2}\left(d_{AC} + d_{BC}\right)\right]$$

5) Le processus d'agglomération se poursuit ainsi jusqu'à ce que toutes les UE soient agglomérées et que les états des caractères soient inférés aux nœuds.

Cette méthode heuristique conduit à un arbre final qui n'est pas nécessairement l'arbre de longueur minimal, même s'il tend à s'en rapprocher. Il est donc utile de la faire suivre de procédures supplémentaires d'optimisation, par exemple en changeant le point de départ de l'agglomération ou en effectuant des réarrangements des branches ou *branch swapping* (paragraphe V.1.2.4).

La distribution des états des caractères aux nœuds est également optimisée de façon à ce que les états impliquent, à nombre égal d'événements, un maximum de synapomorphies et un minimum d'homoplasies.

L'algorithme tel qu'il a été décrit ici ne suppose pas l'existence d'un ancêtre ou l'existence d'une quelconque polarité dans le sens des transformations. Cela peut cependant être fait : il suffit d'introduire l'ancêtre, de le considérer comme une UE et de démarrer l'agglomération à partir de lui. L'algorithme initialement décrit par Kluge et Farris (1969) construit un arbre enraciné en connectant les UE répondant au critère de la plus petite différence, et non de la plus grande comme dans le cas d'un arbre non enraciné (Farris, 1970).

L'étape 3) d'attribution des états des caractères aux UE hypothétiques fonde l'analogie entre la parcimonie de Wagner et l'approche cladistique. Le calcul des longueurs de branches est lié à l'estimation des états des caractères aux nœuds (paragraphe V.1.3). Les méthodes d'agglomération décrites dans le chapitre sur les méthodes phénétiques de construction d'arbres (chapitre VII) présentent quelques analogies avec celles que l'on vient de décrire mais s'en distinguent notamment par l'absence de cette étape d'estimation des états ancestraux. En fait c'est cette étape qui constitue toute la différence entre procédures de parcimonie et analyses de distances.

1.2.3. Algorithmes exacts

Les algorithmes exacts garantissent la solution optimale. Ils le font par analyse exhaustive ou par la technique du *branch and bound*.

La recherche exhaustive est l'évaluation de tous les arbres possibles. Comme il y a plus de $2x10^6$ arbres pour 9 taxons, une telle recherche n'est généralement possible qu'au-dessous de la dizaine de taxons. La recherche exhaustive correspond à la procédure décrite dans le paragraphe 4 du premier chapitre (figure I.8).

La technique du *branch and bound* (Hendy et Penny, 1982) est un algorithme exact qui ne nécessite pas une recherche exhaustive. Il donne la solution optimale pour un nombre de taxons dépassant la dizaine (jusqu'à vingt à trente taxons) selon la qualité de l'implémentation de l'algorithme et la cohérence des données : plus l'homoplasie est importante, plus le temps de calcul est élevé. Sa description détaillée est donnée par Hendy et Penny (1982). La présentation résumée qui est faite ici (figure V.4) est tirée de celle de Swofford et Olsen (1990).

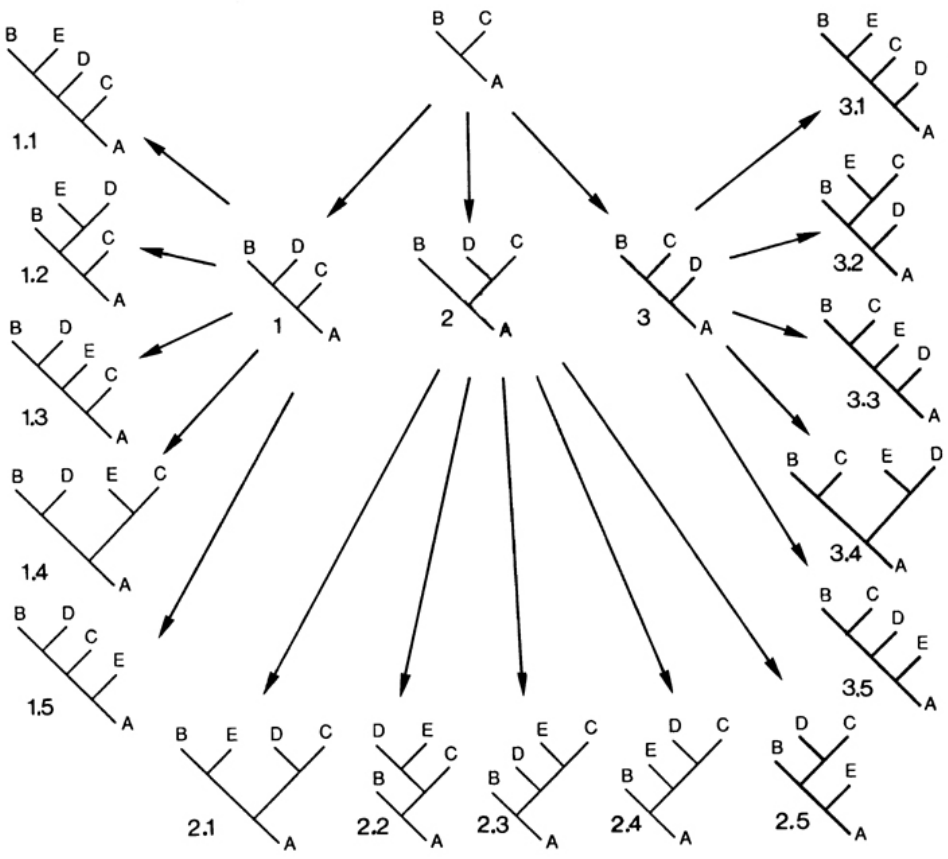


FIGURE V.4. *Description simplifiée de la technique dite du* branch and bound *(d'après Swofford et Olsen, 1990).*

L'originalité de la technique du *branch and bound* est que la recherche exhaustive est contrôlée en référence à un arbre donné, éventuellement pris au hasard, ou calculé par l'algorithme de Wagner ou tout autre algorithme heuristique, arbre dont on calcule le nombre de pas (soit L). Puisqu'un tel arbre existe, l'arbre minimal ne pourra excéder la longueur de cet arbre de référence.

Le point de départ de la recherche est un arbre qui est le seul arbre possible pour les trois premiers taxons A, B et C. On construit ensuite l'un des trois arbres possibles obtenu en insérant le quatrième taxon (D) : arbre 1 de la figure V.4. Puis, on insère sur cet arbre le cinquième taxon – ici le dernier – (E) donnant l'arbre 1.1. Ensuite on retourne à l'étape précédente (arbre 1) et on construit un second arbre qui résulte d'une insertion différente du taxon E sur l'arbre 1 (arbre 1.2).

Quand tous les arbres possibles ont été construits par l'insertion de E sur l'arbre 1 (arbres 1.1 à 1.5), on remonte à l'arbre de départ et on insère le taxon D selon l'arbre 2. Puis l'on construit les cinq arbres possibles par insertion de E sur cet arbre 2 (arbres 2.1 à 2.5). De nouveau on retourne vers l'arbre initial et l'on insère le taxon D selon l'arbre 3 à partir duquel on recommence l'opération d'insertion du taxon E (arbres 3.1 à 3.5). De la sorte ont été construites toutes les topologies possibles (recherche exhaustive).

Lorsqu'on se déplace le long des différents chemins possibles issus de l'arbre initial et décrits ci-dessus, si l'on rencontre un arbre plus long que L, alors on ne progressera pas plus dans ce chemin : on le quitte pour en explorer un autre. Si l'on rencontre un arbre aussi long que L, l'arbre est un possible arbre optimal. Si l'on rencontre un arbre plus court, cet arbre est le meilleur obtenu et devient la nouvelle référence. Si l'on rencontre rapidement un arbre nettement plus court

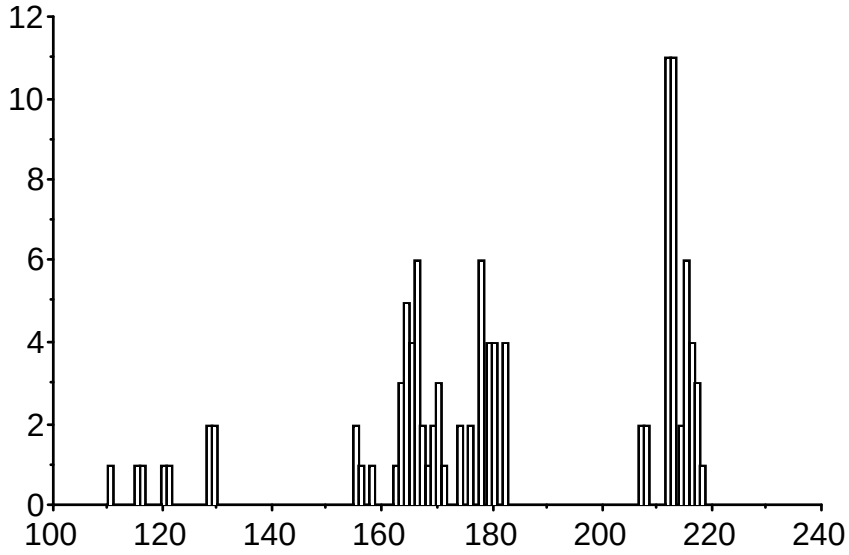


FIGURE V.5. *Histogramme des longueurs de tous les arbres possibles construits à partir de la matrice de la figure III.2 où le macaque et l'atèle sont les extra-groupes. L'arbre minimal fait 110 pas.*

que l'arbre de référence, cela permet de terminer d'autant plus rapidement l'examen des autres chemins. Quand l'ensemble des chemins a été exploré, tous les arbres de longueur minimale – s'il en existe plusieurs – ont été identifiés.

La technique du *branch and bound* est une solution élégante et efficace au problème NP-complet, puisqu'elle garantit la découverte de l'arbre de longueur minimale. Swofford et Olsen (1990) justifient néanmoins l'usage de la recherche exhaustive quand c'est possible, dans la mesure où elle seule permet de connaître le nombre d'arbres plus ou moins proches de l'arbre minimal, ou bien la position d'un arbre de topologie donnée par rapport à l'arbre minimal.

La figure V.5 montre un exemple de distribution des longueurs d'arbres au-delà d'un arbre minimal. Ces arbres ont été construits à partir des données de la matrice de la figure III.2. Seuls les arbres ayant le macaque et l'atèle comme extra-groupes ont été pris en considération. Au-delà de l'arbre le plus parcimonieux (110 pas), les deux arbres les plus proches ont 115 et 116 pas. Les arbres suivants présentent au moins 10 pas de plus que l'arbre le plus parcimonieux.

1.2.4. Algorithmes heuristiques

Lorsque la matrice des données est trop importante pour l'usage d'algorithmes exacts (nombre élevé de taxons et de caractères), des algorithmes heuristiques permettent d'obtenir un résultat en un temps de calcul raisonnable. Mais la découverte de l'arbre optimal n'est pas toujours garantie. Les algorithmes dits d'addition pas-à-pas (*stepwise addition*) sont comparables à la procédure illustrée par la figure V.2. Le résultat est sensible à l'ordre d'introduction des taxons terminaux.

Pour pallier – autant que faire se peut – cette difficulté, il existe plusieurs stratégies heuristiques dites de « réarrangement des branches » (*branch swapping*). L'amélioration d'un arbre initial est opérée par déplacement des branches. Si un réarrangement donne un arbre plus court, ce dernier devient le sujet d'un nouveau réarrangement. A force de réarrangements, l'arbre minimal peut être trouvé. Les options de balayage des branches sont locales ou globales, cette dernière étant plus coûteuse en temps de calcul :

— Le réarrangement local est expliqué par la figure V.6. Il s'agit d'un échange du voisin le plus proche entre 4 taxons X, Y, Z et W. Les taxons X et Y sont permutés selon les trois arbres possibles.

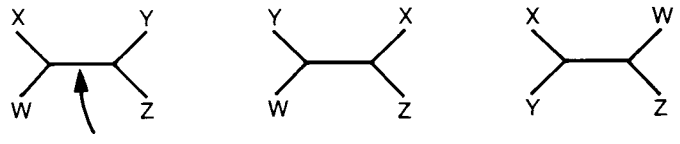


FIGURE V.6. *Réarrangement local. La branche inter-nœuds (désignée par la flèche) définit une configuration locale de trois arbres . Les réarrangements possibles déplacent X vers W, Y vers X ou associent X et Y.*

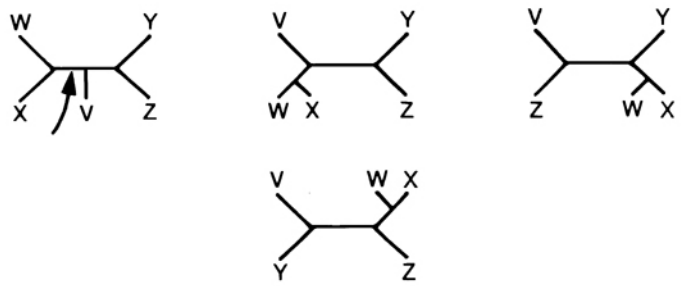


FIGURE V.7. *Réarrangement global. L'ensemble (W,X) est déplacé et connecté à toutes les branches de l'arbre .*

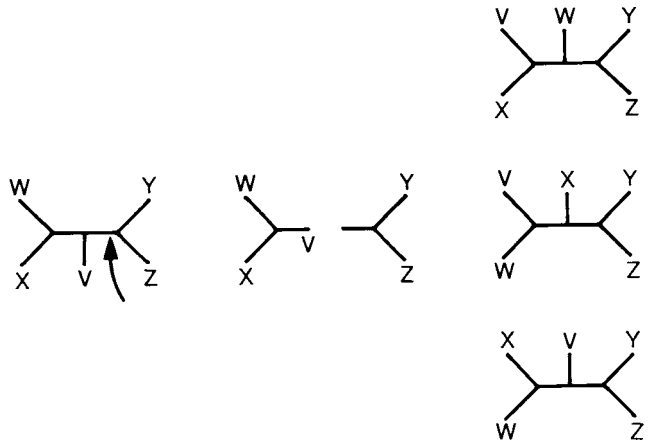


FIGURE V.8. *Réarrangement par bisection et reconnexion. Le groupe (V,W,X) est déplacé puis connecté de telle façon que chaque branche soit connectée sur les branches de l'autre sous-arbre (Y,Z).*

— Le réarrangement global est expliqué par la figure V.7. Il est tel que chaque sous-arbre possible, ici le groupe (W,X), est retiré de l'arbre puis réinséré à toutes les autres positions possibles (Swofford, 1985).

— Le réarrangement par bisection et reconnexion (*tree bisection and reconnection* de Swofford et Olsen, 1990) est expliqué par la figure V.8. L'arbre est découpé en sous-arbres. Chaque sous-arbre est connecté successivement par chacune de ses branches aux autres branches de l'arbre.

L'expérience montre que les algorithmes de réarrangement des branches fonctionnent bien, quoique ne donnant pas toujours tous les arbres de longueur minimale lorsqu'il en existe plusieurs. La difficulté est du même ordre que celle

rencontrée dans l'addition pas-à-pas : si l'obtention d'un arbre minimal par suite de réarrangement nécessite le balayage d'un ancien arbre qui était plus coûteux en pas (et éliminé pour cette raison), l'arbre optimal n'est pas trouvé. Pour éviter cet écueil, il faut que les réarrangements s'appliquent également, au cours de la procédure, aux arbres non parcimonieux.

1.3. Longueur de l'arbre, longueur des branches et optimisation des caractères

L'arbre retenu à l'issue d'une analyse de parcimonie est l'arbre de longueur minimale L, L étant le nombre total de transformations (pas). Les transformations sont distribuées sur les branches internes (branches inter-nœuds) et sur les branches terminales. Les transformations distribuées sur les branches internes sont les synapomorphies des deux groupes frères issus du nœud de rang le plus haut. Sur la figure V.9 les branches ont des longueurs inégales. L'explication cladistique est que les apomorphies ne sont pas distribuées de façon égale sur les branches de l'arbre. La longueur de la branche menant du nœud 2 au nœud 4 correspond au nombre de synapomorphies de (A,B). Les transformations distribuées sur les branches terminales sont les autapomorphies du taxon terminal issu du nœud immédiatement ancestral. La longueur de la branche menant du nœud 2 au taxon terminal C correspond aux autapomorphies de C. Elles sont plus nombreuses que les synapomorphies de (A,B).

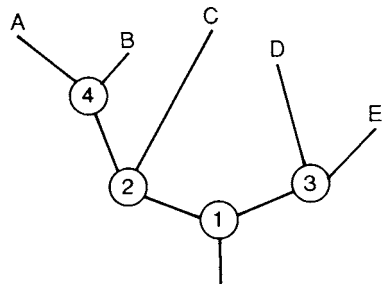


FIGURE V.9. *Arbre où les longueurs des branches correspondent à la quantité de transformations des caractères.*

1.3.1. Localisation des homoplasies

A l'issue d'une analyse de parcimonie il n'est pas rare d'obtenir plusieurs arbres minimaux ayant des configurations différentes. L'homoplasie est responsable d'un tel résultat.

On a vu aussi dans un exemple précédent (figure V.2B-2C) qu'une même configuration parcimonieuse peut être obtenue avec des hypothèses différentes de transformations des caractères. Dans ce cas également l'homoplasie est responsable de cette situation.

CARACTERES :	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
TAXONS																				
A	1	1	1	0	0	0	0	0	0	1	1	1	1	1	1	1	1	0	0	0
B	1	1	1	0	0	0	0	0	0	1	1	1	1	0	0	0	0	1	0	0
C	1	1	1	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	1	1
D	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
E	1	0	0	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0
ANC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

TABLEAU V.3. *Distribution de 20 caractères chez 5 taxons terminaux (A-E) et un ancêtre (anc).*

Dans l'exemple du tableau V.3 et de la figure V.10, le même arbre de longueur minimale (23 pas) dont la topologie est (((A,B)C)(D,E)), correspond à deux des histoires possibles des trois caractères homoplasiques (14, 15 et 16). Ces deux arbres 10A et 10B ne se distinguent que par les longueurs des branches affectées par les transformations de ces caractères, c'est-à-dire les branches reliant les taxons A, B et C et les branches terminales des taxons A et C.

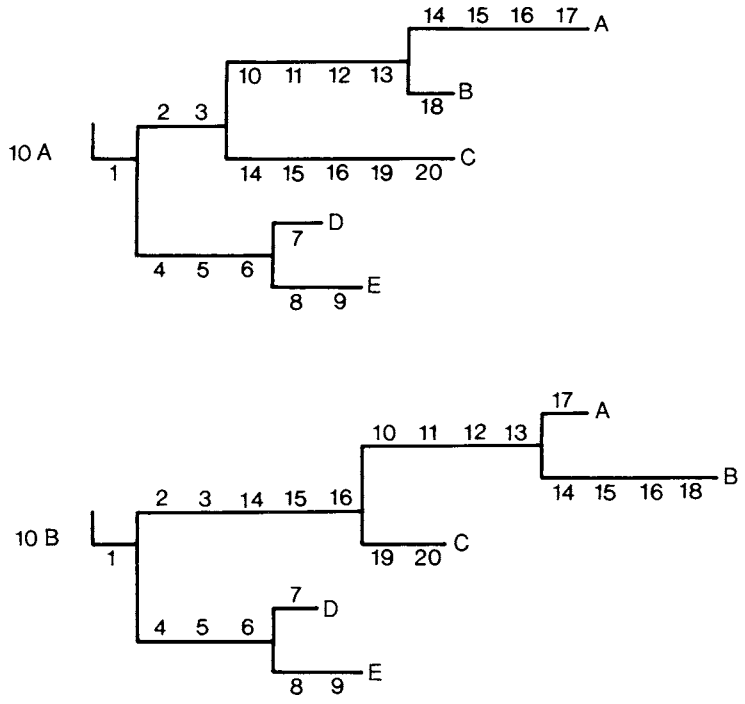


FIGURE V.10. *Les deux arbres parcimonieux (23 pas) de même configuration issus du tableau V.3. Les deux arbres ne diffèrent que par la distribution des caractères 14, 15 et 16 : convergents chez A et chez C (arbre 10A) ou apomorphes pour (A,B,C) puis réverses chez B (arbre 10B). Dans ces deux arbres, les longueurs des branches affectant le groupe ((A,B)C) sont donc différentes.*

Dans une situation telle que celle illustrée par la figure V.10, il est nécessaire d'optimiser la distribution des homoplasies sur les branches si l'on ne s'intéresse pas qu'aux relations de parenté, mais aussi et surtout à l'histoire des événements évolutifs. Ces événements sont les transformations de caractères qui correspondent aux apomorphies distribuées sur l'arbre. Leur nombre représente la quantité d'évolution affectant les différents segments de l'arbre.

Chacun des caractères 14, 15 et 16 se transforme deux fois. Pour chacun d'eux, il peut se produire deux apparitions indépendantes (deux fois 0➞1) (figure V.10A), ou bien une apparition suivie d'une réversion (0➞1➞0 ; figure V.10B). Ces deux cas de figure sont aussi parcimonieux l'un que l'autre.

Dans toutes les situations où un même arbre, ou une portion d'arbre, est compatible avec des évolutions de caractères différentes, on peut choisir systématiquement trois options :

1) privilégier les convergences (figure V.10A) : les traits 14, 15 et 16 sont sur la branche A et sur la branche C. C'est l'option dite *delayed transformation* (« deltran ») (Swofford, 1985).

2) privilégier les réversions (figure V.10B) : les traits 14, 15 et 16 se transforment sur la branche menant à (A(B,C)) (0➞1) puis, par réversion sur la branche B (1➞0). C'est l'option dite « de Farris » ou *accelerated transformation* (« acctran ») (Swofford, 1985). On peut justifier l'optimisation dite de Farris ou « acctran » en soulignant qu'elle renforce le nombre des caractères dus à la descendance puisqu'elle ajoute des caractères à l'ancêtre hypothétique de A, B et C.

3) déplacer les transformations vers les branches terminales, autrement dit maximiser les autapomorphies. C'est l'option minimisant l'indice f de Farris (Farris, 1972) dite « minf » (Swofford, 1985). Ici l'optimisation « minf » donne le même résultat que l'optimisation « deltran » (figure V.10A) : les caractères 14, 15 et 16 sont rejetés vers les branches terminales (autapomorphies respectives de A et de C), plutôt que de figurer à la base du groupe (A,B,C) comme dans la figure V.10B). Ce dernier choix est fondé sur l'idée de ne pas alourdir plus qu'il n'est nécessaire des branches internes, c'est-à-dire les attributs des ancêtres reconstruits : c'est l'option inverse de l'option dite « de Farris » ou « acctran ».

On voit que ces optimisations, sans effet sur la topologie et la longueur globale de l'arbre influent fortement sur les calculs des quantités d'évolution de certaine portions. Sur la figure V.10A, le taxon A a plus évolué (4 pas) que son groupe frère B (1 pas). C'est l'inverse sur la figure V.10B. Le taxon C a plus évolué sur la figure V.10A (5 pas) que sur la figure V.10B (2 pas). A partir de la racine, le groupe ((A,B)C) diverge peu sur la figure V.10A (2 pas), beaucoup plus sur la figure V.10B (5 pas). Les taxons les plus éloignés sont séparés par 15 pas sur la figure V.10A (taxons A et E) et par 18 pas sur la figure V.10B (taxons B et E). Plus qualitativement (contextes biogéographique et écologique, scénarios adaptatifs), l'histoire évolutive du taxon C n'est pas la même s'il a hérité les caractères 14, 15 et 16 de l'ancêtre qu'il partage avec A et B, ou s'il a acquis ces mêmes caractères, indépendamment de A.

On a considéré ici que les caractères 14, 15 et 16 partageaient la même histoire. Mais, pour des raisons *ad hoc*, l'histoire de chacun de ces caractères peut être dissociée des autres et donner autant de combinaisons (ici 8) affectant les longueurs des branches impliquées dans l'origine et la différenciation de A, B et C. Autrement dit, dans les cas tels que ceux illustrés par la figure V.10, les estimations précises des quantités d'évolution fondées sur la transformation des caractères observés obéissent à des modèles évolutifs extérieurs à l'analyse de parcimonie.

2. Les caractères : codage, optimisation, pondération

2.1. Caractères binaires et états multiples

Les caractères, au sens que l'on a donné à ce terme dans le chapitre III, sont codés de telle façon qu'ils puissent donner lieu à des analyses comparatives. La spécification des états plésiomorphe et apomorphe passe par un codage en deux états : 0 — 1, ou a — b, etc., dit codage binaire. L'un des deux états est nécessairement plésiomorphe, l'autre apomorphe. Des états multiples peuvent aussi être codés, tel : 0 — 1 — 2, a — b — c, etc. Quand un caractère est exprimé en états multiples, il renferme déjà une hypothèse phylogénétique qui est celle des relations existant entre ces états.

2.1.1. Caractères binaires

Les caractères binaires codés sous la forme 0 —1 ou a — b etc., n'indiquent pas *a priori* une orientation particulière du morphocline. Dans ce cas, les transformations a➞b *ou* b➞a sont toutes deux également possibles et elles comptent chacune pour un pas.

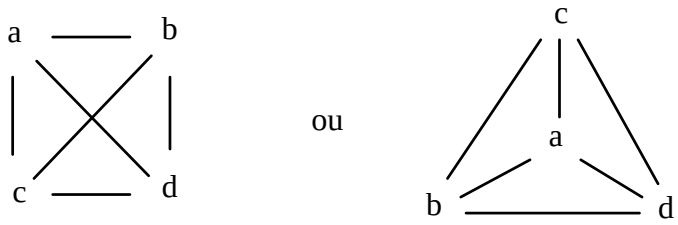
En revanche, si les états a et b sont liés de telle façon que la transformation s'effectue, par exemple, de a vers b , le binôme a ➞b est dit *orienté* ou *dirigé* .

2.1.2. Caractères à états multiples

Les relations entre les états multiples d'un caractère peuvent être de plusieurs types. Elles peuvent être non ordonnées ou, au contraire, être ordonnées. Dans ce dernier cas on parle de « série de transformations » du caractère. Cette série peut être *linéaire* (au sens de sans bifurcation) ou *non linéaire* (avec bifurcation). Les caractères dont les états sont ordonnés seront également appelés additifs.

relations non ordonnées

Chaque état peut se transformer directement en tout autre état, chaque transformation ne comptant que pour un pas. On parle également d'une série *non additive.* Ces relations sont nécessairement non linéaires.



La transformation a ➞ b compte pour un pas, aussi bien que les transformations b ➞ a, a ➞ c ou c ➞ b, etc. Ce cas de figure correspond à la procédure de pondération minimale de Fitch (1971). Il s'applique aux analyses de séquences de protéines ou de nucléotides. Ici les quatre états peuvent être les 4 bases A, C, G, T (ou A, C, G , U), où chacune des bases peut être remplacée de manière équivalente par toute autre.

Relations ordonnées ou additives

Elles correspondent à des séries de transformations linéaires ou non linéaires, orientées ou non orientées :

— série linéaire non orientée :

a — b — c — d

Un caractère à états multiples est dit *linéaire* (donc également additif) quand on peut passer successivement d'un état à un autre. Chaque transformation valant un pas, cela implique nécessairement que le passage d'un état extrême (ici a) à l'autre (ici d) demande autant de pas qu'il y a d'états moins un : le passage de l'état a à l'état d (ou l'inverse) demande ici 3 pas.

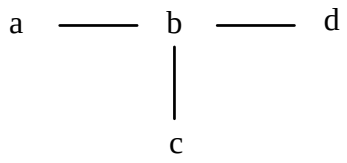
— série linéaire orientée (ou dirigée) :

a ➞ b ➞ c ➞ d

La série linéaire additive orientée correspond au morphocline au sens de Maslin (1952) et à la série de transformations de Hennig (1966) : b est apomorphe par rapport à a et plésiomorphe par rapport à c. Il convient de ne pas faire de confusion entre caractères *ordonnés* (concernant *l'ordre* des transformations) et caractères *orientés* (concernant le *sens* des transformations).

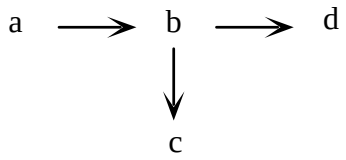
— série non linéaire non orientée:

La série non linéaire présente les relations entre états multiples sous forme d'un arbre. Toutes les transformations ne comptent pas toutes pour un même nombre de pas. Par exemple la série :

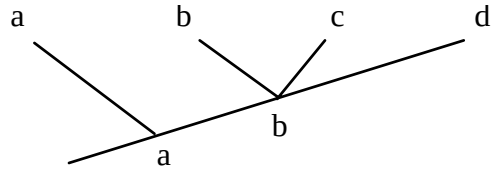


comprend au total 3 transformations, mais elle implique différentes transformations qui comptent 2 pas (transformations de a vers c et c vers a, de a vers d et d vers a, de c vers d et d vers c) tandis que d'autres comptent un seul pas (transformations de b vers a, c ou d et transformations de a, c et d vers b).

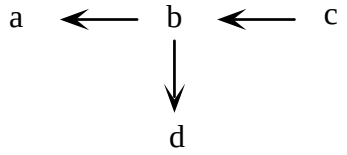
— série non linéaire orientée :



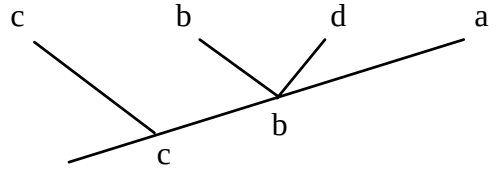
Une telle série peut éventuellement être représentée sous la forme d'un cladogramme de caractère :



Si la série est dirigée sous la forme suivante :



le cladogramme du caractère est :



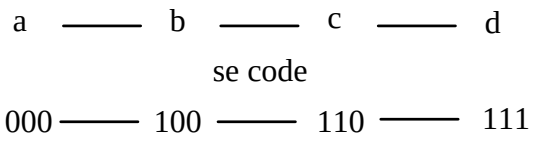
Toutes ces représentations des transformations, ou codages, ne sont pas « neutres », puisque l'on code une hypothèse phylogénétique qui est immédiatement apparente grâce à la forme arborescente du graphe de transformations des états du caractère.

2.1.3. Codage binaire des séries de transformations (factorisation)

Les logiciels d'analyse de parcimonie lisent les caractères binaires, les séries non additives et les séries linéaires additives. Certains d'entre eux nécessitent toutefois un recodage sous forme binaire, « à la main » ou à l'aide d'autres logiciels. En revanche, les séries additives non linéaires nécessitent toujours un recodage, en tout ou partie binaire.

Un tel recodage a pour résultat de faire éclater le caractère à états multiples additifs (linéaires ou non linéaires) en plusieurs caractères à états binaires ou à états multiples linéaires. Le problème à résoudre est celui d'un nouveau codage qui respecte la hiérarchie supposée des états multiples. On appelle « factorisation » le codage sous forme binaire de caractères à états multiples.

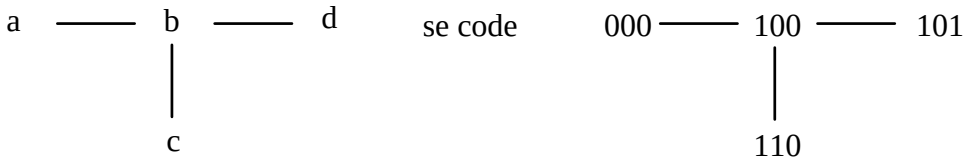
Exemple 1: *La série linéaire additive*



Trois caractères binaires (codés 0—1) suffisent à rendre compte de cette série linéaire de transformation. Le passage 000➞111 (ou 111➞000) compte 3 pas (un par nouveau caractère), comme le faisait le passage de a vers d (ou de d vers a).

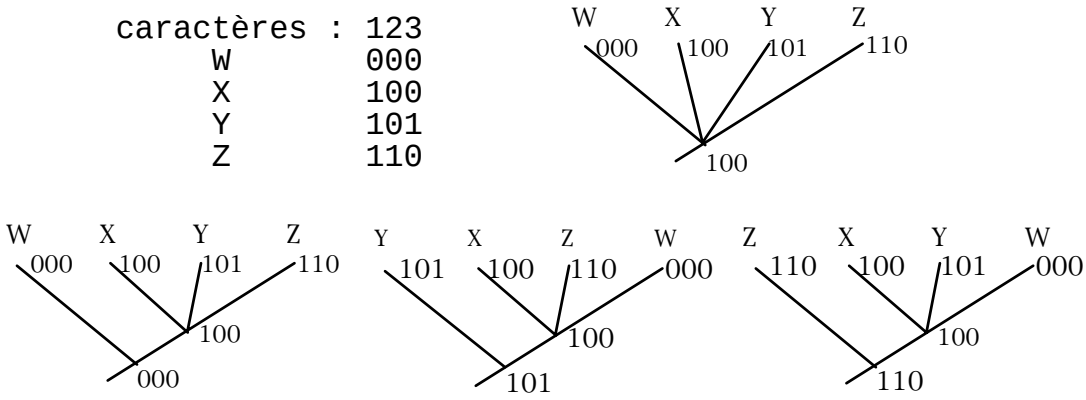
Une telle représentation des transformations, ou codage, n'est pas « neutre », puisque l'on code une hypothèse phylogénétique qui est immédiatement apparente grâce à la forme arborescente du graphe de transformation des états du caractère.

Exemple 2. La série non linéaire additive



Ici le passage 000 vers 101 (anciens états a et d) compte 2 pas.

Si 4 taxons W, X, Y et Z présentent chacun respectivement l'un des quatre états a, b, c et d, codés de la façon précédente, ces états sont introduits dans la matrice des caractères sous la forme de 3 caractères indépendants 1, 2 et 3. Chaque cladogramme obtenu contient 3 pas. Dans le cas où l'ancêtre a les caractères du taxon X (100), l'arbre n'est pas résolu. Dans les trois autres cas, les cladogrammes partiellement résolus restituent les branchements de la série non linéaire additive.



Exemple 3 : caractère morphologique à états multiples

La figure V.11A-E montre un caractère morphologique à états multiples : la région naso-prémaxillaire des primates hominoïdes. Chaque taxon (A-E) présente une morphologie différente. Cette morphologie peut être conçue comme une série d'états multiples (a-e) qui résume les rapports entre maxillaire et prémaxillaire

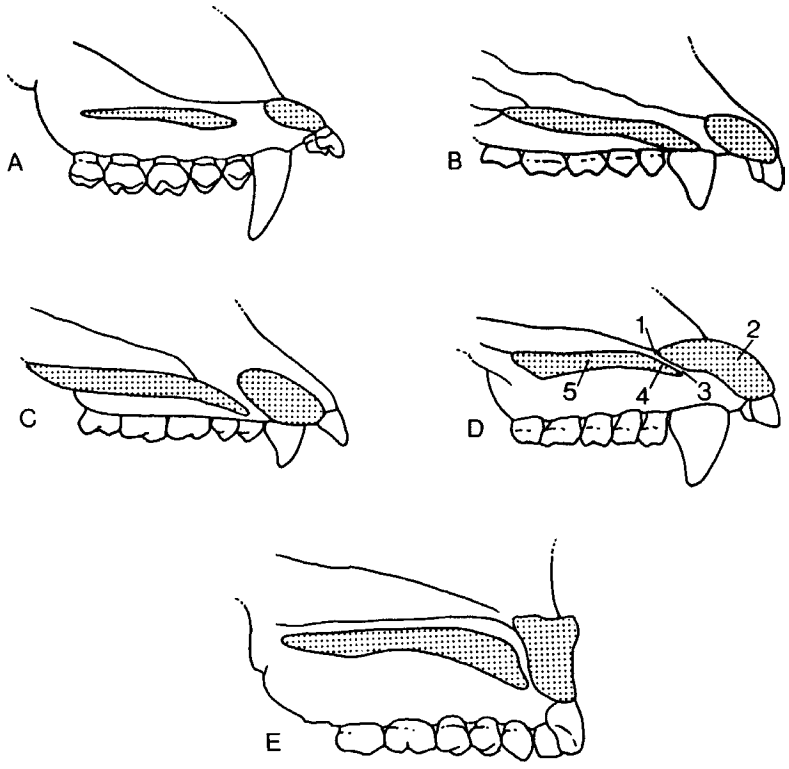
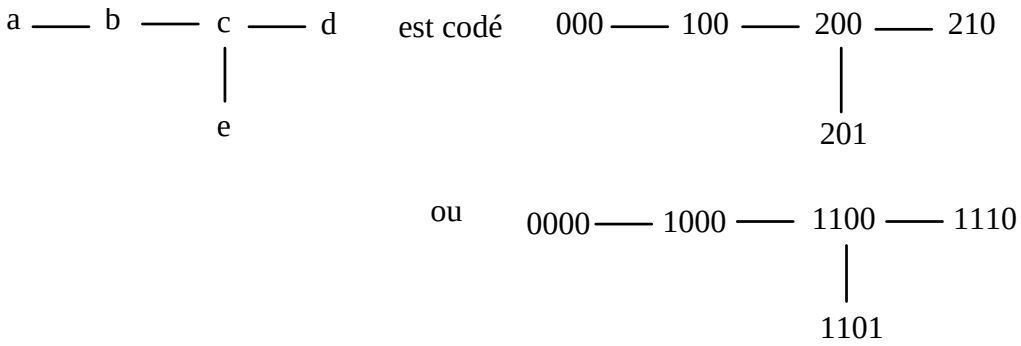


FIGURE V.11. *Sections sagittales de la région naso-prémaxillaire chez les hominoïdes. A:* Hylobates *; B :* Gorilla *; C :* Pan *; D :* Pongo *; E :* Homo*. 1 : fosse incisive ; 2 : processus palatin du prémaxillaire ; 3 : foramen incisif ; 4 : canal incisif ; 5 : processus palatin du maxillaire (d'après Ward et Kimbel, 1983, modifié par Barriel, 1992).*

(traits anatomiques 2 et 5 de la figure V.11) ainsi que la taille et l'orientation du canal incisif (trait anatomique 4) avec la forme de la fosse et du foramen incisifs (traits 1 et 3). Chaque état (a, b, c, d, e) correspond à une combinaison des différents traits anatomiques. Chacun d'eux est spécifique d'un taxon (A-E). La série et ses deux codages possibles se présentent sous la forme suivante :



— une morphologie (a) associant maxillaire et prémaxillaire éloignés au point qu'il n'y a pas de vrai canal incisif (A) est codée 000 (ou 0000) ;
— une morphologie (b) associant maxillaire et prémaxillaire rapprochés avec canal incisif court et large et non vertical (B) est codée 100 (ou 1000) ;
— une morphologie (c) associant maxillaire et prémaxillaire rapprochés avec canal incisif allongé et non vertical (C) est codée 200 (ou 1100) ;
— une morphologie (d) associant maxillaire et prémaxillaire rapprochés avec canal incisif non vertical mais très étroit en raison de l'extension vers l'arrière du prémaxillaire (D) est codée 210 (ou 1110) ;
— une morphologie (e) associant maxillaire et prémaxillaire rapprochés avec canal incisif allongé et orienté verticalement (E) est codée 201 (ou 1101).

L'hypothèse d'additivité du caractère est fondée sur le fait que l'existence du canal commande sa morphologie, qu'il soit allongé ou vertical.

2.1.4. Séries de transformations combinant additivité et non additivité

Certaines séries de transformations multiples ne peuvent être codées, même au moyen des factorisations décrites au paragraphe précédent. C'est le cas de l'exemple suivant, où l'on admet deux possibilités pour le passage de b à d : soit un seul pas, par une transformation directe de b en d, soit deux pas, par une transformation de b en c puis de c en d (figure V.12) :

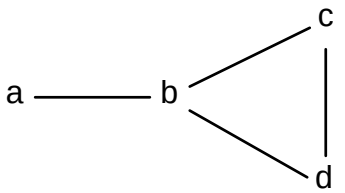


FIGURE V.12. *Exemple de transformations d'un caractère qui ne peuvent être factorisées.*

Une façon particulière d'aborder cette difficulté consiste à décomposer cette série en deux séries S_1 et S_2, qui ont l'avantage de ne pas poser de problème particulier de factorisation (Figure V-13).

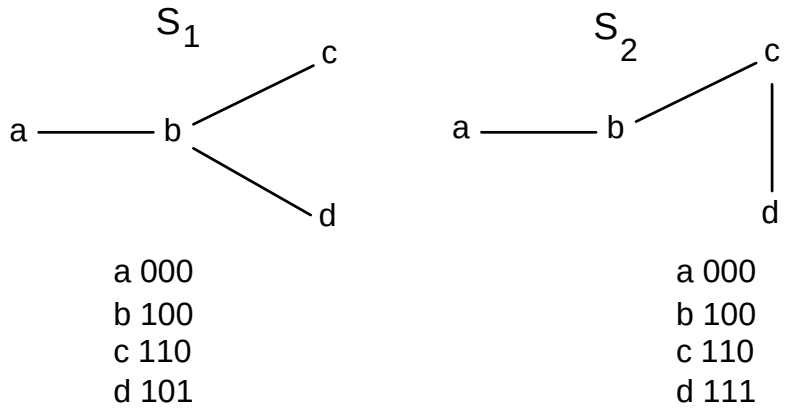


FIGURE V.13. *Exemple d'incertitude entre deux séries de transformations qui s'exprime en codant l'état d : 1?1.*

On voit qu'il n'existe qu'une seule différence de codage entre ces deux séries S_1 et S_2, située sur l'état d qui est codé soit 101 soit 111. Une façon de traiter cette ambiguïté est de remplacer ce codage par un nouveau codage 1?1. Le deuxième caractère du triplet sera alors optimisé *a posteriori* selon la topologie de l'arbre final obtenu par parcimonie à partir de l'ensemble des données. Si le critère de parcimonie impose, dans cet arbre, que ce « ? » = 0, la série de transformations sera la série S_1. Elle sera en revanche la série S_2 si le critère de parcimonie impose « ? » = 1.

Admettons l'existence de quatre UE (W, X, Y et Z), codées respectivement 000, 100, 110 et 1?1. Quinze cladogrammes sont possibles. L'optimisation des états de l'UE Z donnera d = 111 lorsque Y et Z sont en position de groupes frères (trois arbres), produisant la série de transformation S_1. Elle donnera d = 101 dans tous les autres cas, produisant alors à la série de transformation S_2.

La morphologie fournit des exemples fréquents de ces séries alternatives de transformations. C'est le cas quand un *caractère* peut être perdu à partir de différents *états* de ce caractère. Imaginons qu'un caractère soit observé sous trois états a, b et c, comme sur la figure V.13. La perte du caractère peut être la perte de b (comme dans la série S_1) ou la perte de c (comme dans la série S_2).

L'optimisation *a posteriori* des états de caractères proposée ici n'est pas sans affinité avec un mode de codage des caractères à états multiples appelé « T.S.A. » (Mickevich, 1982). Celui-ci fait l'objet du paragraphe suivant.

2.1.5 Analyse des caractères à états multiples selon la méthode du T.S.A

Pour Mickevich (1982), l'analyse cladistique des caractères à états multiples reste un réel problème non encore parfaitement résolu. Pour ce faire, Mickevich propose un mode d'analyse dit « T.S.A. » (*Transformation Series Analysis)* qui a

donné lieu récemment à quelques applications à partir de données morphologiques sur des groupes divers tels les eucaryotes (Lipscomb,1989) ou les lépidoptères tortricidés (Pogue & Mickevich, 1990).

Il s'agit d'une méthode itérative qui a pour but d'établir les transformations entre les états multiples des caractères sous forme d'une série linéaire additive, au moyen du cladogramme construit à l'aide de l'ensemble des caractères de la matrice.

Dans une matrice quelconque, certains caractères sont présents sous deux états et codés de façon binaire, tandis que d'autres peuvent se présenter sous forme d'états multiples. On a vu dans le paragraphe 2.1 que les états multiples d'un caractère s'organisent de manière linéaire ou non linéaire, de façon additive, ou non additive (optimisation dite de Fitch). Le T.S.A. concerne les séries linéaires additives.

L'originalité du T.S.A. est que la série de transformations de chacun des caractères à états multiples de la matrice n'est établie qu'en fonction de son adéquation au cladogramme obtenu par l'analyse de tous les caractères. Le principe du T.S.A. est de se fonder sur le cladogramme pour déduire l'arrangement parcimonieux des états multiples des caractères sous forme d'une série linéaire où les transformations ont lieu sur les nœuds internes du cladogramme (c'est-à-dire chez les ancêtres) ; cette recherche s'effectue de telle sorte que l'arrangement soit celui qui correspond à la hiérarchie du cladogramme (ce que l'on appellera le « cladogramme des états »). Cet arrangement est orienté *in fine* à l'aide du critère de comparaison extra-groupe.

La méthode

La méthode est présentée de façon détaillée dans différents articles dus à Mickevich et collaborateurs (Mickevich, 1982 ; Lipscomb, 1990 ; Mickevich et Lipscomb, 1991). Nous l'illustrons ici brièvement à partir d'un exemple simple.

On observe chez 7 taxons (A, B, C, D, E, F, X ; X étant l'extra-groupe) N caractères à états multiples dont le caractère K qui se présente sous la forme de quatre états : w, x, y, z.

Taxons	Etat du caractère K
X	w
A	w
B	x
C	z
D	y
E	y
F	z

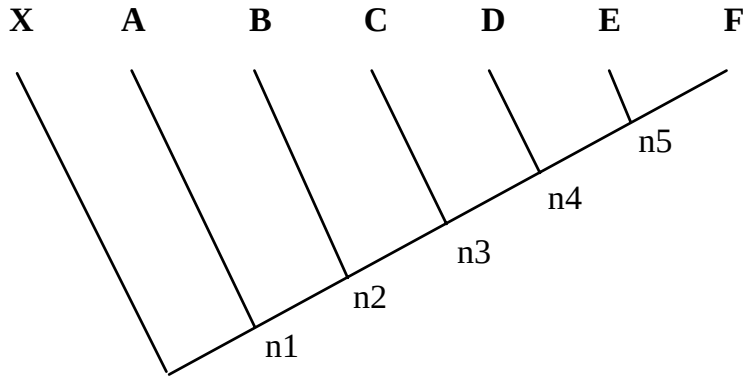
Le T.S.A. est conduit en suivant les étapes décrites ci-après :

1) Une forme linéaire additive quelconque, choisie arbitrairement ou en fonction d'une hypothèse jugée convenable pour toute raison possible, est

attribuée à chacun des caractères à états multiples figurant dans la matrice de données.

2) Une analyse de parcimonie est ensuite effectuée à partir de l'ensemble des caractères dont les formes ont été définies en 1). Il convient de s'assurer que les diverses formes possibles de cette série initiale donnent le même cladogramme, sinon le TSA peut donner des résultats discordants (Pogue et Mickevich, 1990 ; Buckup et Dyer, 1991).

Prenons l'exemple d'un caractère K ayant 4 états et dont la série de transformations est ainsi choisie, au départ de l'analyse : w — x — y — z. Supposons que le cladogramme obtenu par analyse de parcimonie de tous les caractères soit le suivant :



Le cladogramme comporte cinq nœuds internes (n1 à n5).

3) A partir du cladogramme obtenu en 2) on déduit la série de transformations de chacun des N caractères à états multiples de la façon suivante. On établit pour ces caractères la matrice des taxons-voisins et la matrice des états-voisins à partir de laquelle est déduite la série de transformations de chacun des caractères à états multiples. L'exemple donné sera celui du caractère K.

a) Matrice des taxons-voisins

La matrice des taxons-voisins est établie en comptant le nombre de nœuds séparant les taxons pris deux à deux. Par exemple le nombre de nœuds séparant les taxons A et D est de 4 (n1, n2, n3, n4). Les taxons-voisins sont dits adjacents quand ils sont séparés par un nombre minimal de nœuds internes. Dans la matrice, le nombre minimal de nœuds entre taxons-voisins est exprimé en gras. La lecture de la matrice des taxons-voisins permet l'identification des taxons-voisins.

	X	A	B	C	D	E
X						
A	**1**					
B	**2**	**2**				
C	3	3	**2**			
D	4	4	3	**2**		
E	5	5	4	3	**2**	
F	5	5	4	3	**2**	1

Matrice des taxons-voisins

Taxon	Taxons-voisins	Etat	Etats-voisins
X	A	w	w
A	X	w	w
B	A, X et C	x	w,w,z
C	B et D	z	x,y
D	C, E et F	y	z,y,z
E	F	y	z
F	E	z	y

Tableau des taxons-voisins et des états-voisins

b) Les états-voisins

Les états voisins sont les états adjacents, c'est-à-dire séparés par un nombre minimal de nœuds internes lorsque l'on substitue les états des caractères aux taxons qui les portent dans le tableau des taxons voisins. La différence entre le tableau des états-voisins et le tableau des taxons-voisins est qu'un même état peut être porté par différents taxons. On peut donc simplifier ce tableau des états voisins et le représenter sous forme d'une matrice des états-voisins qui indique le nombre de fois ou deux états sont voisins.

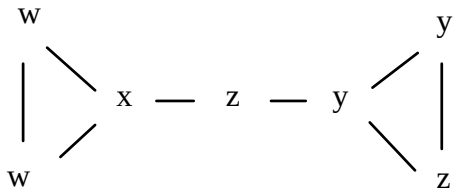
Etat	Etat voisin
w	w
x	w et z
z	x et y
y	z

	w	x	y	z
w	-			
x	1			
y	0	0		
z	0	1	2	

Tableau simplifié des états-voisins et matrice des états-voisins.

La règle suivante doit cependant être appliquée : lorsqu'un état est voisin de lui-même, il est dit homologue. Ainsi l'état voisin de w est w puisque le taxon-voisin de X est A. L'état w de X et l'état w de A sont homologues. De ce fait w et x n'apparaissent qu'une fois états-voisins. De même, x et z sont une fois états-voisins, y et z sont deux fois états-voisins. Les taxons-voisins de D sont C, E et F : les états-voisins de y sont z, y et z, l'état y de E étant homologue à l'état y de D puisque D et E sont voisins. En revanche, l'état z porté par F et l'état z porté par C ne sont pas homologues car F et C ne sont pas des taxons-voisins.

On peut également identifier les états-voisins directement à partir du cladogramme. On construit un réseau où les états remplacent les taxons de telle façon que les états adjacents soient reliés entre eux :



c) La série linéaire additive

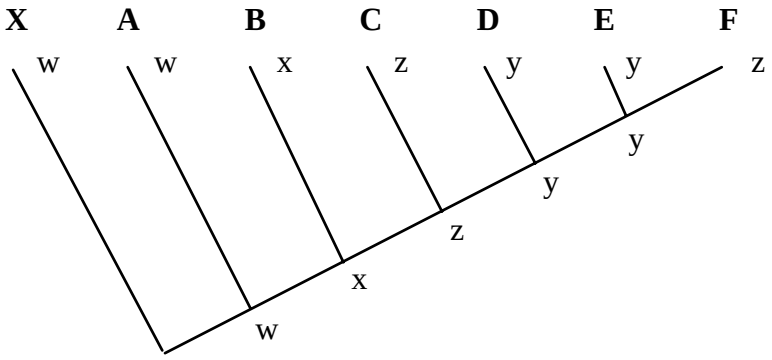
A partir de la matrice des états voisins, il devient possible d'établir une série linéaire : la matrice des états-voisins montre que x est une fois l'état-voisin de w et une fois l'état-voisin de z : tandis que y et z sont deux fois états-voisins. Ce résultat suggère la connexion linéaire additive : w —x — z — y.

Cette série est différente de la série initiale. Elle implique que le passage de l'état w à l'état z compte 2 pas, x étant intermédiaire. Dans la série initiale, le passage de l'état w à l'état z comptait 3 pas et x et y étaient deux intermédiaires entre w et z.

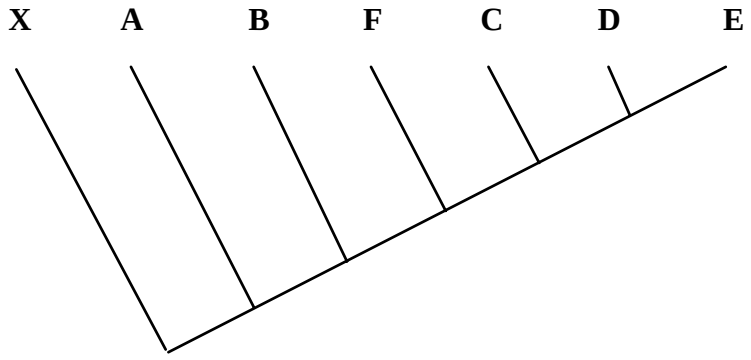
Lue sur le cladogramme, cette série de transformations donne le « cladogramme des états » :

w — x — z — y — z.

Cette opération revient à comprimer le réseau des états-voisins obtenus en 3b), de telle façon que les états-homologues n'apparaissent qu'une fois, ce qui correspond à la série linéaire additive : w — x — z — y — z.

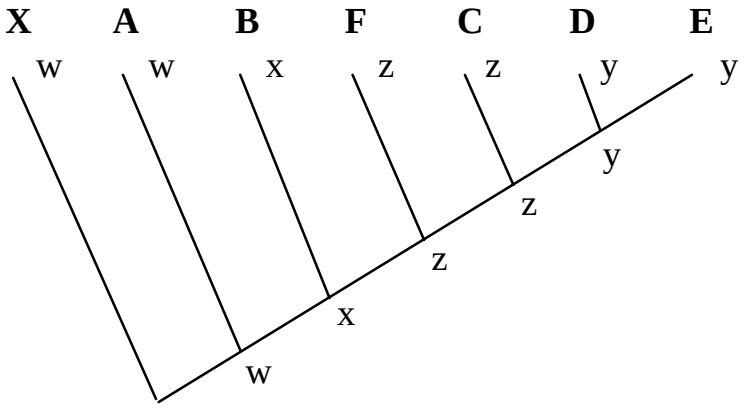


4) On modifie la matrice initiale en donnant aux caractères à états multiples – tel que K – leur nouvelle forme. On analyse ensuite, par parcimonie, l'ensemble de ces données afin d'obtenir un nouveau cladogramme qui peut être identique ou différent du précédent. Admettons qu'il soit différent du cladogramme obtenu en 2).



5) On recommence l'étape 3 à partir de ce nouveau cladogramme afin d'obtenir la série de transformations linéaire de chacun des caractères à états

multiples correspondant à ce cladogramme. Pour le caractère K cette série est la même que précédemment : w — x — z — y. L'optimisation aux nœuds à partir de cette série obéit à la règle selon laquelle les transformations doivent suivre la hiérarchie du cladogramme. La hiérarchie du cladogramme est commandée par la séquence des nœuds ; autrement dit, l'optimisation des états aux nœuds doit correspondre à la série obtenue par T.S.A. Le nombre de pas minimal est ici de 3 :



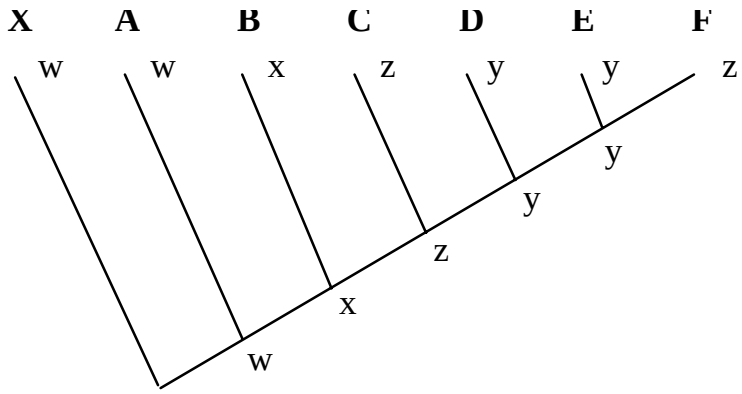
Lue sur le cladogramme, cette série de transformations donne le « cladogramme des états ». Ce cladogramme des états est : w — x — z — y ; il est en accord avec la série obtenue par T.S.A. (un cas de désaccord est détaillé plus loin).

L'orientation de la série de transformations obtenue par T.S.A. est donnée par l'extra-groupe (pôle plésiomorphe). Dans notre exemple, la série orientée serait alors w➞x➞z➞y.

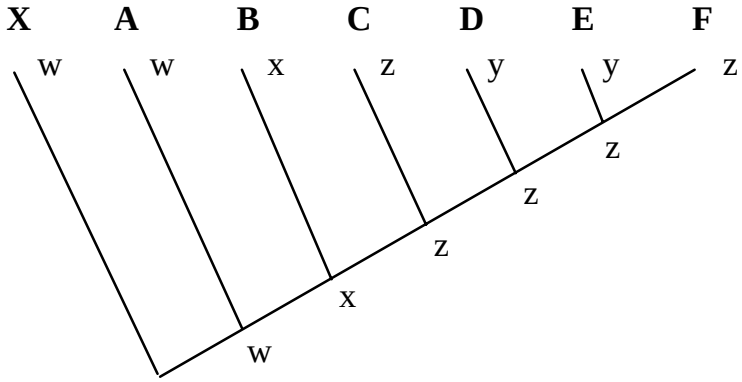
A ce stade, pour d'autres caractères à états multiples, les séries peuvent être à nouveau différentes. Dans ce cas, on recommence l'étape 3 avec introduction pour ces caractères de leur nouvelle série de transformations. Le processus itératif se poursuit jusqu'à ce que les séries de transformations lues sur le cladogramme soient identiques aux séries introduites lors de l'étape précédente : le T.S.A. est alors achevé pour tous les caractères à états multiples. Les séries de transformations retenues pour tous les caractères à états multiples sont celles qui correspondent à ce cladogramme final.

Les cas d'homoplasie peuvent être résolus par le T.S.A. L'homoplasie implique que, sur le cladogramme, plusieurs séries de transformations soient possibles (elles comptent pour un même nombre de pas).

On a vu que le T.S.A. permet de construire une série de transformations linéaire en accord avec la hiérarchie du cladogramme (la séquence des nœuds), c'est-à-dire avec le « cladogramme des états ». Reprenons l'exemple précédent. Admettons cette fois que le cladogramme obtenu en fin de TSA soit identique à celui obtenue à l'étape 2, mais que la série de transformations de K soit celle issue de l'étape 4 (w — x — z — y). Lue sur le cladogramme, cette série implique une homoplasie : le « cladogramme des états » est en effet : w — x — z — y — z. L'état z porté par F et l'état z porté par C ne sont pas homologues. La matrice des taxons-voisins construite à l'étape 3 indique en effet que les taxons-voisins de C sont B et D, mais non F.



Or il existe une autre possibilité d'optimiser les états aux nœuds à partir de la série w — x — z — y, illustrée par le cladogramme ci-dessous (sans mentionner les autres optimisations possibles mais différentes de la série obtenue par T.S.A.).

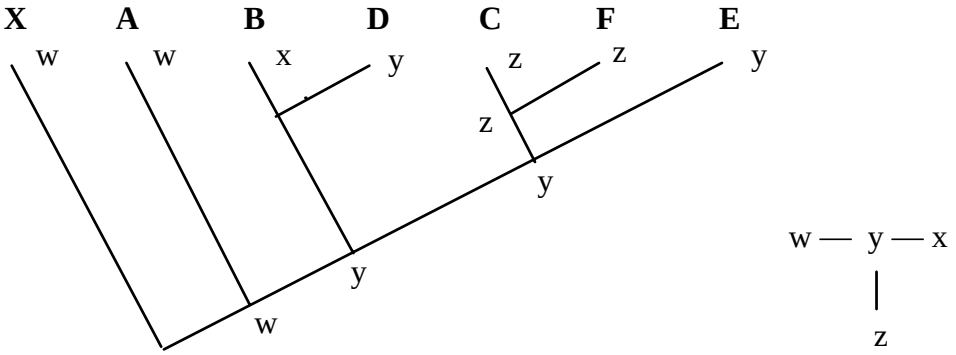


Cette solution aussi parcimonieuse implique une transformation du caractère K qui, lue sur le cladogramme, ne correspond pas à la hiérarchie (à la séquence des nœuds) ; le « cladogramme des états » est le suivant :

w — x — z — y
|
y

Dans ce cas, ce serait l'état z porté par F et par C qui serait homologue ; au contraire, y apparaît deux fois, sur les taxons terminaux D et E. Or cette hypothèse est rejetée par le T.S.A. puisque F et C ne sont pas des taxons-voisins : l'homoplasie se porte sur z, non sur y. Cette élimination revient à privilégier la transformation compatible avec la hiérarchie du cladogramme : le deuxième « cladogramme des états » ne reflète pas la hiérarchie du cladogramme des taxons. Rejeter les deux transformations z — y chez D et chez E, c'est rejeter les transformations non situées aux nœuds du cladogramme.

On a choisi ici un exemple simple. Mais des situations plus complexes peuvent être envisagées où différentes séries sont possibles. De tels cas, et les cas où les transformations ne peuvent être résolues sous forme linéaire, sont discutés en détail par Mickevich et Weller (1990). L'exemple ci-dessous donne un autre cladogramme final qui conduit à envisager, pour le caractère K, une série non linéaire (3 pas) plutôt qu'une série linéaire moins parcimonieuse (4 pas).



Cette méthode, dont l'usage est encore marginal et les présupposés sujets à controverses, n'est implémentée sur aucun des logiciels actuellement disponibles, même si Mickevich et Lipscomb (1991) ont publié une marche à suivre dans le cadre du logiciel Hennig86 dû à Farris (1988). Or, si le nombre de caractères à états multiples est important, l'opération peut être laborieuse. Pour ce qui est des caractères morphologiques, il reste que cette méthode qui considère les séries linéaires additives de transformations de caractères en accord avec la structure du cladogramme, est potentiellement compétitive aussi bien vis-à-vis des options de non-additivité des états (parcimonie de Fitch) que des séries construites *a priori*, parfois intuitivement. En morphologie, l'option Fitch, pour sa part, peut masquer des étapes de transformations et, par conséquent, de possibles homologies. Néanmoins, son application peut soulever de nombreuses difficultés. Le fait que le choix de la série initiale des états multiples utilisée pour commencer l'analyse puisse influer sur le cladogramme de départ, et donc sur le résultat du T.S.A., est un obstacle non négligeable d'un point de vue pratique. Enfin, d'un point de vue plus général, rechercher seulement une série de transformations linéaire compatible avec le cladogramme plutôt qu'une série aussi parcimonieuse mais non linéaire, est un choix, qui, en tant que tel, peut se discuter.

2.2. Polymorphisme

Le polymorphisme est un cas particulier car il implique que les états plésiomorphe et apomorphe (0 et 1) soient présents dans un même taxon.

Les caractères morphologiques polymorphes peuvent être codés de façon discrète sous la forme de séries linéaires. Dans ce cas, on admet que le passage d'un état 0 à un état 1 (ou l'inverse) se fait nécessairement par l'intermédiaire d'une combinaison 01 caractérisant l'état polymorphe. Autrement dit, le caractère est polymorphe chez un taxon dont certains membres ont le morphe 0 et d'autres le morphe 1. Le codage de la série revient à un codage d'un caractère à états multiples : 0 – 1 – 2 (où 1 est l'état polymorphe = intermédiaire). La fréquence des morphes 0 ou 1 à l'intérieur du taxon n'intervient pas.

La question du codage du polymorphisme se pose avec acuité dans le traitement des fréquences alléliques (électromorphes). La plupart du temps, les données électrophorétiques se présentent sous forme de fréquences et sont analysées par des méthodes de distances (voir chapitre VI). Mais le codage des allèles sous forme de caractères discrets est le seul qui soit traitable de façon

A

Taxons	Fréquences des allèles						Allèles indépendants					
	a	b	c	d	e	f	a	b	c	d	e	f
U	0	0	1	0	0	0	0	0	1	0	0	0
V	0	0	1	0	0	0	0	0	1	0	0	0
W	0	0,5	0	0,5	0	0	0	1	0	1	0	0
X	0,1	0	0	0	0	0,9	1	0	0	0	0	1
Y	0,2	0	0	0	0	0,8	1	0	0	0	0	1
Z	0,3	0	0	0	0,7	0	1	0	0	0	1	0

B

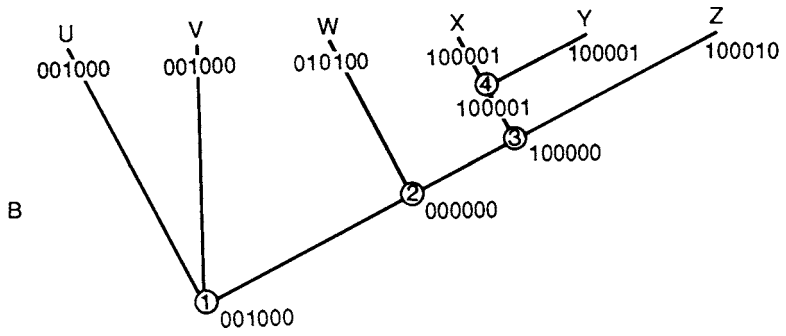


FIGURE V.14. *A : Fréquences des allèles pour les taxons U à Z, avec, à droite, le codage sous le modèle des allèles indépendants. B : Cladogramme issu de A, à partir du modèle des allèles indépendants. On remarque que l'ancêtre (2) du groupe (W((X,Y)Z)) est dépourvu d'allèles.*

cladistique. Une telle approche revient à ne tenir compte que de la présence ou de l'absence des allèles. C'est le modèle des « allèles indépendants » (Mickevich et Johnson, 1976; Mickevich et Mitter, 1981). Ce modèle a donné lieu à des applications satisfaisantes (Patton et Avise 1983), mais il peut aboutir à des situations biologiquement aberrantes. Il suffit de considérer le cas de la figure V.14. A la suite d'une réversion, l'ancêtre représenté par le nœud 2 est dépourvu d'allèles. Sous ce modèle de présence/absence, les taxons X et Y, qui ne diffèrent que par les fréquences des allèles partagés a et f, sont identiques et identiques à leur ancêtre commun (nœud 4). Selon Mickevich et Mitter (1981, 1983) l'application aux données électrophorétiques du T.S.A. (voir paragraphe précédent) permet d'éviter ce genre de problèmes. Quant à la perte d'information qu'implique l'abandon des fréquences, elle est considérée comme négligeable par ces auteurs dans la mesure où les fréquences alléliques sont très facilement modifiées, par exemple par dérive, et n'apportent pas nécessairement des informations phylogénétiques pertinentes. En outre, dans le cas d'échantillons de tailles très différentes, les estimations des fréquences n'ont pas nécessairement la même précision. Cet argument joue néanmoins aussi pour le codage en présence/absence. Le fait qu'un allèle n'a pas été observé (par exemple codé 0 par

opposition à 1) n'est peut-être dû qu'à sa fréquence basse dans la population au point qu'un petit échantillon de celle-ci n'a pas permis de l'y détecter.

Le traitement automatique du polymorphisme par les logiciels de parcimonie sont des expédients qui ne résolvent pas véritablement le problème. Dans certains cas (logiciel MIX (option P) de *Phylip,* l'algorithme ajoute au nombre de transformations contenues dans l'arbre minimal autant de pas qu'il y a d'états déclarés polymorphes. Cet ajout est automatique et indépendant du critère de parcimonie. Dans d'autres (DOLLOP de *Phylip* le traitement dépend du modèle de parcimonie de Dollo (seules les réversions sont admises). Dans ce cas, des ancêtres peuvent apparaître comme polymorphes, avec perte ultérieure d'un morphe chez les descendants ; mais les taxons terminaux ne peuvent pas être polymorphes.

2.3. Pondération des caractères et des transformations

Toute matrice de données contient des caractères qui, à l'issue de l'analyse phylogénétique, vont se révéler être des synapomorphies ou des homoplasies. Le but de la pondération est de privilégier, lors de la reconstruction de l'arbre, les informations phylogénétiques pertinentes au détriment du « bruit » occasionné par les homoplasies. Dans certains cas, cette pondération s'appuie sur les observations elles-mêmes et implique les caractères. Le plus souvent elle nécessite l'introduction d'hypothèses ou d'informations extrinsèques aux données.

On abordera dans ce paragraphe les questions de pondération en amont de l'analyse. La « pondération successive » qui s'apparente, d'une certaine manière, aux comparaisons d'arbres, et qui est une pondération en aval de l'analyse, est discutée au paragraphe 4.3.

La procédure de pondération peut se concevoir à deux niveaux différents selon que l'on vise les caractères ou bien leurs transformations.

1) *Pondérer un caractère* c'est, lui donner *a priori* une plus ou moins grande importance lors de la recherche de l'arbre le plus parcimonieux. C'est, éventuellement, l'éliminer en lui donnant un poids nul. Si on attribue, par exemple, un poids de 2 à un caractère, cela revient à le répéter deux fois dans la matrice de données. Le résultat phylogénétique prévisible d'une telle pondération est d'augmenter les chances pour que les taxons partageant ce caractère artificiellement dupliqué se trouvent étroitement apparentés. En choisissant de la sorte d'attribuer plus ou moins de poids à tel ou tel caractère, on peut obtenir pratiquement le résultat phylogénétique que l'on souhaite. Les critères du choix des pondérations sont donc particulièrement importants à définir préalablement à toute analyse phylogénétique.

Rappelons que la reconstruction phylogénétique est fondée sur l'hypothèse que les caractères évoluent indépendamment les uns des autres. Lorsque ce n'est pas le cas, lorsque deux caractères évoluent donc de manière concertée pour des raisons non liées à la parenté (comme le font par exemple les mutations compensées au niveau de l'ARN ribosomique), alors la recherche de l'arbre le plus court s'effectue avec une pondération implicite de ces caractères.

2) *Pondérer une transformation* c'est estimer *a priori* que, pour un caractère donné, la transformation d'un état en un autre est plus ou moins difficile – ou rare – selon les états concernés. L'option d'additivité ou de non–additivité des états multiples vue au paragraphe 2.1 est une forme de pondération des transformations. Par exemple, pour un caractère ayant trois états différents non additifs (a, b et c), le passage de l'état a à l'état c constitue, dans l'option de non-additivité, une transformation de poids égal à 1, tandis que lorsque le caractère est additif (série a—b—c), la transformation a—c constitue une transformation de poids égal à 2. Dans le cas de caractères à états multiples additifs, toutes les transformations n'ont donc pas le même poids.

2.3.1. Caractères morphologiques

L'analyse cladistique attribue en principe un poids égal à tous les caractères. Cette option est néanmoins controversée en raison d'une tradition bien ancrée qui consiste à construire de manière intuitive des arbres phylogénétiques en sélectionnant, *a priori* ou sur la base de l'expérience des spécialistes, les « bons » caractères au détriment des « mauvais ». En fait, cette sélection revient à pondérer les caractères en donnant un poids nul aux « mauvais » caractères. Une telle pratique est d'autant plus compréhensible, même si elle n'est pas justifiée, que la phylogénie des organismes ne reste décelable qu'au travers des caractères : il est donc tentant de postuler que des caractères judicieusement choisis doivent refléter la phylogénie des organismes.

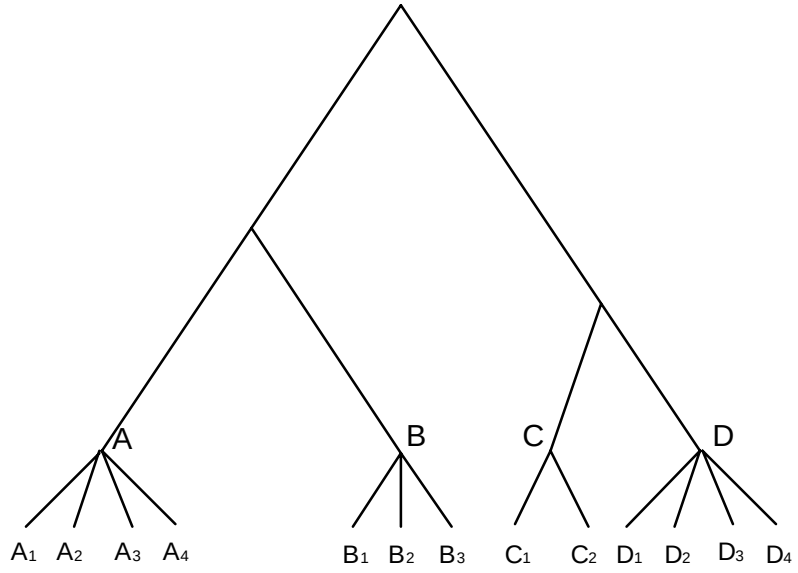
La pondération des caractères morphologiques reste un sujet de discorde entre phylogénéticiens, notamment en ce qui concerne les caractères adaptatifs souvent considérés *a priori* comme de mauvais caractères car trop facilement soumis aux phénomènes d'homoplasie : les mêmes pressions de sélection peuvent en effet aboutir à des morphologies comparables chez des organismes non apparentés. Mais la question de l'identification des caractères adaptatifs ou non adaptatifs est aussi anciennement débattue et controversée que celle de la pondération des caractères. La valeur adaptative prêtée à un caractère est souvent conjecturale et, inversement, l'hypothèse de non-adaptativité est parfois un aveu d'ignorance.

Par ailleurs, on attache souvent une grande importance à la signification fonctionnelle d'un caractère ou d'un ensemble de caractères. Pour de nombreux anatomistes (Szalay, 1981a et b), des caractères bien définis du point de vue fonctionnel devraient peser d'un poids plus lourd que de simples observations dont l'importance biologique est faible. Cependant il reste que des caractères parfaitement intégrés dans une fonction particulière peuvent également être soumis à homoplasie. On pourrait même avancer, à l'inverse, qu'un caractère ayant une fonction importante peut avoir de plus forte chance d'être le résultat adaptatif de pressions sélectives et donc d'être soumis à homoplasie. On peut penser tout aussi bien que des caractères fonctionnellement neutres d'un point de vue adaptatif, ont moins de chance d'être homoplasiques et devraient donc avoir un poids plus important. La compréhension de la fonction d'un caractère n'est donc pas garante de sa signification phylogénétique.

En morphologie, la question de la pondération des caractères surgit notamment en considérant la distribution de leurs états selon les différents taxons, indépendamment même de la construction de l'arbre, c'est-à-dire indépendamment

de l'obtention de l'arbre le plus court. Certains caractères passent, à juste titre, pour apparaître facilement dans diverses lignées. On sait par exemple que dans l'évolution des mammifères la perte de prémolaires, ou bien, inversement, la molarisation des prémolaires, sont des phénomènes à apparitions multiples. Cependant, nécessairement, l'apparition de ces traits caractérise des groupes monophylétiques, c'est-à-dire qu'ils sont bien, pour ces groupes, hérités d'un ascendant. Plutôt que d'être supprimés, de tels traits pourraient être pondérés « en baisse », par rapport à des caractères rencontrés rarement.

A cette fin, la prise en compte de la variation intra-taxon et de la variation inter-taxons des caractères permet d'effectuer une telle pondération.



Prenons l'exemple de 4 taxons A, B, C et D. Chacun de ces taxons est constitué de plusieurs sous-taxons connus (A_1, A_2, ...B1, B2 ...). Considérons un caractère K_1 dont la distribution est telle que les variations intra-taxons soient faibles : tous les taxons A et B (sauf A_4) ont l'état 1 de ce caractère et tous les taxons C et D ont l'état 0. La variation intra-taxon est donc faible pour ce caractère K_1. En revanche, elle est très forte pour le caractère K_2 qui prend aussi bien l'état 0 ou 1 à l'intérieur de chacun des taxons A, B, C et D. Il est clair que les caractères de type K_1 donneront *a priori* une meilleure information phylogénétique que les caractères de type K_2. Pour ces derniers, on est en effet obligé d'admettre qu'ils sont extrêmement variables, puisqu'ils changent même à l'intérieur d'un taxon et donc que cette variabilité résulte d'événements homoplasiques.

Le caractère K_2, très variable au niveau intra-taxon, devrait donc se voir attribuer un poids plus faible que le caractère K_1, peu ou pas variable. Une pondération envisageable pour ces caractères serait donc l'inverse de la variation intra-taxon (Kluge et Farris, 1969).

Il est également possible de pondérer en fonction de la variation inter-taxons. Un caractère présent sous le même état dans tous les taxons aura une variabilité inter-taxon nulle. De même, un caractère présent sous deux états dont l'un ne serait propre qu'à un taxon terminal conduirait à une variabilité inter-taxons faible. De tels caractères qui n'apportent aucune ou très peu d'information sur les parentés peuvent donc se voir attribuer un poids faible.

Associant ces deux types de pondérations, on peut proposer une pondération unique qui soit fonction du rapport entre la variance inter- et la variance intra-taxons (Farris, 1966 ; Goodman, 1969 ; Sneath et Sokal, 1973).

Il faut insister sur le fait qu'une telle pondération n'est possible qu'à la condition de posséder des informations sur le polymorphisme des caractères *à l'intérieur* des taxons. De plus, il faut considérer *a priori* que les hypothèses de parenté à l'intérieur des taxons (dans l'exemple A, B, C et D) ne sont pas remises en cause. Ce type de pondération peut être mis en place aussi bien pour des caractères morphologiques que moléculaires.

2.3.2. Caractères moléculaires

La pondération des caractères moléculaires vise à donner un poids plus ou moins important aux différents sites où l'on observe des transformations, qu'il s'agisse de nucléotides ou d'acides aminés. La pondération des transformations vise à relativiser les différentes transformations les unes par rapport aux autres.

Acides nucléiques

Pour les séquences alignées d'ADN (ou d'ARN), le site est considéré comme un caractère et les 4 nucléotides A, C, G et T (ou U) sont les quatre états possibles de ce caractère.

La *pondération des caractères* revient donc ici à pondérer les sites.

- Les sites où se sont exercées de multiples transformations peuvent être localisés par l'étude de la répartition des différents nucléotides en ces sites. Par exemple, lorsqu'un site montre une répartition des 4 nucléotides qui correspond à la répartition de chacun d'eux dans l'ensemble des sites étudiés et sur l'ensemble des UE étudiés, ou qui correspond à une répartition « aléatoire », il peut être justifié de considérer ce site comme n'introduisant que du bruit phylogénétique. On dit qu'il est « saturé ». La pondération d'un tel site aura comme but d'en diminuer l'impact lors de la reconstruction phylogénétique, au point parfois d'être amené à l'ignorer (poids nul).

— Les sites à mutations compensées (lorsqu'une certaine mutation en un site s'accompagne d'une mutation particulière en un autre site) ont une pondération dépendant de leur nombre. Ils apportent en effet tous la même information phylogénétique.

— En raison de la dégénérescence du code génétique, il est envisageable de pondérer différemment les sites en fonction de leur place dans le codon. Il est clair que les sites en troisième position contiennent une information phylogénétique moindre dans la mesure où les mutations y sont *a priori* plus nombreuses. En revanche, elles peuvent aussi être considérées comme plus

« neutres » dans la mesure où elles sont le plus souvent silencieuses. La pondération sur ce critère de position est donc particulièrement délicate.

— La pondération différente d'un gène ou d'un pseudogène, d'une séquence codante ou non codante, constitue aussi une forme de pondération des caractères puisqu'elle permet de privilégier, dans la reconstruction phylogénétique, une source d'information plutôt qu'une autre. On peut envisager en effet que les parties codantes qui ont une fonction connue peuvent présenter des mutations convergentes adaptatives, comme cela peut être le cas en morphologie. A l'inverse, chez un pseudogène non codant où les mutations seraient « neutres », des mutations partagées par différents taxons ont plus de chances d'être héritées d'un même ancêtre.

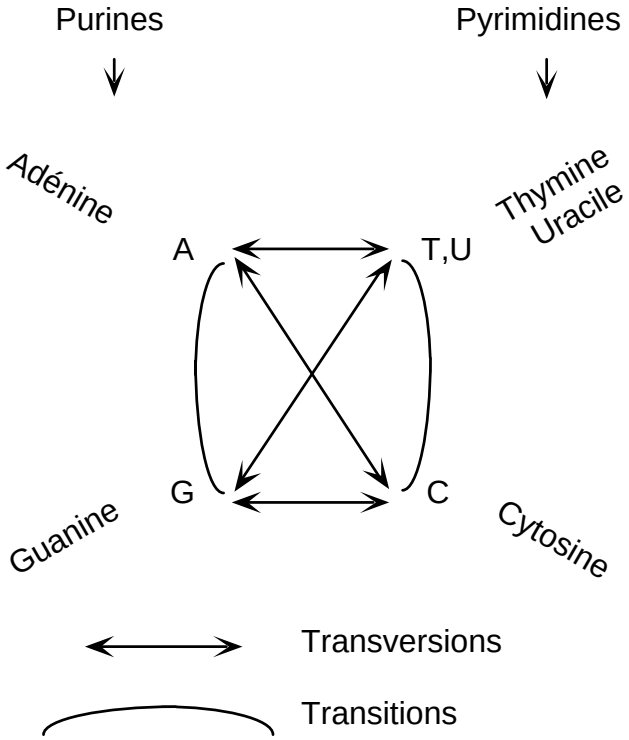


FIGURE V.15. *Différentes transformations entre bases puriques (Adénine et Guanine) et pyrimidiques (Thymine ou Uracile et Cytosine) de l'ADN ou de l'ARN. Les transitions se font entre deux purines ou deux pyrimidines et les transversions entre purine et pyrimidine.*

La *pondération des transformations* (figureV.15) s'effectue à partir d'une distinction entre les différents types possibles de transformations (voir notamment Sankoff et Cedergen, 1983).

— Puisqu'il y a 4 états possibles (A, C, G, T ou U), 12 transformations différentes peuvent être observées. En première approximation, on peut réduire ces 12 types à 6 types seulement si l'on considère que le sens de la transformation n'est pas pertinent (même type de transformation lorsque A se change en G ou G en A par exemple).

— En deuxième approximation, on peut se contenter de distinguer entre les *transitions* qui sont des transformations d'une purine en purine ou d'une

pyrimidine en pyrimidine et les *transversions* qui sont les changements d'une purine en pyrimidine et inversement. Ceci renvoie aux « invariants de Lake », notamment (paragraphe V.5.2).

— Enfin un autre type de transformation peut être pris en compte, l'« indel » : c'est-à-dire l'insertion et la délétion d'un site, indépendamment ou non de la nature des nucléotides insérés.

La pondération peut s'effectuer, par exemple, en raison inverse de la fréquence des différents types de transformations que l'on vient de décrire. L'exemple le plus classique se fonde sur l'observation, issue des comparaisons de certaines séquences alignées, que les transitions sont plus nombreuses que les transversions (Brown *et al*, 1982). Dans ce cas il est possible de donner un poids plus élevé aux transversions qu'aux transitions. On peut considérer en effet que les transversions, plus rares, apportent une information phylogénétique plus solide que les transitions dont la trop grande fréquence d'apparition finit par ne produire que du « bruit ». Cette conclusion doit cependant être relativisée par le niveau hiérarchique des UE que l'on compare. En effet lorsque deux UE ont divergé récemment, les événements de type transition restent informatifs alors que peu ou pas de transversions seront effectivement observées entre ces deux UE. En revanche lorsque deux UE ont divergé très tôt, seules les transversions seront vraiment informatives, les transitions n'étant que du bruit. Il serait donc pertinent de développer une méthodologie qui permette de modifier les pondérations en fonction du niveau de hiérarchie des UE comparées.

La pondération dite « parcimonie des transversions » (*transversion parsimony sensu* Swofford et Olsen, 1990) ignore simplement les transitions. Les 4 nucléotides sont codés de telle façon qu'il n'existe que 2 états: R (purine) et Y (pyrimidine). Une telle approche revient à donner un poids zéro aux transitions. Une part d'information est donc éliminée *a priori*, ce qui reste un choix discutable puisqu'il revient à considérer que les transitions ne sont que du bruit.

Une autre approche plus pragmatique consiste à rechercher les arbres de longueur minimum ou proche du minimum, en ne spécifiant pas *a priori* de pondération particulière pour les transformations. Implicitement cela revient en fait à admettre un poids identique pour tous les types de transformations. Il convient ensuite de comparer les différents arbres retenus et d'examiner les différents types de transformations que chacun d'eux implique. Si l'un de ces arbres, T_1 par exemple, requiert 8 transformations et un autre, T_2, en demande 10, il est clair que le premier est plus « parcimonieux » en terme de nombre de transformations. Cependant il n'est pas indifférent de connaître la proportion de transitions et de transversions parmi ces transformations : supposons que l'arbre T_1 exige 6 transversions et 2 transitions, alors que l'arbre T_2 exige seulement 2 transversions et 8 transitions, on pourrait conclure que T_2 est, de fait, plus « parcimonieux » que le second.

Acides aminés

Dans le cas de séquences alignées de protéines, chaque position d'acide aminé de la chaîne polypeptidique constitue un caractère et les différents états de ce caractère correspondent aux 20 acides aminés possibles constituant les protéines.

La *pondération des caractères* revient ici à donner à certaines positions de la chaîne polypeptidique une importance plus grande qu'à d'autres.

— On peut par exemple imaginer que les zones près du site actif d'une enzyme ou les zones qui déterminent sa structure (ponts disulfures par exemple) sont constituées d'acides aminés dont l'importance phylogénétique est plus grande que ceux des zones périphériques pour lesquels les fonctions sont moins claires. En ce sens les discussions développées à propos des caractères morphologiques, en ce qui concerne les convergences adaptatives et la neutralité, peuvent être transposées ici.

— Les positions sur lesquelles on observe une grande diversité d'acides aminés sont manifestement le résultat de transformations multiples dont il ne ressort que du bruit. Une telle remarque a été effectuée à propos des séquences d'ADN ou d'ARN. Une différence importante tient cependant au fait que l'on a ici 20 états différents au lieu de 4 seulement pour l'ADN. La saturation est donc théoriquement plus lente.

En ce qui concerne la *pondération des transformations*, ici la substitution d'un acide aminé à un autre, plusieurs approches sont possibles :

— Considérer que toutes les substitutions ont un poids identique. Cela revient à appliquer simplement aux acides aminés la parcimonie de Wagner avec optimisation de Fitch (caractères non additifs) (voir paragraphes 1.1.1. et 2.1.2.). Cette simplification ignore la facilité relative de substitution qui dépend des acides aminés impliqués. Elle ignore également le nombre de changements de nucléotides nécessaires pour observer une substitution d'un acide aminé par un autre.

— Pondérer en raison inverse de la fréquence de la substitution. Cette méthode revient à prendre en compte la fréquence avec laquelle un acide aminé se transforme en un autre, avec l'idée d'attribuer un poids important aux substitutions rares et un poids faible aux substitutions très courantes. L'estimation de telles fréquences de substitution peut reposer sur l'exploitation de l'ensemble des séquences de protéines connues, telles qu'elles figurent dans les banques de données, issues elles même de l'« *atlas of protein sequence and structure* » de Dayhoff (1972). Ces auteurs proposent d'ailleurs une matrice de fréquence de substitution basée sur l'observation. Cette matrice carré 20x20, si on la suppose symétrique, ne comprend pas moins de 190 pondérations différentes. Même en effectuant des simplifications, par exemple en regroupant les acides aminés en fonction de leurs propriétés physico-chimiques, ce système de pondération reste difficile à mettre en place. Il repose, de plus, sur des estimations fondées sur un ensemble de données dont il n'est pas toujours facile d'apprécier la représentativité.

— pondérer en fonction du nombre de nucléotides impliqués dans la substitution d'un acide aminé en un autre. Cette pondération est donc fondée sur le code génétique. La difficulté principale est celle de l'inférence du nombre exact de substitutions de nucléotides impliqués par la substitution d'acides aminés. En raison de la dégénérescence du code génétique, cette inférence reste difficile.

Deux procédures au moins sont possibles. La première, celle de Moore *et al.* (1973), Moore (1976) et Goodman *et al.* (1978) consiste à ne pas effectuer une véritable pondération, mais à rechercher directement l'arbre qui minimise le

nombre de transformations de nucléotides, en inférant donc simultanément l'arbre et les différents nucléotides présent aux nœuds. La figure V.16 montre un exemple de reconstruction de la séquence nucléotidique ancestrale à partir de l'observation de la distribution des acides aminés. La figure V.17 est une analyse des parentés de seize des dix-huit ordres de *Mammalia* actuels à partir de sept protéines ; ces protéines ne sont néanmoins pas toutes séquencées chez chacun des taxons terminaux et les données manquantes sont optimisées. Cependant, en raison des multiples combinaisons possibles, l'efficacité de cette procédure fait l'objet de nombreuses controverses. Voir notamment des discussions dans Kimura (1981a), Allard (1990), Goodman (1990).

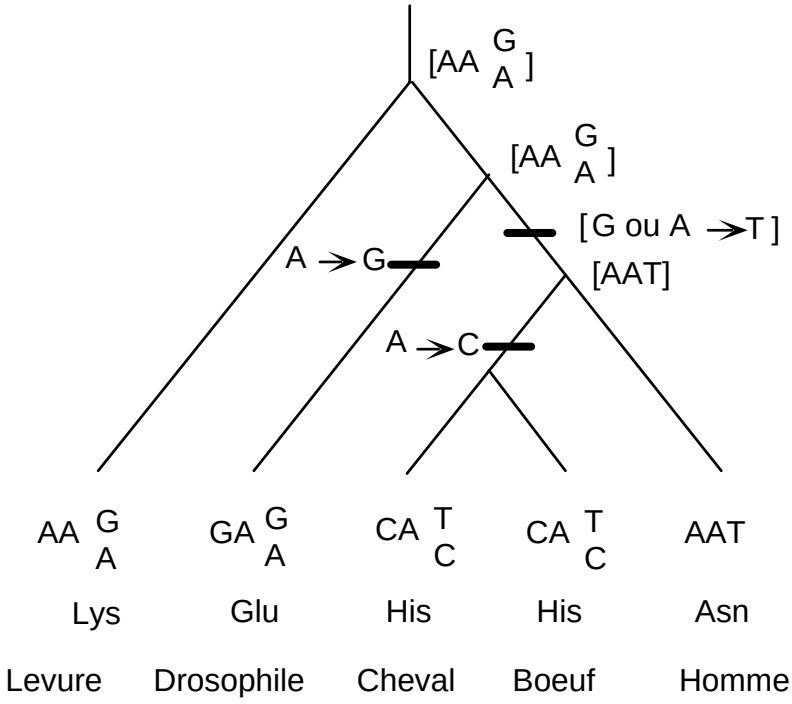


FIGURE V.16. *Distribution du site 19 de la Super-oxyde-dismutase (Lee* et al., *1985). A partir de la distribution des acides aminés, on peut tenter d'inférer l'arbre* et *les séquences de nucléotides aux nœuds qui nécessitent le nombre minimal de changement de nucléotides. Dans ce cas 3 changements de nucléotides suffisent. La levure est, ici, considérée comme extra-groupe. La séquence de l'Homme a été établit par Sherman* et al. *(1983).*

La seconde procédure consiste à effectuer une pondération basée sur le nombre minimum de changements de nucléotides nécessaires pour passer d'un acide aminé à un autre. Dans ce cas, l'arbre minimum est recherché en n'inférant aux nœuds que l'un ou l'autre des acides aminés présents dans les UE qui en descendent, à la différence donc de la procédure de Moore. Le nombre de pas tenant compte du code génétique est estimé ensuite. Considérons le cas simple de trois taxons de la figure V.16 : la drosophile, le cheval et le bœuf. Le cheval et le bœuf partagent une histidine (CAT ou CAC) et forment, ensemble, un groupe monophylétiques. La drosophile possède un acide glutamique (GAG ou GAA). Le nombre de transformations *minimum* pour passer de l'acide aminé ancêtre de

FIGURE V.17. *Relation de parenté de 16 ordres de Mammifères actuels (arbre de consensus de Adams, voir paragraphe 4.2.2.). L'arbre est construit à partir de 7 protéines (*α *et*β *hémoglobines, myoglobine, protéine* α*A du cristallin, fibrinopeptides A et B, cytochrome C). D'après Miyamoto et Goodman (1986).*

ces trois taxons à une histidine (chez le cheval et le bœuf) est de deux, l'une survenant sur le premier nucléotide et l'autre sur le troisième ; pour passer de la l'acide aminé ancêtre à un acide glutamique (chez la drosophile) il faut une transformation sur le premier nucléotide.

De façon générale, on conçoit que la substitution d'acides aminés la plus coûteuse que l'on puisse envisager correspond à celle demandant le changement de 3 nucléotides, comme par exemple celle d'un tryptophane (UGG) en asparagine (GAU ou GAC).

Il existe cependant des cas plus complexes, comme le remplacement d'une phénylalanine par une glutamine : UUC (ou UUU) ➞ CAA (ou CAG). Dans ce cas, au niveau des codons, il y a trois substitutions de nucléotides. En fait, en passant par un intermédiaire, on peut ne compter que deux pas si l'on considère que la substitution silencieuse en troisième position est très « facile » par rapport aux autres : UUC (Phe) ➞ CUC (leucine) ➞ CUG (leucine) ➞ CAG (glutamine). Si l'on suppose que la deuxième transformation CUC (leucine) ➞ CUG (leucine) a un poids nul, alors le remplacement d'une phénylalanine en glutamine ne demande effectivement que deux transformations et non trois. Cette approche est notamment celle proposée par J. Felsenstein (*Phylip,* programme *Protpars*) et applicable également dans PAUP version 3 de Swofford (1990).

3. L'enracinement de l'arbre

Les algorithmes de parcimonie, qu'ils soient exacts ou heuristiques, construisent des arbres enracinés ou non enracinés. La racine de l'arbre – le point de départ – peut être donnée par l'introduction d'un ancêtre dont, par définition, les états des caractères sont plésiomorphes. La racine de l'arbre peut aussi être indiquée par l'introduction d'un ou plusieurs extra-groupes.

3.1. Racine et ancêtre

Dans les cas où la racine de l'arbre est donnée par un ancêtre, celui-ci est toujours un ancêtre reconstruit. On admettra en effet que l'identification d'un ancêtre véritable préalablement à toute enquête phylogénétique est une rareté. La polarité des caractères étudiés pour les n taxons terminaux d'un groupe dont on cherche à reconstruire la phylogénie est donnée par l'opérateur : l'ancêtre reconstruit n'a que des caractères plésiomorphes par rapport aux taxons du groupe étudié. Cette option implique généralement que la monophylie du groupe étudié est admise par l'opérateur et n'a pas à être contrôlée.

3.2. Racine et extra-groupe(s)

Dans un arbre non enraciné, le taxon pris comme extra-groupe donne une orientation aux transformations des caractères et un ordre de lecture de la succession des branchements de l'arbre. La racine se situe sur la branche menant à l'extra-groupe et l'arbre se déploie à partir de la racine (figure V.18 construite à partir du tableau V.4).

Le choix de l'extra-groupe est crucial. On a pris comme exemple l'analyse de la phylogénie des mammifères au niveau des trois sous-classes : monotrèmes, marsupiaux et placentaires. La raison de ce choix est que ce groupe est familier et que le statut de groupe naturel des Mammalia n'est guère sujet à polémique. Ajoutons un quelconque sauropside à l'analyse. On constate sur la figure V.18 qu'à partir de la topologie de l'arbre non enraciné (figure V.18A), les monotrèmes ne sont le groupe frère des marsupiaux et des placentaires (figure V.18B)

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
saur	0	0	0	0	0	0	0	0	0	0	0	0	0
mono	0	1	0	1	0	1	1	1	1	0	1	0	1
mars	1	1	1	2	1	1	1	1	1	1	0	1	1
plac	1	1	1	3	2	1	1	1	1	1	0	0	0

TABLEAU V.4. *Matrice de caractères pour l'analyse de la phylogénie des mammifères (saur : sauropsides ; mono : monotrèmes ; mars : marsupiaux ; plac : placentaires; 1-13 : caractères).*

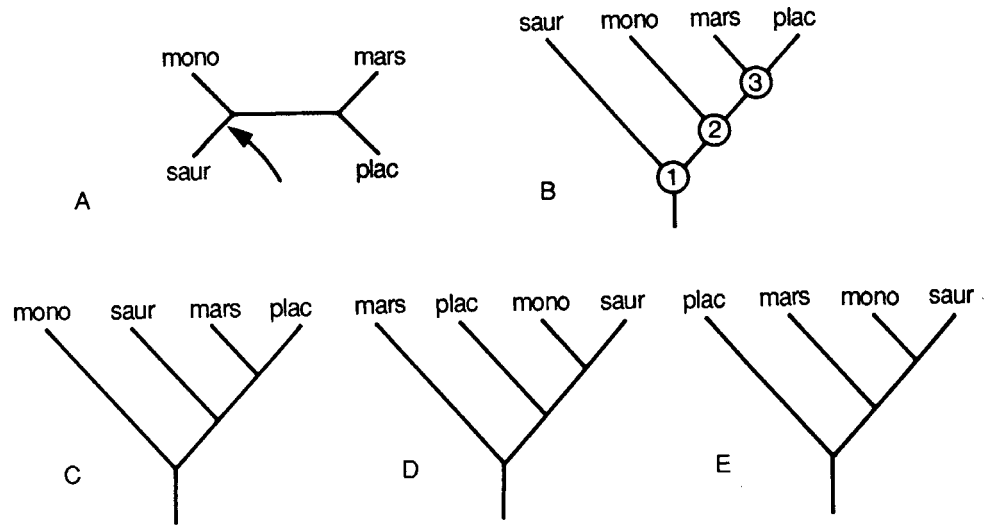


FIGURE V.18. *A : arbre non enraciné correspondant à l'analyse du tableau V.4. La flèche indique la racine de l'arbre B où les sauropsides sont pris comme extra-groupe. C-E : arbres enracinés où les monotrèmes, marsupiaux et placentaires sont respectivement pris comme extra-groupes.*

(autrement dit : les Mammalia sont monophylétiques) que si la racine de l'arbre est placée selon la flèche de la figure V.18A, c'est-à-dire si le sauropside (« saur ») est choisi comme extra-groupe. Chacun des quatre taxons terminaux peut être pris comme extra-groupe ce qui définit des parentés différentes (figure V.18B-E) qui restent toutes compatibles avec la topologie de l'arbre non enraciné.

Si l'on construit un arbre (figure V.19) où le sauropside est remplacé par un taxon hypothétique (« hyp » du tableau V.5) dont les caractères sont totalement différents de ceux de « saur » de la figure V.18, la topologie de l'arbre non enraciné est différente. Par exemple, marsupiaux et monotrèmes sont deux groupes frères si « hyp » est pris comme extra-groupe : ils ne le sont pas dans la figure V.18.

Cet exemple simple montre que 1) le choix de l'extra-groupe détermine, par ses caractères, la structure de l'arbre le plus parcimonieux, et 2) on ne peut identifier les groupes frères que si le point de départ de l'arbre est connu.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
hyp	1	1	1	3	2	1	1	1	1	1	1	1	1
mono	0	1	0	1	0	1	1	1	1	0	1	0	1
mars	1	1	1	2	1	1	1	1	1	1	0	1	1
plac	1	1	1	3	2	1	1	1	1	1	0	0	0

TABLEAU V.5. *Matrice de caractères pour l'analyse de la phylogénie des mammifères avec un extra-groupe hypothétique « hyp ».*

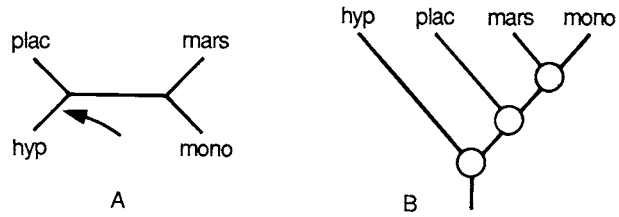


FIGURE V.19. *A : arbre non enraciné correspondant au tableau V.5. La flèche indique la racine de l'arbre B. B : arbre enraciné à partir de « hyp » comme extra-groupe.*

3.3. Racine : dichotomie et trifurcation

Dans la plupart des cas, les analyses où un seul extra-groupe est introduit impliquent que la monophylie du groupe étudié est considérée comme acquise. Par exemple, dans le cas de la figure V.18, on sait que les mammifères sont monophylétiques et l'on introduit un extra-groupe non mammalien (un lézard, ou un crocodile, ou un oiseau, ou encore une tortue, autrement dit un sauropside) afin de mettre en évidence les parentés entre les trois taxons terminaux, c'est-à-dire les branchements à l'intérieur des mammifères. Mais il n'y a pas ici de test de la monophylie du groupe étudié : les Mammalia. Les synapomorphies au nœud 2 de la figure V.18B ne sont pas identifiés avec certitude. L'analyse de parcimonie ne fait que proposer deux solutions. L'état dérivé des caractères 2, 4, 6, 7, 8, 9 (tableau V.4) peut effectivement définir le nœud 2 : l'état dérivé est 1 ((sens de la transformation 0 ➞ 1). L'état dérivé des caractères 2, 4, 6, 7, 8, 9 peut tout aussi bien définir (autapomorphies) le taxon « saur » : l'état dérivé est 0 (sens de la transformation 1 ➞ 0). Autrement dit, à partir des seules données du tableau V.4, la polarisation de ces caractères est impossible. Pour cette raison l'analyse de parcimonie de ce tableau conduit à la figure V.20 où la monophylie des Mammalia n'est pas attestée, bien que « saur » soit choisi comme extra-groupe. C'est pourquoi la racine est représentée sous forme d'une trifurcation.

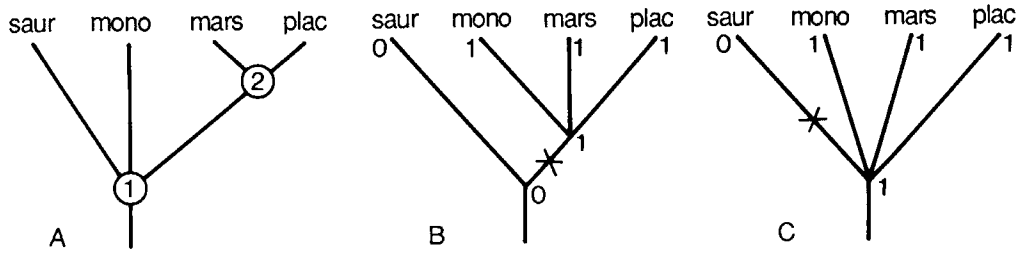


FIGURE V.20. *A : arbre enraciné construit à partir du tableau V.4. « saur » étant choisi comme extra-groupe, la racine est une trifurcation. B et C : les deux cladogrammes possibles du caractère 2 (un pas chacun).*

Si un seul extra-groupe est introduit, la trifurcation sur la figure V.20A souligne l'ambiguïté de la définition du groupe étudié. Enraciner l'arbre par une trifurcation explique l'impossibilité de choisir entre les figures V.20B et V.20C pour les caractères 2, 4, 7, 8, 9, où le monotrème « mono » ne se place pas à côté des autres mammifères. L'absence de résolution phylogénétique est ainsi résumée par la trifurcation basale. Par exemple, sur les figures V.20B et C, l'analyse de parcimonie du caractère 2 produit deux arbres (un pas). Sur l'arbre global où la racine est une trifurcation (figure V.20A), l'analyse de parcimonie du caractère 2 favorise en effet la transformation 1 ➞ 0 chez l'extra-groupe « saur » (un pas) plutôt que chez « mono » et au nœud 2 (deux pas).

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Saur 1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Saur 2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Mono	0	1	0	1	0	1	1	1	1	0	1	0	1	0	0
Mars	1	1	1	2	1	1	1	1	1	1	0	1	1	0	0
Plac	1	1	1	3	2	1	1	1	1	1	0	0	0	0	0

TABLEAU V.6. *Matrice des caractères pour l'analyse de la phylogénie des mammifères (« saur 1 » et « saur 2 » sont deux sauropsides).*

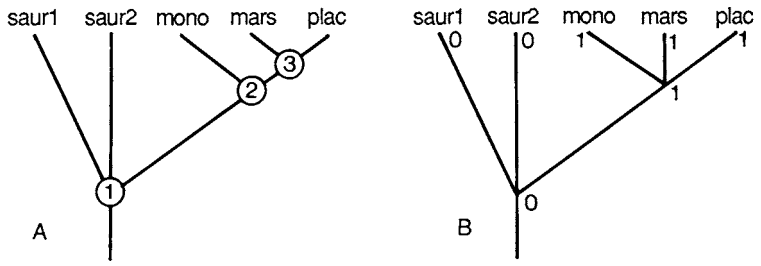


FIGURE V.21. *A : arbre enraciné au moyen d'une trifurcation, construit à partir du tableau V.6, « saur 1 » et « saur 2 » étant choisis comme extra-groupes. B : le cladogramme du caractère 2 (un pas).*

Si l'on veut contrôler la monophylie d'un groupe pris comme sujet d'étude (ici les *Mammalia*), l'application pure et simple du principe de parcimonie exige l'introduction de plusieurs extra-groupes (voir chapitre IV, paragraphe 4.1.1.) Les extra-groupes ne doivent pas être introduits dans l'analyse comme formant un taxon monophylétique, sinon, étant groupes frères, ils ne formeraient qu'un seul extra-groupe et nous serions ramenés à l'exemple précédent. Si deux sauropsides qui ne diffèrent que par deux caractères (tableau V.6), sont introduits comme extra-groupes, formant un groupe paraphylétique, le contrôle de la monophylie des *Mammalia* est positif (figure V.21). Dans la figure V.21, le caractère 2 est présent à l'état dérivé (qui est 1) au nœud 2. En effet, la phylogénie du caractère 2

soutient cette fois une proche parenté des trois groupes de mammifères (figure V.21A et B). La figure ne détaille que la distribution du caractère 2 : la situation est comparable pour les autres caractères 4,6,7,8,9 (une transformation 0 ➞ 1 pour chacun de ces caractères au nœud 2).

L'introduction de plusieurs extra-groupes paraphylétiques impliquant une trifurcation à la racine permet de contrôler la monophylie du groupe étudié. Elle permet de découvrir éventuellement lequel des extra-groupes est le groupe frère du groupe étudié et permet d'éviter des erreurs dues à un choix malheureux d'un seul extra-groupe qui serait trop divergent ou bien qui appartiendrait en fait au groupe étudié sans qu'on l'ait discerné préalablement. Le choix des extra-groupes est donc déterminant. Les extra-groupes supposés *a priori* comme tels, peuvent apparaître, du point de vue de la parcimonie, comme intérieurs au groupe étudié. Dans le cas d'une analyse où un seul taxon est choisi comme extra-groupe une racine dichotomique indique que la monophylie de l'ensemble des autres taxons est postulée.

4. Mesures de l'homoplasie et comparaisons d'arbres

Dans un ensemble de données, les synapomorphies, parce qu'elles définissent les parentés entre les UE, constituent l'information phylogénétique nécessaire à la construction d'un arbre. De ce fait, la quantité d'information phylogénétique contenue dans un ensemble de données peut être évaluée par la fréquence relative de ces synapomorphies. En revanche, plus la proportion d'homoplasie est importante, moins l'information phylogénétique est de qualité car elle se trouve alors noyée dans le « bruit » constitué par ces homoplasies. En l'absence totale d'homoplasie, c'est-à-dire lorsqu'un arbre particulier rend compte parfaitement de la distribution de tous les caractères, sans qu'il y ait de conflit, il est clair que l'information phylogénétique est maximale. Dans de multiples occasions, il s'avère utile de mesurer avec précision la quantité d'information phylogénétique contenue dans un ensemble de données.

Une autre préoccupation, liée à la précédente, concerne la comparaison de deux ou plusieurs arbres. En effet, deux arbres (ou plus) peuvent différer par leur structure, c'est-à-dire par les groupements monophylétiques qui les composent. Dans quelle mesure peut-on dire qu'une représentation est meilleure qu'une autre, de combien et pourquoi ? Telles sont les questions auxquelles ce chapitre se propose de répondre.

4.1. Mesures de l'homoplasie

La quantité de synapomorphies et, corrélativement, le degré d'homoplasie, sont habituellement estimés à l'aide de l'indice de cohérence *I.C.* (*consistency index* de Kluge et Farris, 1969). L'indice de cohérence *I.C.* d'un arbre est égal au rapport entre le nombre minimum (R) de transformations qui sont nécessaires pour expliquer les états de tous les caractères et le nombre effectif de transformations (L) dans l'arbre considéré. S'il s'agit de l'arbre le plus parcimonieux, L représente la longueur minimum de l'arbre.

Soit un caractère c qui peut se présenter sous s états distincts. L'amplitude r_c de ce caractère est égale à s – 1. Elle représente le nombre de transformations qui sont nécessaires pour rendre compte de tous les états du caractère. Par exemple, pour un caractère présent sous trois états additifs (0—1—2), r = 3 – 1 = 2.

L'amplitude totale *R*, estimée sur l'ensemble de K caractères d'une matrice de données est :

$$R = \sum_{c=1}^{K} r_c$$

si *L* est la longueur de l'arbre exprimée en nombre de transformations, l'indice global de cohérence *I.C.* est :

$$I.C. = \frac{R}{L}$$

Un arbre pour lequel l'indice de cohérence est strictement égal à 1 est donc dépourvu d'homoplasie. La différence ($L - R$) représente simplement le nombre d'homoplasies et l'inverse de l'*I.C.* constitue le nombre moyen de transformations par caractère, dans le cas de caractères uniquement binaires (H* de Sokal, 1983).

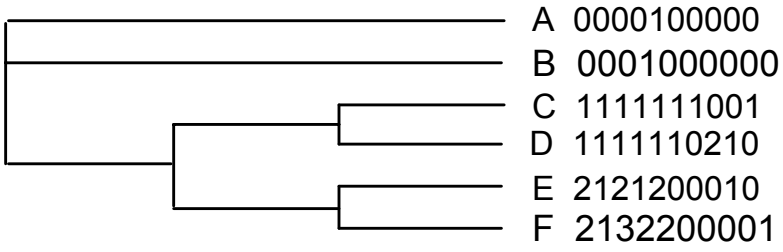
Un tel indice de cohérence *I.C.* présente quelques inconvénients. En effet il converge vers 1 quand le nombre d'autapomorphies augmente. De plus, la valeur minimale de *I.C.* n'est pas 0 mais égale au rapport entre *R* et la valeur maximale *L*. Elle se situe généralement autour de 0.2 lorsque les données sont « randomisées », c'est-à-dire lorsque les états des caractères sont distribués au hasard sur les UE. Des arbres ayant un *I.C.* autour de 0.25 ou 0.30 seront donc particulièrement riches en homoplasie et l'information phylogénétique que contiennent les données est donc faible. Dans ces conditions, il serait difficile, lors de la comparaison de deux arbres, de privilégier un arbre ayant un *I.C.* de 0.25 par rapport à un arbre ayant un *I.C.* de 0.29. Il existe par ailleurs une corrélation inverse entre l'indice de cohérence *I.C.* et le nombre d'UE, même lorsque la quantité d'homoplasie est tenue constante (Archie, 1989).

En conséquence, l'*I.C.* ne permet pas d'évaluer correctement le degré d'homoplasie d'un arbre. Pour pallier ces inconvénients, on peut corriger l'*I.C.* en ne comptant pas les transformations autapomorphiques, ou bien en excluant les caractères non informatifs (*C.I. excluding uninformative characters* de Swofford, 1989). Cela a pour conséquence naturelle de diminuer la valeur de l'indice, tout en reflétant mieux la proportion véritable d'homoplasie.

Un autre indice a été proposé par Archie (1989a et b). Il tente de rendre compte du « bruit de fond » sans signification phylogénétique qui contribue à donner à l'indice de cohérence une valeur toujours supérieure à 0. Il s'agit d'un indice d'excès relatif d'homoplasie (*homoplasy excess ratio*). Il se définit ainsi :

$$HER = \frac{M - L}{M - R}$$

Pour obtenir *M*, il faut d'abord transformer la matrice de données en distribuant au hasard les états des caractères dans les UE, tout en respectant cependant les proportions des états observés par caractère. On recherche ensuite la longueur de l'arbre le plus parcimonieux sur ces données ainsi randomisées.



CARACTERES	r_c	l_c	g_c
1	2	2	4
2	1	1	2
3	3	3	5
4	2	2	2
5	2	2	3
6	1	1	2
7*	1	1	1
8*	2	2	2
9	1	2	2
10	1	2	2
Total	R = 16	L = 18	G = 25

* = caractères non informatifs
I.C. = R/L = 0.889
I.C. sans * = (R — 3)/(L — 3) = 0.867
I.R. = (G — L)/(G —R) = 0.778

TABLEAU V.7. *Distribution de 10 caractères additifs sur 6 UE (A, B, C, D et E). L'arbre figuré est celui de longueur minimale (18 transformations).* r_c, l_c *et* g_c *sont respectivement l'amplitude du caractère c, le nombre observé de transformations et le nombre maximal de transformations pour le caractère c. I.C.et I.R. sont les indices de cohérence et de rétention.*

L'opération est répétée un grand nombre de fois afin d'estimer une longueur moyenne *M* qui représente donc le nombre moyen de transformations observées sur des arbres de longueur minimum obtenus par randomisation des données.

Il est possible d'obtenir une approximation de cet indice, sans avoir à effectuer la procédure décrite précédemment. On peut en effet remplacer *M* par *G*, le maximum de transformations requis par les données pour construire un arbre quel qu'il soit. *G* peut être calculé à partir du nombre d'UE, du nombre de caractères et du nombre d'états par caractère. En fait *G* correspond au nombre de pas qui serait nécessaire si tous les changements d'états ne survenaient que le long des branches terminales de l'arbre. L'indice que l'on obtient est alors (*Homoplasy excess ratio maximum*, Archie, 1989a et b) :

$$HERM = \frac{G - L}{G - R}$$

G est toujours supérieur à *M*. De ce fait *HERM* est une surestimation de *HER* et sous-estime donc l'homoplasie.

Dans cette dernière formulation, *HERM* représente le rapport entre le nombre d'homoplasies observables et le nombre d'homoplasies observées : c'est l'indice de rétention *I.R.* (*retention index* de Farris (1989), calculé par son logiciel Hennig86). Les discussions sur les qualités respectives de ces indices peuvent se trouver dans Archie (1990) et Farris (1990, 1991).

Exemple :

Soit la matrice constituée de 6 UE et 10 caractères additifs (tableau V.7). L'arbre non enraciné le plus parcimonieux est donné dans le tableau V.7. Les caractères 7 et 8 présentent uniquement des transformations autapomorphes. Ils sont donc non informatifs. Examinons le cas du caractère 3 codé (0 — 1— 2— 3). Le nombre minimum de transformations, c'est-à-dire l'amplitude du caractère 3, est $r_3 = 3$. Le nombre de transformations nécessaires pour expliquer la distribution des quatre états sur l'arbre est également de 3 (donc pas d'homoplasie). Le nombre maximum de transformations est celui que l'on observerait si toutes les transformations survenaient sur les branches terminales. Pour déterminer ce nombre, il faut partir d'un état ancestral qui soit le plus représenté et le plus central dans l'ordre des états, afin de *minimiser* ce nombre *maximum* de transformations. Dans le cas présent c'est l'état 1, et le nombre de transformations maximum est $g_3 = 5$: il faut en effet 2 pas pour observer 0 chez A et B à partir de l'état 1, 1 pas pour observer l'état 2 chez E et 2 pas pour observer l'état 3 chez F. Remarquons que si l'on supposait que l'état ancestral était l'état 0 ou 2, on aurait $g_3 = 7$.

4.2. Les arbres de consensus

Dans ce paragraphe, on n'abordera que le problème de la comparaison de deux ou plusieurs arbres, généralement de même longueur. Seuls les tests sur le paramètre « structure de l'arbre » est donc traité ici, tandis que les tests concernant les autres paramètres (longueur des branches par exemple), seront traités au niveau de chacune des méthodes de reconstruction (chapitre VII pour les méthodes de distance ; chapitre VIII pour les méthodes de vraisemblance).

Lorsqu'un même ensemble de données conduit à l'obtention de plusieurs arbres d'une longueur totale équivalente, il n'est généralement pas de critères permettant de déterminer si l'un de ces arbres est meilleur qu'un autre, sauf à faire appel à d'autres critères extrinsèques ou intrinsèques (comme la pondération successive (paragraphe 4.3). C'est pourquoi il est recherché une représentation de ces arbres telle que leurs parties concordantes apparaissent clairement par rapport aux parties discordantes. Cette représentation est appelée *arbre de consensus*. Il en existe principalement deux : l'« arbre de consensus strict » (Sokal et Rohlf, 1962, 1981) et l' « arbre de consensus d'Adams » (Adams 1972).

4.2.1. L'arbre de consensus strict

Cet arbre (*strict consensus tree* de Sokal et Rohlf (1962, 1981) ou *general cladogram* de Nelson (1979), dit encore « arbre de Nelson ») est construit en ne retenant des arbres comparés que les groupements de taxons qui sont identiques dans tous les arbres. Les points de conflits sont représentés par des multifurcations.

Prenons l'exemple simple de la figure V.22 représentant deux arbres (ou portions d'arbre) entièrement dichotomiques (T1 et T2) composés de 4 taxons. Le seul groupement de taxons qui soit commun à T1 et T2 est celui formé par A d'un côté et l'ensemble (B,C,D) de l'autre. Aucun autre groupe ne se retrouve dans les deux arbres. Dans ces conditions l' « arbre de consensus strict » est celui donné en mettant B, C et D au même niveau hiérarchique. Il faut bien préciser que cette représentation ne signifie pas une spéciation triple, c'est-à-dire un point à partir duquel les trois UE auraient évoluées indépendamment. Son but n'est que de rendre compte de l'impossibilité de conclure quant aux relations de parenté entre B, C et D. Cette représentation simplifie même plus qu'il n'est nécessaire. En effet elle ne permet plus de remarquer que la combinaison (B,D) monophylétique n'est pas observée. Malgré cet inconvénient, l'arbre de consensus strict est celui qui est le plus utilisé lors des opérations de comparaison d'arbres.

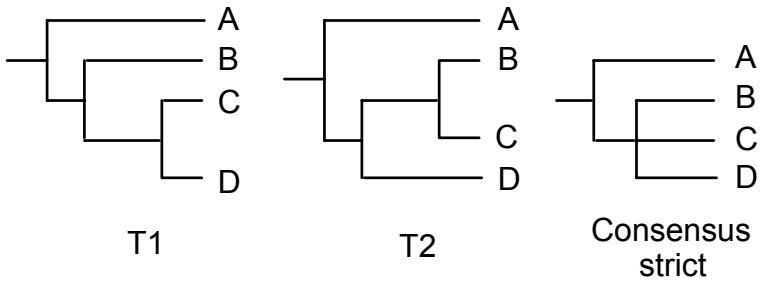


FIGURE V.22. *Construction d'un arbre de consensus strict.*

4.2.2. L'arbre-consensus d'Adams

L'arbre-consensus d'Adams discuté ici est plus connu sous le nom d'«arbre-consensus Adams-2 ». Il correspond à l'arbre de consensus proposé par Adams (1972) à partir d'arbres entièrement dichotomiques et où seule la racine de l'arbre est supposée connue, les autres nœuds étant seulement déduits de la reconstruction. Cette méthode consiste, en partant de la racine, à comparer, entre deux ou plusieurs arbres, les deux sous-ensembles de taxons qui découlent d'une dichotomie. S'il existe un recouvrement entre ces sous-ensembles observés sur les différents arbres, ce recouvrement constitue un groupement « consensus » de taxons.

L'exemple simple de la figure V.23 permet de comprendre le processus de construction. L'arbre (ou portion d'arbre) T1 identifie, *à partir de sa racine*, deux sous-ensembles de taxons, (A) et (B,C,D,E), tandis que l'arbre T2 identifie deux autres sous-ensembles : (A,B,C) et (D,E). En croisant les sous-ensembles de l'un des arbres avec les sous-ensembles de l'autre, on effectue des intersections d'ensembles qui définissent trois sous-ensembles : (A), (B,C) et (D,E) qui

permettent de construire l'arbre de consensus d'Adams-2 représenté sur la figure V.23. Cet arbre montre une trifurcation qui, comme dans le cas du consensus strict, ne s'interprète pas comme telle. Si B, C ou (D,E) étaient eux-même des ensembles de taxons montrant une parenté différente d'un arbre à l'autre, le processus décrit précédemment serait de nouveau appliqué à partir de la racine de ces ensembles.

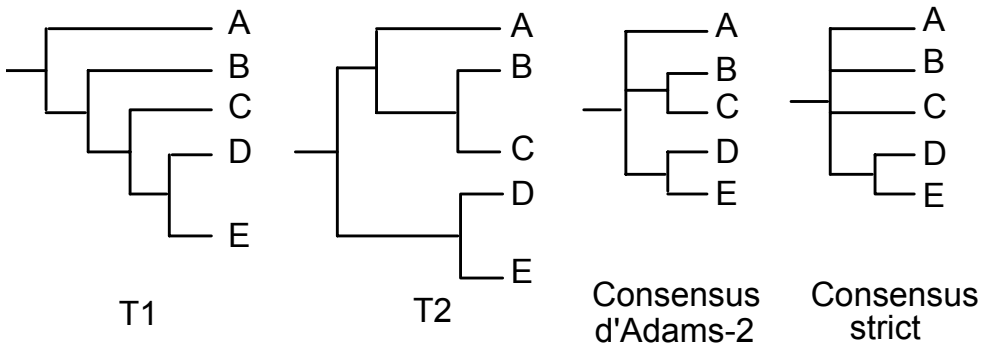


FIGURE V.23. *Construction d'un arbre de consensus d'Adams-2.*

Cette représentation a l'avantage de souligner que B et C ont un ancêtre commun qui est distinct de la racine de l'arbre (ou de la portion d'arbre). L'inconvénient de cette représentation est de considérer parfois comme monophylétiques des taxons qui ne le sont pas nécessairement sur tous les arbres comparés. Ainsi C est étroitement apparenté à D dans l'arbre T1 alors qu'il est étroitement apparenté à B dans l'arbre T2. Le groupement (B,C) retenu dans l'arbre de consensus d'Adams ne peut évidemment pas s'interpréter comme s'il s'agissait d'un consensus strict, puisque le groupe (B,C) n'est pas monophylétique dans T1. Quand une monophylie est observée sur un arbre de consensus strict, elle se retrouve également sur un consensus d'Adams-2, alors que l'inverse n'est pas exact.

Dans l'exemple de la figure V.22, l'arbre de consensus d'Adams-2 est identique à l'arbre de consensus strict. En revanche, l'arbre de consensus strict de la figure V.23 est différent de l'arbre de consensus d'Adams-2.

4.2.3. L'arbre de consensus majoritaire (Majority rule consensus tree)

Dans la comparaison de plusieurs arbres présentant des topologies différentes, il est possible de rechercher les groupes monophylétiques qui se rencontrent le plus fréquemment parmi l'ensemble des arbres comparés (Margush et McMorris, 1981). Ainsi, dans la figure V.24, parmi les trois arbres comparés (T1, T2 et T3), on observe deux fois sur trois le groupe monophylétique formé de C et D. Puisque ce groupe (C,D) est majoritaire, il est représenté dans l'arbre de consensus. On peut également choisir de représenter l'arbre de consensus seulement à partir des groupes monophylétiques présents dans au moins 50% des arbres ou dans au moins 75% ou tout autre pourcentage. Cette méthode de construction d'arbre de consensus est celle généralement utilisée dans les méthodes de ré-échantillonnage décrites plus loin.

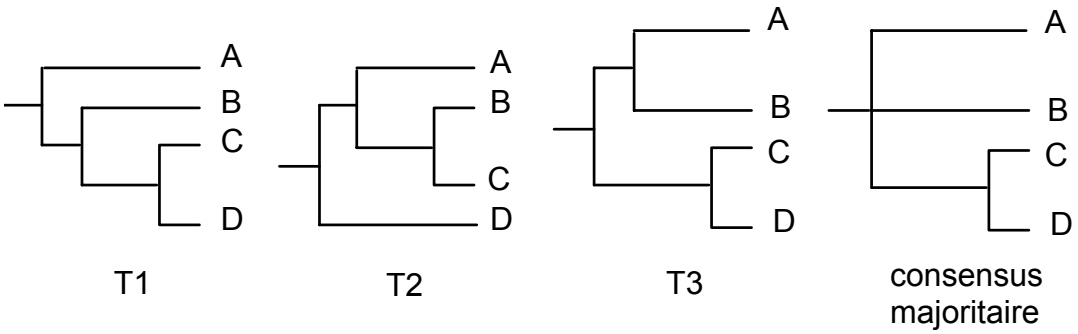


FIGURE V.24. *construction d'un arbre de consensus majoritaire.*

4.3. Pondération successive

Considérons le cas de données pour lesquelles l'analyse de parcimonie classique, sans pondération particulière des caractères, conduit à plusieurs arbres différents de même parcimonie. Que faire d'un tel résultat ? On peut discuter les incidences de tous ces arbres, choisir l'un de ces arbres pour des raisons extérieures à l'analyse elle-même, ou bien considérer un arbre de consensus. La méthode de pondération successive (*successive weighting* de Farris, 1969) offre une autre possibilité.

Les pondérations de caractères dont il a été question dans le paragraphe précédent, étaient des pondérations *a priori*, effectuées avant l'analyse de parcimonie. La pondération successive est, au contraire, une pondération *a posteriori,* effectuée après l'analyse de parcimonie.

L'idée centrale est qu'il est préférable de choisir, parmi tous les arbres ayant le même nombre minimal de pas, celui qui donne le moins de poids aux caractères homoplasiques.

Pour cela, on pourrait pondérer chaque caractère par son indice de cohérence. Un caractère non homoplasique (I.C.=1) aurait un poids plus élevé qu'un caractère présentant de l'homoplasie (I.C.< 1) et une analyse de parcimonie avec de tels caractères pondérés reviendrait à minimiser l'impact des caractères homoplasiques sur la longueur totale de l'arbre (voir chapitre VII sur la compatibilité). Cependant, comme l'indice de cohérence varie entre 1 et une valeur supérieure à 0, Farris (1989) propose d'utiliser plutôt un indice qui soit strictement compris entre 1 et 0 et qui est, en fait, le produit entre l'indice de cohérence I.C. et l'indice de rétention I.R. (*rescaled consistency index*).

Les indices de cohérence et de rétention de chacun des caractères sont d'abord calculés pour chacun des arbres de même parcimonie. La valeur de l'indice retenue par caractère sera soit la valeur la plus élevée obtenue sur l'ensemble des arbres de même parcimonie (option maximale), soit la valeur moyenne des indices de ces arbres (option moyenne). Par exemple, lorsqu'un caractère est cohérent avec tous les arbres, c'est-à-dire qu'il n'est pas homoplasique, il lui sera attribué un poids de 1. En revanche, si un caractère présente un indice de cohérence de 0.33 pour un arbre, de 0.50 pour un autre et 0.66 pour un troisième, l'indice de cohérence retenu sera la valeur la plus élevée observée (0.66 : option maximale) ou sa valeur moyenne (0.50 : option moyenne).

GOMPHOTHERIUM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ELEPHAS	1	1	1	1	1	1	1	1	1	1	1	1	2	?	?	?	?	1	0	0	0	1	1	2	1	1	1	1	1	1	1	1	0	0	1	?
MAMMUTHUS	1	1	1	1	1	1	1	1	1	1	1	0	2	?	?	?	?	1	0	0	0	1	1	2	1	1	1	1	1	1	1	1	0	0	1	?
LOXODONTA	1	1	1	1	1	1	1	1	1	1	1	1	2	?	?	?	?	1	0	0	0	1	1	2	1	1	1	1	1	1	1	0	0	0	1	?
PRIMELEPHAS	1	?	?	1	1	1	?	?	?	?	1	1	?	?	?	0	0	1	?	0	0	0	1	?	1	1	1	1	0	0	1	?	0	0	1	1
STEGODIBELODON	?	?	?	?	1	1	1	1	?	?	1	0	2	?	?	0	0	1	?	0	0	0	1	?	1	1	1	1	1	0	0	?	0	0	1	?
STEGOTETRABELODON	1	?	1	1	1	1	1	1	1	?	0	1	2	1	1	0	0	1	0	0	0	0	1	?	1	1	1	1	0	0	1	?	0	0	1	1
MALUVALENSIS	?	?	?	?	?	1	?	?	?	?	1	0	2	?	?	0	?	1	?	0	1	0	0	?	?	0	0	1	?	0	0	?	0	0	1	1
GIGANTOROSTRIS	?	?	?	?	?	1	?	?	?	?	0	0	2	1	1	0	0	1	?	0	0	0	1	?	?	0	0	0	0	0	0	?	2	0	1	1
PARATETRALOPHODON	1	1	1	1	1	?	1	0	0	?	?	?	?	?	?	?	?	1	0	0	0	0	1	1	0	0	0	0	?	0	?	0	2	0	1	?
STEGOLOPHODON	1	0	1	1	1	1	0	1	1	1	1	0	2	?	?	0	?	0	0	1	1	0	0	?	0	0	0	1	?	0	0	0	0	0	1	1
STEGODON	1	1	1	1	1	1	1	1	1	1	1	0	2	?	?	?	?	0	1	1	1	0	1	0	1	1	0	1	1	0	1	0	0	0	1	?
TETRALOPHODON	0	0	0	0	1	0	1	0	0	0	0	1	1	1	0	0	0	1	0	0	0	0	0	?	0	0	0	0	0	0	0	0	1	1	1	1
ANANCUS	0	1	1	0	1	1	1	1	0	0	1	0	2	?	?	?	?	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	2	1	?

TABLEAU V.8. *Distribution de 36 caractères chez 14 taxons faisant partie des Proboscidiens (D'après Tassy et Darlu, 1987).*

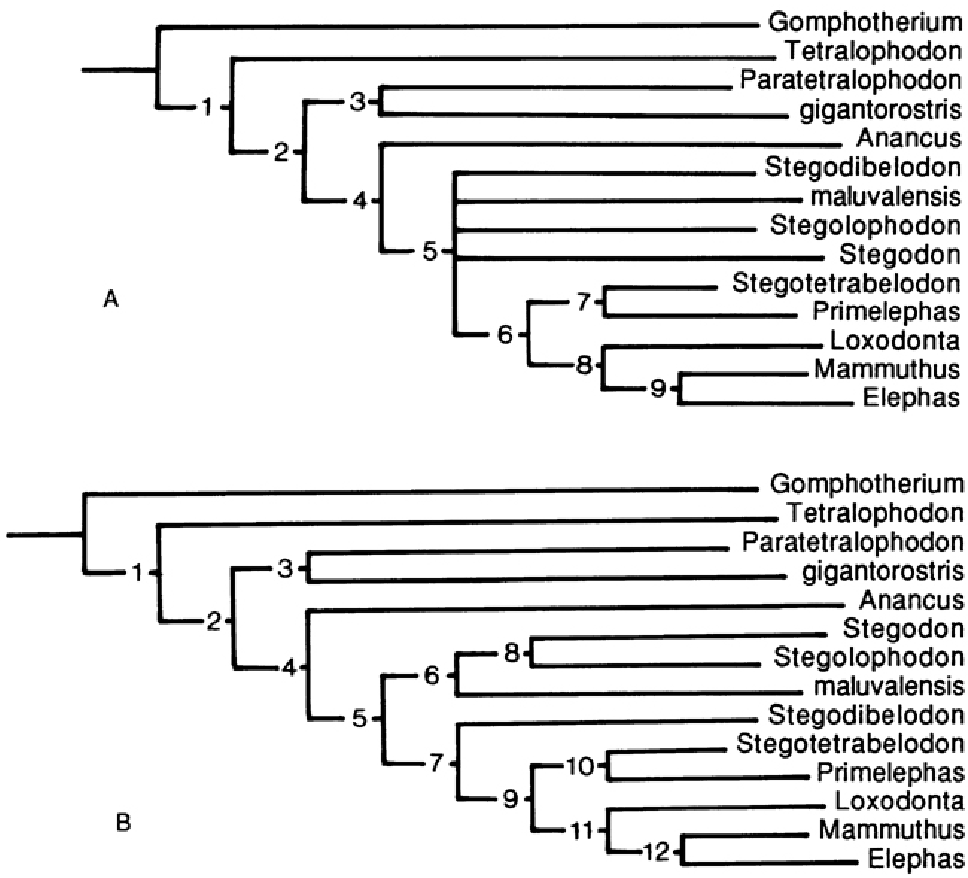


FIGURE V.25. *Arbres obtenus à partir du tableau* V.8. *A : arbre de consensus strict ; B : arbre unique entièrement dichotomique obtenu après pondérations successives.*

De même pour l'indice de rétention. En définitive, le poids attribué à un caractère homoplasique sera d'autant plus faible, donc inférieur à 1, que ses indices de cohérence et de rétention seront faibles.

L'analyse de parcimonie est ensuite effectuée en utilisant cette pondération. Si plusieurs arbres de même parcimonie, bien qu'en nombre inférieur à celui de l'analyse initiale, sont encore obtenus après cette pondération, une nouvelle pondération est calculée et une nouvelle analyse de parcimonie utilisant cette nouvelle pondération est alors effectuée. Ce processus itératif est stoppé lorsque l'on ne peut réduire davantage le nombre d'arbres.

Exemple 1 :

L'analyse de tableau V.8 fournit 7 arbres de même parcimonie (53 pas, C.I. = 0.69 ; R.I. = 0.82). L'arbre consensus strict (figure V.25A) montre que la parenté des taxons terminaux *Stegodibelodon*, *maluvalensis* , *Stegolophodon* et *Stegodon* n'est pas résolue. La pondération successive à partir de ces 7 arbres permet de sélectionner un seul arbre entièrement dichotomique (figure V.25B). L'analyse détaillée de trois caractères permet d'illustrer la méthode (tableau V.9). Dans le cas présent, l'extra-groupe *Gomphotherium* a été dédoublé afin d'asseoir la monophylie des 13 autres taxons (nœud 1).

CARACTERES	1	3	12
pas	2	1	3
I.C.	0.5	1	0.3
I.R	0.6	1	0.5
POIDS = IC∗IR	0.3	1	0.15

TABLEAU V.9. *Pondération des caractères. Le meilleur indice de cohérence du caractère 1 sur l'ensemble des 7 arbres de même parcimonie est de 0.5, le meilleur indice de rétention est de 0.6. Le poids donné à ce caractère dans l'analyse de parcimonie sera de 0.3 (option maximale). Les poids inférieurs attribués aux caractères 1 et 12 par rapport au caractère 3, qui est toujours cohérent, et le poids supérieur attribué au caractère 1 par rapport au caractère 12 permettent de sélectionner un seul arbre parmi les 7 arbres initiaux.*

Exemple 2 :

Un autre exemple peut être fourni par la comparaison des deux arbres de même parcimonie obtenus à partir de la matrice du tableau V.10. Ce tableau correspond à la phylogénie illustrée par la figure V.26.

Cette phylogénie est considérée comme vraie : elle servira d'exemple à de multiples reprises en permettant de comparer les résultats obtenus par différentes méthodes d'analyse phylogénétique. Cette phylogénie se caractérise par des quantités de transformations évolutives différentes dans les groupes frères (vitesses d'évolution inégales) et une homoplasie importante (plus du tiers des caractères) – qui n'implique aucune réversion mais seulement des convergences.

```
          1 1111111112 2222222223 3333333334 444444444
   1234567890 1234567890 1234567890 1234567890 123456789

a  1000000000 0100000011 1000000000 0000000100 100000000
b  0100000000 0011000011 0100001000 0111100100 100000000
c  0010000000 0011111111 0000101000 0000000101 100000000
d  0001100000 0011111111 0000001000 0000000100 100000000
e  0000100000 0000000011 0000000000 0000000000 100000000
f  0000010000 0000000000 1000000001 1000000000 100000000
g  0000001000 0000000000 1111000000 0000000000 100000000
h  0000000100 0000000000 1111110000 0000000000 100000000
i  0000100010 0000000000 1111111100 0000011010 100000000
j  0000100001 1000000000 1111111110 0000000110 100000000
k  1000100000 1000000000 1111111110 0000000110 100000000
l  1000100000 0100000000 1111111110 0000000110 100000000
m  0000000000 0100000000 0000000110 0000000000 001111000
n  0000001000 0000000000 0000011000 0000000000 010000111
```

TABLEAU V.10. *Matrice des caractères construite à partir de l'arbre de la figure V.25. Une analyse de parcimonie de cette matrice donne deux arbres résumés par l'arbre de consensus strict de la figure V.26.*

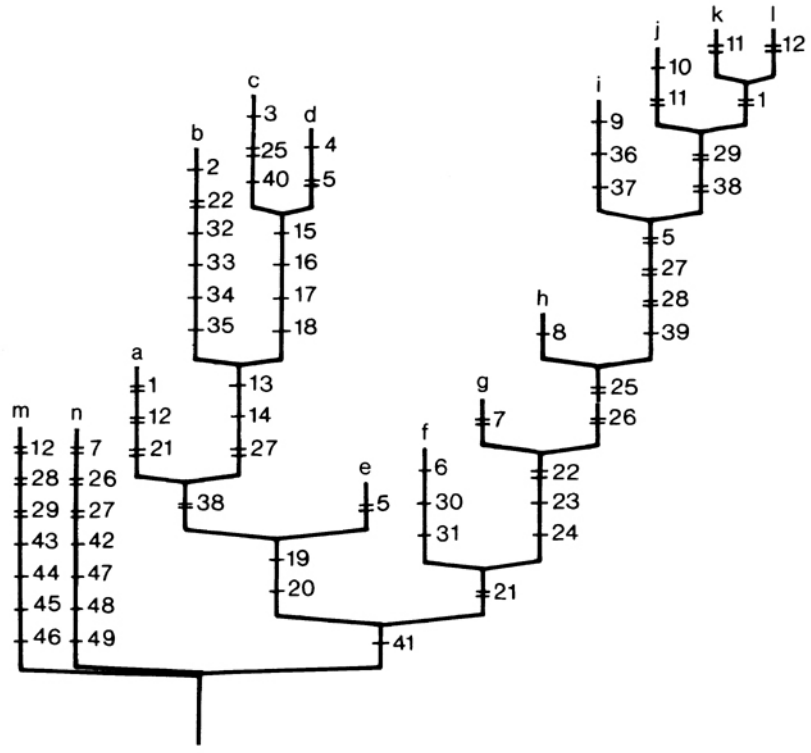


FIGURE V.26. *Phylogénie choisie pour tester différentes méthodes de reconstruction. Les états de caractères sont donnés dans le tableau V.10. Un double trait indique la localisation des homoplasies (convergences). Les taxons m et n sont des extra-groupes.*

L'analyse de parcimonie fournit deux arbres (voir l'arbre de consensus strict, figure V.27) dont l'un correspond à la phylogénie théorique illustrée par la figure V.26.

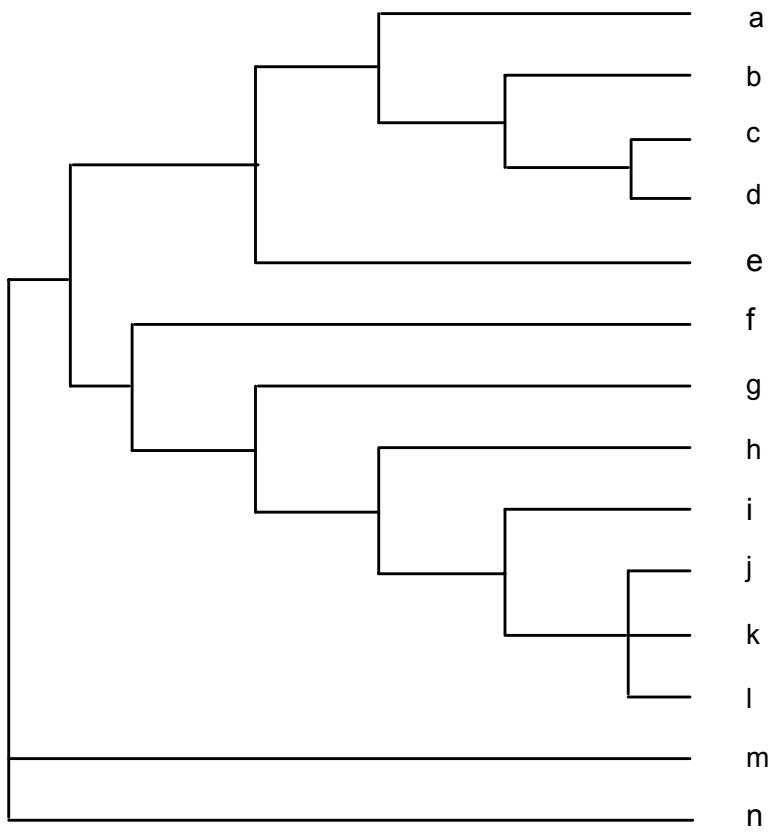


FIGURE V.27. *Arbre de consensus strict obtenu par parcimonie à partir du tableau V.10 (66 pas ; I.C. = 0.74 ; I.R. = 0.78). Cet arbre résume deux arbres dichotomiques de 65 pas (I.C. = 0.75 ; I.R. = 0.80).*

Les deux arbres diffèrent par l'organisation des taxons j, k et l. Le premier (arbre 1) donne le groupement (j,(k,l)) et le deuxième (arbre 2) donne le groupement (l,(j,k)). La différence tient à l'interprétation des caractères 1 et 11, comme le résume le tableau V.11, dans lequel sont détaillés les éléments permettant de calculer la pondération (option maximale).

La pondération successive conduit à préférer l'arbre 2 (l,(j,k)) parce que cet arbre est la solution qui possède un caractère non homoplasique (caractère 11, poids de 1), associé à un caractère très homoplasique (caractère 1, poids nul). Par opposition, l'arbre 1 associe deux caractères modérément homoplasiques (caractères 1 et 11, I.C. = 0.5) mais dont l'un (caractère 11) a un poids nul. L'option moyenne conduit à la même conclusion. La solution de l'arbre 2 ne correspond pas à la phylogénie théorique (voir figure V.26) : celle-ci implique en effet que les deux caractères 1 et 11 sont homoplasiques.

		Caractère 1	Caractère 11
	R	1	1
Arbre 1	L	2	2
(j,(k,l))	G	3	2
	I.C.	0.5	0.5
	I.R.	0.5	0.0
	POIDS	0.25	0.0
	R	1	1
Arbre 2	L	3	1
(l,(j,k))	G	3	2
	I.C.	0.33	1.0
	I.R.	0.0	1.0
	POIDS	0.0	1.0

TABLEAU V.11. *Calcul de la pondération des caractères 1 et 11 de la matrice du tableau V.10 pour les deux arbres d'égale parcimonie représentés figure V.27 sous forme de consensus. La pondération successive (option maximale) permet de privilégier l'arbre 2 par rapport à l'arbre 1. Pour la signification de R, L, G, se reporter au paragraphe 4.1.*

4.4. Les méthodes de ré-échantillonnage

Un problème difficile est celui de l'évaluation de la confiance que l'on peut avoir en un arbre, en un groupe monophylétique, en une longueur de branche.

La statistique classique n'est pas armée pour répondre à cette question, essentiellement parce que les distributions de probabilité des paramètres à estimer sont généralement inconnues (ainsi en est-il de l'arbre ou des longueurs des branches) ou ne peuvent s'exprimer en termes simples.

Une façon de contourner la difficulté consiste à faire appel aux méthodes de ré-échantillonage (*resampling methods*) développées par Efron (1979, 1982), c'est-à-dire les méthodes de *Jackknife* et de *Bootstrap*. Un aperçu général de ces méthodes appliquées aux données phylogénétiques a été présenté par Felsenstein (1988). Toutes ces méthodes supposent que les caractères ont évolués indépendamment les uns des autres et suivent tous une même loi de distribution. Ces restrictions importantes posent le problème de leur applicabilité aux données morphologiques pour lesquelles aucune hypothèse plausible n'est formulable en termes de probabilité. Enfin, ces méthodes sont applicables quelle que soit la façon dont les arbres sont obtenus, que ce soit par des méthodes phénétiques (Chapitre VII) ou par parcimonie.

4.4.1. Le Jackknife

Cette méthode a été appliquée aux problèmes de phylogénie par Mueller et Ayala (1982). Supposons une matrice de données constituée de K caractères. Le

Jackknife consiste à effectuer K reconstructions phylogénétiques différentes, chacune d'elles ayant été obtenue en supprimant un caractère différent.

Par exemple, si les données sont constituées de K fréquences géniques, on calcule d'abord une matrice de distances en omettant la première fréquence, puis une deuxième en supprimant la deuxième fréquence etc. jusqu'à la $K^{ième}$ matrice calculée en supprimant la dernière fréquence K. Un arbre est reconstruit à partir de chacune de ces K matrices différentes. La perturbation de la matrice de données par l'abandon d'une seule fréquence parmi les K disponibles est généralement très faible, lorsque K est assez grand.

La procédure à suivre est la même si l'arbre est reconstruit par parcimonie. Dans ce cas, on construit autant d'arbres qu'il y a de caractères dans la matrice, chacun de ces arbres étant construit par suppression de l'un des K caractères de la matrice.

Lorsque les structures des arbres obtenus sur ces K différentes matrices sont les mêmes, on peut alors tester une certaine longueur de branche L de la façon suivante.

Si L est la longueur estimée en utilisant simultanément les K caractères de la matrice de données et L^* celle obtenue en utilisant K–1 caractères, l'estimation de la longueur $\hat{L}$ est donnée par :

$$\hat{L} = nL - (n-1)L^*$$
$$\hat{L} = n(L - L^*) + L^*$$

Des tests peuvent être effectués (t de Student par exemple) pour savoir si $\hat{L}$ est significativement différente de 0.

4.4.2. Le Bootstrap

Cette méthode (Efron, 1979 ; Felsenstein, 1985b) consiste à tirer au hasard avec remise un ensemble de K caractères parmi les K caractères constituant les données. Ce tirage se faisant avec remise, cela signifie que le nouvel échantillon, constitué, lui aussi, de K caractères, peut contenir des caractères présents plusieurs fois, car retirés après remise, et, au contraire, d'autres caractères absents, n'ayant jamais été tirés. Cela revient à pondérer les caractères de manière aléatoire.

Le nouvel échantillon fait ensuite l'objet d'une analyse phylogénétique (par méthode cladistique ou phénétique) conduisant à l'obtention d'un arbre.

Cette procédure de ré-échantillonnage peut être effectuée N fois, suivie chaque fois par une recherche d'arbre. En fin de *bootstrap*, on est en possession de N arbres qui peuvent, éventuellement, être différents.

Si l'on souhaite tester l'existence d'une monophylie particulière (ici définie comme un ensemble d'UE, quelle que soit l'organisation phylogénétique interne à cet ensemble), il suffit de dénombrer combien de fois on la retrouve parmi les N arbres. Si l'on donne la valeur 1 à la présence et 0 à l'absence de la monophylie que l'on souhaite tester, le paramètre testé est l'occurrence de la monophylie. Par exemple, une monophylie retrouvée dans 95% des échantillons signifie qu'il y a 5 chances sur 100 de se tromper en disant que la monophylie n'existe pas. Le cas de tests multiples est développé par Felsenstein (1985b, 1988).

Par ailleurs, le nombre de tirages aléatoires doit être aussi élevé que possible, le nombre minimum étant dépendant du nombre de caractères et du degré d'homoplasie.

Cette méthode s'applique aussi bien aux méthodes phénétiques qu'aux méthodes de parcimonie.

Exemple 1

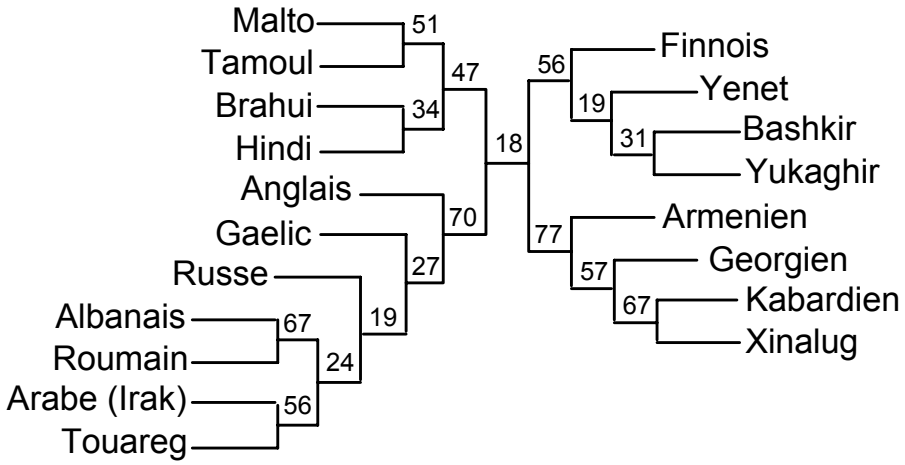


FIGURE V.28. *Arbre de consensus (*majority rule*) obtenu par la comparaison des arbres différents obtenus à l'issue de 100 ré-échantillonnages (*bootstrapping*) des données, chacun d'eux étant suivi de la recherche de l'arbre le plus parcimonieux (méthode de Wagner) (Darlu* et al.*, 1990; Darlu, 1992). Il s'agit ici de données linguistiques de natures phonologique (181 caractères), syntaxique (16), et grammaticale (77). Elles ont été relevées sur 18 langues eurasiennes (Ruhlen, 1976). La valeur attachée à un nœud indique combien de fois (sur 100) se retrouve la dichotomie qu'il occasionne sur l'ensemble des 18 langues. Par exemple, sur 100 ré-échantillonnages, le groupe constitué de l'Albanais, du Roumain, de l'Arabe et du Touareg ne se retrouve que 24 fois, alors que le groupe constitué de l'Albanais et du Roumain se retrouve 67 fois.*

Exemple 2

Les données du tableau V.10 ont été analysé par la méthode du *bootstrap*. Il a été effectué 100 ré-échantillonnages qui ont produit 100 arbres de longueur minimale. Cet arbre de consensus (figure V.29) reproduit la solution de parcimonie retenue à l'issu d'une pondération successive (tableau V.11). On observe une corrélation étroite entre le nombre de synapomorphies définissant les groupes monophylétiques dans l'analyse de parcimonie et le nombre de fois où se retrouvent ces groupes monophylétiques sur 100 ré-échantillonnages. Ainsi la valeur la plus élevée (98) correspond aux 4 synapomorphies du groupe (c,d) tandis que la valeur la plus faible (42) correspond à la seule synapomorphie définissant (j,k).

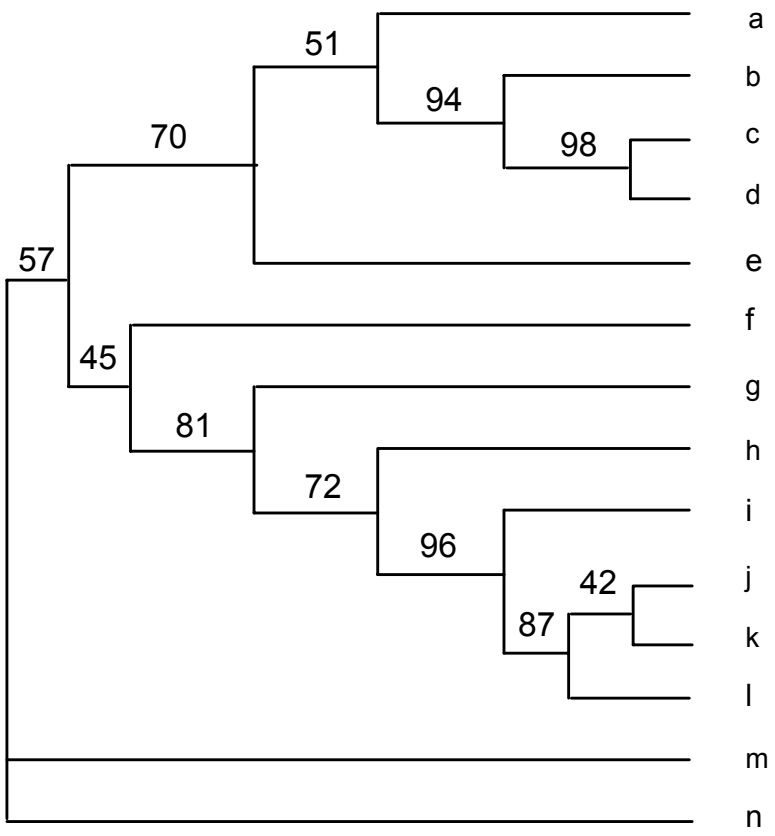


FIGURE V.29. Arbre de consensus (majority rule) *obtenu par la comparaison des arbres différents obtenus par la méthode de parcimonie, à l'issue de 100 ré-échantillonnages (*bootstrappings*) des données du tableau V.10.*

Certains auteurs ont proposé que le ré-échantillonnage s'effectue sur les UE plutôt que sur les caractères (Lanyon, 1985). Cependant, dans ce cas, la condition nécessaire d'applicabilité évoquée plus haut et postulant que les tirages doivent être indépendants, n'est évidemment pas satisfaite puisque les UE n'évoluent pas de manière indépendante, sinon l'arbre de parenté n'existerait pas. D'un point de vue pragmatique, cette approche peut cependant donner des informations intéressantes, même si elles ne sont pas interprétables en terme de théorie des tests.

5. Les invariants

La question de la pondération des transformations (entre deux états, 0 et 1, ou entre quatre états, comme les quatre nucléotides) a été vue sous l'angle de la parcimonie au paragraphe V.2.3. Cette question peut aussi être abordée à partir des méthodes probabilistes traitées au chapitre VIII.

Les approches de parcimonie et les méthodes probabilistes présentent des relations qui seront discutées plus loin (Chapitre VIII). L'une d'elle est la dépendance vis-à-vis d'hypothèses *a priori* sur les probabilités de changement des

caractères, et sur le degré d'égalité dans les *vitesses* de changement le long des différentes branches de l'arbre. C'est pour s'affranchir de telles contraintes que Cavender (1978, 1981, 1989), Cavender et Felsenstein (1987), Lake (1987a et b), Sankoff (1990) ont proposé diverses méthodes fondées sur la recherches d'« invariants ». Il s'agit de trouver une relation entre les différentes distributions possibles des états des caractères des différentes UE, relation qui ne dépende que de la structure de l'arbre et qui ne soit vérifiée que pour une structure particulière d'arbre et non pour les autres.

5.1. Les invariants de Cavender

Prenons l'exemple, qui sera repris dans le chapitre VIII, de quatre UE (A, B, C, D) et d'un ensemble de N caractères présents sous deux états 0 ou 1. Chacun de ces N caractères peut se répartir sur les 4 UE selon 16 combinaisons possibles (0000, 1000, 1100, 1110, 0100, 0110, ...1111), en notant les états d'un caractère dans l'ordre des UE (A, B, C et D). Ces 16 combinaisons peuvent se regrouper dans les 8 catégories suivantes (*p* et *n* représentant respectivement la probabilité et le nombre de caractères parmi N qui présentent cette combinaison) :

0000 et 1111 p_1 n_1
1000 et 0111 p_2 n_2
0100 et 1011 p_3 n_3
0010 et 1101 p_4 n_4
0001 et 1110 p_5 n_5
1100 et 0011 p_6 n_6
1010 et 0101 p_7 n_7
1001 et 0110 p_8 n_8

Comme on le sait, quatre UE peuvent s'organiser selon les 3 arbres non enracinés différents T_1, T_2 et T_3 de la figure suivante :

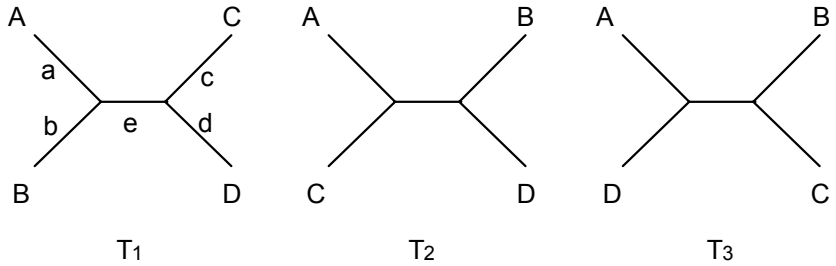


FIGURE V.30. *Les trois arbres non enracinés possibles pour quatre taxons A, B, C et D. a, b, c, d et e représentent les branches de l'arbre.*

Si l'arbre véritable est de structure T_1, la probabilité pour que les UE A et B possèdent toutes les deux le même état (0 ou 1) pour un caractère donné ne dépend évidemment que des événements situés sur la branche a et la branche b et non des événements qui peuvent se produirent ailleurs dans

l'arbre. Cette probabilité est $p_1+p_4+p_5+p_6$. Le même raisonnement tient également pour les UE C et D : la probabilité pour que l'état en C soit identique à l'état en D est égale à $p_1+ p_2+p_3+p_6$. Par ailleurs ces deux événements (A=B) et (C=D) sont des événements indépendants étant donnée la structure de l'arbre T_1.

Il est donc possible de tester si toutes ces conditions sont bien remplies, simplement en vérifiant ces hypothèses d'indépendance par un test χ^2 effectué sur le tableau de contingence suivant, obtenu à partir du décompte, dans les données analysées, des 8 différentes catégories définies plus haut :

	C = D	C ≠ D	
A = B	n_1+n_6	n_4+n_5	$n_1+n_4+n_5+n_6$
A ≠ B	n_2+n_3	n_7+n_8	$n_2+n_3+n_7+n_8$
	$n_1+n_2+n_3+n_6$	$n_4+n_5+n_7+n_8$	N

Si les conditions d'indépendance sont satisfaites, l'invariant L_1 de l'arbre T_1 est nul :

$$L_1 = (n_1 + n_6)(n_7 + n_8) - (n_2 + n_3)(n_4+ n_5) = 0$$

Deux autres tableaux de contingence similaires peuvent être construits et deux autres invariants calculés L_2 et L_3, l'un pour l'arbre T_2 et l'autre pour l'arbre T_3.

La structure de l'arbre pour laquelle les hypothèses d'indépendance ne seraient pas rejetées à un seuil de signification donné (5% par exemple) sera retenue.

Cette approche repose sur l'hypothèse que les probabilités de changement d'état obéissent à un processus de Markov : les conditions d'équilibre de ce processus, qui sont celles adoptées par la méthode, supposent une symétrie des changements d'états (même probabilité de changer 0 en 1 ou 1 en 0). Elle suppose également que la probabilité d'observer un état de caractère chez une UE est la même pour toutes les UE. Enfin l'hypothèse doit également être faite que tous les caractères changent, de manière indépendante, avec la même vitesse. Remarquons qu'il n'est fait aucune hypothèse sur les valeurs des probabilités de changement d'état (sauf qu'elles doivent être inférieures à 0.5), ni sur les variations possibles de cette probabilité selon les branches de l'arbre.

Il est possible d'étendre cette approche à des caractères ayant plus de deux états (par exemple les 4 nucléotides possibles en un site) (Felsenstein, 1983) et à plus de 4 UE (Sankoff, 1990).

5.2. Les invariants de Lake

Cette méthode, développée par Lake (1987a et b), est aussi appelée « *Evolutionary parsimony method* ». Elle s'applique à des données nucléotidiques, ADN ou ARN. A la différence des invariants de Cavender où

seuls étaient considérés deux états par caractère, ici chaque caractère (le site) peut se trouver sous 4 états différents, les 4 nucléotides. Comme dans les invariants de Cavender, cette méthode se propose de tester les 3 arbres non enracinés que l'on peut construire à partir de 4 UE (voir figure V.30). Elle se fonde sur l'observation des fréquences des 256 combinaisons possibles de nucléotides pour 4 UE (au lieu des 16 dans le cas des invariants de Cavender).

Cette méthode fait l'hypothèse que les transversions (figure V.15) sont plus rares que les transitions. Pour cette raison, *seules les transversions sont considérées comme pertinentes pour estimer la structure de l'arbre*. Les transitions sur les branches ne constituent donc que du « bruit » qui masque éventuellement l'information phylogénétique des transversions. L'hypothèse est également faite que les différentes transitions sont équiprobables (A⇔G, T ou U⇔C) comme le sont les différentes transversions (A⇔T, A⇔C, T⇔G, G⇔C). Enfin tous les sites doivent évoluer indépendamment les uns des autres. En revanche, à la différence des invariants de Cavender, il n'est pas nécessaire qu'ils évoluent à la même vitesse.

Le but des invariants de Lake est d'estimer le nombre d'événements de type « transversion » qui ont pu survenir sur la branche centrale de l'arbre T_1, T_2 ou T_3.

Considérons donc 4 UE (A, B, C, D). Pour représenter la distribution chez ces quatre UE des nucléotides d'un site donné, on utilise la règle suivante : Le chiffre 1 est donné, arbitrairement, à A (A=1). Si B possède le même nucléotide que A, alors B=1. Si B diffère de A par une transition, alors B=2. Si B diffère de A par une transversion, B=3. Par ailleurs, C peut avoir le même nucléotide que A ou B. Il est alors codé comme eux. Il en est de même pour D. Si A=B et que C (ou D) diffère de A par une transition, C (ou D) sera codé 2 et s'il diffère par une transversion, il sera codé 3. Enfin si C ou D possèdent un nucléotide qui différe des autres par une transversion non identique à une transversion déjà observée, il sera codé 4.

Exemples : CGGC : 1331 ; UGAU : 1341 ; UCGA : 1234 ; GUAU : 1323 etc.

Imaginons maintenant que l'arbre véritable soit de type T_1 (Figure V.31), et intéressons-nous aux sites codés 1133. Du point de vue de la parcimonie, il est clair que ces sites codés 1133 plaident tous en faveur de l'arbre T_1 (Figure V.31.I). En effet la substitution de deux nucléotides identiques chez A et B aux deux nucléotides identiques chez C et D implique au minimum une transversion sur la branche centrale.

Lake montre que cette conclusion peut être erronée, en raison des différentes transitions ou transversions qui peuvent survenir sur les branches. En effet, tous

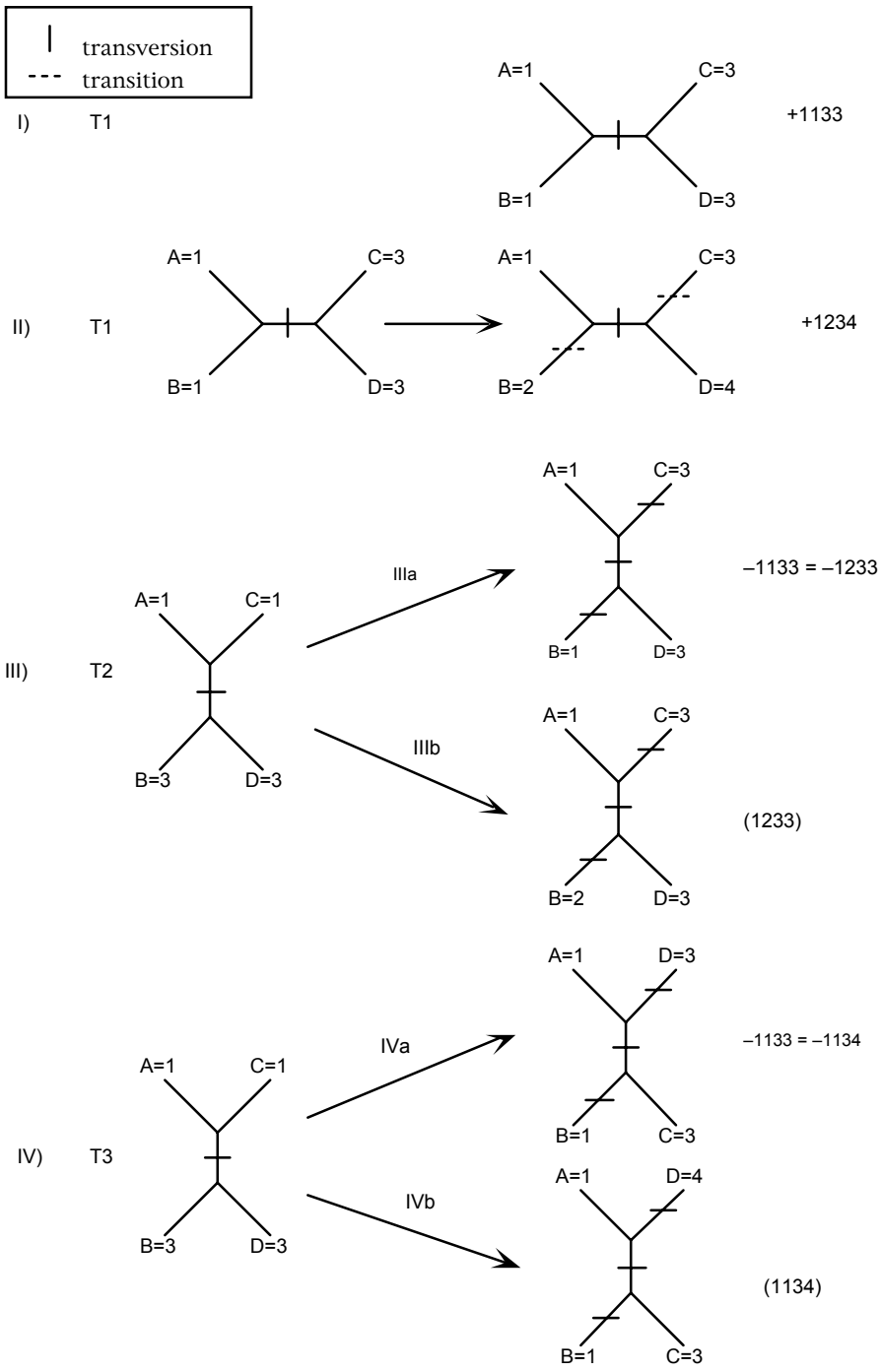


FIGURE V.31. *Schéma montrant comment des transitions survenant sur les branches externes de l'arbre T_1 (II) ou des transversions survenant sur les branches de l'arbre T_2 ou T_3 conduisent à des conclusions erronées concernant le nombre de transversions survenant sur la branche centrale de l'arbre T_1 (I). En raison de l'équiprobabilité des transversions, la fréquence des distributions conduisant à IIIa et IIIb sont identiques. De même pour IVa et IVb.*

les sites 1133 ne plaident pas *exclusivement* en faveur de T_1 car certains de ces sites codés 1133 peuvent fort bien provenir d'un arbre qui serait de type T_2 et où deux branches externes porteraient chacune une transversion supplémentaires (figure V.31.III). Ces substitutions supplémentaires, survenant sur les branches terminales de l'arbre T_2 « miment » un arbre que la parcimonie identifierait comme étant de type T_1 (Figure V.31.IIIa). Autrement dit, s'il survient des événements de transversions « parallèles », l'arbre inféré par la parcimonie n'est plus le bon. On comprend bien que ces sites 1133 qui ne sont pas en faveur de T_1 mais de T_2 ne doivent pas être pris en compte pour démontrer, à partir de l'observation des combinaisons 1133, que l'arbre est bien T_1. Il faut donc les décompter. La question se pose de savoir comment. Il se trouve que l'on peut en avoir une estimation. En effet, en raison de l'hypothèse effectuée (l'équiprobabilité des événements substitutifs de même nature, transition ou transversion), il y a autant de sites 1133 issus d'un arbre T_2, avec transversions sur les branches terminales (figure V.31.IIIa), qu'il y a de sites codés 1233, d'où le signe « = » entre 1133 et 1233 dans la figure V.31.IIIa. On peut donc corriger en négatif le nombre de sites 1133 plaidant en faveur de T_2 par la quantité de sites 1233, d'où le signe négatif de la figure V.31.IIIa.

De la même façon, certains sites codés 1133 peuvent aussi bien provenir d'arbres de type T_3 où les branches terminales auraient subi une transversion (Figure V.31.IVa). On peut avoir une estimation de ces derniers dans la mesure où ils sont en principe aussi nombreux que les sites codés 1134 (Figure V.31.IVb).

Enfin, des transitions survenant en parallèle sur les branches terminales conduisent à « camoufler » un certain nombre de sites codés 1133 dans l'hypothèse où l'arbre est bien T_1 (Figure V.31.II). On peut en avoir une estimation en comptant les sites codés 1234.

En conclusion, le nombre $n(1133)$ de sites de type 1133 qui sont exclusivement en faveur de la structure T_1, après correction des artefacts dus à des transitions ou transversions non informatives, constitue l'invariant L_1 propre à la structure T_1 :

$$L_1 = n(1133)+n(1234) - n(1233) - n(1134)$$

Les deux autres invariants, L_2 et L_3, définissant les structures T_2 et T_3 s'écrivent, respectivement :

$$L_2= n(1313)+n(1324) - n(1323) - n(1314)$$
$$L_3= n(1331)+n(1342) - n(1332) - n(1341)$$

En tenant ce raisonnement, Lake démontre que, finalement, seules les 12 combinaisons de nucléotides (sur les 256 possibles) qui permettent de calculer les invariants sont véritablement informatives pour choisir entre les structures T_1, T_2 et T_3.

Lorsque l'arbre T_1 est l'arbre véritable, on s'attend à une valeur de L_1 différente de 0 (indiquant qu'il y a des transversions sur la branche centrale reliant A et B d'une part à C et D de l'autre), tandis que les deux autres invariants L_2 et L_3 seront tous deux égaux à 0.

Les règles de décisions en faveur d'un arbre sont donc :

T_1 si $L_1 > 0$ et $L_2 = L_3 = 0$
T_2 si $L_2 > 0$ et $L_1 = L_3 = 0$
T_3 si $L_3 > 0$ et $L_1 = L_2 = 0$

Comme l'échantillon de sites observés dans une telle comparaison entre quatre UE est nécessairement limité, se pose le problème de la signification statistique de ces égalités et inégalités à zéro. Il est possible d'effectuer un χ^2 (Lake, 1987a) ou un test binomial exact (Holmquist *et al.*, 1988).
Dans le cas de la structure T_1 par exemple, ce dernier test consiste à comparer $n^+ = n(1133) + n(1234)$ et $n^- = n(1233) - n(1134)$, qui doivent être égaux.
On calcule donc la probabilité pour que $n^+/(n^+ + n^-)$ soit différent de 1/2 à un seuil donné.

Comme on l'a vu, cette méthode des invariants de Lake repose sur l'idée que seules les transversions sont pertinentes pour reconstruire une phylogénie. Ce point peut être discuté, particulièrement lorsque l'on s'intéresse à la phylogénie d'UE qui se sont différenciées depuis peu (12 à 14 millions d'années) et pour lesquelles le nombre de transversions est nécessairement faible (voir l'exemple 2 où l'on n'observe que 6 transversions significatives sur plus de 10000 sites, pour la phylogénie des Primates). Dans de telles conditions, les substitutions de nucléotides de type « transition » ne sont certainement pas dépourvues d'information phylogénétique et ne doivent donc pas être négligées.

Pour être valable, cette méthode des invariants suppose également que les différentes transitions sont équivalentes, tout comme le sont les différentes transversions. Ce point reste également discutable. Enfin cette méthode, comme bien d'autres d'ailleurs, ne prend pas en compte les délétions ni les insertions qui, dans bien des cas pourtant, peuvent être déterminantes dans la reconstruction phylogénétique. En revanche elle ne nécessite pas, à la différence des invariants de Cavender, de faire l'hypothèse que les vitesses d'évolution soient les mêmes pour tous les sites. Par ailleurs cette méthode permet également d'effectuer des estimations des longueurs de branches en comptant les événements qui s'y sont produits (Lake, 1987b). Cette application particulière de la méthode de Lake ne sera pas développée ici. On peut en trouver un exemple dans Holmquist *et al*. (1988) sur la phylogénie des Primates.

Exemple 1

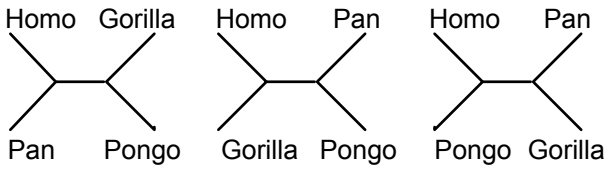
Les séquences d'une portion de l'ADN ribosomique 28S ont été comparées chez 4 espèces : La souris (*Mus musculus*), le riz (*Oryza sativa*), un champignon (*Saccharomyces cerevisiae*) et un procaryote (*Escherichia coli*). Ces séquences ont été alignées en même temps que 16 autres séquences d'autres espèces (Baroin *et al*., 1988).

T_1 = (*M. musculus,O. sativa*) (*S. cerevisiae,E. coli*) $L_1 = 5$; P = 0.09
T_2 = (*M. musculus,S. cerevisiae*) (*O. sativa,E. coli*) $L_2 = 2$; P = 0.38
T_3 = (*M. musculus,E. coli*) (*O. sativa,S. cerevisiae*) $L_3 = -8$; P = 0.99

A partir de ces données, le choix entre les différentes structures n'est donc pas possible. Si l'on acceptait cependant un risque de se tromper de 9%, on pourrait conclure à la structure T_1.

Exemple 2

Les séquences nucléotidiques de la ψ–η globine (Miyamoto *et al.*, 1987) (6901 sites), d'une région entre la ψ–η globine et la δ globine (Maeda *et al.*, 1988) (3145 sites) et de l'ADN mitochondrial (Brown *et al.,* 1982) (893 sites) ont été analysées par la méthode des invariants de Lake par Holmquist *et al.* (1988). Aucune séquence n'est à elle seule décisive (au seuil de 5%) dans le choix d'un arbre. Seul le cumul des données permet d'atteindre un seuil de signification raisonnable. Dans la figure V.32, les nombres sous chaque arbre sont les valeurs des invariants (P<0.03 d'erreur en rejetant l'hypothèse T_1).



sites				
6901	Myamoto *et al.* (1987)	3	0	0
3145	Maeda *et al*. (1988)	0	1	0
893	Brown *et al.* (1982	3	–1	0
10939	Tous	6 P<0.03)	0	0

FIGURE V.32. *Application de la méthode des invariants de Lake à diverses séquences nucléotidiques chez les hominoïdes (Holmquist* et al. *1988).*

6. L'évolution est-elle parcimonieuse ?

La nécessité logique du principe de parcimonie consiste à ne pas envisager plus de changements évolutifs qu'il n'est nécessaire pour construire une hypothèse de parenté. L'usage de ce principe a-t-il des implications sur l'inférence du processus évolutif lui-même ? Est-il indépendant de toute considération sur les probabilités de ces changements ? Dans un contexte phylogénétique la question- titre de ce paragraphe peut être formulée autrement : les homoplasies sont-elles rares ? Si la réponse est négative, si l'on admet que l'évolution n'est pas parcimonieuse, doit-on en conclure que le principe de parcimonie nous induit en erreur ? Nous amène-t-il à reconstruire des arbres et à attribuer des états de caractères aux nœuds qui soient erronés ?

Ces questions sont la source de vives controverses. Selon Felsenstein (1978b) par exemple, l'usage des méthodes de parcimonie implique que l'on admet que l'évolution est parcimonieuse : les événements évolutifs doivent être rares et, *a fortiori,* les homoplasies encore plus rares. Au contraire, selon Farris (1983)

l'application du principe de parcimonie ne renvoie à aucun modèle évolutif particulier et n'exige pas que l'homoplasie soit rare : la méthode est jugée libre de toute contrainte.

Il n'existe pas de réponse simple à ces questions car elles revêtent plusieurs aspects. L'un concerne la signification du cladogramme, l'autre concerne la question de la définition d'un modèle évolutif.

Le cladogramme peut-il être tenu pour un arbre phylogénétique sans connaissance préalable des modalités d'évolution des caractères ? Selon le point de vue cladistique la réponse est positive, à la condition toutefois d'admettre que toutes les observations se valent, que les caractères évoluent indépendamment et d'admettre le postulat de la descendance avec modification. Selon ce postulat, l'information phylogénétique (signal) est intelligible en terme d'hypothèse d'homologie (synapomorphie). Toute hypothèse de non-homologie (homoplasie) est une hypothèse non phylogénétique (bruit), une hypothèse *ad hoc.*

L'approche hypothético-déductive vise à privilégier l'information phylogénétique, autrement dit à minimiser le nombre des hypothèses *ad hoc.* Cela revient à maximiser le nombre des hypothèses d'homologie (à leur niveau de synapomorphies). C'est sur cette opération de maximisation des synapomorphies et de minimisation des homoplasies que reposent des points de vue contradictoires.

Selon certains, une seule contradiction (une seule homoplasie) réfute le système : un cladogramme serait infirmé s'il renfermait ne serait-ce qu'une homoplasie. Autrement dit, le système phylogénétique ne pourrait fonctionner qu'en l'absence totale de bruit. Cette position se retrouve, sous une forme moins catégorique, dans les méthodes de compatibilité qui recherchent l'arbre pour lequel le *nombre* de caractères sans homoplasie est *maximal*. Cela revient simplement à rejeter les caractères homoplasiques, considérés comme du « bruit », et à ne s'intéresser qu'aux autres. Mais, même si cette approche (voir chapitre VI) rejette le bruit, elle n'en reconnaît pas moins son existence.

Un autre point de vue est fondé sur un raisonnement probabiliste concernant le processus évolutif lui-même. Ce raisonnement consiste à attribuer une probabilité aux changements d'état des caractères. Comme dans l'approche cladistique, le partage de caractères dérivés entre deux UE peut être un signe de parenté, mais peut également survenir « par hasard », à la suite de deux événements indépendants, constituant ainsi une homoplasie. Ce qui différencie l'approche probabiliste de l'approche cladistique dans leur recherche des parentés est la façon de considérer l'homoplasie. En raison des hypothèses qu'elle pose sur les probabilités de changement, l'approche probabiliste donne la possibilité d'estimer la part des caractères dérivés qui peuvent être partagés par hasard par deux groupes frères de ceux qui sont partagés en raison d'un ancêtre commun. Ce point de vue sera discuté plus loin sous un autre aspect (Chapitre VIII).

Si le processus évolutif produit en réalité une quantité importante d'homoplasies, l'application du principe de parcimonie – la maximisation des synapomorphies, c'est-à-dire la minimisation des homoplasies - ne nous induit-elle pas en erreur ?

Farris (1983) a proposé un exemple devenu un cas d'école, commenté favorablement par Sober (1985, 1988) et Tassy (1991), tendant à démontrer que

l'application du principe de parcimonie n'implique pas que l'homoplasie soit rare. Soit l'observation de dix caractères, chacun sous deux états (0 primitif et 1 dérivé) chez trois UE : A, B et C.

A 1111111110
B 1111111111
C 0000000001

Le raisonnement de Farris est le suivant. La distribution des caractères 1 à 9 suggère l'arbre ((A,B)C). La distribution du caractère 10 suggère l'arbre (A(B,C)). L'application du principe de parcimonie permet d'opter pour le premier arbre. Il nécessite 9 transformations synapomorphiques et 2 transformations par convergence (soit 11 pas) tandis que l'arbre (A(B,C)) nécessite 1 transformation synapomorphique et 18 transformations par convergence (19 pas). L'hypothèse nulle : pas de parenté, implique 20 transformations par convergence (20 pas).

Admettons maintenant que l'évolution ne soit pas parcimonieuse en ce sens que les homoplasies ne sont pas rares. Supposons donc qu'un seul caractère, parmi les 10, présente un état dérivé partagé d'origine généalogique (c'est-à-dire qu'il existe une seule homologie sur 10 caractères). S'il en est ainsi, lequel des dix caractères n'est pas homoplasique ? L'homologie, à son niveau de synapomorphie, a plus de chances d'être parmi les 9 caractères dérivés partagés par A et B que d'être le seul caractère partagé par B et C. Autrement dit, dans une situation où l'homoplasie est fréquente, le choix de l'arbre ((A,B)C) reste le meilleur pari. Ce pari correspond à la solution la plus parcimonieuse. Ce raisonnement rentre dans le cadre d'une réflexion statistique sur l'échantillonnage des caractères.

Or, il est possible de contourner l'argument de Farris en suivant précisément une approche probabiliste. Telle est la démonstration apportée par Forster (1986). Ce dernier fait remarquer, en effet, que l'on ne peut affirmer que les 9 caractères dérivés partagés supportant l'arbre ((A,B)C) sont *tous* des synapomorphies, puisque, parmi eux, peuvent se trouver des homoplasies. C'est pourquoi Forster souhaite pouvoir distinguer entre les caractères dérivés partagés hérités d'un ancêtre commun, par définition les synapomorphies, et les caractères dérivés partagés seulement « par hasard », constituant donc des homoplasies (figure V.33).

Pour réaliser concrètement cette distinction, d'un point de vue probabiliste, on a besoin de disposer d'une estimation de la fréquence des caractères dérivés par taxon et de la fréquence des synapomorphies entre deux taxons. Le raisonnement est le suivant.

Considérons trois taxons A, B et C sur lesquels sont observés N caractères sous deux états (0, plésiomorphe et 1, apomorphe).

Soit f(A=1) et f(B=1) les fréquences des caractères apomorphes respectivement chez A et B. Ces fréquences sont estimées à partir de l'observation des N caractères chez A et B. Cette estimation n'incorpore aucune hypothèse de descendance ou d'ancêtre. On raisonne simplement ici en terme de caractères dérivés observés dans chacune des UE étudiées. Les différentes fréquences des apomorphies chez A, B et C sont donc des paramètres propres à A, B et C respectivement.

Soit *f*(A=1 ; B=1) la fréquence des apomorphies rencontrées simultanément chez A et B.

Le nombre *S*(A,B) d'apomorphies partagées par A et B *et* héritées d'un ancêtre commun (synapomorphies) est simplement la différence entre le nombre total des apomorphies partagées par A et B, *Nf*(A=1 ; B=1), et l'estimation du nombre d'apomorphies partagées du seul fait du hasard, *Nf*(A=1)*f*(B=1). Ce dernier est, en effet, *N* fois le produit des fréquences estimées des apomorphies chez A et chez B.

$$S(A,B) = Nf(A=1; B=1) - Nf(A=1)f(B=1)$$

Autrement dit, plus les apomorphies sont fréquentes dans deux UE, plus grandes sont les chances de trouver, par hasard, des apomorphies partagées par ces deux UE. Pour Forster, seule *S*(A,B), qu'il appelle la covariance entre A et B, nécessite une « cause commune », puisqu'elle est débarrassée de toute « cause aléatoire ». Cette cause commune est alors assimilée à la parenté.

Le meilleur arbre est évidemment celui qui optimise les *seules* synapomorphies, c'est-à-dire les caractères dérivées partagés « par ascendance », et non les caractères dérivées partagés par « hasard ». Dans l'esprit de Forster, c'est donc bien l'hypothèse de parenté que l'on cherche à optimiser, mais après avoir pris en compte la ressemblance due à des causes ou des processus *aléatoires*. Tout ce qui ne peut s'expliquer par le hasard peut *alors* s'expliquer en terme de parenté. Cette position est cohérente dans le contexte de l'analyse probabiliste des processus évolutifs que propose Forster. En revanche, elle n'a

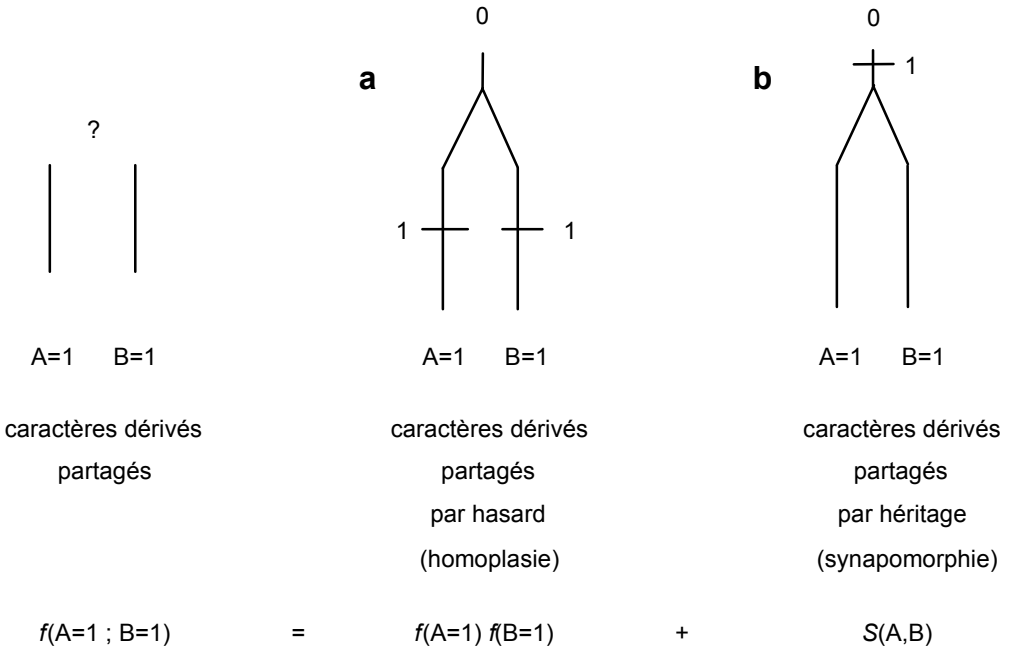


FIGURE V.33 : *Le partage par A et B de l'apomorphie « 1 » peut résulter : a) de l'existence d'une homoplasie provenant de deux événements indépendants survenant au hasard et dont la fréquence peut être estimée par le produit des fréquences des apomorphies chez A et B ; b) de l'héritage à partir d'un ancêtre commun (synapomorphie).*

évidemment plus de raison d'être si aucune probabilité ne peut raisonnablement être attachée aux transformations des caractères.

Reprenons l'exemple de Farris développé précédemment. On a vu que dans son raisonnement, la fréquence des apomorphies dans les différentes UE n'était pas prise en compte. Or cette fréquence est particulièrement élevée chez A comme chez B, alors qu'elle est faible chez C :

$$f(A=1) = 0.9 \; ; f(B=1) = 1 \; ; f(C=1) = 0.1$$

De ce seul fait, on peut s'attendre à ce que les apomorphies partagées « par hasard » entre A et B soient également très fréquentes, en dehors de toute hypothèse de parenté. En conséquence, les nombres $S(A,B)$ et $S(B,C)$ de synapomorphies (dues au seul partage d'un état de caractère hérité d'un ancêtre commun), corrigés donc des apomorphies dont le partage est d'origine homoplasique et qui surviennent au « hasard », s'écrivent :

$$S(A,B) = 10(0.9 - 0.9x1.0) = 0$$
$$S(B,C) = 10(0.1 - 1.0x0.1) = 0$$

Cette fois, le nombre estimé d'homologie (synapomorphie) ne permet pas d'effectuer un choix entre les deux arbres ((A,B),C) et ((B,C),A). Le fait que les apomorphies soient fréquentes chez A et B amène à contredire la solution de parcimonie. Dans ce cas, le hasard explique tout aussi bien la distribution des caractères partagés que l'une et l'autre des deux hypothèses de parenté qui n'ont donc plus besoin d'être posées.

En résumé, on peut dire que le raisonnement de Farris conduit à affirmer : n'y aurait-il qu'une seule synapomorphie entre A et B, on aurait plus de chance de la rencontrer parmi les 9 caractères partagés par A et B ; tandis que le raisonnement de Forster conduit à la conclusion : la synapomorphie n'a pas plus de chances d'être l'un des 9 caractères apomorphes partagés par A et B que d'être le seul caractère partagé par B et C.

Pour bien comprendre la raison de la divergence entre l'approche de Forster et celle de Farris, supposons que l'on rajoute 90 caractères, tous à l'état plésiomorphes (état 0) chez A, chez B et chez C, dans le seul but de modifier les fréquences estimées des apomorphies. On a :

$$S(A,B) = 100(0.09 - 0.09x0.1) = 8.1$$
$$S(B,C) = 100(0.01 - 0.1x0.01) = 0.9$$

Dans ces conditions, l'arbre ((A,B),C) est effectivement celui qui montre le plus de synapomorphies. Cependant, une fois que l'on obtient un tel résultat, se pose la question de savoir quelles sont véritablement les caractères synapomorphes et quels sont les caractères homoplasiques ? Dans cet exemple, on peut dire que 8 caractères dérivés partagés entre A et B, parmi les 9, sont de véritables synapomorphies, sans que l'on sache pour autant clairement identifier les 8 caractères synapomorphes du seul caractère homoplasique.

Cet exemple n'a pour but que de souligner les conditions dans lesquelles les hypothèses de parenté sont fondées. Il montre clairement, quand on admet que les événements évolutifs sont de nature probabiliste, que la méthode cladistique fait

implicitement l'*hypothèse que les apomorphies sont rares par rapport aux plésiomorphies*. De ce point de vue, l'application du principe de parcimonie implique que les changements évolutifs, c'est-à-dire les transformations de caractères, sont rares.

La mise en pratique de la démarche de Forster comporte un point faible, celui de l'estimation des fréquences d'apomorphies chez les UE. En effet, si les caractères qui ne changent pas chez tous les taxons étudiés ne sont pas introduits dans la matrice de données – pour la raison précise qu'ils ne changent pas – ce qui est le cas de la plupart des analyses morphologiques, voire moléculaires lorsque ne figurent que les sites dits informatifs, l'estimation de ces fréquences sera manifestement biaisée. Par exemple, si l'on compare les affinités de deux marsupiaux par rapport à un placentaire, il est possible de conclure à l'absence de parenté des deux marsupiaux si la matrice de caractères n'inclut pas les très nombreux caractères d'amniotes non mammaliens, tous plésiomorphes pour les marsupiaux et les placentaires.

Elle comporte également un préalable, celui de l'identification des états plésiomorphes et apomorphes des caractères. On a vu dans le chapitre IV que cette identification s'effectue sur la base du principe de parcimonie, même dans le cas de pratiques intuitives ou d'observation directe telle l'analyse de données ontogéniques. La démarche de Forster ne s'effectue donc pas en dehors de ce principe.

On peut aussi faire remarquer que les caractères qui changent dans une partie de l'arbre peuvent être stables dans les autres parties de cet arbre pendant que l'inverse est observé pour d'autres caractères. Dans ce cas, la rareté des changements n'a pas de valeur universelle : elle peut être globalement vérifiée pour un caractère si l'on considère l'arbre dans son ensemble, mais ne pas l'être sur un sous-ensemble restreint de cet arbre. C'est l'« évolution en mosaïque » des évolutionnistes, appelée par Hennig « hétérobathmie des caractères », un concept de base de la démarche cladistique. De son côté, le raisonnement probabiliste admet qu'il existe une probabilité de changement définie pour *l'ensemble* de l'arbre et que seule la *réalisation* de cette probabilité peut entraîner des disparités locales dans les fréquences observées de changement.

Restons-en donc à cette réponse simple à la question-titre de ce chapitre. L'application du principe de parcimonie ne fournit pas de solution erronée si l'évolution est, effectivement, parcimonieuse, c'est-à-dire si les changements évolutifs sont « rares ». La quantification de cette rareté reste le fruit d'une approche empirique liée à chaque cas concret offert par les analyses phylogénétiques de différents organismes. Dans cette perspective, la perception des limites au-delà desquelles les solutions de parcimonie peuvent être erronées constitue une question à laquelle les modèles probabilistes tentent, avec leurs propres hypothèses, de donner des réponses (Chapitre VIII).

CHAPITRE **VI**

LA MÉTHODE DE COMPATIBILITÉ

La méthode de compatibilité repose essentiellement sur les travaux de Le Quesne (1969, 1972) et d'Estabrook et collaborateurs (Estabrook 1972, Estabrook *et al.*, 1976, 1977). Ce sont, d'une certaine manière, des variantes des méthodes de parcimonie, en ce sens qu'elles utilisent également le principe de parcimonie. Ce ne sont cependant pas des méthodes cladistiques au sens strict. En effet, bien qu'elles soient fondées sur le principe de congruence des caractères, les notions d'apomorphie et de plésiomorphie ne résultent pas de l'application de la méthode de compatibilité elle-même.

On a vu au chapitre précédent que le caractère qui nous renseigne le mieux sur la phylogénie est celui qui s'est transformé une seule fois au cours de l'évolution : les taxons portant l'état transformé d'un tel caractère forment ensemble une communauté de descendance, autrement dit un groupe monophylétique non ambigu. Les caractères qui se transforment plusieurs fois indépendamment sont des homoplasies : ce sont eux qui brouillent l'image phylogénétique. C'est par le traitement de l'homoplasie que l'analyse de compatibilité se distingue fondamentalement de l'analyse cladistique et des analyses de parcimonie décrites précédemment.

1. La méthode

Des caractères sont dits mutuellement compatibles quand il existe un arbre qui rende compte des changements d'états de ces caractères sans nécessiter d'hypothèses d'homoplasie. Ainsi, lorsque les caractères sont codés 0 et 1, cela signifie que l'on n'observe qu'une seule transformation par caractère dans l'arbre en question. De tels caractères sont définis par Le Quesne (1972) comme des « caractères dérivés uniques ». L'ensemble des caractères mutuellement compatibles est appelé une « clique » (Estabrook *et al.* , 1977). Les caractères non compatibles sont donc homoplasiques.

La méthode de compatibilité consiste simplement à rechercher l'arbre pour lequel la clique est la plus nombreuse. Cet arbre est construit sans caractères homoplasiques.

Un exemple dû à Felsenstein (1984b), à peine modifié ici, illustre l'approche de compatibilité (tableau VI.1). Un traitement classique de ces données par

parcimonie indique un taux élevé d'homoplasie. Il existe en effet pas moins de 7 arbres différents, chacun d'une longueur minimale de 10 pas. L'arbre de consensus-strict (figure VI.1) montre que le problème des relations de parenté entre les sept taxons A-Y n'est pas résolu avec une telle matrice de caractères. Les taxons X et Y sont choisis ici comme extra-groupes.

	CARACTERES						
	1	**2**	**3**	**4**	**5**	**6**	**7**
TAXONS							
A	1	1	0	1	1	0	1
B	1	1	0	0	0	1	1
C	1	0	0	1	1	0	1
D	0	0	1	0	0	0	1
E	0	0	1	1	1	0	1
X	0	0	0	0	0	0	0
Y	0	0	0	0	0	0	0

TABLEAU VI.1. *Un exemple de distribution des états de 7 caractères chez 7 taxons.*

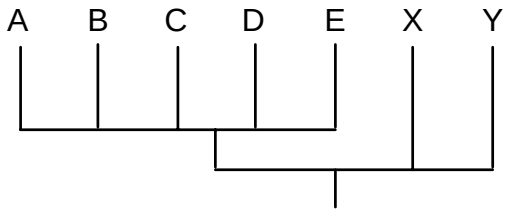


FIGURE VI.1. *Arbre de consensus -strict construit à partir des 7 arbres également parcimonieux (10 pas) obtenus du tableau VI.1.*

Le tableau VI.2 indique les différentes partitions de l'ensemble de taxons que l'on peut effectuer à partir de chacun des 7 caractères. Ainsi, le caractère 1 permet de regrouper les taxons A, B et C en un premier sous-ensemble, et les taxons D, E, X et Y en un autre. Le caractère 4 permet de reconnaître les sous-ensembles A, C et E d'une part et B, D, X et Y de l'autre.

	ARBRE
CARACTERES	
1	(ABC) (DEXY)
2	(AB) (CDEXY)
3	(ABCXY) (DE)
4	(ACE) (BDXY)
5	(ACE) (BDXY)
6	(B) (ACDEXY)
7	(ABCDE) (XY)

TABLEAU VI.2. *Les sept combinaisons de taxons définies par chacun des sept caractères et construites à partir des distributions du tableau VI.1.*

Ce tableau VI.2 montre également que les caractères 1 et 2 sont mutuellement compatibles dans le sens défini plus haut, c'est-à-dire qu'ils définissent chacun deux sous-ensembles de taxons qui peuvent se combiner sans contradiction de telle façon que soit définie une partition qui est la suivante : ((A,B)C)(D,E,X,Y) et qui permet de construire un arbre, non enraciné ici. De la même façon, le caractère 3 définit une partition en deux sous-ensembles de taxons compatibles simultanément avec ceux définis par le caractère 1 et par le caractère 2. En revanche, les caractères 4 et 5 opposent les sous-ensembles (A,C,E) et (B,D).

Le tableau VI.3 représente la matrice des caractères mutuellement compatibles. Le symbole « c » de ce tableau, à l'intersection d'un caractère en ligne et d'un caractère en colonne, indique que ces deux caractères sont mutuellement compatibles. Le symbole « . » signifie qu'ils ne le sont pas.

CHARACTERES	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1	c	c	c	.	.	c	c
2	c	c	c	.	.	c	c
3	c	c	c	.	.	c	c
4	.	.	.	c	c	c	c
5	.	.	.	c	c	c	c
6	c	c	c	c	c	c	c
7	c	c	c	c	c	c	c

TABLEAU VI.3. *Matrice de compatibilité des caractères du tableau VI.1. Le symbole « c » indique que le caractère défini en ligne et celui en colonne sont mutuellement compatibles, le symbole « . » qu'ils ne le sont pas.*

Le tableau VI.3 montre clairement que les caractères 4 et 5, qui sont compatibles entre eux, sont, en revanche, incompatibles avec les caractères 1, 2, 3, 6 et 7. De leur côté, les caractères 6 et 7 sont compatibles avec tous les autres. On est donc en présence de deux cliques possibles : l'une comprenant les caractères 4, 5, 6 et 7 (clique I), l'autre comprenant les caractères 1, 2, 3, 6 et 7 (clique II). Puisque la méthode de compatibilité consiste à retenir l'arbre qui correspond à la clique la plus nombreuse, il s'agit donc dans cet exemple de l'arbre qui est défini par les caractères de la clique II (5 caractères). Cet arbre (figure VI.2), qui est entièrement résolu à la différence des solutions données par la méthode de parcimonie, ne figure pas parmi les 7 arbres les plus parcimonieux construits à partir des 7 caractères.

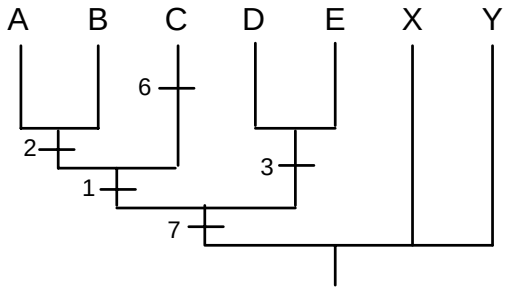


FIGURE VI.2. *Arbre de compatibilité obtenu à partir des données du tableau VI.1 et VI.3 et position des changements d'état des caractères compatibles 1, 2, 3, 6 et 7. X et Y sont les extra-groupes.*

Notons qu'à une même clique peut ne pas correspondre nécessairement un arbre unique quand les caractères ne sont pas orientés (Fitch, 1975).

2. Compatibilité et parcimonie

La méthode de compatibilité peut être considérée comme une méthode utilisant le principe de parcimonie, en ce sens qu'elle retient comme arbre celui qui *minimise* le *nombre* des caractères rejetés parce qu'ils sont homoplasiques. En revanche elle se distingue des méthodes de parcimonie dans la mesure où elle ne cherche pas à minimiser la *quantité* elle-même d'événements homoplasiques. De ce fait elle ne permet pas de localiser les homoplasies dans l'arbre, puisque les caractères homoplasiques ne sont pas pris en considération en tant que tel.

La méthode de compatibilité présente des analogies avec la méthode de parcimonie quand cette dernière est utilisée en pondérant les caractères. En effet, en parcimonie, si l'on décide de donner aux caractères qui changent deux, trois ou quatre fois (ou plus) un poids plus faible qu'à ceux qui ne changent qu'une fois, ces caractères très variables, quoique non éliminés de l'analyse, auront une influence moins grande que les autres caractères dans la recherche de l'arbre le plus parcimonieux, et cela d'autant plus qu'on leur attribue un poids faible. A la limite, l'analyse de compatibilité est donc comparable à une analyse de parcimonie où les caractères homoplasiques (changeant plus d'une fois) auraient un poids nul, autrement dit seraient éliminés de la recherche de l'arbre le plus court (Felsenstein, 1981a).

Comme les logiciels d'analyse de compatibilité ne traitent pas les caractères manquants ou non observés, un moyen de contourner cette difficulté, tout en restant dans la perspective de compatibilité, consiste à effectuer une analyse de parcimonie, qui accepte les caractères manquants, mais attribuant une pondération très faible aux caractères homoplasiques (Felsenstein, 1981a).

Le rejet des caractères homoplasiques lors de la recherche d'un arbre phylogénétique pose différents problèmes. En effet, il peut arriver que, dans la réalité, un caractère soumis à convergence ou réversion, donc un caractère homoplasique, soit néanmoins diagnostique d'un groupe monophylétique situé à l'intérieur du groupe étudié. Le supprimer reviendrait donc à perdre une information phylogénétique utile. Par exemple, l'homéothermie est classiquement considérée comme une synapomorphie des oiseaux d'une part, et une synapomorphie des mammifères d'autre part. Ce caractère serait donc apparu deux fois, par convergence. Si cette distribution est correcte, le caractère serait éliminé d'une analyse de compatibilité parmi les amniotes.

Dans le cas de la figure VI.2, les méthodes de compatibilité et de parcimonie ne donnent pas le même résultat. En effet, l'arbre qui est construit à partir de la clique formée des cinq caractères 1, 2, 3, 6 et 7 nécessite, d'un point de vue de la méthode de parcimonie classique, 11 pas, soit un de plus que l'arbre de longueur minimal donné par la parcimonie.

Dans d'autres cas, l'analyse de compatibilité donne les mêmes résultats que l'analyse de parcimonie. Il en est ainsi des données paléontologiques de 5 proboscidiens de la figure VI.3.

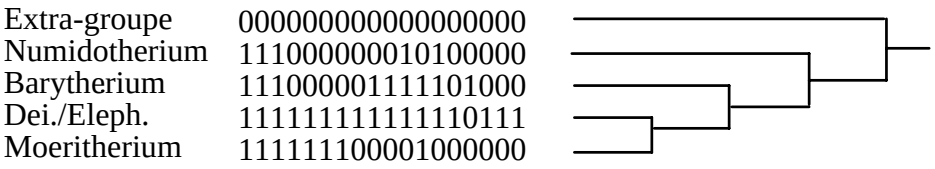


FIGURE VI.3. *Distribution des états de 18 caractères chez 5 proboscidiens (d'après Tassy, 1988) et arbre donné par les méthodes de parcimonie (22 pas, I.C. = 0.8 ; I.R. = 0.7) et de compatibilité (clique : 1, 2, 3, 4, 5, 6, 7, 8, 12, 14, 15, 16, 17, 18).*

Un autre exemple peut être tiré des données de la phylogénie de la figure V.26 dont les caractères sont dans le tableau V.10. L'analyse de compatibilité du tableau V.10 conduit à construire un arbre (Figure VI.4) à partir d'une clique incluant 37 caractères parmi les 49 caractères (excluant les caractères : 1, 5, 7, 12, 21, 22, 25, 26, 27, 28, 29, 38). L'enracinement de l'arbre est effectué à partir des taxons m et n.

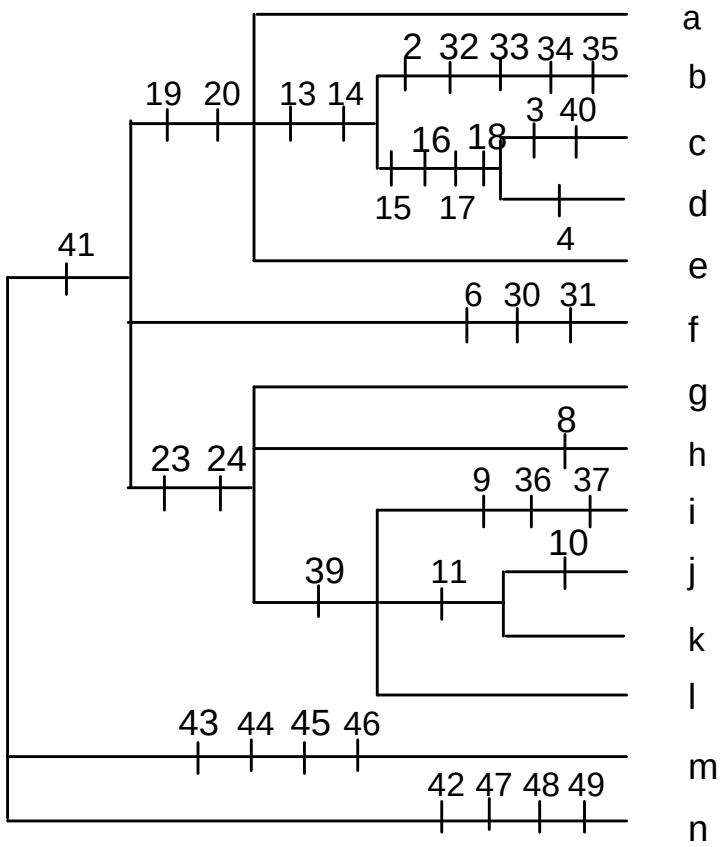


FIGURE VI.4. *Arbre de compatibilité construit à partir du tableau V.10. Les caractères portés sur cette figure sont ceux définissant une clique de 37 caractères.*

Les caractères 13 et 14 définissent le groupe monophylétique (b,c,d), le caractère 39 le groupe (i,j,k,l), les caractères 19 et 20 le groupe (a,b,c,d,e). Indiquons qu'il existe une clique comportant un caractère de moins (36 caractères), une clique comportant deux caractères de moins (35 caractères), une de 34 caractères, 7 cliques de 33 caractères etc.

Comme on le voit sur la figure VI.4, l'arbre n'est pas entièrement résolu. Ainsi, les caractères 19 et 20 définissent bien le groupe (a,b,c,d,e) mais aucun caractère ne permet de préciser l'ordre de branchement de a, de e et du groupe (b,c,d). Il en est de même pour les caractères 23 et 24 et les taxons (g,h,i,j,k,l).

L'absence d'information sur la répartition des traits homoplasiques est dommageable. En effet, l'analyse de compatibilité ne donne que la liste des caractères homoplasiques : sont homoplasiques ceux qui n'appartiennent pas à la clique. Mais il n'est pas possible d'apprécier quantitativement le nombre d'événements homoplasiques ni de donner leur répartition sur les nœuds ou les taxons terminaux. Ces informations ne peuvent être obtenues que par une analyse de parcimonie, ou bien en ajoutant sur l'arbre de compatibilité les caractères éliminés lors de l'analyse tout en minimisant le nombre de leurs transformations. Dans l'exemple de la figure VI.2, c'est ce qui conduit à compter 11 pas. Cette option repose néanmoins sur une contradiction : la distribution de caractères éliminés ne peut être logiquement évaluée à partir d'un arbre dont la construction est conditionnée par la suppression desdits caractères.

Comme on l'a vu, un arbre construit en rejetant les caractères homoplasiques fournit une information appauvrie – sinon erronée – pour toute interprétation évolutive, par exemple une explication des convergences dans un contexte écologique en termes d'adaptation. La méthode de compatibilité conduisant à éliminer les caractères homoplasiques, elle ne peut être invoquée pour expliquer les modalités d'apparition de ces caractères. Reprenons l'exemple précédent (Figure VI.4). Certains des caractères homoplasiques supprimés par l'analyse de compatibilité définissent en fait des monophylies (voir figures V.26 et V.27) : par exemple le caractère 21 définit le groupe (f,g,h,i,j,k), en même temps qu'il est un caractère autapomorphe de a. De même le caractère 5 définit le groupe (i,j,k,l) en même temps qu'il se trouve être une autapomorphie de e et de d etc... Cet exemple montre bien que si des caractères homoplasiques définissent des groupes monophylétiques, leur suppression appauvrit donc l'information phylogénétique.

3. Compatibilité et cladisme

Les concepteurs de la méthode de compatibilité et les concepteurs des procédures de parcimonie se réclament de Hennig (voir Duncan et Stuessy, 1984). Selon les uns et les autres, le fondateur de la systématique phylogénétique aurait préconisé ou la compatibilité ou la parcimonie (deux termes non utilisés par Hennig). On a vu que le concept de « congruence » ne se conçoit que dans le cadre du principe de parcimonie. Les méthodes de compatibilité font également appel au principe de congruence : ne sont retenus que les caractères congruents, ceux qui construisent le même arbre. Cette construction se fait néanmoins au prix

de l'élimination des caractères non congruents, qui changent plus d'une fois. Or l'approche cladistique vise à minimiser les homoplasies mais non à les éliminer. Les schémas théoriques dus à Hennig qui illustrent des contradictions sont rares, mais ils existent. Ils sont réservés aux cas des analyses ontogéniques. En fait, plutôt que de rejeter les caractères incompatibles (les homoplasies), Hennig préconise le « retour aux caractères » afin de vérifier si les caractères dérivés dus à des transformations indépendantes sont véritablement les mêmes caractères, ce qu'ils ne sont effectivement pas, généalogiquement parlant. On peut en conclure qu'il y a tout au plus une analogie entre le cladisme hennigien et la méthode de compatibilité mais non une homologie.

CHAPITRE **VII**

LES MÉTHODES PHÉNÉTIQUES

Les méthodes phénétiques se proposent de reconstruire des arbres en partant des ressemblances observées entre chaque paire d'unités évolutives (UE). Cette ressemblance est une ressemblance *globale* établie à partir du maximum d'observations disponibles. Ces observations doivent cependant constituer un ensemble de nature homogène, par exemple un ensemble de fréquences alléliques, un ensemble de caractères morphologiques codés présents ou absents, la séquence nucléotidique d'un même gène etc. Parfois ces méthodes s'imposent quand seule la ressemblance globale est directement appréhendée (degré d'hybridation d'ADN entre deux UE par exemple).

Pour ces méthodes, plus la ressemblance globale entre deux UE est importante, plus leurs liens de parenté sont étroits. Puisque c'est la ressemblance globale qui est en cause, il est clair que la parenté est fondée sut tous les caractères, non seulement les synapomorphies, mais aussi les caractères plésiomorphes, autapomorphes et homoplasiques. Ce point important distingue les méthodes phénétiques des autres méthodes. Il sera donc abondamment discuté.

Dans ce chapitre, après une brève introduction retraçant l'historique de ces méthodes, le concept de distance sera discuté. C'est lui qui permet en effet de quantifier la ressemblance globale. On insistera tout particulièrement sur les propriétés des distances qui sont essentielles dans la perspective phylogénétique (additivité, métricité). Trois méthodes principales seront ensuite évoquées : les méthodes d'agglomération qui se proposent de rapprocher les unes des autres les UE, en partant des plus ressemblantes pour aller jusqu'aux moins ressemblantes ; les méthodes d'ajustement qui recherchent l'arbre et les longueurs de branches qui expliquent au mieux l'ensemble des ressemblances existant entre toutes les UE prises deux à deux ; les méthodes de parcimonie qui recherchent un arbre dont la somme des longueurs de branches serait la plus faible. Les méthodes de vraisemblance qui analysent des matrices de ressemblance entre UE prises deux à deux seront traitées à part dans le chapitre VIII à propos des méthodes probabilistes. Parmi toutes ces méthodes, on peut distinguer celles qui résultent de l'utilisation d'une certaine procédure algorithmique (comme par exemple la méthode dite UPGMA) de celles qui recherchent, par optimisation d'un certain critère, l'arbre qui ajuste au mieux les données de la matrice de distance (méthodes d'ajustement). Cette distinction reste cependant assez arbitraire dans la mesure où certaines méthodes procèdent à la fois de l'une et de l'autre de ces approches (comme la méthode du plus proche voisin ou *neighbor joining*). Une

conclusion développera enfin les problèmes et les limites de ces méthodes, particulièrement liés à la gestion de l'homoplasie et à l'enracinement de l'arbre.

1. Historique

Rappelons brièvement le contexte dans lequel se sont élaborées les méthodes phénétiques de construction d'arbre.

Elles trouvent leur origine dans les méthodes de la taxinomie numérique conçues dès 1957 par C. D. Michener et R. R. Sokal. Au cours des années soixante, ces méthodes dites phénétiques ou numériques s'opposèrent aux pratiques de systématiques de l'école évolutionniste dont les chefs de file étaient le zoologiste E. Mayr et le paléontologue G. G. Simpson en ce sens qu'elles se voulaient libres de toute spéculation phylogénétique. Les techniques employées sont d'abord des techniques de *classification* d'organismes sur la base de la *similitude globale*. Ce n'est qu'accessoirement, à l'aide d'autres critères, que des inférences phylogénétiques peuvent être tirées des taxons ainsi construits.

Les concepts de base et les méthodes de la systématique phénétique sont clairement présentés dans le premier chapitre de la nouvelle édition de *Numerical Taxonomy* par Sneath et Sokal (1973: 3-10) :

— les relations entre taxons fondées sur la similitude globale sont des « relations phénétiques » (« *phenetic relationships* » de Cain et Harrison, 1960) et non des relations phylogénétiques ;

— plus grand est le nombre de caractères étudiés meilleure sera la classification des taxons ainsi construits ;

— les caractères ont *a priori* le même poids, bien qu'une pondération puisse parfois s'effectuer sur la base de critères « opérationnels » ;

— la ressemblance est calculée entre chaque paire d'unités taxinomiques et s'exprime par des coefficients de similitude qui forment les éléments d'une matrice de similitude ;

— la restitution des relations taxinomiques s'effectue à partir de la matrice de similitude au moyen de techniques numériques variées (*cluster analysis*) ;

— la représentation de la construction taxinomique peut se faire au moyen de schémas (les *phénogrammes*) indiquant les relations phénétiques ;

— les inférences phylogénétiques s'effectuent en dernier en intégrant des hypothèses sur l'histoire et sur les mécanismes évolutifs ;

— les mesures de similitude phénétique entre les organismes appartenant à différentes époques géologiques fournissent une information objective sur la vitesse et la direction de l'évolution.

L'approche phénétique de la phylogénie n'est donc pas comparable à l'approche cladistique fondée sur l'analyse de parcimonie des caractères, même si la reconstruction de l'arbre phylogénétique à partir des données de la matrice de similitude nécessite généralement une procédure de minimisation. Le phénogramme exprime des degrés de similitude : ce n'est pas un cladogramme. « Nous croyons, écrivent néanmoins Sneath et Sokal (1973 : 313) que les phylogénies sont déduites nécessairement des relations phénétiques ». Les inférences phylogénétiques sont subordonnées aux relations phénétiques, elles-

même fondées sur la similitude globale, et non à l'analyse des caractères au sens hennigien du terme : « l'arrangement des états de caractères dans une séquence évolutive est, au mieux, une procédure difficile et, au pire, peut être grossièrement trompeuse » (Sneath et Sokal, 1973 : 320).

Que la problématique phénétique soit étrangère à la problématique phylogénétique ne fait aucun doute dans l'esprit de Sneath et Sokal. Ces derniers soulignent d'ailleurs que « le principe de base des taxinomistes numériques est la stricte séparation entre les spéculations phylogénétiques et les procédures taxinomiques ». Ils affirment ailleurs qu' « on ne devrait pas demander à un systématicien de tenir une classification phénétique pour une véritable phylogénie des organismes. Mais on devrait lui demander de la tenir pour une classification *phénétique* » (1973 : 420). Ils concluent, en répondant à une critique du cladiste américain J. Cracraft, qu' « à l'évidence une approche non-évolutionniste n'a pas besoin de donner une image correcte des événements du passé » (1973 : 420).

Pourquoi donc introduire des méthodes phénétiques dans un ouvrage consacré aux reconstructions phylogénétiques ? La réponse est triple. D'abord il est fréquent de rencontrer dans la littérature phylogénétique des phénogrammes interprétés comme des phylogénies. Ensuite, et surtout, des constructions phénétiques peuvent être assimilées à des arbres phylogénétiques à la condition qu'un certain nombre d'hypothèses soient posées, hypothèses concernant les processus évolutifs des caractères. Les méthodes phénétiques sont donc des *méthodes dont la nature phylogénétique n'apparaît qu'à la condition d'y introduire des hypothèses évolutives extrinsèques*, de telle manière que la similitude globale puisse être interprétée en termes de filiation. Enfin certaines sources d'information (données immunologiques, hybridation d'ADN) ne peuvent être interprétées qu'au moyen de méthodes phénétiques.

2. Similitude et distance

Le concept de base des méthodes phénétiques est celui de *similitude globale* : plus la ressemblance entre deux UE est importante, plus la parenté entre elles a des chances d'être proche.

La ressemblance s'établit à partir d'informations biologiques de nature très variée :

— *alternative* : présence ou absence de particularités morphologiques ou génétiques ;

— *qualitative* : séquences d'acides aminés ou d'acides nucléiques, états multiples de caractères morphologiques ; il est parfois possible, comme on l'a vu au chapitre V.2, de transformer ce type de données qualitatives en données alternatives ;

— *quantitative* : fréquences géniques, génotypiques ou phénotypiques, mesures morphométriques, etc.

Par ailleurs la similitude entre deux UE i et j, σ_{ij}, peut être obtenue à partir de l'observation d'*une seule* variable, comme le degré d'hybridation entre leurs ADN respectifs par exemple. Cette variable résume directement la proportion d'identité

entre les ADN de deux UE. Notons que, dans cette situation, les méthodes phénétiques de reconstruction sont les seules envisageables.

La similitude peut également résulter de la prise en compte simultanée de *nombreuses variables*, par exemple un ensemble de fréquences alléliques ou de données morphométriques. Dans ce cas, l'indicateur de la similitude, σ_{ij}, est une certaine combinaison de ces variables.

2.1. La notion de similitude et de distance

La notion de *distance* découle naturellement de celle de similitude : plus la similitude σ_{ij} entre deux UE i et j est forte, plus la distance δ_{ij} entre elles est faible.

Les distances sont définies de la façon suivante. Soit trois UE i, j et k. La distance δ_{ij} entre i et j obéit aux propriétés suivantes :

$\delta_{ij} > 0$ si $i \neq j$ (positivité) ;
$\delta_{ij} = 0$ si $i = j$ (la distance de l'UE à elle-même est nulle) ;
$\delta_{ij} = \delta_{ji}$ (commutativité).

On admettra aisément qu'une distance soit un nombre positif. Certaines distances sont bornées à 1, d'autres peuvent aller jusqu'à l'infini.

Ces distances peuvent avoir une propriété dite de « l'inégalité triangulaire » :

$$\delta_{ij} \leq \delta_{ik} + \delta_{jk}$$

L'inégalité triangulaire signifie simplement qu'il est plus court de passer directement de i à j que de passer par un intermédiaire k. Les distances répondant à cette définition sont dites *distances métriques*.

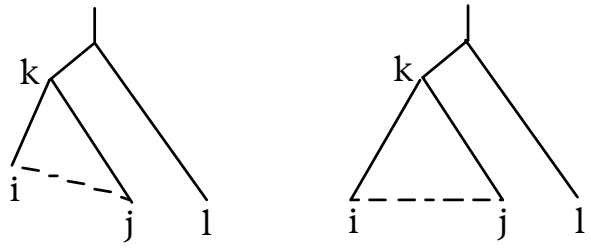
Lorsque, de plus :

$$\delta_{ij} \leq \max(\delta_{ik}, \delta_{jk})$$

la distance est dite *ultra-métrique*. Cela signifie que les deux plus grandes distances sont égales. Donc : $\delta_{ik} = \delta_{jk}$, ou $\delta_{ij} = \delta_{ik}$ ou $\delta_{ij} = \delta_{jk}$.

En revanche la stricte égalité $\delta_{ij} = \delta_{ik} + \delta_{jk}$ signifie que la plus courte distance pour aller de i à j passe nécessairement par k. On parle alors de distance *additive*.

On peut résumer ces différentes propriétés à partir du schéma suivant, en les appliquant à une situation d'arbre qui sera celle rencontrée dans cet ouvrage :



— la *métricité* signifie qu'il est plus court de passer de i à j directement (en tirets) que de passer par k ;

— l'*ultra-métricité* signifie que la distance entre i et k d'une part et j et k de l'autre sont égales (comme sur le schéma de droite) ;

— *l'additivité* signifie que la distance entre i et j est égale à la somme des distances reliant i à k et k à j.

Il existe deux façons de situer deux UE l'une par rapport à l'autre : la similitude, σ_{ij}, et la distance δ_{ij}. Quand la distance augmente, la similitude diminue. La relation entre ces deux indicateurs peut prendre différentes formes, par exemple :

$\delta_{ij} = 1 - \sigma_{ij}$ $(0 \leq \sigma_{ij} \leq 1)$;

$\delta_{ij} = (1 - \sigma_{ij}) / \sigma_{ij}$ $(0 \leq \sigma_{ij} \leq 1)$;

$\delta_{ij} = -\alpha \, \mathrm{Ln}(\beta - \gamma\sigma_{ij})$, α, β et γ étant des constantes $(0 \leq \beta - \gamma\sigma_{ij})$.

Les formes algébriques et les propriétés des différents indices de similitude ou de distance sont extrêmement variées. Elles ne seront pas détaillées ici de façon exhaustive. Pour cela voir Sokal et Sneath (1963), Jacquard (1973), Smith (1977), Rao (1980), Lalouel (1980), Jorde (1985), Gregorius (1978).

Les quelques indices donnés ici à titre d'exemple se classent en deux catégories : les indices fondés sur des attributs ou des données qualitatives et les indices fondés sur des données quantitatives.

D'une façon générale et quelle que soit la formulation retenue pour le calcul de la distance, la totalité de l'information est contenue dans une matrice carrée symétrique, dite matrice de distances, où figure l'ensemble des distances entre UE prises deux à deux, les valeurs de la diagonale étant nulles.

2 .2. Indices de similitude et de distance fondés sur des attributs

Considérons un caractère pouvant se présenter sous les états distincts a, b, c, ...,h..., s et deux UE i et j. Les états de ce caractère peuvent être concordants chez l'UE i et l'UE j. Cet état est alors l'un des s états possibles. Mais ils peuvent aussi être discordants : il existe $s^2 - s$ différentes combinaisons possibles de discordances. Lorsque l'on observe chez i et j un ensemble de K caractères pouvant tous se présenter sous les mêmes s états possibles, on peut comptabiliser combien de fois se rencontre chacune des combinaisons d'états possibles de ces caractères chez i et j. C'est ce que représente le tableau VII.1 où la somme des occurrences :

$$n_{aa}+n_{ab}+n_{ac}+ \ldots+n_{ah}+\ldots n_{as}+n_{ba}+\ldots+ n_{bs}+\ldots +n_{ss} = K$$

Par exemple, dans des comparaisons de gènes, on peut être amené à considérer plusieurs situations :

1) 20 états correspondant aux 20 acides aminés (s = 20) ;

2) 4 états correspondant aux quatre bases (A, T ou U, C, G) (s = 4) ;

3) 2 états correspondant aux 2 types de bases : purique, pyrimidique (s = 2) ;

		UE j						
		a	b	c	...	h	...	s
	a	n_{aa}	n_{ab}	n_{ac}		n_{ah}		n_{as}
	b	n_{ba}	n_{bb}	n_{bc}		n_{bh}		n_{bs}
	c	n_{ca}	n_{cb}	n_{cc}		n_{ch}		n_{cs}
	.							
	.							
	.							
UE i	h	n_{ha}	n_{hb}	n_{hc}		n_{hh}		n_{hs}
	.							
	.							
	.							
	s	n_{sa}	n_{sb}	n_{sc}		n_{sh}		n_{ss}

TABLEAU VII.1. *Distribution des combinaisons des s états différents d'un caractère entre deux UE i et j, observée sur un ensemble de K caractères. n_{hc} signifie que l'on observe n_{hc} caractères dans l'état h chez l'UEi et dans l'état c chez l'UE j. La somme des valeurs n de ce tableau est égal à K*

Dans chacune de ces situations, un état supplémentaire peut être ajouté pour rendre compte des *gaps* (ou *indels* : insertions et délétions).

Pour un caractère morphologique, les états peuvent être :

— la présence ou l'absence de ce caractère (s = 2);

— la présence sous l'état ancestral ou sous l'état dérivé (s = 2).

Il paraît difficile d'envisager des cas plus complexes. En effet il est rare que plusieurs caractères morphologiques puissent être présents sous des états dont la nature et/ou le nombre soient comparables d'un caractère à un autre. Par exemple il n'y a pas de comparaison à faire entre les états observés sur un radius (réduction ou non de l'apophyse styloïde) et sur un fémur (présence ou non d'un troisième trochanter).

2.2.1. Caractères où deux états seulement sont comparés

Les deux états peuvent être la présence ou l'absence du caractère (codé 0 et 1 par exemple) ou sa présence sous deux formes distinctes (a et b). K est le nombre de caractères. On se borne à calculer les effectifs n_{aa}, n_{ab}, n_{ba}, n_{bb} du tableau VII.1.

Le cas plus complexe où les caractères présentent plus de 2 états peut, bien souvent, se ramener à la situation simple à 2 états, lorsque, par exemple, on ne s'intéresse qu'à la concordance ou à la discordance entre les états de caractères de

deux UE : seuls sont décomptés le nombre de caractères concordants et le nombre de caractères discordants, sans se préoccuper de la nature de ces caractères.

Les similitudes et les distances *observées* seront notées s_{ij} et d_{ij} respectivement.

L'indice de similitude de Jaccard (1908) :

$$s_{ij} = \frac{n_{bb}}{K - n_{aa}}$$

La présence conjointe de l'un des deux états (a par exemple) chez i et j est considérée comme non informative. Il peut s'agir, par exemple, de l'absence partagée du caractère chez i et j.

L'indice de concordance simple de Sokal et Michener (1958) :

$$s_{ij} = \frac{n_{aa} + n_{bb}}{K}$$

L'indice de similitude est ici la proportion de caractères qui sont dans le même état à la fois chez i et chez j.

Une transformation de cette similitude en distance $d_{ij} = 1 - s_{ij}$ est souvent effectuée pour obtenir l'indice de divergence entre deux séquences de protéines ou de nucléotides : d_{ij} est alors la proportion de sites (acides aminés ou nucléotides) dont les états sont différents entre les séquences de i et de j.

Une autre transformation est possible, lorsqu'il y a deux états possibles (a et b) par caractère et que la probabilité de changement du caractère est indépendante du sens de ce changement (c'est-à-dire qu'elle est identique pour une transformation de a vers b et pour une transformation de b vers a). Dans ce cas, la similitude peut être transformée en distance de la façon suivant (voir paragraphe VII.3.2.2) :

$$d_{ij} = -\frac{1}{4} \ln \left(1 - 2(1 - s_{ij})\right)$$

Cette distance suppose que s_{ij} est supérieur à 0.5, c'est-à-dire que les caractères discordants sont moins nombreux que les caractères concordants.

La distance de Jukes et Cantor (1969) :

Lorsqu'un caractère peut se présenter sous quatre états différents (les quatre acides nucléiques par exemple) et que les probabilités de changement d'état sont toutes égales entre elles, l'indice de concordance s_{ij} de Sokal et Michener défini plus haut peut être transformé en distance, en suivant pour cela un raisonnement analogue à celui décrit paragraphe VII.3.2.2 :

$$d_{ij} = -\frac{3}{4} \ln \left(1 - \frac{4}{3}(1 - s_{ij})\right)$$

Il est nécessaire que s_{ij} soit supérieur à 0.25.

La distance Manhattan :

$$d_{ij} = K - (n_{aa} + n_{bb})$$

Cette distance est nulle lorsque la concordance est totale entre les états de caractères observés chez i et chez j. Sa valeur maximale, K, s'obtient lorsque tous les caractères sont discordants entre i et j. Elle correspond donc au nombre de changements d'états qui est nécessaire pour passer de l'UE i à l'UE j. Il s'agit d'une distance métrique et additive.

2.2.2. Caractères où plusieurs états sont comparés

L'indice de la différence symétrique (Renyi, 1966; Jacquard, 1973)

$$d_{ij} = \frac{\sum_{h=1}^{s} \lambda_h [p_{ih}(1 - p_{ih}) + p_{jh}(1 - p_{jh})]}{\sum_{h=1}^{s} \lambda_h}$$

La distance entre deux UE i et j est la probabilité pour qu'un état choisi au hasard soit présent chez i et absent chez j ou réciproquement. Une pondération λ_h peut être affectée à chacun des *s* états. Les probabilités peuvent être remplacées par les fréquences d'occurrence des différents états : $p_{ih} = n_{ih}/K$.

L'indice de divergence moléculaire :

$$d_{ij} = \frac{t_s + t_v + id}{K + id}$$

Dans le cas de comparaisons de séquences de nucléotides, la divergence entre i et j peut être calculée comme la somme relative des divergences dues à des transitions (t_s), à des transversions (t_v) et à des insertions/délétions (indels : *id*). K est ici le nombre total de sites.

L'indice de Kimura (1980) :

$$d_{ij} = -\frac{1}{2} \ln \left[(1 - 2P - Q) \sqrt{(1 - 2Q)} \right]$$

où *P* et *Q* sont respectivement les fréquences des transitions et des transversions entre les deux séquences i et j d'ADN ou d'ARN homologues. Dans cette formule, les sites où l'on observe des insertions/délétions (« indels ») ne sont pas comptabilisés.

2.3. Indices de distances fondées sur des données quantitatives

Les données quantitatives peuvent être, par exemple, des mensurations de caractères morphologiques ou bien des fréquences alléliques. Dans ces cas, chaque UE i se définit généralement par le vecteur X_i constitué des moyennes des K caractères mesurés. Si ces caractères ne sont pas indépendants, il est également possible de calculer leur matrice de variance-covariance S.

La distance de Mahalanobis (1936) :

$$d_{ij}^2 = (X_i - X_j)' S^{-1} (X_i - X_j)$$

X_i et X_j sont les vecteurs des moyennes des K caractères chez i et j, S est la matrice de variance-covariance entre ces caractères. Cette distance tient compte des corrélations pouvant exister entre les caractères.

Lorsque la matrice de covariance est une matrice diagonale, on obtient la distance euclidienne pondérée :

$$d_{ij}^2 = \sum_{h=1}^{K} (x_{ih} - x_{jh})^2 / s_{hh}$$

avec x_{ih} la valeur du caractère h dans l'UE i et s_{hh} la variance du caractère h.

Lorsque les éléments de la diagonale sont, de plus, égaux à 1 (s_{hh} = 1 pour tout h), on retrouve la distance euclidienne simple.

$$d_{ij}^2 = \sum_{h=1}^{K} (x_{ih} - x_{jh})^2$$

Dans tous les cas, on peut aussi bien utiliser cette distance, qui est métrique, que sa racine carré d_{ij}.

L'estimation de la matrice des covariances entre caractères pose un problème. En effet on ne peut utiliser une standardisation à partir des covariances *entre* taxons, puisque la reconstruction est justement fondée sur l'interprétation des covariations en terme de parenté. Les supprimer reviendrait à jeter le bébé avec l'eau du bain.

La matrice de covariances doit donc être estimée à partir d'observations effectuées à l'*intérieur* des UE. Il reste nécessaire de s'assurer que les matrices de covariances intra-taxons obtenues pour chaque UE ne sont pas significativement différentes les unes des autres.

Les données quantitatives peuvent également se présenter sous forme de fréquences. Chaque UE est alors définie par un vecteur de fréquences. Ces dernières présentent la particularité d'être comprises entre les bornes 0 et 1.

La distance de Cavalli-Sforza et Edwards (1967) :

$$d_{ij}^2 = \frac{1}{K} \sum_{1}^{K} \frac{2}{\pi} \cos^{-1} \left[\sum_{h=1}^{s} \sqrt{x_{ih} x_{jh}} \right]$$

x_{ih} est la fréquence de l'allèle h, parmi les s allèles possibles en un locus donné, dans l'UE i. Cette distance est calculée à partir de l'ensemble des allèles présents en chacun des K loci considérés. La transformation angulaire a pour but de rendre la variance des fréquences transformées indépendante de la fréquence elle-même. L'hypothèse est faite que les fréquences alléliques varient exclusivement de manière aléatoire (dérive génique). Dans ces conditions la distance est fonction du temps exprimé en nombre de générations séparant i et j de leur ancêtre commun, mais aussi des variations des effectifs efficaces des populations i et j depuis cet ancêtre commun. Lorsque les fréquences alléliques sont corrélées entre elles, on peut appliquer la distance définie par Balakrishnan et Sanghvi (1968).

La distance de Nei (1972) :

$$d_{ij} = -\ln\left(\frac{\sum_{1}^{K}\sum_{h=1}^{s} x_{ih}x_{jh}}{\sqrt{\sum_{1}^{K}\sum_{h=1}^{s} x_{ih}^2 \sum_{1}^{K}\sum_{h=1}^{s} x_{jh}^2}}\right)$$

Les notations sont identiques à celles utilisées pour la distance précédente. Les sommes arithmétiques sur les K loci peuvent être remplacées par des produits géométriques (Nei, 1972). Comme l'ont observé de nombreux auteurs, les distances de Nei ne sont pas métriques (Farris, 1981). Cette distance d_{ij} mesure l'accumulation de différences alléliques par locus. Si le taux de substitution génique est constant par unité de temps, alors la distance de Nei varie linéairement avec le temps de divergence entre l'UE i et l'UE j.

La distance de Czekanowski (1909) :

$$d_{ij} = \frac{1}{K}\sum_{1}^{K}\frac{1}{s}\sum_{h=1}^{s}|x_{ih} - x_{jh}|$$

Cette distance est la transposition de la distance Manhattan décrite plus haut à propos des caractères qualitatifs. La présentation qui en est faite ici s'applique à des fréquences géniques (K systèmes ayant s allèles). Cette distance a été reprise et généralisée par Sanchez-Mazas *et al.* (1986) sous le terme de « PIG », *percentage of isoactive genes*.

Autres distances et comparaisons entre elles

Sur le thème particulier des distances, qui n'est pas le propos de ce livre, la littérature est particulièrement abondante. On consultera avec profit les articles ou ouvrages suivants : Nei (1987), Felsenstein (1985), Jacquard (1973), Gregorius (1978), Rao (1980), Smith (1977), Lalouel (1980), Jorde (1985).

3. Distances patristique, observée, estimée

Dans ce chapitre, il sera fait uniquement référence aux matrices de distance $D = \{d_{ij}\}$ dont l'élément d_{ij} représente la distance entre l'UE i et l'UE j. La transformation de la similitude globale en distance se fait par l'une des méthodes précisées plus haut (paragraphe 2.1).

3.1.Distance patristique ou phylétique

Considérons l'exemple simple de trois UE (l'ancêtre k et ses deux UE filles i et j) et d'un seul caractère pouvant prendre trois états non additifs : a, b et c. L'ancêtre k est dans l'état a. La distribution des caractères chez i et j peut se présenter sous quatre types différents en raison de la nature et de la localisation des transformations (a➞b ou a➞c) :

I : transformation entre k et i ;

II : transformation entre k et j ;

III : transformations identiques entre k et i d'une part et k et j de l'autre ;

IV : transformations différentes entre k et i d'une part et k et j de l'autre.

(Le cas où i, j et k sont sous le même état n'est pas représenté).

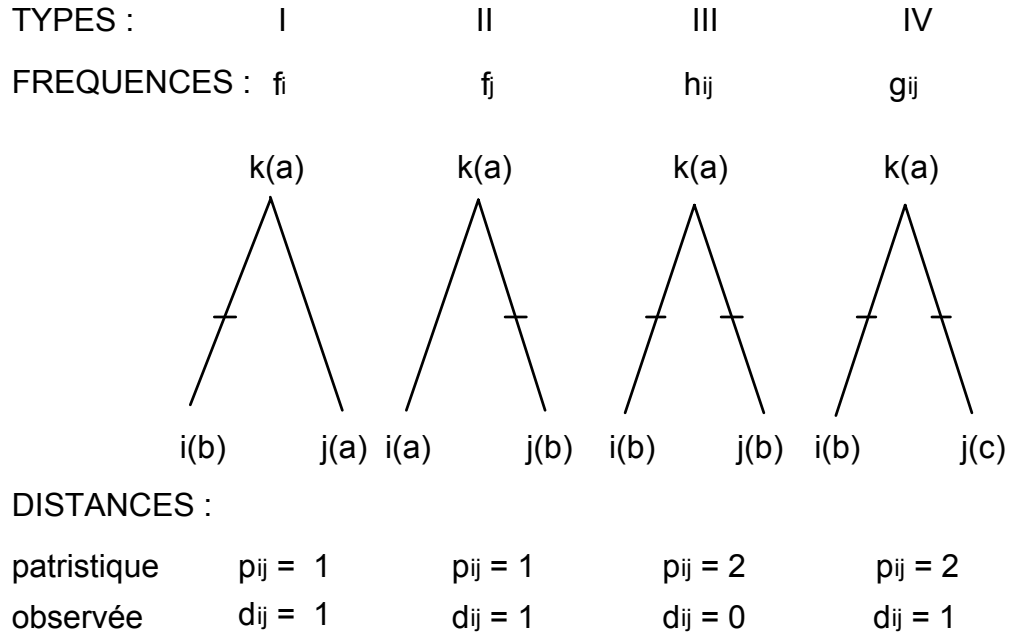


FIGURE VII.1. *Différents types de changements survenant entre les états de caractères de 2 UE et de leur ancêtre. Les distances patristique et observée diffèrent selon les types.*

La distance patristique (Farris, 1967) dite aussi distance phylétique (Fitch, 1984) entre i et j est donnée par la somme des événements de type « changement d'état » survenus entre k et i d'une part, k et j de l'autre. Il faut noter que Nei (1987) ne donne pas la même définition de la distance patristique. Pour lui, il s'agit en fait d'une distance estimée (voir plus loin).

Si, dans un ensemble de K caractères, les types de différences observables entre i et j sont dans les proportions indiquées sur la figure VII.1 (f_i, f_j, h_{ij}, g_{ij}), alors la distance patristique p_{ij}, sachant que l'état de l'ancêtre est a, est :

$$p_{ij} = K(f_i + f_j + 2h_{ij} + 2g_{ij})$$

Dans cette formulation, une proportion avec un indice simple (f_i et f_j) indique l'existence d'un seul changement (soit vers i, soit vers j), tandis qu'un indice double indique une homoplasie (h_{ij}) ou un double changement (g_j). La distance patristique est une *distance métrique* puisqu'elle satisfait l'inégalité triangulaire. Il s'agit également d'une distance *additive* puisque la distance patristique entre deux UE est strictement égale (et non pas seulement inférieure) à la somme des distances reliant i à j en passant par toutes les UE intermédiaires (dans l'exemple précédent k seulement). Par ailleurs, elle ne dépend que des transformations survenant le long des branches.

La distance patristique est, de toute évidence, celle qui intéresse le phylogénéticien, puisque c'est elle qui informe sur le nombre véritable d'événements survenant entre deux UE et leur ancêtre. La difficulté vient de ce que de telles transformations ne sont pas accessibles à l'observation. Elles doivent donc être déduites à partir des états de caractères observés sur les UE. Cette inférence se fait nécessairement sur la base de modèles incluant, de manière implicite ou explicite, des hypothèses sur les processus évolutifs eux-mêmes. Le but est d'obtenir les meilleures estimations possibles de ces distances patristiques ou phylétiques, pour toutes les branches.

Il faut noter que la distance patristique n'est définissable qu'à la condition de pouvoir donner une signification précise aux événements modifiant les caractères, c'est-à-dire à la condition que les transformations de caractères soient assimilées à des événements évolutifs identifiables qualitativement. Une telle signification n'est pas toujours évidente, particulièrement lorsqu'il s'agit d'événements se traduisant par des variations continues de caractères, des modifications de fréquences géniques par exemple.

3.2. La distance observée

La distance observée d_{ij} entre deux UE est celle donnée par l'application d'une certaine fonction aux données observées sur i et j. Quelques-unes de ces fonctions ont été décrites dans le paragraphe précédent (VII.2). Dans un premier temps, le problème des relations entre distance observée et distance patristique sera posé, à partir d'exemples simples. Puis, dans un deuxième temps, on montrera comment on peut tenter de résoudre ce problème, en faisant appel à des modèles d'évolution de caractères.

3.2.1. Distance patristique et distance observée

Reprenons l'exemple de la figure VII.1, en supposant que l'état de l'ancêtre k est connu.

Choisissons dans cet exemple la distance Manhattan : la distance observée, pour un caractère donné, est égale à 1 lorsque i et j sont dans des états différents et égale à 0 dans le cas inverse. Ainsi, dans la situation III de la figure VII.1, la distance observée entre i et j est nulle puisque les deux UE sont dans l'état b. Appliquant le principe de parcimonie, on en déduirait, faussement, que l'ancêtre k est, lui aussi, dans l'état b. En revanche, dans la situation IV où i et j sont dans deux états différents, la distance observée est égale à 1, alors que la distance réelle est de 2. Par ailleurs, en appliquant toujours le principe de parcimonie, c'est l'état b ou c qui serait, de manière erronée, attribué à l'ancêtre k.

La distance d_{ij} observée sur l'ensemble des K caractères est donc égale à la somme des distances observées sur chacun d'eux :

$$d_{ij} = K(f_i + f_j + g_{ij}) = p_{ij}$$

Ainsi la distance observée est-elle une *sous-estimation* de la distance patristique. Plusieurs événements, qui ont pu éventuellement se produire, ne sont en effet pas pris en considération dans son calcul : les deux changements d'état survenant entre k et i et entre k et j dans la situation III et l'un de ceux survenant dans la situation IV.

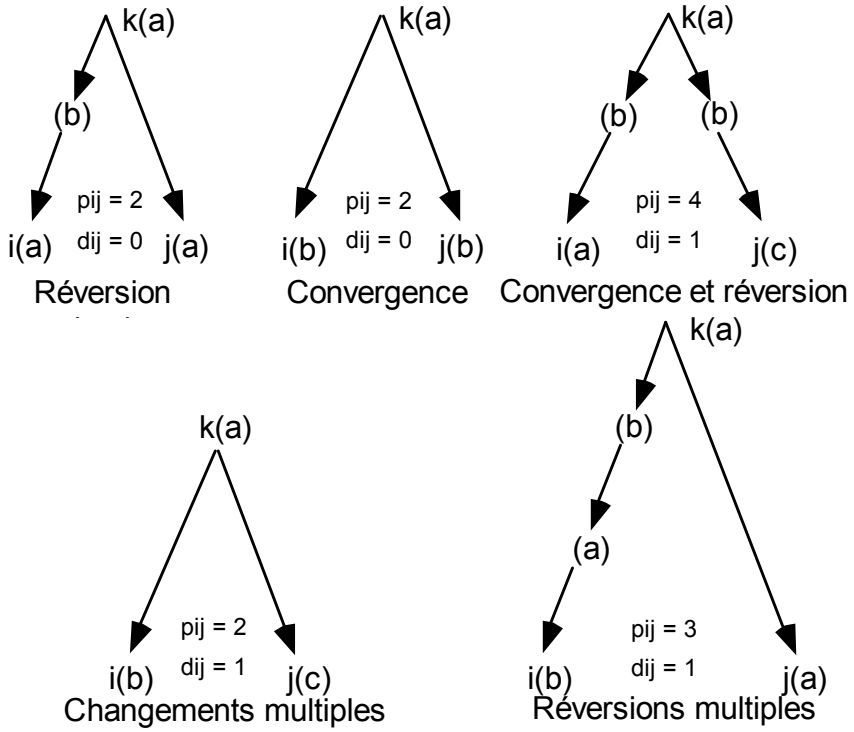


FIGURE VII.2 *Exemples de l'effet des convergences, des réversions simples et multiples et des changements multiples sur les distances observées et phylétiques.*

La figure VII.2 reprend ces différents changements : réversion simple, réversions multiples, convergence, changements multiples. Ils ne sont pas exclusifs les uns des autres et peuvent évidemment se combiner. De ce fait, on conçoit aisément que plusieurs types d'événements puissent rester dissimulés à l'observation.

Par ailleurs, il est clair que la distance observée ne donne une bonne approximation de la distance patristique que lorsque les homoplasies et les changements multiples sont rares, c'est-à-dire lorsque les fréquences h_{ij} et g_{ij} sont négligeables par rapport à f_i et f_j ; quand elles sont nulles, on a $d_{ij} = p_{ij}$.

3.2.2. Correction des distances observées

Plusieurs fonctions peuvent être mises en application pour tenter de prendre en compte tous les événements survenant le long des branches mais cachés à l'observation. Leur finalité est de trouver une relation entre distance patristique et distance observée. Toutes ces fonctions nécessitent des hypothèses sur les probabilités des événements et/ou leur occurrence en fonction du temps. Du bien-fondé de ces hypothèses évolutives, parfois difficile à démontrer, et des propriétés de la distance utilisée (en particulier métricité et additivité) dépend la qualité de l'inférence des distances patristiques à partir des distances observées.

Un exemple simple de correction peut être donné à partir de la distance déduite de l'indice de similitude de Sokal et Michener (paragraphe 2.2.1).

Supposons que la probabilité d'observer une différence d'état entre k et i soit égale à π. Soit f la probabilité *a priori* pour que k soit dans l'état a et $(1 - f)$ dans l'état b. La probabilité d'observer simultanément l'état a chez i et chez j est calculée ainsi :

— Ou bien l'ancêtre k est dans l'état a (probabilité f). Alors il ne faut pas observer de changements entre k et i d'une part (probabilité $1 - \pi$) ni entre k et j (probabilité $1 - \pi$).

— Ou bien l'ancêtre est dans l'état b (probabilité $1 - f$). Dans ce cas il faut observer indépendamment une différence entre k et i (probabilité π) et entre k et j (probabilité π). D'où :

$$p_{aa} = f(1 - \pi)^2 + (1 - f)\,\pi^2$$

et, par un raisonnement analogue :

$$p_{bb} = (1 - f)(1 - \pi)^2 + f\,\pi^2$$

De même la probabilité d'observer l'état a chez i et l'état b chez j (ou l'inverse) est donnée par :

$$p_{ab} = p_{ba} = \pi\,(1 - \pi)$$

Ces probabilités p_{aa}, p_{bb}, p_{ab} et p_{ba} peuvent être estimées par les diverses fréquences de combinaisons des deux états a et b entre i et j, ces combinaisons étant observées sur un ensemble de K caractères suivant tous la même probabilité de changement f. On a donc :

$$p_{aa} = n_{aa}/K \; ; \; p_{ab} = n_{ab}/K \; ; \; p_{bb} = n_{bb}/K \; ; \; p_{ba} = n_{ba}/K$$

Si la probabilité *m* de changement de a en b par unité de temps est constante, alors la probabilité *p(r)* d'observer *r* changements d'états dans le temps *t* est donnée par :

$$p(r) = \frac{(mt)^r}{r!} e^{-mt}$$

Dans ces conditions, π, la probabilité pour que k et i soient dans deux états différents, est égale à la somme :

$$\pi = p(1) + p(3) + p(5) + p(7) + \ldots = (1+e^{-2mt})/2$$

Il s'ensuit que la distance d_{ij} s'écrit :

$$d_{ij} = \frac{n_{ab} + n_{ba}}{K} = 2\pi(1-\pi) = \frac{1}{2}(1 - e^{-4mt})$$

d'où la valeur de la distance corrigée d'_{ij} :

$$d'_{ij} = 2mt = -\frac{1}{2}\ln\left(1 - 2d_{ij}\right)$$

On prendra donc pour nouvelle distance entre i et j la valeur 2*mt*, fonction de d_{ij}, qui est proportionnelle au temps et tient cette fois compte des événements multiples survenant selon un processus aléatoire le long des branches menant de i et j à leur ancêtre commun. Cette distance corrigée, d'_{ij}, n'est définie qu'à la condition que $d_{ij} = 0.5$, c'est-à-dire que le nombre de discordances entre i et j soit inférieur au nombre de concordances.

Cet exemple peut être généralisé au cas de transformations entre plusieurs états, par exemple les transformations entre les quatre bases de l'ADN ou de l'ARN, avec des probabilités différentes pour passer d'un état à un autre. Telles sont les distances de Jukes et Cantor (1969), Kimura et Ohta (1972), Tajima et Nei (1984). D'une façon générale on peut écrire que la distance corrigée d'_{ij} s'écrit:

$$d'_{ij} = 2mt = -b\ln\left(1 - 2\frac{d_{ij}}{b}\right)$$

où *b* est la valeur attendue de d_{ij} après un long temps d'évolution, indépendante de i et j. La distance d_{ij} est la proportion de caractères discordants entre l'UE i et l'UE j estimée sur K caractères. La valeur de *b* dépend de la nature des séquences. La variance de d'_{ij} est égale à :

$$V(d'_{ij}) = \frac{b^2 d_{ij}(1 - d_{ij})}{K(b - d_{ii})^2}$$

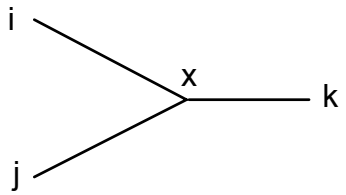
3.3. Distance estimée

La distance estimée e_{ij} entre deux UE i et j est celle déduite de l'analyse phénétique elle-même. Comme plusieurs formulations sont envisageables pour la distance observée et que plusieurs choix sont possibles pour l'analyse, la distance estimée sera évidemment dépendante de ces choix.

En règle générale, deux problèmes se posent à propos des distances, celui de la métricité et celui de l'additivité des distances. Ils seront exposés à partir de quelques exemples.

Examinons le cas simple de 3 UE i, j et k et les distances observées entre elles d_{ij}, d_{jk}, d_{ki}.

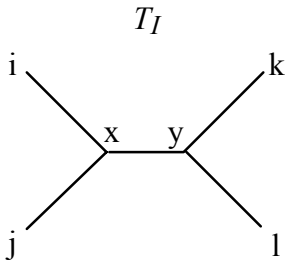
Les distances estimées entre l'UEH (Unité évolutive hypothétique) x et chacune des UE sont données sans ambiguïté par les équations suivantes :



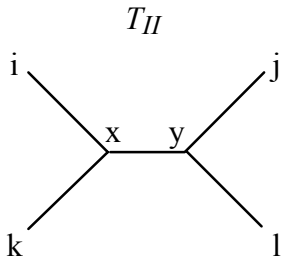
$$e_{ix} = (d_{ij} + d_{ik} - d_{jk})/2$$
$$e_{jx} = (d_{ij} + d_{jk} - d_{ik})/2$$
$$e_{kx} = (d_{ik} + d_{jk} - d_{ij})/2$$

Lorsque la distance observée n'est pas métrique, c'est-à-dire lorsque l'on peut avoir $d_{ij} > d_{ik} + d_{kj}$ par exemple, alors l'une des distances estimées (e_{kx}) aura une valeur négative. Dans cette situation, la question se pose de l'interprétation phylogénétique de distances négatives.

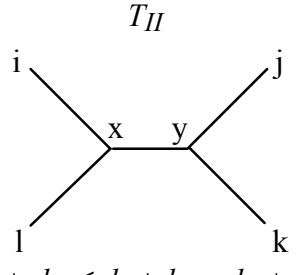
Considérons maintenant le cas de 4 UE, i, j, k et l, les deux nœuds internes étant x et y. Hormis le cas où la distance entre x et y est nulle par construction, trois arbres non enracinés différents sont possibles : T_I, T_{II}, T_{III}.



$$d_{ij} + d_{kl} \leq d_{il} + d_{kj} = d_{ik} + d_{jl}$$



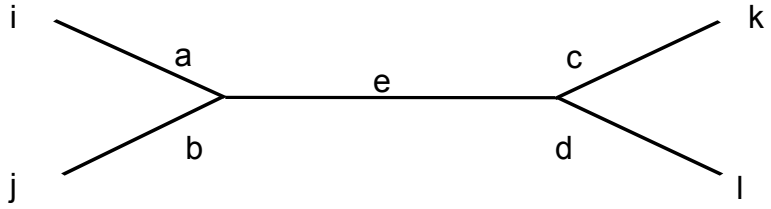
$$d_{ik} + d_{jl} \leq d_{il} + d_{kj} = d_{ij} + d_{kl}$$



$$d_{il} + d_{jk} \leq d_{ik} + d_{jl} = d_{ij} + d_{kl}$$

Si les conditions d'additivité reliant entre elles les distances observées sont remplies, l'une des équations suivantes est vérifiée (condition dite des 4 points) et permet donc d'en déduire l'arbre non enraciné compatible avec les distances observées.

A titre d'exemple, supposons que l'arbre non enraciné ait la structure T_I .Dans ce cas, 5 distances sont à estimer (a,b,c,d,e) à partir des six équations du tableau VII.2. Si les distances sont réellement additives, une des six équations est superflue et le système d'équations se résout sans difficulté. En revanche, lorsque ce n'est pas le cas, l'estimation des distances peut être obtenue en utilisant, par exemple, la méthode des moindres carrés (voir paragraphe VII.4.2). Trois exemples sont proposés.



	I	II	III	III'
d_{ij} = a+b	5	5	3	3
d_{ik} = a+e+c	11	11	12	12
d_{il} = b+e+c	12	12	14	14
d_{jk} = b+e+c	7	9	8	8
d_{jl} = b+e+d	8	8	10	10
d_{kl} = c+d	6	6	6	6
Estimations :				
a	4.50	4.00	3.50	3.50
b	0.50	1.00	–0.50	0.00
c	2.50	3.00	2.00	2.00
d	3.50	3.00	4.00	4.00
e	4.00	4.50	6.50	6.25

TABLEAU VII.2. *Trois exemples de distances observées entre 4 UE. Dans l'exemple I les distances sont additives et les estimations sont obtenues sans ambiguïté. Dans l'exemple II, les distances ne sont pas additives mais métriques. Dans l'exemple III, les distances sont additives mais non métriques : l'estimation d'une distance est négative (III) ou contrainte à être positive ou nulle (III').*

L'exemple I est celui de distances parfaitement additives. L'arbre T_I est choisi et les longueurs estimées de a, b, c, d et e (colonne I) permettent d'expliquer sans ambiguïté les distances observées.

Dans l'exemple II, la condition d'additivité des 4 points n'est pas remplie. La première inégalité est bien satisfaite : $d_{ij} + d_{kl} \leq d_{il}+d_{kj}$ et $d_{ij} + d_{kl} \leq d_{ik} + d_{jl,}$

mais pas l'égalité suivante, car $d_{il}+d_{kj} > d_{ik} + d_{jl}$. Il est impossible de satisfaire parfaitement les 6 équations du tableau VII.2 quelle que soit la structure de l'arbre (T_I, T_{II} ou T_{III}). Comme c'est la structure de l'arbre T_I qui minimise, par la méthode des moindres carrés, les écarts entre la matrice de distances observées et la matrice de distances estimées (paragraphe VII.4.2), c'est donc l'arbre T_I qui est choisi, avec les ajustements indiqués dans le tableau VII.2. Dans une telle situation de non additivité qui est certainement l'une des plus courantes, il peut arriver que les longueurs estimées soient négatives (voir l'exemple du tableau VII.2 et de la figure VII.9.

Dans l'exemple III, la condition d'additivité est bien remplie. On choisit donc l'arbre T_I. Mais on constate une absence de métricité. En effet, on a $d_{il} \geq d_{ij} + d_{jl}$. Dans ces conditions la méthode des moindres carrés (comme la méthode du *neighbor joining*, paragraphe VII.4.1.2), appliquée sans poser de contrainte sur les longueurs des branches, conduit à l'estimation d'une longueur négative, b, (Tableau VII.2, colonne III), avec un ajustement parfait en ce sens que l'on peut retrouver les distances observées à partir des distances estimées. En revanche, si l'on impose que les distances estimées soient positives, les estimations de b et e sont modifiées et l'ajustement est de moindre qualité (Tableau VII.2, colonne III'). On voit donc que la méthode reste applicable même en l'absence de métricité, bien que l'interprétation de longueurs négatives devienne problématique en terme de phylogénie.

Au total, lorsque les distances observées sont bien métriques et que les conditions d'additivité sont satisfaites pour tous les quatrets que l'on peut constituer à partir de toutes les UE, alors le choix de l'arbre non enraciné et l'estimation des longueurs de branches se font, en théorie, sans ambiguïté. Cette situation idéale n'est malheureusement pas la règle. Les difficultés surgissent du fait que les distances utilisées peuvent ne pas être nécessairement métriques (c'est le cas par exemple de la distance de Nei) et du fait que les distances observées ont peu de chances d'être additives, en raisons de l'homoplasie des caractères (changements multiples, réversions, convergences...) qui ne se répartit pas nécessairement de façon aléatoire le long des branches. Même si cette homoplasie se répartissait ainsi, les fluctuations dues à un échantillonnage des caractères insuffisant ou biaisé pourraient conduire à observer des distances non additives ou non métriques.

En conclusion, l'application d'une méthode de reconstruction qui suppose métricité et additivité dans des conditions où cette supposition n'est pas vérifiée ne peut conduire qu'à des résultats *a priori* contestables.

4. Méthodes phénétiques de construction d'arbres

Les différentes méthodes phénétiques peuvent être regroupées en plusieurs catégories : les méthodes agglomératives, les méthodes d'ajustement, les méthodes de parcimonie et les méthodes de vraisemblance. Chacune d'elles diffère à la fois par les hypothèses évolutives qu'elles impliquent et par les algorithmes qu'elles utilisent. Cette classification n'est cependant pas parfaitement rigide dans la mesure où certaines méthodes font appel à la fois à des procédures d'agglomération et à des procédures d'ajustement. Les méthodes de vraisemblance seront traitées dans le chapitre réservé à cette approche (chapitre VIII)

4.1. Les méthodes agglomératives

Il s'agit de regrouper ensemble les UE qui se ressemblent le plus et de situer les différents niveaux de hiérarchie entre elles sur la base de l'intensité de leur ressemblance. Une « taxinomie » de ces méthodes taxinomiques a été décrite par Sneath et Sokal (1973). Certaines méthodes contraignent les distances estimées à être ultra-métriques. C'est le cas des classifications hiérarchiques. Parmi ces dernières, on peut également distinguer les méthodes dites combinatoires qui se bornent à combiner les éléments de la matrice des distances au cours des séquences d'agglomération et les méthodes non combinatoires qui nécessitent de recalculer de nouvelles distances à chacune des étapes d'agglomération, à partir des données initiales elles-mêmes. Ces méthodes ne seront pas évoquées ici.

4.1.1. Les classifications hiérarchiques combinatoires.

De telles méthodes conduisent à l'estimation de distances ultra-métriques. Si les distances patristiques ou phylétiques ne sont pas elles-même ultra-métriques, on peut s'attendre à des distorsions importantes aussi bien dans l'estimation de la structure de l'arbre que dans l'estimation des longueurs des branches. L'hypothèse évolutive qui satisfait l'ultra-métricité de distances patristiques consiste à poser que les taux de mutation (pour des caractères qualitatifs) ou les vitesses de changements (pour des traits quantitatifs) sont identiques sur toutes les branches de l'arbre et donc que la distance phylétique est proportionnelle au temps évolutif. Cette hypothèse est souvent décrite sous le terme d'« horloge moléculaire » (*molecular clock*).

Il apparaît donc clairement que les méthodes de classification qui nécessitent la formulation d'une telle hypothèse ne peuvent être considérées comme des méthodes de reconstruction phylogénétique qu'en admettant le bien-fondé de l'hypothèse.

Le fait de retenir une telle hypothèse sur le processus évolutif a pour conséquence de permettre une localisation sans ambiguïté de la position de l'ancêtre : l'arbre est nécessairement enraciné. Encore une fois, si l'hypothèse d'ultra-métricité ou d'horloge moléculaire n'est pas fondée, la position de l'ancêtre risque aussi d'être totalement erronée.

Dans une première étape, on recherche les deux UE i et j les plus ressemblantes. On les regroupe alors en une UEH (unité évolutive hypothétique) résultante. Une nouvelle matrice de distances est alors calculée, par combinaison de distances dans laquelle l'UEH résultante remplace les deux UE i et j qu'elle fusionne. Le mode de calcul de cette nouvelle matrice varie d'une méthode à une autre (Lance et Williams, 1967). Le processus est continué jusqu'à ce que toutes les UE soient regroupées en une seule UEH.

Soit x une UE (ou une UEH) résultant de la fusion de r UE, dont l'UE i, et y une autre UE (ou UEH) résultant de la fusion de s UE, dont l'UE j.

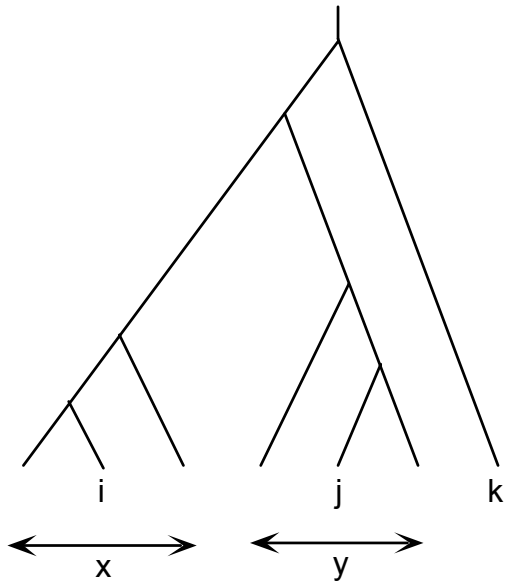


FIGURE VII.3 *(voir texte)*

La façon de calculer la distance d_{xy} entre les deux UE (ou UEH) x et y et celui de la distance $d_{k(xy)}$ entre l'UEH (xy), résultant de l'agglomération de x et y, et une autre UE (ou UEH) k définissent différentes méthodes de classification que l'on peut résumer ainsi (Lance et Williams, 1967) :

$$d_{k(xy)} = \alpha_x d_{xk} + \alpha_y d_{yk} + \gamma \left| d_{xk} - d_{yk} \right|$$

La distance d_{xy} et les coefficients α, β et γ prennent les valeurs suivantes :

1) Simple lien (*single linkage*) ou voisin le plus proche (*nearest neighbor*) :

$$d_{xy} = min\{d_{ij}\} \; ; \; \alpha_x = \alpha_y = 1/2 \; ; \; \gamma = -1/2$$

2) Lien complet (*complete linkage*) ou voisin le plus éloigné (*furthest neighbor*) :

$$d_{xy} = max\{d_{ij}\} \; ; \; \alpha_x = \alpha_y = 1/2 \; ; \; \gamma = 1/2$$

3) Lien moyen (*average linkage*) :

— UPGMA (*unweighted pair-group method of arithmetic averages*) :

$$d_{xy} = \frac{1}{rs}\sum_{i=1}^{r}\sum_{j=1}^{s} d_{ij} \; ; \alpha_x = \frac{r}{r+s}, \; \alpha_y = \frac{s}{r+s}, \; \gamma = 0$$

r et s sont les nombres d'UE qui sont comprises dans x et y respectivement.

— WPGMA (*weighted pair-group method of arithmetic averages*) :

$$d_{xy} = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{1}{2c_i}\frac{1}{2c_j} d_{ij} \; ; \alpha_x = \alpha_y = \frac{1}{2}, \; \gamma = 0$$

où c_i et c_j sont les nombres d'étapes précédant l'étape d'agglomération de x et y.

Parmi ces méthodes agglomératives, la méthodes UPGMA est la plus fréquemment utilisée (Sneath et Sokal, 1973).

Exemple I

Les distances de Kimura (deux paramètres, paragraphe 2.2.2) ont été calculées entre 5 espèces de primates à partir des séquences du gène de la ψ-η Globine (Tableau VII.3). Cette matrice de distances a été analysée par la méthode de l'UPGMA. L'homme et le gorille se retrouvent groupés (figure VII.4), alors que l'application d'autres méthodes (figure VII.7) contredit ce résultat.

	Hsa	Ptr	Ggo	Ppy	Mmu
Ptr	1.46				
Ggo	1.45	1.82			
Ppy	2.96	3.37	3.32		
Mmu	6.94	7.41	7.10	7.23	
Age	10.12	10.70	10.29	10.45	11.73

TABLEAU VII.3. *Comparaison deux à deux des séquences de la ψ-η Globine d'*Homo sapiens *(Hsa),* Pan troglodytes *(Ptr),* Gorilla gorilla *(Ggo),* Pongo pygmaeus *(Ppy),* Macaca Mulatta *(Mmu) et* Ateles geoffroyi *(Age), (Barriel et Darlu, 1990). La distance est celle de Kimura (1980).*

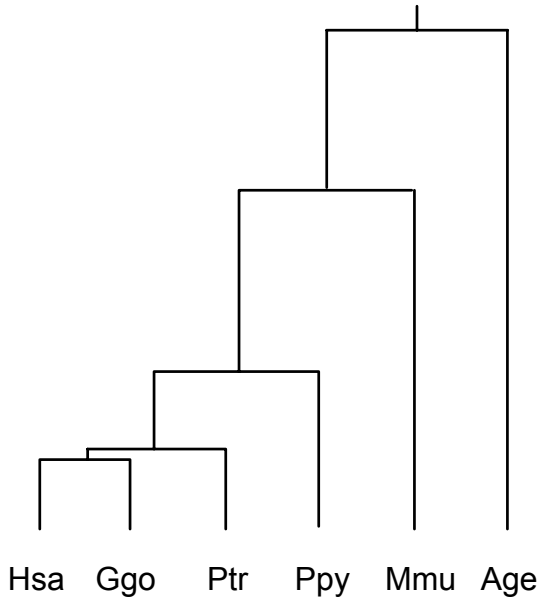


FIGURE VII.4. *Représentation de la matrice du tableau VII.3 par la méthode de l'UPGMA. L'homme et le gorille sont groupés ensemble.*

Exemple II

L'exemple de la figure VII.5 reprend les données du tableau V10. La distance choisie est la distance Manhattan, en raison de ses propriétés d'additivité et de

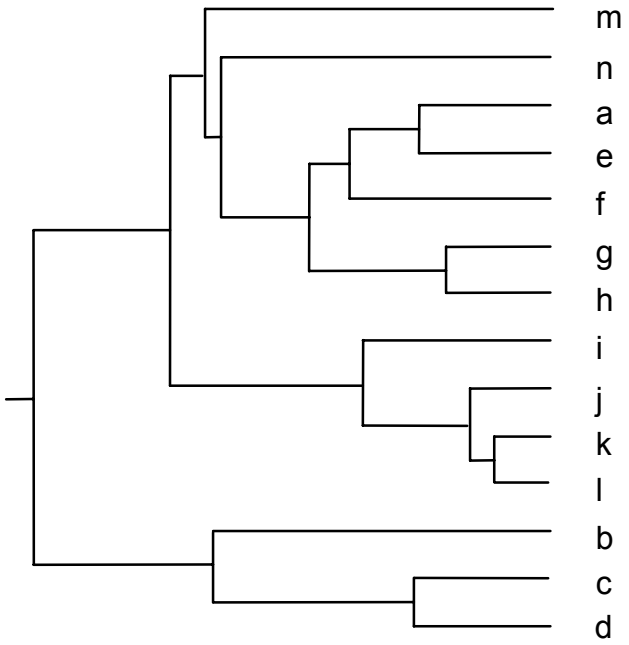


FIGURE VII.5.*Représentation des relations phénétiques entre 14 UE utilisant la matrice de distances Manhattan calculée à partir des données du tableau V.10. La méthode est celle de l'UPGMA.*

métricité. La distance entre deux UE est donc ici le nombre de caractères présent sous des états différents chez l'une et l'autre de ces UE. On peut remarquer que seuls les groupes monophylétiques (i,j,k,l) et (b,c,d) de la figure V.26 sont identifiés. Les groupes frères de ces groupes et toutes les autres combinaisons de taxons sont erronées par rapport à la figure V.26. La raison en est que la quantité d'évolution est très inégale selon les branches. Comme la méthode UPGMA ne tient pas compte de ce fait, il n'est pas surprenant que les taxons a, e et f se regroupent ensemble dans la mesure où ils ont tous peu divergés par rapport à leur ancêtre commun.

Exemple III

La séquence des acides aminés de la super-oxyde dismutase a été obtenue par Lee *et al.* (1985) pour la levure, la drosophile, le bœuf, le cheval et l'homme. Après avoir aligné les séquences, le nombre minimum de substitutions de nucléotides entre espèces prises deux à deux a été estimé puis transformé en distance par la formule de Jukes et Cantor (paragraphe 2.2.1). Le tableau VII.4 donne la matrice des distances et la figure VII.6 l'arbre obtenu par UPGMA. Il faut noter que cette représentation n'associe pas le cheval et le bœuf dans le même groupe, alors qu'ils sont classiquement regroupés dans les ongulés.

	H	C	B	D	L
Homme	0				
Cheval	0.100	0			
Bœuf	0.077	0.082	0		
Drosophile	0.237	0.249	0.234	0	
Levure	0.253	0.232	0.239	0.260	0

TABLEAU VII.4. *Distances de Jukes et Cantor calculées sur le nombre minimum de substitutions de nucléotides entre deux séquences alignées de la super-oxyde dismutase (d'après Lee* et al.*, 1985).*

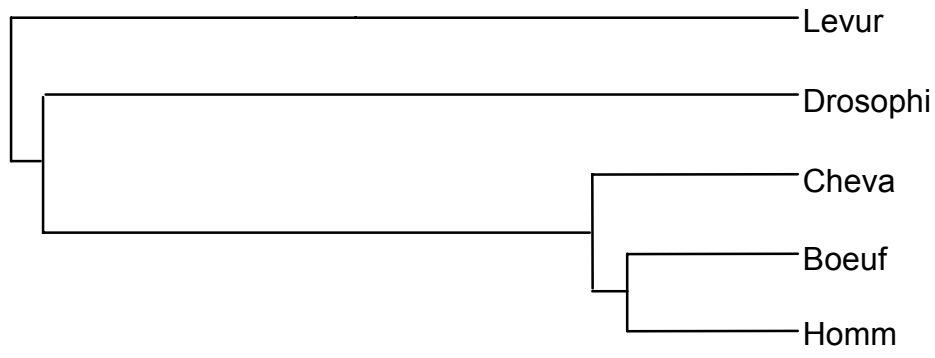


FIGURE VII.6. *Représentation des relations phénétiques entre la levure, la drosophile, le cheval, le bœuf et l'homme à partir de la matrice de distance du tableau VII.4 par la méthode de l'UPGMA.*

4.1.2. La méthode dite du «Neighbor-joining » (NJ) (Saitou et Nei, 1987)

Cette méthode, inspirée de celle proposé par Fitch et Margoliash (1967), Sattath et Tvsersky (1977) et Fitch (1981) se fonde sur une stratégie différente d'agglomération. A la différence des méthodes précédentes, elle n'impose pas aux distances estimées d'être ultra-métriques. L'hypothèse évolutive d'« horloge moléculaire » n'est donc pas posée. En revanche les distances doivent être métriques et additives (satisfaire les conditions des 4 UE, voir paragraphe VII.3.3) pour avoir l'assurance d'obtenir l'arbre de longueur minimum, c'est-à-dire l'arbre dont la somme des longueurs estimées soit minimale, et de permettre une estimation correcte des longueurs des branches.

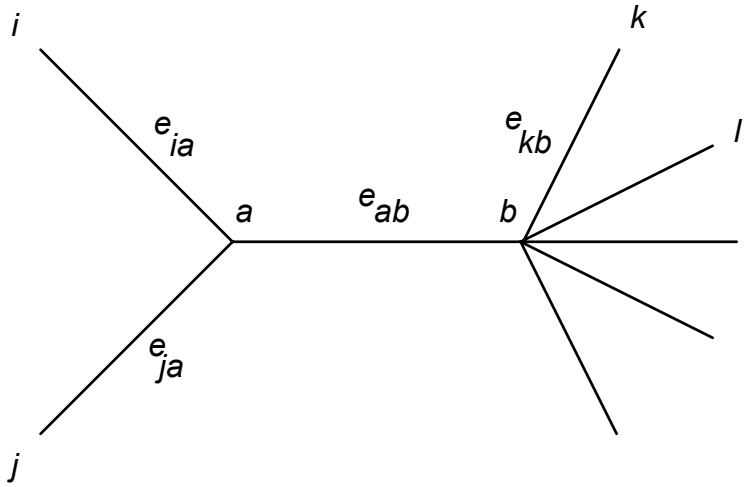


FIGURE VII.7. *Schéma représentant les branches reliant entre elles plusieurs UE (i, j, k, l...) ou UEH (a, b) et dont on cherche l'estimation e.*

On recherche, dans un premier temps, les deux UE i et j les plus «proches voisines» parmi N UE, c'est-à-dire les UE qui sont plus proches entre elles qu'elles ne le sont de toutes les autres UE. Comme l'ont démontré Saitou et Nei (1987) et Studier et Keppler (1988), les deux UE les plus proches voisines sont celles qui donnent la plus petite valeur de S_{ij} qui corresponde à la somme totale des distances *estimées* (longueur totale) d'une configuration comme celle de la figure VII.7, où k est l'une des UE branchées sur b.

$$S_{ij} = e_{ia} + e_{ja} + e_{ab} + \sum_{K \neq i \neq j}^{N} e_{kb}$$

Posant :

$$R_i = \sum_{k=1}^{N} d_{ik} \; ; R_j = \sum_{k=1}^{N} d_{jk} \; ; R = \sum_{i=1}^{N} \sum_{j>i}^{N} d_{ij}$$

où les *d* sont les distances observées entre UE.

On peut exprimer la valeur S_{ij} en fonction des distances observées, sachant que l'on peut démontrer que :

$$S_{ij} = \frac{1}{2(N-2)} \left[2R + (N-2) d_{ij} - (R_i + R_j) \right]$$

On peut ainsi choisir les deux UE i et j qui minimisent S_{ij}. De même, on peut calculer les longueurs e_{ia}, e_{ja} et e_{ab}, en remarquant que :

$$R_i = (N-1)e_{ia} + e_{ja} + e_{ab} + \sum_{k \neq i \neq j}^{N} e_{kb} = (N-2)e_{ia} + S_{ij}$$

$$R_j = (N-1)e_{ja} + e_{ia} + e_{ab} + \sum_{k \neq i \neq j}^{N} e_{kb} = (N-2)e_{ja} + S_{ij}$$

Les distances estimées e_{ia} et e_{ja} sont données en combinant les équations précédentes :

$$e_{ia} = \frac{d_{ij}}{2} + \frac{(R_i - R_j)}{2(N-2)}$$

$$e_{ja} = d_{ij} - e_{ia}$$

Une nouvelle matrice de distances est ensuite calculée, après avoir retiré les UE i et j pour les remplacer par l'UEH a. Les distances de cette UEH a aux autres UE sont celles préconisées par Fitch et Margoliash (1967) (paragraphe VII.3.3):

$$d_{ak} = \frac{1}{2}(d_{ik} + d_{jk} - d_{ij})$$

Sur cette nouvelle matrice, on recherche de nouveau les deux UE (ou UE et UEH) qui minimisent la longueur totale de l'arbre exprimée par l'équation donnant S_{ij}. Le processus est poursuivi ainsi, jusqu'à ce que toutes les UE soient agglomérées et les longueurs estimées.

Exemples

Les données du tableau V.10 ont été transformées en matrice de distances Manhattan. L'application de la méthode du *neighbor joining* conduit à un arbre qui est discuté et comparé à celui donné par la méthode des moindres carrés (paragraphe 4.2.2 ; figure VII.11).

Les données du tableau VII.3 ont été analysées par la méthode du *neighbor joining* (figure VII.7). En comparant ce résultat à celui donné par la méthode de l'UPGMA (Figure VII.4), on s'aperçoit que les espèces ne sont pas regroupées de la même façon (cette fois l'homme et le chimpanzé sont regroupés), mais également que l'égalité des longueurs de branches à partir d'un ancêtre commun n'est pas respectée (par exemple une longueur de 94 entre l'ancêtre et *Pan troglodytes* contre 52 seulement entre l'ancêtre et *Homo sapiens)*. La question se pose donc de la validité de l'hypothèse d'horloge moléculaire. Cet exemple sera repris plus loin.

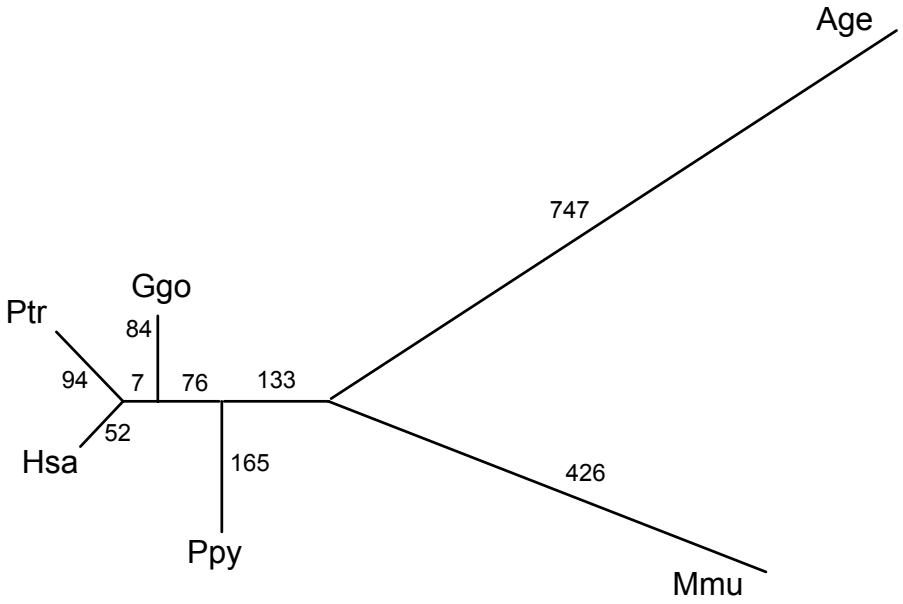


FIGURE VII.7. *Relations entre les Hominoidea obtenues à partir des pourcentages de divergence (tableau* VII.3*) en appliquant la méthode du* neighbor-joining *de Saitou et Nei (1987). Contrairement à la figure* VII.4*,* Homo sapiens *est groupé avec* Pan troglodytes.

La méthode du NJ donne l'arbre non enraciné de longueur minimale si les conditions d'additivité des distances observées (conditions des 4 points, paragraphe VII.3.3) sont satisfaites. Ces distances peuvent ne pas être additives en raison, par exemple, de la présence d'homoplasie ou parce que les corrections effectuées pour tenir compte des changements multiples ne l'ont pas été sous des hypothèses évolutives correctes. Remarquons enfin que l'estimation de la distance e_{ia} s'effectue en corrigeant une estimation de distance qui serait ultra-métrique ($e_{ia} = d_{ij}/2 = e_{ja}$) par un facteur représentant la différence entre la divergence moyenne de i (R_i) et la divergence moyenne de j (R_j) chacune d'elles étant estimées sur l'ensemble des UE. Il apparaît donc qu'un arbre non enraciné qui présenterait une très large variabilité de longueurs des branches ne pourrait qu'augmenter l'imprécision du facteur de correction.

4.2. Les méthodes d'ajustement

Ces méthodes consistent à rechercher l'arbre non enraciné et à estimer les longueurs des branches qui donnent le meilleur ajustement à la matrice des distances observées. Le choix de l'arbre et l'estimation de la longueur des branches se font, simultanément ou séparément, sur la base d'un critère ou d'une fonction à minimiser qui peut varier d'une méthode à l'autre. La méthode des moindres carrés est souvent utilisée à cette fin d'ajustement, et c'est elle qui sera décrite ici. Les méthodes par maximum de vraisemblance sont également utilisables (chapitre VIII).

Le plus souvent ces méthodes ne permettent pas de situer la place de l'origine (l'ancêtre) sur l'arbre non enraciné qu'elles infèrent. D'autres procédures sont habituellement proposées pour cela (extra-groupe, par exemple) ou bien il est nécessaire de faire l'hypothèse que l'évolution le long des branches est constante (hypothèse dite de l'« horloge »).

4.2.1. Le modèle

Un modèle statistique est posé comme fondement de cette méthode (Felsenstein, 1984a; pour une discussion lire Farris, 1981, 1985, 1986 et Felsenstein, 1986 ; Bulmer, 1991). On suppose que la distance entre deux UE i et j, d_{ij} , suit une distribution normale dont la valeur attendue est e_{ij}, et la variance ω_{ij}.

Les distances attendues e_{ij} doivent être *additives*, c'est-à-dire que la distance attendue entre i et j doit être égale à la somme des différentes distances attendues formant le chemin reliant i à j. Eventuellement, la distance observée doit être « corrigée » afin de la situer sur une échelle de distance additive. Comme il a déjà été souligné (paragraphe VII.3) cette méthode estime des distances e_{ij} qui ne sont pas les distances patristiques recherchées. Elles ne le sont que lorsque la distance entre deux UE n'est pas due (ou de façon négligeable) à l'homoplasie ou bien lorsque cette homoplasie a pu être prise en compte au moyen d'un modèle d'évolution admis et vérifié par ailleurs (paragraphe VII.3.2.b). La discussion reste vive sur le fait de savoir quelle signification peuvent avoir des estimations de longueurs de branches lorsque les distances observées ne sont pas elles-mêmes additives.

Ce modèle suppose également que les erreurs statistiques sur les différentes distances sont indépendantes. Si tel n'est pas le cas, il est préférable d'en tenir compte en calculant, si possible, la matrice de covariance entre distances et en la prenant en compte lors des estimations (Cavalli-Sforza et Edwards 1967 ; Chakraborty, 1977 ; Farris, 1981; Bulmer, 1991).

Sous ce modèle, les distances négatives sont concevables : elles sont interprétées alors comme une simple conséquence des fluctuations aléatoires de part et d'autre de la distance attendue. En fait, ces distances négatives peuvent également résulter d'une absence de métricité ou d'additivité des distances (tableau VII.2 et figure VII.9). Il reste parfois difficile de trancher entre ces deux explications, la deuxième conduisant à admettre que les estimations obtenues ne sont pas interprétables en terme de quantités d'évolution le long d'une branche.

A propos des distances estimées négatives, plusieurs attitudes sont possibles :

— ne pas considérer les arbres produisant des distances négatives ;

— rechercher l'arbre minimisant le critère d'ajustement (même s'il possède des distances négatives) et ajuster les longueurs de branches en posant que ces distances sont nulles ;

— rechercher l'arbre optimal en contraignant les distances négatives à être nulles.

Quelle que soit la stratégie retenue, il est conseillé de ne pas opter pour une approche qui conduirait à feindre d'ignorer la présence de distances négatives alors qu'elles existent. La méthode des moindres carrés appliquée à la reconstruction d'arbre (Kidd et Sgaramella-Zonta, 1971 ; Chakraborty, 1977 ;

Bulmer, 1991) peut être présentée de la façon suivante, en définissant successivement :

— le vecteur colonne *D* des distances observées entre s UE, comportant donc s(s – 1)/2 distances (ici r est la s –1$^{\text{ième}}$ UE) :

$$D = (d_{12}, d_{13}, d_{14}, ..., d_{23}, d_{24}, ...d_{rs})'$$

— le vecteur colonne correspondant aux distances estimées *E* est :

$$E = (e_{12}, e_{13}, e_{14}, ..., e_{23}, e_{24}, ...e_{rs})'$$

— la matrice $\boldsymbol{W}$ dont l'élément $\omega_{ij,kl}$ représente la covariance entre la distance d_{ij} et la distance d_{kl}. Les variances des distances se situent sur la diagonale de cette matrice.

$$\boldsymbol{W} = \begin{vmatrix} \omega_{12}^2 & \omega_{12,13} & \omega_{12,14} & \cdots & \omega_{12,rs} \\ \omega_{13,12} & \omega_{13}^2 & \omega_{13,14} & \cdots & \omega_{13,rs} \\ \omega_{14,12} & \omega_{14,13} & \omega_{14}^2 & \cdots & \omega_{14,rs} \\ \cdots & \cdots & \cdots & \cdots & \\ \omega_{rs,12} & \omega_{rs,13} & \omega_{rs,14} & \cdots & \omega_{rs}^2 \end{vmatrix}$$

— le vecteur colonne des 2*s* – 3 longueurs de branches :

$$L = (a_1, a_2, a_3, a_4, a_h ...)'$$

— la matrice de « passage », $\boldsymbol{A}$, avec les distances observées en lignes (dimension s(s – 1)/2) et les différentes branches en colonnes (dimension 2s – 3). Cette matrice $\boldsymbol{A}$ définit précisément la forme de l'arbre. Pour expliquer simplement cette matrice, considérons l'exemple d'arbre suivant, comportant 4 UE et 6 longueurs de branches (a_1 à a_6).

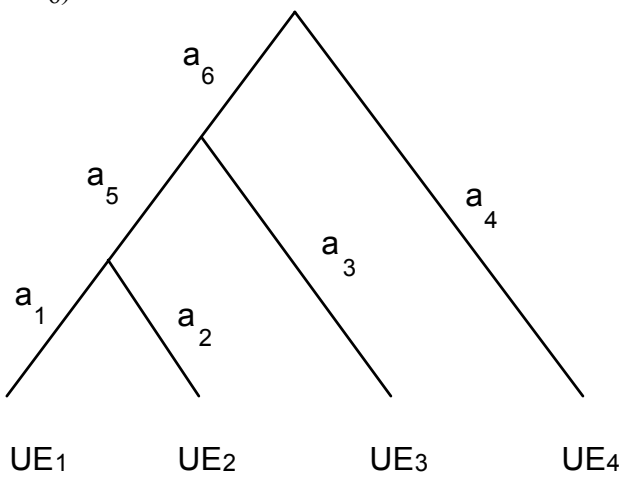


FIGURE VII.8. *Schéma d'arbre à six branches reliant quatre UE.*

Un élément de cette matrice prend la valeur 1 lorsque le chemin passant de l'UE i à l'UE j, et définissant la ligne d_{ij}, passe par l'une des branches définie en colonne. Il prend la valeur 0 autrement. Par exemple, l'élément à l'intersection de

la ligne d_{13} et de la colonne a_5, prend la valeur 1 puisque le chemin passant de l'UE 1 à l'UE 3 passe par la branche a_5. La matrice $\boldsymbol{A}$, de dimensions 6x6 (Tableau VII.5), est une façon de représenter la structure de l'arbre de la figure VII.8.

		a1	a2	a3	a4	a5	a6
	d12	1	1	0	0	0	0
	d13	1	0	1	0	1	0
A =	d14	1	0	0	1	1	1
	d23	0	1	1	0	1	0
	d24	0	1	0	1	1	1
	d34	0	0	1	1	0	1

TABLEAU VII.5. *Matrice définissant la structure de l'arbre de la figure VII.8.*

On a donc : $E = L\boldsymbol{A}$

L'estimation du vecteur L par la méthode des moindres carrés est donnée en minimisant la somme pondérée des carrés des écarts (SCE) suivante :

$$SCE = (D - E)' \boldsymbol{W}^{-1} (D - E)$$

d'où

$$SCE = (D - L\boldsymbol{A})' \boldsymbol{W}^{-1} (D - L\boldsymbol{A})$$

et, extrayant le vecteur estimé L :

$$L = (\boldsymbol{A}'\boldsymbol{W}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{W}^{-1}D$$

La variance du vecteur L est :

$$V(L) = (\boldsymbol{A}'\boldsymbol{W}^{-1}\boldsymbol{A})^{-1}$$

L'exemple précédent inclut, dans la matrice $\boldsymbol{A}$, un vecteur a_6 dont les éléments sont tous égaux à ceux du vecteur a_4.

Dans ces conditions, la matrice $(\boldsymbol{A}'\boldsymbol{W}^{-1}\boldsymbol{A})^{-1}$ ne peut évidemment pas être inversée. Cet exemple n'est donné que pour illustrer l'impossibilité de déterminer la racine de l'arbre par cette méthode : la longueur estimée sera donc la somme (a_4+a_6), sans pouvoir distinguer entre les deux. De ce fait, la racine disparaît et l'on obtient donc un arbre non enraciné.

4.2.2. La construction de l'arbre

La méthode de reconstruction phylogénétique se propose donc ici d'estimer la structure de la matrice $\boldsymbol{A}$ (donc la structure de l'arbre non enraciné) et les longueurs des branches qui minimisent la somme des carrés des écarts (*SCE*).

Cette dernière peut prendre différentes formes selon les valeurs que l'on donne à la matrice de variance-covariance $\boldsymbol{W}$.

1) *Méthode des moindres carrés ordinaires*. Dans cette méthode proposée par Cavalli-Sforza et Edwards (1967), la matrice $\boldsymbol{W}$ est une matrice diagonale où tous les éléments sont égaux. La variance des distances, ω_{ij}, est supposée être indépendante de i et de j et est donc constante, signifiant par là que l'erreur sur l'estimation de la distance est indépendante de la distance elle-même. On pose donc $\omega_{ij} = 1$:

$$SCE = \sum_{i=1}^{n} \sum_{j=1}^{i} \frac{(d_{ij} - e_{ij})^2}{d^2_{ij}}$$

avec (n = s(s – 1)/2)

2) *Méthode des moindres carrés pondérés.* Fitch et Margoliash (1967) proposent une pondération telle que l'erreur sur la distance soit proportionnelle à la distance e_{ij} ou, en première approximation, à la valeur d_{ij} qui s'en approche. Dans ce cas l'erreur sur la distance est d'autant plus grande que cette distance est importante. La matrice $\boldsymbol{W}$ est donc la matrice où les éléments de la diagonale sont les variances des distances et où les covariances entre distances sont nulles. On pose donc $\omega^2_{ij} = d^2_{ij}$.

Fitch et Margoliash proposent par ailleurs le pourcentage de déviation standard *(%SD)* pour estimer la qualité de l'ajustement :

$$\%SD = 100 \left[\sum_{i=1}^{n} \sum_{j=1}^{i} \left(\frac{d_{ij} - e_{ij}}{d_{ij}}\right)^2 / \left(\frac{n(n-1)}{2}\right) \right]^{\frac{1}{2}}$$

3) *Méthode des moindres carrés généralisées*. Cette méthode suggérée par Cavalli-Sforza et Edwards (1967), Chakraborty (1977), Farris (1981) et développée par Bulmer (1991) tient compte cette fois des covariances existant entre les distances. L'estimation de ces covariances dépend de la nature du matériel utilisé (séquences d'acides aminés, de nucléotides par exemple) et des hypothèses évolutives retenues. Cette méthode présente l'intérêt de tenir compte du fait que les distances ne sont pas nécessairement indépendantes. En effet, dans l'exemple de la figure VII.8, les distances d_{13} et d_{23} ont une partie commune (donc une covariation commune) formée de la longueur a_5.

4) *En marge de ce modèle* des moindres carrés, Farris (1972) propose de comparer distances observées et distances estimées par un indice f tel que :

$$f = \sum_{i=1}^{n} \sum_{j=1}^{i} |d_{ij} - e_{ij}|$$

Dans ce cas l'arbre retenu sera celui minimisant la somme des différences absolues entre distances estimées et distances observées.

Exemple 1

A partir de la matrice du tableau VII.3, l'arbre obtenu par la méthode des moindres carrés pondérés en utilisant le programme Fitch de *Phylip* ne diffère pas de celui obtenu par la méthode du NJ (figure VII.7). Les longueurs de branches sont identiques par les deux méthodes. La valeur ω_{ij} choisie ici est estimée par la distance observée d_{ij} puisque la distance utilisée (indice de Kimura à deux paramètres, Chapitre VII.2.2) varie linéairement avec sa variance (Kimura, 1980 ; Nei, 1987).

Exemple 2

La matrice des distances (tableau VII.4) a été analysée par la méthode des moindres carrés pondérés par l'inverse du carré de la distance. La figure VII.6 montre que l'arbre non enraciné donnant le meilleur ajustement présente une longueur de branche estimée négative (figure VII.9a). On peut remarquer d'ailleurs que les distances ne satisfont ni l'égalité ni l'inégalité de la condition d'additivité des 4 points :

$$d_{ch} + d_{db} < d_{bce} + d_{hd} = d_{hb} + d_{cd}$$
$$0.334 > 0.319 \neq 0.326$$

Si l'on contraint les distances estimées à être positive, l'homme et le bœuf se retrouvent groupés (figure VII.9b)

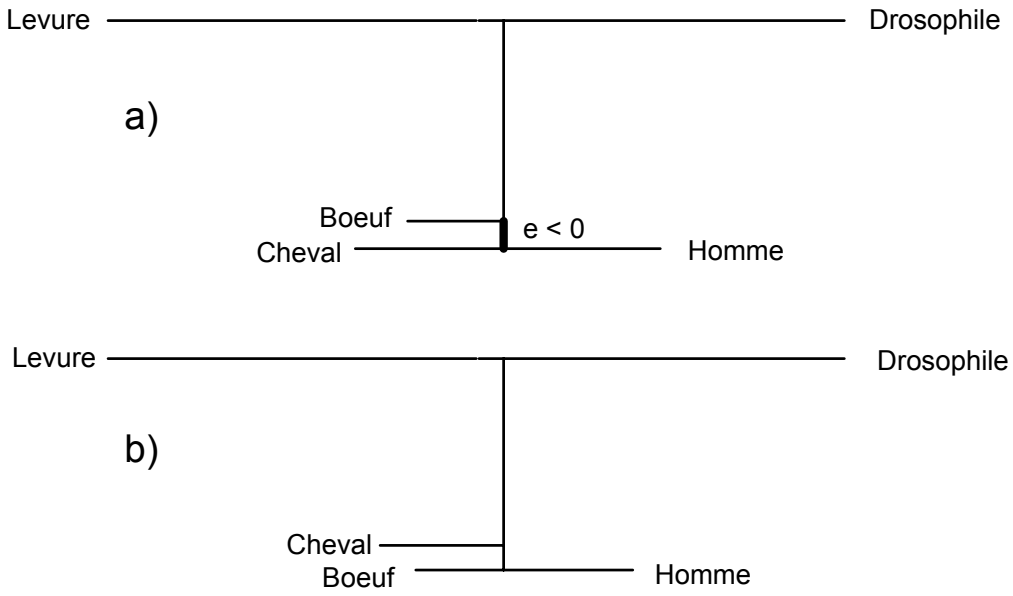


FIGURE VII.9. *Arbres non enracinés reconstruits par la méthode des moindres carrés pondérés (Fitch et Margoliash, 1967) à partir de la matrice du tableau VII.4, en admettant (a) ou non (b) des distances estimées négatives (e).*

Exemple 3

Les données suivantes portent sur le polymorphisme génétique de différentes populations humaines. Ce polymorphisme a été observé sur 100 sites répartis sur l'ensemble des chromosomes. Les distances (Reynolds *et al.*, 1983) ont ensuite été calculées entre les populations, puis un arbre non enraciné a été obtenue par la méthode des moindres carrés ordinaires. On peut remarquer que les distances ne satisfont pas les conditions d'additivité. En effet, comme le montre le tableau VII.6, on a :

$$d_{ac} + d_{de} < d_{ae} + d_{cd} \neq d_{ad} + d_{ce}$$

C'est la raison pour laquelle un arbre ayant une branche de longueur négative ajuste légèrement mieux les données. Lorsque de telles longueurs ne sont pas admises, l'arbre obtenue est celui de la figure VII.10. Il montre une longueur estimée pour les européens qui est très courte et qui peut s'interpréter, selon les auteurs, comme le résultat d'un mélange survenu entre deux sortes de populations : celle à l'origine des Chinois actuels et celle à l'origine des populations pygmées actuelles.

	a	b	c	d
a - Pygmées (RCA)	0.000			
b - Pygmées (Zaïre)	0.043			
c - Européens	0.141	0.142		
d - Chinois	0.235	0.235	0.093	
e - Mélanésiens	0.242	0.265	0.148	0.171

TABLEAU VII.6. *Matrice de distances (F_{st}, Reynolds et al., 1983), obtenue à partir de 100 marqueurs d'ADN sur 5 populations (Bowcock et al, 1991).*

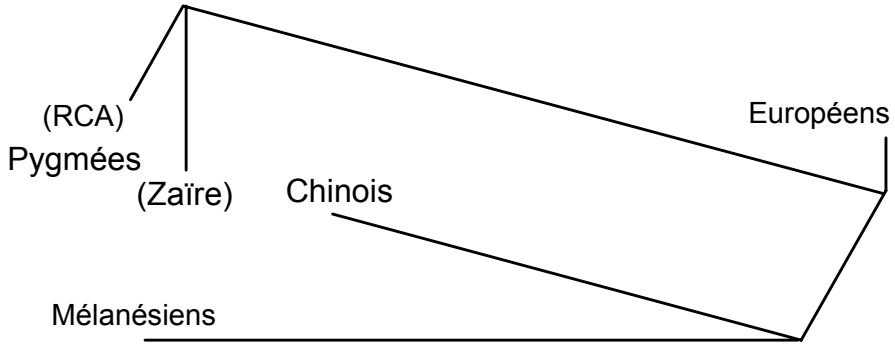


FIGURE VII.10. *Arbre non enraciné reconstruit par la méthode des moindres carrés (Cavalli-Sforza et Edwards, 1967) à partir de la matrice du tableau VII.6, en contraignant les longueurs des branches à être positives. La courte branche conduisant aux Européens laisse supposer qu'ils résultent d'un mélange (Bowcock* et al. *1991).*

Exemple 4

Cet exemple reprend les données du tableau V.10 à partir duquel les distances Manhattan entre les UE prises deux à deux ont été calculées. L'arbre non enraciné de la figure VII.11 a été obtenu par la méthode des moindres carrés ordinaires contraignant les longueurs estimées à être positives. Cet arbre non enraciné est à comparer à celui de la figure VII.5 obtenue par la méthode de l'UPGMA, à celui de la figure V.27 obtenue par la méthode de parcimonie et à celui obtenu par la méthode de compatibilité (figure VI.4).

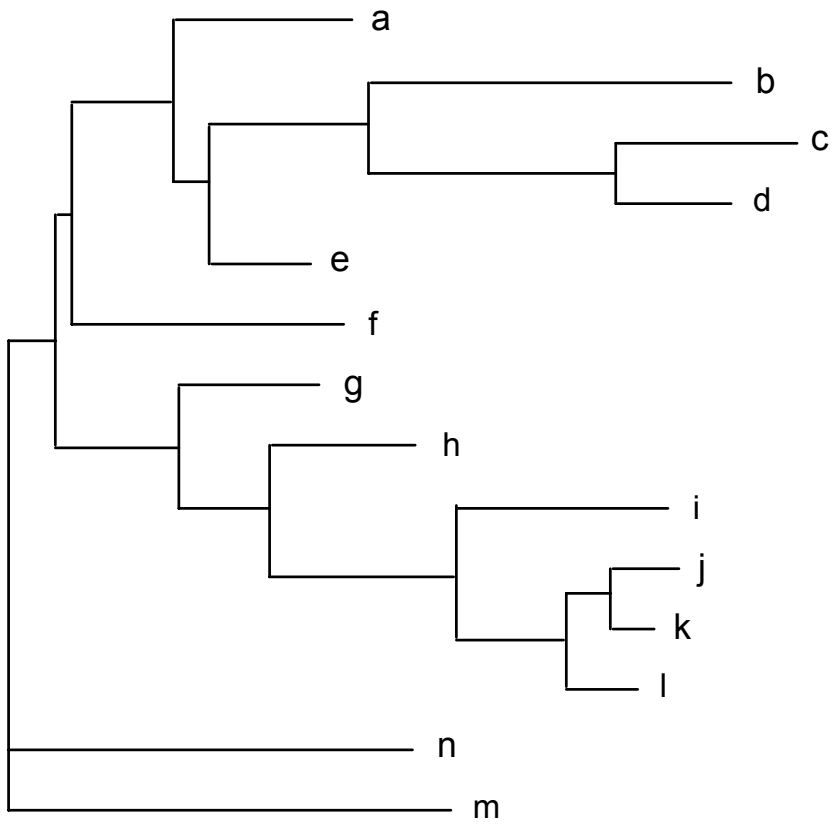


FIGURE VII.11 *: Représentation des relations (arbre non enraciné) entre 14 UE utilisant la matrice de distances Manhattan calculée à partir des données du tableau V.10. La méthode employée est celle des moindres carrés (Cavalli-Sforza et Edwards, 1967), contraignant les longueurs à être positives.*

Si l'on considère m et n comme les extra-groupes, le groupe (g(h(i(j(k,l))))) est ici identifié correctement, ainsi que les relations de parenté à l'intérieur de ce groupe. En revanche f est considéré à tort comme le groupe frère de (a(e(b(c,d)))) et non comme le groupe frère de (g(h(i(j(k,l))))). L'hypothèse de monophylie de (e(b(c,d))) est également erronée.

L'arbre obtenu par la méthode du NJ place correctement f en position de groupe frère de (g(h(i(j(k,l))))) mais maintient l'erreur de monophylie du groupe (e(b(c,d))).

4.2.3. Quelques tests statistiques

Test F

Comme on l'a vu, les méthodes d'ajustements consistent à rechercher l'arbre qui minimise la somme des carrés des écarts (*SCE*) entre distances observées et distances estimées. Cette somme est obtenue sans poser de contrainte *a priori* sur les longueurs de branches. Lorsque le nombre d'UE est de n, le nombre de longueurs estimées est de $2n - 3$. Ces estimations sont tirées de l'observation de $N = n(n - 1)/2$ distances.

On peut cependant imposer, *tout en maintenant la même topologie*, que les longueurs de branches entre chaque ancêtre et les deux UE qui en descendent soient égales (figure VII.12). Dans ce cas, la somme des carrés des écarts (SCE_0) obtenue mesure la qualité de l'ajustement lorsque l'on fait l'hypothèse d'une « horloge » identique sur toutes les branches ; seulement $n - 1$ longueurs sont alors estimées.

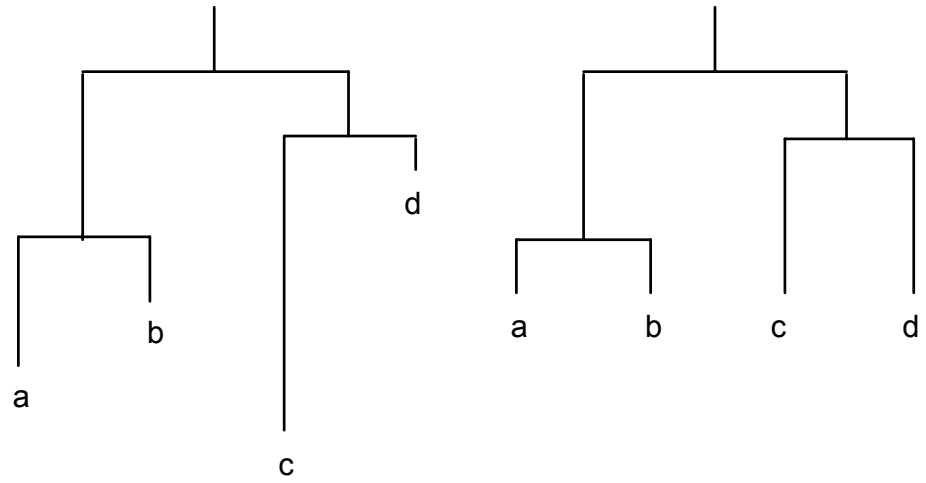


FIGURE VII.12. *Deux arbres de structure identique ne différant que par les longueurs des branches. A gauche : arbre obtenu sans poser de contrainte sur les longueurs ; à droite : la contrainte d' « horloge » est posée, c'est-à-dire l'égalité des branches des UE depuis leurs ancêtres communs.*

La différence entre SCE_0 et *SCE* représente l'excès de variation due à l'hypothèse d'égalité des longueurs de branches, tandis que *SCE* représente la variation due aux erreurs résiduelles. Si cette différence, $SCE_0 - SCE$, n'est pas plus grande que l'erreur résiduelle *SCE*, l'hypothèse d'« horloge » ne sera pas rejetée. Le rapport suivant :

$$F = \frac{\dfrac{SCE_0 - SCE}{n - 2}}{\dfrac{SCE}{N - (2n - 3)}}$$

se distribue comme un F de Fisher à $(n - 2)$ et $(N - (2n - 3))$ degrés de libertés. L'hypothèse d'« horloge » sera rejetée lorsque la probabilité d'observer une certaine valeur de ce rapport dépasse un seuil choisi (5% ou 1% par exemple).

Exemple

L'ajustement de l'arbre de la figure VII.7 (Primates) donne un $SCE = 0.015$ et un $SCE_0 = 0.469$ pour $n = 6$ UE. La valeur de F est donc 44.6. Si l'hypothèse d'égalité des longueurs de branches depuis l'ancêtre était fondée, la probabilité d'observer une telle valeur de F serait inférieure à $P < 0.001$. On est donc amené raisonnablement à rejeter cette hypothèse.

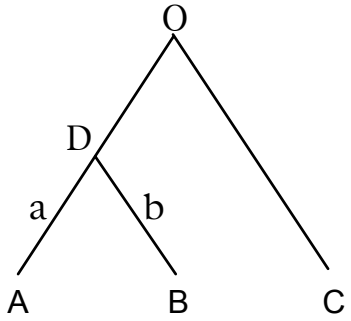
Ce test proposé par Felsenstein (1984a ; 1985a) dans le cadre de la méthode des moindres carrés ordinaires, n'est valable que sous certaines hypothèses, rarement vérifiées, en particulier que les distances observées sont indépendantes, que chacune d'elles est obtenue à partir de données différentes, et que la distance varie linéairement avec le temps. Comme le remarque justement Felsenstein (1988), les distances obtenues à partir de données moléculaires ne satisfont pas ces exigences. Il faut noter également que l'on teste ici la présence d'une horloge qui s'exercerait simultanément sur *toutes* les branches.

Dans le contexte de la méthode des moindres carrés généralisée, certaines de ces contraintes peuvent être prises en compte (linéarisation de la relation entre distances et temps, interdépendance des distances) (Bulmer, 1991). Cependant il faut insister sur une importante limitation : une telle approche dépend de la validité du modèle évolutif que l'on a retenu pour expliquer les changements d'états des caractères.

Le test du taux relatif

A la différence du précédent, ce test (Sarich et Wilson, 1973) se propose, dans un premier temps, de ne tester que l'égalité ou la différence de deux branches.

L'application des formules élaborées paragraphe VII.3.3 (Fitch et Margoliash, 1967) à l'arbre de la figure VII.13 montre que si l'on veut tester l'égalité des longueurs de branches a et b, il suffit de comparer leurs estimations qui s'expriment en fonction des distances observées :



$$a = (d_{AC} + d_{AB} - d_{BC})/2$$
$$b = (d_{AB} + d_{BC} - d_{AC})/2$$

FIGURE VII.13. *Le test du taux relatif revient à comparer a et b.*

La différence entre a et b est nulle lorsque les longueurs de branches sont égales. Cette différence $(a - b)$ peut être estimée par $d_{AC} - d_{BC}$.

Un test χ^2 a été proposé par Fitch (1976). Il consiste à calculer :

$$\chi^2_{ddl=1} = \frac{(a-b)^2}{a+b} = \frac{(d_{AC} - d_{BC})^2}{d_{AB}}$$

Ce test peut se généraliser dans la mesure où il est possible de calculer un nombre considérable de χ^2 pour un seul et même arbre. On peut ensuite en étudier les distributions (Sherer, 1989). Ce test a longuement été discuté, commenté et critiqué, par Fitch lui-même (1976).

Variance des longueurs de branches

Plusieurs auteurs ont proposé des méthodes pour donner un intervalle de confiance à l'estimation des longueurs de branches ou des points de branchement. Chakraborty (1977) se fonde sur la théorie de l'estimation des moindres carrés pour obtenir les variances des longueurs de branches (voir le vecteur *V(L)* paragraphe VII.4.2). Il suppose l'existence d'une horloge et un processus de Poisson expliquant la substitution des acides aminés. Nei *et al.* (1985) dérivent plusieurs formules donnant, cette fois, l'estimation de la variance du temps de branchement en fonction de la nature de l'information retenue pour calculer les distances (séquences d'acides aminés, de nucléotides ou sites de restriction). La méthode de reconstruction est ici l'UPGMA. Elle ne s'applique donc que si l'hypothèse d' « horloge moléculaire » est vérifiée. Li (1989) présente une méthode d'estimation des variances des branches qui ne nécessite pas une telle horloge.

Enfin, en utilisant une méthode analogue au test F, il est possible de tester si la longueur d'une branche est significativement différente de 0 (autrement dit, on teste l'existence d'une trifurcartion). En effet il suffit de comparer, pour un arbre donné, la valeur SCE obtenue en estimant cette longueur à celle obtenue en supposant que cette longueur est nulle. Pour plus de détails, voir Felsenstein (1986) et Bulmer (1991).

4.3. Les méthodes de parcimonie

Ces méthodes (Farris, 1972 ; Tateno *et al.*, 1982 ; Faith, 1985) se proposent de trouver, à partir d'une matrice de distances, un arbre non enraciné minimisant globalement la part des distances due aux homoplasies. Elles recherchent donc l'arbre le plus court (en nombre d'événements évolutifs) et c'est en ce sens qu'elles sont dites méthodes de parcimonie.

La méthode est fondée sur la résolution de deux problèmes distincts :

1) Quelle UE choisir parmi toutes celles qui ne sont pas encore intégrées à l'arbre non enraciné, pour l'insérer à son tour sur l'arbre ?

2) Comment estimer les longueurs des branches créées par cette adjonction ?

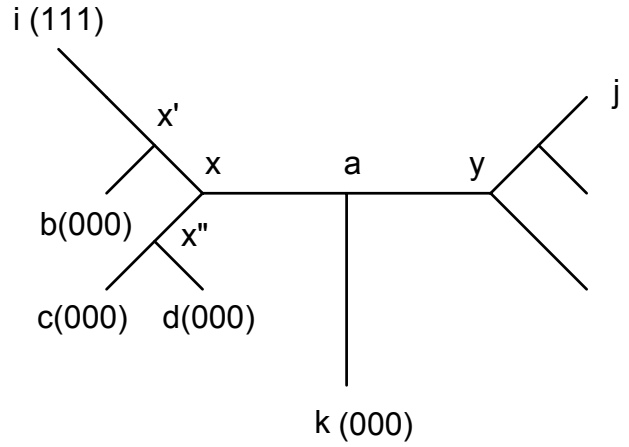


FIGURE VII.14. *Schéma d'adjonction d'une UE sur un arbre non enraciné préexistant (voir texte).*

Farris (1972) définit d'abord la « similitude spéciale » entre une UE k et la branche reliant deux UE (ou UEH) x et y (figure VII.14) par :

$$s_{k(xy)} = \frac{1}{2}(d_{kx} + d_{ky} - d_{xy})$$

Dans ce contexte, la distance utilisée est la distance Manhattan qui décompte simplement le nombre de caractères partagés entre deux UE (paragraphe VII.2.2.1) et pour laquelle l'interprétation d'une longueur de branche se fait sans ambiguïté. La similitude spéciale peut être considérée comme la distance estimée entre k et la branche reliant x à y, c'est-à-dire la distance e_{ka}. En fait elle n'est pas utilisée à cette fin mais *seulement* comme critère d'agglomération. En effet l'UE (ou UEH) k est insérée entre les deux UE (ou UEH) x et y pour lesquelles la similitude spéciale $S_{k(xy)}$ est la plus faible. C'est une solution parcimonieuse en ce sens que ce choix minimise les changements d'états de caractères (et donc la distance s'il s'agit d'une distance Manhattan) qui sont exigés par l'implantation de l'UE k dans l'arbre entre x et y. Pour que l'estimation des pas supplémentaires nécessités par cette insertion soit correcte, il faut supposer que les événements homoplasiques sont inexistants ou rares dans les segments reliant x, y et k. Cela impose donc que les UE (ou UEH) x, y et k soient très proches pour supprimer toute possibilité d'apparition d'homoplasie.

L'arbre est construit par agglomérations successives sur la base d'une minimisation du critère de similitude spéciale. Cependant la simple insertion d'une nouvelle UE dans l'arbre précédemment obtenu ne permet pas de remettre en cause les relations établies dans les étapes antérieures, si bien que le résultat final est dépendant de l'ordre d'introduction des UE au cours de la construction de l'arbre : si chacune des étapes peut être considérée comme la plus parcimonieuse, le résultat final ne l'est pas nécessairement. Des stratégies supplémentaires de permutations de branches (« *branch swapping* ») sont donc nécessaires pour obtenir une meilleure optimisation. Il faut remarquer également que cette méthode nécessite de calculer, à chaque étape, des longueurs de branches qui sont utilisées à l'étape suivante d'agglomération.

La deuxième question est celle de l'estimation de la longueur des branches. La méthode de Farris consiste à attribuer aux longueurs reliant k à x et y les valeurs maximales suivantes :

$$e_{kx} = max_i [e_{ki} - e_{xi}]$$

$$e_{ky} = max_j [e_{kj} - e_{yj}]$$

où i est l'une des r UE, ou l'un des nœuds (UEH) branchés sur x et j l'une des s UE, ou l'un des nœuds (UEH) branchés sur y.

La distance estimée entre k et a est simplement équivalente à la similitude spéciale, en remplaçant les distances observées par les distances estimées :

$$e_{ka} = \frac{1}{2}(e_{kx} + e_{ky} - e_{xy})$$

La distance e_{xy} est la distance observée d_{xy} entre x et y quand i et j sont des UE et non des UEH (c'est le cas au départ de l'algorithme).

Farris justifie cette procédure par le fait que la distance patristique ne peut être que *supérieure* à la distance observée, en raison des homoplasies. Pour minimiser l'écart entre distance patristique et distance observée, il est donc amené à choisir la distance observée maximale. Une conséquence indésirable de cette procédure est de ne pas être sensible à l'introduction de plusieurs UE (par exemple sur x) dont les distances à k sont faibles : en effet cela ne modifie pas l'estimation des distances entre k, a et x, alors même que ces UE supplémentaires peuvent être très voisines ou même parfaitement ressemblantes à x et k. Ainsi dans la figure VII.14, trois UE b, c et d, ont été branchées sur l'UE x. Trois caractères codés chacun 0 ou 1 sont observés pour chaque UE. Ces trois UE présentent les mêmes états de caractères que k (000), tandis que i possède les états de caractères (111). La distance entre k et x reste égale à 3, sans que la présence de b, c ou d, ne la modifie en aucune façon alors qu'aucun événement n'intervient entre b, c, d et k (distances égales à 0).

Pour pallier cette difficulté, Tateno *et al.* (1982) proposent une modification de la procédure de Farris, en ne choisissant pas les estimations fondées sur des maximums mais celles fondées sur des moyennes :

$$e_{kx} = \frac{1}{r} \sum_i [d_{ki} - d_{xi}]$$

$$e_{ky} = \frac{1}{s} \sum_j [d_{kj} - d_{yj}]$$

d'où, l'estimation de la distance entre k et la branche (xy), $s_{k(xy)}$ étant la similitude spéciale, i et j étant ici exclusivement des UE :

$$e_{ka} = \frac{1}{rs} \sum_i \sum_j s_{k(ij)}$$

Cette modification a une conséquence également indésirable dans la mesure où elle ne tient pas compte de la proximité ou de l'éloignement des nœuds sur lesquels se branchent les UE par rapport à k. Si le nombre d'UE branchées sur x est très élevé, certaines très proches de x d'autres très éloignées, les premières auront le même poids que les secondes dans l'estimation des longueurs de branches. Ainsi, reprenons la figure VII.14 et supposons maintenant que i est une UEH regroupant 12 UE sous les mêmes états que i (111). Dans ce cas de figure, en raison de la présence des états (000) chez b, c et d, la distance e_{kx} entre k et x sera égale à (12*3+3*0)/15, soit 36/15, une estimation proche de 2, alors que la présence des UE b, c, et d conduirait plus logiquement à une distance proche de 0.

Pour répondre à cette objection, Faith propose une variante qui a l'avantage de tenir compte du nombre de nœuds entre k et les UE. Il définit la similitude spéciale de Farris en remplaçant les distances, comme d_{kx}, par leur expression en terme de similitude spéciale, comme $s_{k(x'x'')}$, x' et x" étant les deux UE ou UEH entre lesquelles se branche x (figure VII.14). Ce processus de remplacement récursif conduit à exprimer la similitude spéciale $s_{k(xy)}$ sous la forme :

$$s_{k(xy)} = \sum_i \sum_j \frac{1}{2^{N_i} + 2^{N_j}} s_{k(ij)}$$

où N_i et N_j représentent les nombres de nœuds entre a et i et entre a et j.

Cette nouvelle formulation montre que l'agglomération peut se faire sans avoir à calculer des longueurs de branches : dans la partie droite de l'équation n'interviennent en effet que des distances observées entre UE (i, j et k). Cela permet d'éviter la recherche de l'arbre le plus court par l'emploi d'une quelconque méthode qui nécessiterait le calcul de longueurs de branches. Cela est préférable dans la mesure où il n'est pas impossible que l'arbre le plus court ne soit pas celui que produirait le meilleur ajustement local ou global à une matrice de distances. En revanche, les longueurs des branches peuvent être calculées *in fine*, une fois l'arbre reconstruit, éventuellement par des méthodes d'ajustement.

5. Remarques et conclusions à propos des méthodes phénétiques

L'utilisation des distances pour inférer des phylogénies soulève un certain nombre de problèmes qui ont été abordés tout au long de ce chapitre. Certains de ces problèmes se rencontrent également à propos d'autres méthodes de reconstructions. Dans ce chapitre ils sont brièvement résumés en guise de conclusion.

5.1. Similitude globale et caractères

Le concept essentiel sur lequel se fondent les reconstructions phénétiques est celui de *similitude globale*. De ce fait la ressemblance entre deux UE peut être due aussi bien au partage de caractères hérités d'un ancêtre commun immédiat (apomorphies) qu'au partage de caractères hérités d'un ancêtre plus lointain

(plésiomorphies), ou qu'à l'identité de caractères due à d'autres causes (homoplasies : convergences, parallélismes, réversions, partage d'apomorphies dues au seul hasard). Les méthodes phénétiques ne permettent pas de discriminer entre ces explications de la ressemblance bien que les méthodes de parcimonie introduites par Farris (1972) soient une tentative de solution. En réalité les méthodes phénétiques évacuent totalement le concept de caractère et d'état de caractère pour ne plus s'intéresser qu'à celui d'UE. Ceci est le résultat de la transformation de la matrice de caractères en matrice de distances. Autrement dit, le phénogramme n'est pas un cladogramme puisque les nœuds n'y représentent pas les états ancestraux des caractères mais seulement les degrés de similitude entre les UE qui en dérivent.

Implicitement l'application des méthodes phénétiques revient donc à postuler un certain nombre de propriétés concernant les caractères :

— Les caractères doivent être choisis de manière non biaisée parmi l'ensemble des caractères soumis à évolution. En effet si le choix des caractères se portait, malencontreusement, sur ceux soumis à de fortes pressions sélectives, les groupements d'UE (« *clusters* ») obtenus à l'issue de l'analyse phénétique reflèteraient davantage des groupes partageant les mêmes réponses adaptatives plutôt que des groupes partageant les mêmes ancêtres. Il faut donc pouvoir considérer, comme dans d'autres méthodes d'ailleurs, que chaque caractère a une histoire évolutive qui est le reflet de l'histoire évolutive réelle de l'ensemble des UE et que les événements homoplasiques sont l'exception.

— Tous les caractères utilisés dans la constitution de la distance sont généralement considérés comme ayant le même *poids*, c'est-à-dire qu'ils participent de manière égale à la ressemblance globale. Dans certains cas on peut être amené à pondérer les caractères, par exemple par leur fréquence relative observée à l'intérieur d'un groupe d'UE (*cluster*) par rapport à ce qu'elle est entre *clusters* (paragraphe V.2.3.1). Cependant il reste parfois difficile de justifier une reconstruction phylogénétique fondée sur une pondération des caractères, quand cette pondération est elle-même établie à partir d' *a priori* sur l'organisation des UE que l'on cherche justement à préciser. Ce type de pondération est extrêmement dépendant de la façon dont les UE ont été échantillonnées.

— Les caractères sont supposés évoluer indépendamment les uns des autres de telle façon que la présence conjointe de deux caractéristiques particulières dans deux UE puisse être interprétée comme le résultat de leur héritage concomitant d'un même ancêtre et non comme le résultat d'une liaison entre elles, fonctionnelle par exemple, où la présence d'une des caractéristiques impliquerait nécessairement la présence de l'autre. Dans cette dernière hypothèse cela reviendrait à attribuer une pondération aux caractères liés. Cette objection n'est évidemment pas propre à l'approche phénétique.

— En travaillant non sur les caractères eux-mêmes mais sur un indice, il est clair que l'on perd une certaine information, en ce sens que l'on ne peut généralement pas restituer sans ambiguïté la matrice des caractères à partir de la seule matrice constituée des indices de distance (Penny, 1982 ; Steel *et al.*, 1988 ; Fitch, 1984). Un arbre construit à partir d'une matrice de distances ne donne donc pas d'information sur l'état des caractères aux nœuds ni sur le sens d'évolution des

caractères et ne permet généralement pas de préciser si un caractère est primitif ou dérivé. De ce fait la matrice de distances ne donne pas non plus d'indications ni sur la quantité d'homoplasie qui est jugée *a priori* négligeable par rapport à l'information phylogénétique ni sur sa localisation qui est supposée répartie de manière homogène sur l'ensemble de l'arbre.

5.2. Arbre : racine et branches

La plupart des méthodes phénétiques ne permettent pas de situer la position de l'ancêtre. Quand elles le font (par exemple UPGMA), cette position découle seulement des hypothèses évolutives qu'implique la méthode elle-même, celles que l'on résume sous l'expression d'« horloge moléculaire ». Il suffit que cette hypothèse ne soit pas vérifiée pour que la position de la racine soit contestable.

La procédure la plus utilisée pour situer l'origine consiste à rechercher l'arbre non enraciné en intégrant une UE dont on sait qu'elle représente un extra-groupe. Un seul extra-groupe est suffisant pour enraciner l'arbre. Cependant le choix de cet extra-groupe peut influencer la forme de l'arbre non enraciné dans la mesure où il est inclus parmi l'ensemble des UE pour établir l'arbre non enraciné. Aussi est-il recommandé d'estimer l'arbre non enraciné sans l'extra-groupe, puis de rechercher l'insertion la plus parcimonieuse, la plus ajustée ou la plus vraisemblable (chapitre VIII) de cet extra-groupe sur l'arbre constitué préalablement à son insertion.

L'interprétation des longueurs des branches en terme de « nombre d'événements évolutifs » n'est pas évidente. D'abord parce que la distance peut avoir des propriétés qui ne la permettent pas (non métricité, non additivité...) ensuite parce que les tentatives pour estimer les distances patristiques ou phylétiques à partir des distances observées ne valent qu'à la condition de disposer d'un modèle testé et vérifié qui rende compte correctement de l'homoplasie, ou qu'à la condition d'avoir des raisons objectives pour penser que les homoplasies sont négligeables.

CHAPITRE **VIII**

LES MÉTHODES PROBABILISTES

L'inférence de l'histoire évolutive des espèces ou des populations présentée dans ce chapitre repose sur une méthodologie différente des précédentes en ce sens qu'elle repose sur un raisonnement probabiliste. Cette méthode suppose en effet que les événements évolutifs, essentiellement les transformations de caractères, obéissent à certaines lois de probabilité définies *a priori.* C'est une particularité de cette méthode que de nécessiter la définition préalable d'un modèle *explicite* d'évolution des caractères, qu'il s'agisse de caractères quantitatifs, comme des fréquences géniques, ou de caractères qualitatifs, comme les acides nucléiques de séquences d'ADN. Une fois cette démarche accomplie, il devient possible d'exprimer la probabilité pour qu'un arbre évolutif particulier aboutisse aux observations que l'on peut effectuer sur un ensemble de caractères et de taxons. De la même façon que l'on choisit l'arbre le plus parcimonieux dans les méthodes cladistiques, de même on optera, compte tenu des observations et du modèle, pour l'arbre et pour les longueurs de branches les plus probables.

Pour que cette méthode soit bien comprise, il nous a paru nécessaire de donner en introduction quelques indications sur le cadre conceptuel dans lequel elle se situe, cadre qui a été très largement développé dans le domaine de la statistique par Fisher dès les années 1920 et qui conduit aux méthodes d'estimation dites du maximum de vraisemblance (Edwards, 1972). Historiquement, une des premières tentatives d'application de cette méthode aux problèmes de phylogénie est due à Edwards et Cavalli-Sforza (1964).

A la suite de ces généralités introductives, on développera un exemple simple qui permettra de souligner les particularités de la méthode. Deux parties suivront, décrivant les modèles d'évolution les plus couramment proposés, l'un pour des données quantitatives, l'autre pour des données qualitatives. A chaque occasion, il sera montré comment s'intègre le facteur temps dans ces modèles de reconstruction.

Dans une dernière partie, la question de la différence entre reconstructions phylogénétiques par parcimonie et par vraisemblance sera évoquée. En particulier on montrera quelles sont les hypothèses que sous-entendent les méthodes de parcimonie quand elles sont considérées comme une application particulière des méthodes probabilistes.

1. Introduction

1.1. Généralités

Un *modèle* est constitué d'un ensemble de *paramètres* θ sur lesquels on peut formuler différentes *hypothèses* et qui constituent les hypothèses du modèle. Ainsi un modèle d'évolution comporte-t-il plusieurs hypothèses concernant des paramètres comme les probabilités d'événements évolutifs (spéciations, changements d'états des caractères, taux de mutation...), ou comme la structure hiérarchique des différentes UE, c'est-à-dire l'arbre, lui aussi considéré comme un paramètre.

Soit $P(E_i)$ la probabilité d'un ensemble E_i d'hypothèses sur les paramètres θ constitutifs d'un modèle explicite d'évolution, *M*.

Soit *D* l'ensemble des données observées sur lequel s'appuie l'inférence. Il s'agit par exemple de séquences alignées d'ADN ou d'une série d'observations codées présence/absence sur plusieurs UE.

Le problème consiste à évaluer, dans le contexte *exclusif* du modèle *M*, la probabilité conditionnelle, $P(E_i|D)$, de l'ensemble d'hypothèses E_i, sachant que l'on a observé les données *D*. Le théorème de Bayes permet d'écrire :

$$P(E_i|D) = \frac{P(E_i).P(D|E_i)}{\sum_r P(E_r).P(D|E_r)}$$

La sommation des termes du dénominateur s'effectue sur toutes les *r* hypothèses évolutives alternatives formant un ensemble complet d'hypothèses, c'est-à-dire un ensemble tel que la somme des probabilités de chacune d'elles soit égale à 1. Il peut s'agir, par exemple, de l'ensemble des arbres possibles. Il est clair que les probabilités *a priori* $P(E_r)$ sont, en règle générale, inconnues. On peut par exemple supposer qu'elles sont toutes égales entre elles et donc égales à $P(E_i)$. Cependant il n'est pas toujours nécessaire de faire des hypothèses sur les probabilités *a priori*. En effet, si l'on se borne à rechercher l'ensemble E_i d'hypothèses évolutives rendant le mieux compte des données *D,* on peut tout aussi bien calculer $P(E_i|D)$ que $P(D|E_i)$ qui lui est proportionnelle. On dit que la vraisemblance (*Likelihood*) *L* de E_i sachant les données *D* est proportionnelle à $P(D|E_i)$. On écrira :

$$L(E_i|D) = P(D|E_i)$$

La démarche inférentielle consiste donc à rechercher la vraisemblance des données *D* sous différentes hypothèses évolutives E_i d'un modèle *M* et à retenir les hypothèses qui rendent cette vraisemblance maximum. Cela revient à rechercher les valeurs des paramètres $\theta_1, \theta_2, \ldots, \theta_j, \ldots$, pour lesquelles les dérivées partielles de la vraisemblance s'annulent :

$$\frac{\delta L(E_i|D)}{\delta \theta_j} = 0$$

et pour lesquelles la dérivée seconde est positive.

Dans cette démarche, il n'est pas question de tester le modèle *M* lui-même. La vraisemblance obtenue est en effet conditionnée au modèle sans l'existence duquel aucune inférence n'est possible. Par exemple le modèle peut stipuler que l'évolution se fait par dichotomies successives. Les hypothèses du modèle, celles qui constituent l'ensemble *E,* seront alors la structure dichotomique de l'arbre, les modalités d'évolution des caractères le long des branches et les longueurs des branches elles-mêmes. Si le modèle invoque des réticulations, la vraisemblance qu'il donnera des données ne sera pas comparable à la vraisemblance obtenue sous un modèle strictement dichotomique (sauf à paramétrer les réticulations elles-mêmes). De même ne pourra-t-on pas comparer deux modèles intégrant un nombre différent d'UE.

Il est important pour la suite d'insister sur le fait que des modèles différents impliquent nécessairement des paramètres différents, et pas seulement des valeurs différentes des mêmes paramètres. On ne peut donc juger des qualités respectives de divers modèles mais seulement rechercher les meilleures hypothèses concernant le même ensemble de paramètres, pour un modèle donné *M*.

On peut distinguer plusieurs classes de paramètres : les paramètres de structure, les paramètres d'incidence et les paramètres de nuisance (Goldman, 1990).

Considérons d'abord l'ensemble {X} des variables aléatoires X_1, X_2, X_3,...X_i, définissant la réalisation du modèle aboutissant aux données observées.

— L'ensemble { θ } est constitué des paramètres de **structure** θ_1, θ_2,…, θ_j,…, ceux qui apparaissent dans la loi de probabilité de la totalité des éléments de l'ensemble{X}. La structure de l'arbre peut être considérée comme un paramètre commun à toutes les variables observées qui varient le long des branches.

— L'ensemble { ζ } est constitué des paramètres d'**incidence** ζ_1, ζ_2,…, ζ_k,…, qui n'apparaissent que dans la loi de probabilité d'un sous-ensemble d'éléments de {X}. Les états des caractères en un nœud particulier sont parfois considérés comme de tels paramètres, puisque les états aux nœuds, en tant que paramètres, ne sont pas communs à toutes les variables X.

— L'ensemble { η } des paramètres de **nuisance** η_1, η_2, η_3,… η_l,..., est constitué des différents paramètres de structure ou d'incidence dont on ne juge pas l'estimation intéressante. Ainsi, les états intermédiaires des caractères dans l'arbre peuvent-ils être considérés comme des paramètres de nuisance (Felsenstein, 1973a,b).

La vraisemblance des données est calculée et sa maximisation recherchée après avoir pris en compte les paramètres de nuisance. Pour ce faire, deux techniques sont possibles :

— attribuer à ces paramètres de nuisance ou bien des probabilités *a priori* ou bien des probabilités conditionnées par les paramètres de structure ;

— les maximiser en même temps que tous les autres paramètres de structure.

Une propriété de l'estimation par maximum de vraisemblance est la *consistance* :: quand le nombre des données augmente, les estimations des paramètres du modèle convergent vers leurs vraies valeurs, sans que l'on puisse pour autant tirer des conclusions sur la validité du modèle lui-même. Cette propriété de consistance n'est cependant vérifiée que pour les paramètres de structure, mais non pour les paramètres d'incidence sauf lorsque ceux-ci suivent tous une loi de distribution identique et qu'ils sont indépendants. C'est pourquoi il est préférable de traiter les paramètres d'incidence comme des paramètres de nuisance et d'essayer de les supprimer de l'inférence, c'est-à-dire ne pas chercher à les estimer, selon des méthodes identiques à celles qui permettent de traiter les paramètres de nuisance.

1.2. Exemple

Imaginons l'histoire évolutive de 3 UE (i, j et k). Parmi les 3 arbres possibles, admettons que seuls les arbres de la figure VIII.1 doivent être envisagés : T_1 et T_2. La question que l'on se pose est donc celle du choix entre ces deux « histoires », sur la base de données (notées D) qui ne sont constituées ici que d'un seul caractère codé 0 et 1.

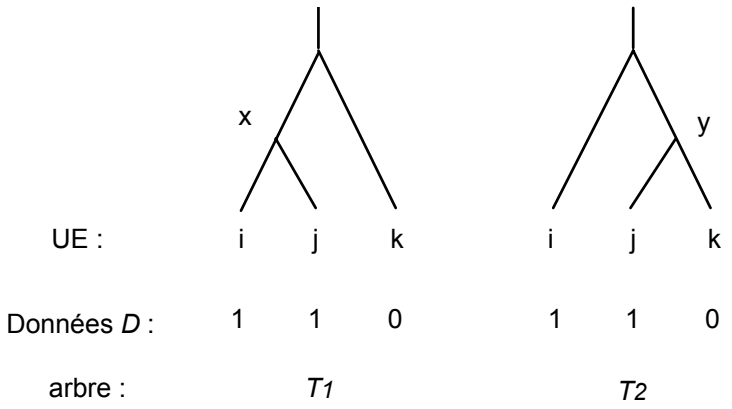


FIGURE VIII.1. *Exemple de la distribution d'un caractère dans 3 UE dont les relations phylétiques peuvent être de deux formes différentes : T_1 et T_2.*

La démarche peut se comparer à celle que suivrait un parieur : tirer le meilleur parti des observations disponibles (D) pour retenir l'arbre le plus probable, celui qui a le plus de chance d'être le bon. La question peut donc se formuler ainsi : quelle est la probabilité de l'arbre T_1 compte tenu des observations D effectuées sur les 3 UE ? Formellement cela s'écrit simplement : $p(T_1|D)$.

Puisque seulement deux alternatives sont possibles, T_1 et T_2, on peut écrire (théorème de Bayes) :

$$p(T_1|D) = \frac{p(T_1)p(D|T_1)}{p(T_1)p(D|T_1) + p(T_2)p(D|T_2)}$$

où $p(T_1)$ et $p(T_2)$ sont les probabilités *a priori* des arbres T_1 et T_2, celles que l'on peut posséder antérieurement à l'analyse des données D. Les probabilités d'observer les données D quand l'arbre est T_1 et T_2 sont, respectivement, $p(D|T_1)$ et $p(D|T_2)$.

Une difficulté inhérente à cette méthode réside dans la quantification objective de telles probabilités *a priori* de l'arbre T_1 et T_2. Pour contourner ce problème, on remarquera que l'on peut aussi bien effectuer notre choix entre T_1 et T_2, non pas en comparant $p(T_1|D)$ et $p(T_2|D)$, mais en comparant $p(D|T_1)$ et $p(D|T_2)$ qui leur sont proportionnelles et que l'on appelle les vraisemblances des données sachant que l'arbre est T_1 et T_2, respectivement $L(T_1|D)$ et $L(T_2|D)$. On va donc s'intéresser au rapport de vraisemblance :

$$F = \frac{L(T_1|D)}{L(T_2|D)}$$

Lorsque ce rapport F est supérieur au rapport des probabilités *a priori* :

$$F_O = \frac{p(T_2)}{p(T_1)}$$

on choisira de préférence l'arbre T_1 ; lorsqu'il est inférieur, on choisira plutôt l'arbre T_2. Si les probabilités *a priori* sont égales, ce rapport F_0 est naturellement égal à 1.

Pour calculer ce rapport de vraisemblance F, il est indispensable de définir au préalable un modèle d'évolution et de préciser les différents paramètres qui le composent.

1.2.1. Le modèle d'évolution et les paramètres

— Le premier paramètre est constitué par la structure S de l'arbre. Deux hypothèses sont possibles pour ce paramètre : $S = T_1$ ou $S = T_2$.

— Les probabilités des événements, c'est-à-dire les transformations des caractères de 0 vers 1 et de 1 vers 0 constituent d'autres paramètres du modèle. Pour simplifier, considérons ici que la probabilité est la même pour passer de 0 à 1 et de 1 à 0. Désignons cette probabilité par ρ et par $\pi = 1 - \rho$ la probabilité qu'il n'y ait pas de changement le long d'une branche : ρ (ou π) constitue donc un paramètre dont les valeurs peuvent aller de 0 à 1. Supposons, de plus, que, pour un caractère donné, *un seul* changement par branche soit possible. Il s'agit là d'une contrainte supplémentaire imposée au modèle qui sera discutée plus loin (paragraphe VIII.4 et 5).

— Un autre paramètre est constitué par la probabilité attribuée à l'état du caractère chez l'ancêtre de i, j et k. Cet état est 0 avec une probabilité égale à f ou bien il est 1 avec une probabilité égale à $(1 - f)$.

Reste enfin à exprimer les probabilités $p(D|T_1)$ et $p(D|T_2)$, sous ce modèle, en fonction des différents paramètres. Pour cela il faut examiner *toutes* les situations possibles pour les états du caractère aux nœuds de l'arbre T_1 , de l'arbre T_2 et de l'ancêtre.

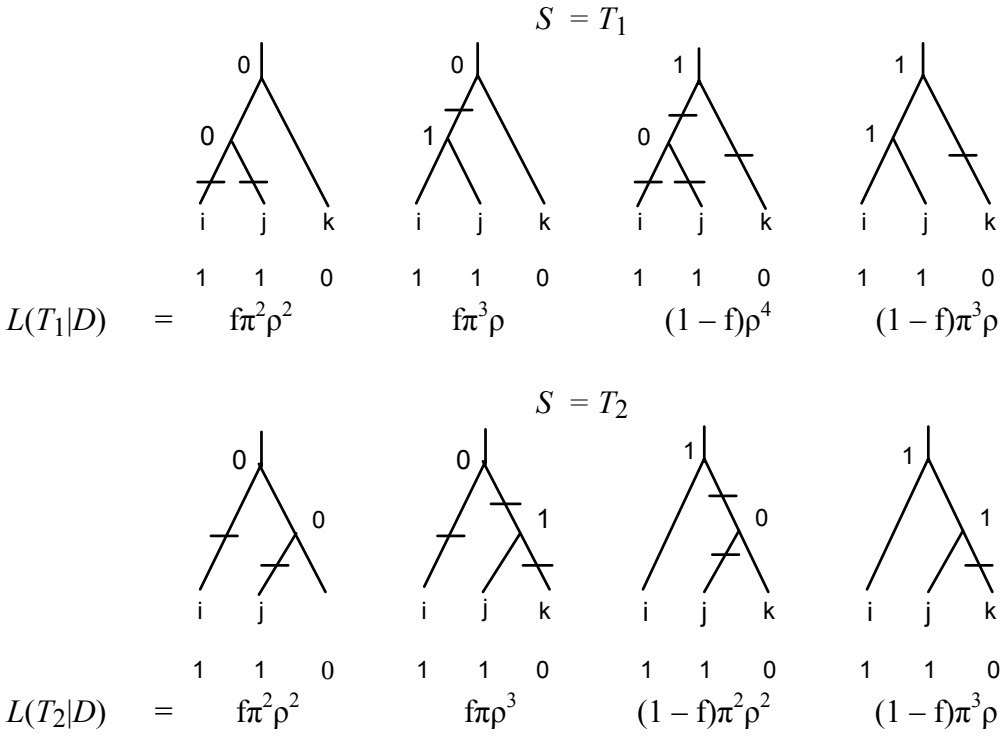


FIGURE VIII.2. *Probabilités des différents événements qui ont pu conduire aux données de la figure VIII.1, pour les arbres de structure S = T_1 et S = T_2, l'état de l'ancêtre étant O ou 1. Un seul événement par branche, de probabilité ρ, est ici pris en considération.*

La figure VIII.2 montre toutes les combinaisons possibles et leurs probabilités respectives. Puisqu'il y a 4 branches dans chaque arbre, chacune d'elles est affectée d'une probabilité ρ ou π selon qu'il y survient un changement d'état ou non. La probabilité d'une configuration est donc le produit des probabilités attachées à chacune des 4 branches, multipliée par la probabilité *f* ou (1 – *f*) de l'état de l'ancêtre.

1.2.2. Le calcul des vraisemblances, de leurs variations et de leur rapport

A l'aide de la figure VIII.2, on peut calculer les vraisemblances des arbres T_1 et T_2 en fonction des deux paramètres *f* et π (ou ρ = 1 – π). On a en effet :

$$L(T_1|D) = f\pi^2\rho + (1-f)\rho(1-3\pi\rho) \text{ et } L(T_2|D) = f\pi\rho^2 + (1-f)\pi^2\rho$$

Les variations du rapport $F=L(T_1|D)/L(T_2|D)$, ou de son logarithme, en fonction des deux paramètres ρ (ou π) et *f* sont représentées dans la figure VIII.3. Elles permettent de formuler un choix entre T_1 et T_2 en fonction des hypothèses retenues pour les paramètres du modèle, compte tenu des données disponibles, et en fonction du rapport des probabilités *a priori* des arbres T_1 et T_2.

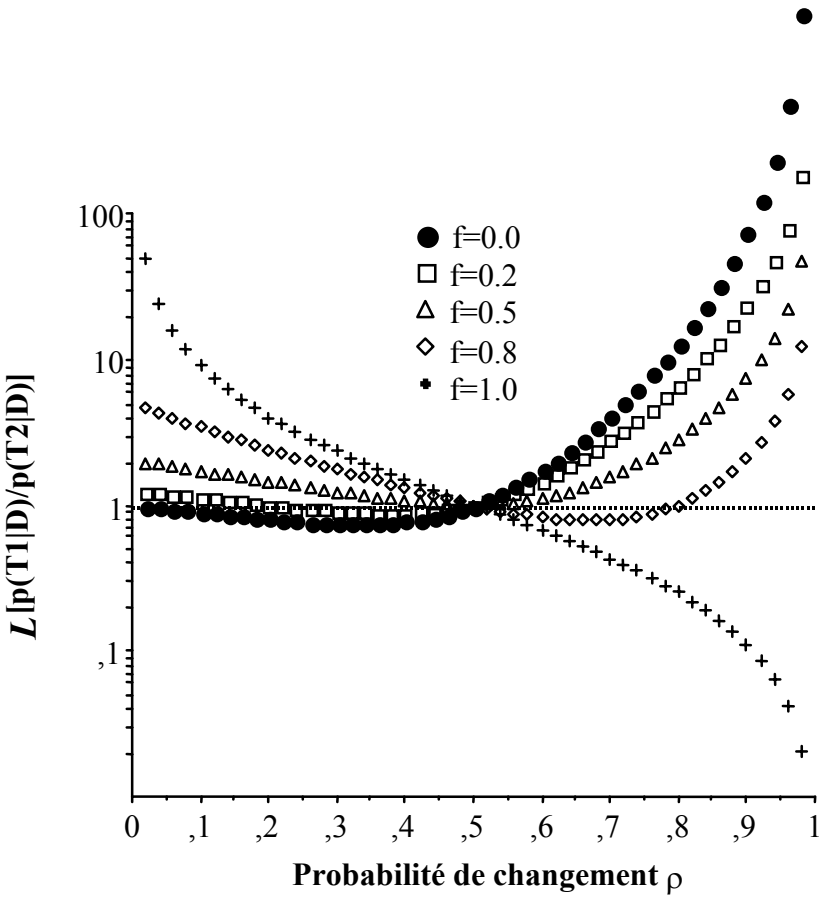


FIGURE VIII.3. *Variation du rapport entre la vraisemblance de l'arbre T_1 et celle de l'arbre T_2 en fonction des variations des paramètres* ρ *(probabilité de changement) et f (probabilité que l'état ancestral du caractère soit 0).*

Supposons que les probabilités *a priori* des arbres T_1 et T_2 soient égales, signifiant par là que ces deux arbres ont autant de possibilité, avant l'analyse, d'être le bon. Dans cette situation, on choisira l'arbre T_1 si le rapport des vraisemblances *F* est supérieur à 1. On choisira l'arbre T_2 dans la situation inverse (Figure VIII.3). Lorsque l'ancêtre a plus de chance d'être dans l'état 0 que dans l'état 1 ($f > 0.5$), l'arbre retenu est toujours T_1 aussi longtemps que la probabilité de changement est plus faible que la probabilité de non changement ($\rho < 0.5$). Il peut être T_2 ou T_1 lorsque $\rho > 0.5$, selon la valeur de f.

La situation est plus complexe lorsque l'ancêtre a plus de chance d'être dans l'état 1 ($f < 0.5$). L'arbre T_2 peut alors donner une meilleure vraisemblance dans certaines conditions : par exemple lorsqu'on a simultanément $f = 0.2$ et $0.2 < \rho < 0.5$. Lorsque la probabilité de changement se situe au dessus de 0.5 et que l'ancêtre est supposé être dans l'état 0 ($f = 1$), l'arbre le plus vraisemblable est encore T_2 , de même lorsque $f = 0.8$ et que $0.5 < \rho < 0.8$.

Dans cet exemple, la méthode du maximum de vraisemblance nous a servi à n'estimer que le paramètre *S* correspondant à la structure de l'arbre, en fonction

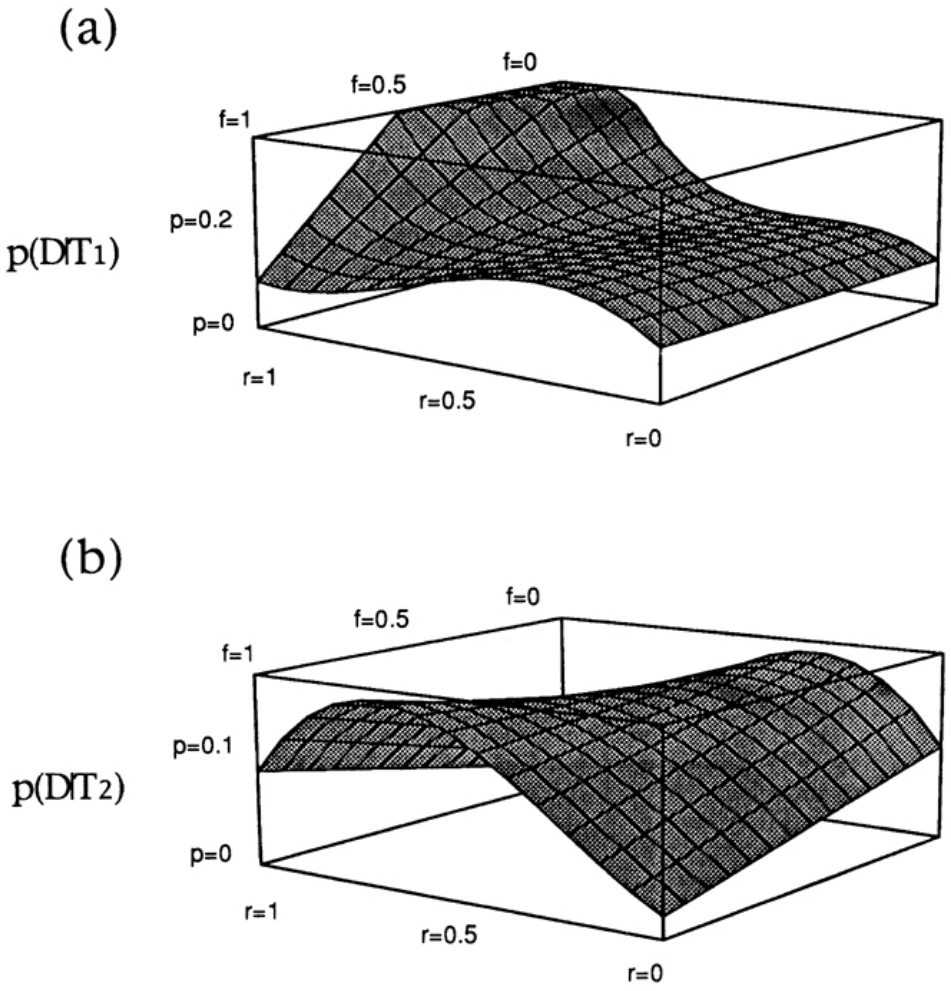


FIGURE VIII.4. *Surfaces de vraisemblance de l'arbre T_1 (a) et de l'arbre T_2 (b) en fonction des valeurs des paramètres ρ (probabilité de changement sur une branche) et f (probabilité que l'ancêtre soit sous l'état 0).*

des autres paramètres (la probabilité de changement d'état ou la probabilité de l'état ancestral).

Supposons maintenant que l'arbre pris en considération soit l'arbre T_1. Les estimations des paramètres f et ρ s'obtiennent en cherchant les valeurs de ces paramètres qui maximisent la vraisemblance $L(T_1|D)$. La figure VIII.4a montre que la surface de vraisemblance est telle que le maximum s'obtient dans une situation très particulière, quand le paramètre f tend vers 0 et que ρ tend vers 1, c'est-à-dire que la vraisemblance est maximale quand l'ancêtre est dans l'état 1 et qu'il y a un changement sur chaque longueur de branche. En revanche lorsque l'ancêtre est à l'état 0 ($f = 1$), la vraisemblance maximale s'obtient quand la

probabilité de changement sur chacune des branches est de une chance sur 3 ($\rho = 1/3$).

Si l'on admet maintenant que l'arbre est T_2, les variations de la vraisemblance exprimées en fonction de ρ et de f peuvent également être calculées. Ces variations permettent de situer les valeurs maximales de la vraisemblance (figure VIII.4b). L'ancêtre étant dans l'état 0 ($f = 1$), la probabilité de changement peut être estimée à $\rho = 2/3$, car c'est la valeur de ρ conduisant à la vraisemblance maximale. En revanche quand l'ancêtre est 1, la valeur estimée de ρ est 1/3.

1.3. Conclusions

Au travers de ces exemples simples, on a pu voir que la méthode de vraisemblance consiste d'abord à définir un *modèle* à l'aide d'un certain nombre de *paramètres*. Plusieurs *hypothèses* sont ensuite formulées à propos de ces paramètres, hypothèses qui reviennent à leur attribuer des valeurs particulières. On retient comme valeurs estimées des paramètres celles qui rendent la vraisemblance maximale.

Remarquons que l'on ne fait aucune inférence sur les états des caractères aux nœuds sauf en terme de probabilité. Si l'arbre choisi est T_1 par exemple, alors la probabilité pour que l'état de x (figure VIII.1) soit 0 dépend de f et de ρ. Elle est donnée par (figure VIII.2) :

$$p(x = 0 | T_1, D) = \frac{f\pi^2\rho^2 + (1-f)\rho^4}{f\pi^2\rho + (1-f)\rho(1-3\pi\rho)}$$

Une analyse de parcimonie de la distribution des caractères de la figure VIII.1 conduirait, plus directement en apparence, aux conclusions suivantes : si l'ancêtre est dans l'état 0, alors l'arbre le plus parcimonieux est l'arbre T_1 puisqu'une seule transformation est nécessaire, entre l'ancêtre et x. En revanche quand l'ancêtre est dans l'état 1, le choix, sur la base des données disponibles, entre l'arbre T_1 et l'arbre T_2 est impossible en ce sens que l'unique transformation peut aussi bien survenir sur la branche entre l'ancêtre et k dans l'arbre T_1 qu'entre le nœud y et k dans l'arbre T_2.

Cependant, si l'arbre T_1 est celui qui est retenu, alors il n'y a plus d'ambiguïté dans l'attribution des caractères aux nœuds dans l'approche par parcimonie : x est nécessairement dans l'état 1. Tel n'est pas le cas dans une approche par maximum de vraisemblance : l'état du caractère au nœud x n'est connu qu'en probabilité.

Cet exemple sera repris plus loin de manière plus complète lors de la comparaison entre méthodes probabilistes et méthodes de parcimonie. Il souligne bien l'une des différences fondamentales entre vraisemblance et parcimonie.

2. Modèle d'évolution de caractères quantitatifs

Soit X_i, X_j, X_k... X_o les valeurs du caractère X dans les taxons, UE ou UEH i, j, k, o, aux temps t_i, t_j, t_k, t_o (figure VIII.5). L'origine de l'arbre T est constituée par la population ancestrale o au temps t.

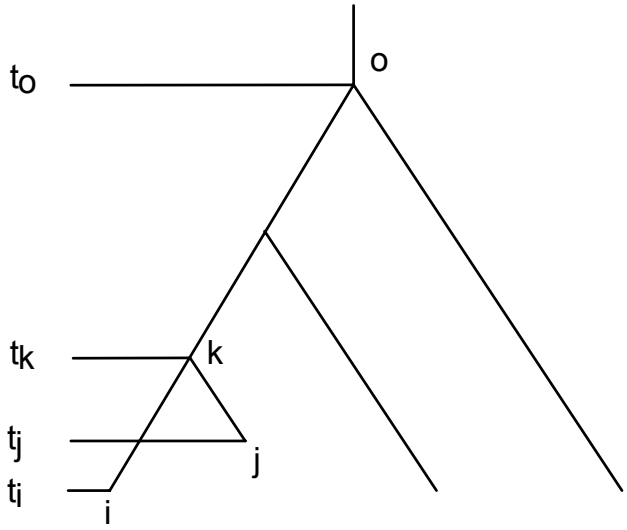


FIGURE VIII.5. *Temps de division dans un arbre phylogénétique.*

On suppose que X se distribue normalement et évolue avec le temps selon un processus aléatoire, analogue à un mouvement brownien, de telle façon que l'espérance de X est constante au cours du temps :

$$E(X_i) = E(X_0) \quad \forall \text{i},$$

et que la variance du caractère chez le taxon i est égale à sa variance chez le taxon k qui lui est ancestral, augmentée d'une quantité proportionnelle au temps écoulé dt entre t_i et t_k.

$$V(X_i) = V(X_k) + \alpha_{ik} dt,$$

Le facteur de proportionnalité α_{ik} mesure la vélocité du mouvement brownien pour le caractère X durant le temps dt. On suppose de plus que les caractères évoluent indépendamment les uns des autres et que l'évolution le long d'une branche se fait indépendamment de l'évolution sur les autres branches. On peut supposer également que le coefficient α_{ik} varie d'une branche à l'autre de l'arbre (c'est pourquoi il est indicé ik). On parle alors de taux d'évolution variable (TEV). En revanche si α est identique pour toutes les branches de l'arbre, le taux d'évolution est alors constant (TEC). Si l'on se place dans un contexte intra-

spécifique, ce coefficient α_{ik} est lui-même proportionnel à l'effectif efficace N_e de la population entre le temps t_i et t_k (Thompson, 1975).

Il n'est généralement pas possible de distinguer les deux termes du produit $\alpha_{ik}dt$. Pour cela, il faudrait en effet ou bien connaître correctement les temps auxquels les taxons se sont différenciés, ou bien avoir des arguments pour pondérer, selon les branches, les vitesses d'évolution. En l'absence de telles informations, on parle généralement de quantité d'évolution e_{ik} pour représenter le produit $\alpha_{ik}dt$.

La covariance entre X_i et X_j est simplement égale à la variance de leur ancêtre commun k :

$$Cov(X_i,X_j) = V(X_k).$$

Dans les conditions définies précédemment, la différence entre la réalisation de la variable X, pour le caractère s, dans l'UE i et k, respectivement $x_{i,s}$ et $x_{k,s}$ est égale à :

$$d_{ik,s} = (x_{i,s} - x_{k,s})$$

et se distribue normalement avec une espérance nulle et une variance égale à la quantité d'évolution e_{ik} entre i et son ancêtre k. La vraisemblance des observations faites sur la branche entre i et k s'écrit comme le produit de c lois normales, c étant le nombre total de caractères évoluant indépendamment les uns des autres :

$$L_{ik} = \prod_{s=1}^{c} \frac{1}{(2\pi e_{ik})^{1/2}} \exp\left[\frac{(d_{ik,s})^2}{2e_{ik}}\right]$$

La vraisemblance globale d'un arbre T peut ensuite être calculée comme le produit des vraisemblances attachées à chacune des branches de T :

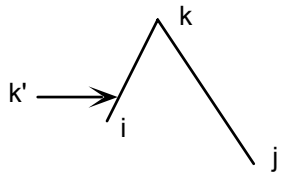
$$L = \prod_{i,k} L_{ik}$$

Il est ensuite possible d'obtenir par dérivation les estimations des quantités d'évolution, en résolvant les équations différentielles du type suivant :

$$\frac{\delta L}{\delta e_{ik}} = 0$$

En revanche, il n'est pas possible d'estimer par dérivation les valeurs des variables $x_{k,s}$ à chacun des nœuds, à moins de les considérer comme des paramètres et non comme des réalisations d'une variable aléatoire sous un certain modèle. Dans ce cas il s'agirait cependant de paramètres d'incidence puisque propres à chacun des nœuds, et leur estimation n'aurait pas nécessairement, de ce fait, la propriété de consistance.

Par ailleurs un problème fondamental se pose à la racine de l'arbre dans la mesure où la surface de vraisemblance présente une singularité. En effet, supposons que l'ancêtre soit k et que i et j soient les deux UE filles :



Il ressort clairement de la formule précédente de la vraisemblance que si l'on choisit la position de k telle que e_{ik} soit aussi petit que l'on souhaite (cela revient par exemple à prendre k' comme racine, proche de i), la vraisemblance égale à $L_{ik}L_{jk}$ (et donc la vraisemblance totale L) tend vers une valeur infiniment grande (Cavalli-Sforza et Edwards, 1966). La position de la racine n'est donc pas localisable en même temps que sont estimées les quantités d'évolution le long des branches. Plusieurs solutions peuvent être envisagées pour contourner ces difficultés.

2.1. La solution de Felsenstein (1973b)

Felsenstein (1973b) fait remarquer que la distance d_{ij} entre deux UE se distribue normalement avec une espérance nulle et une variance e_{ij} égale à la somme des quantités d'évolution qui les séparent de k :

$$e_{ij} = e_{ik} + e_{jk}$$

Il est possible de remplacer les deux UE i et j par une nouvelle UE hypothétique (ij) dont la valeur des caractères serait une combinaison linéaire des valeurs observées chez i et j, les pondérations des caractères se faisant en fonction des quantités e_{ik} et e_{jk}. Ces valeurs sont celles données par la dérivation suivante de la fonction de vraisemblance :

$$\frac{\delta L}{\delta x_k} = 0.$$

Ainsi la valeur x du caractère *s* chez la nouvelle UEH (ij) est-elle égale à :

$$x_{(ij),s} = \left(\frac{e_{jk}}{e_{ij}}\right)x_{i,s} + \left(\frac{e_{ik}}{e_{ij}}\right)x_{j,s}$$

La vraisemblance d'une branche reliant cette nouvelle UEH (ij) à une autre UE k, $L_{(ij)k}$, se calcule facilement en utilisant les formules du paragraphe précédent. La vraisemblance totale de l'arbre non enraciné incluant les UE i, j et k est alors égale au produit $L_{ij}L_{(ij)k}$. Cette méthode revient à considérer les valeurs des caractères pour les UE hypothétiques (ou ancestraux) comme conditionnées par les UE qui en découlent : dans ce cas les paramètres d'incidence sont traités comme des paramètres de nuisance et pris en compte en leur attribuant des probabilités conditionnelles.

Par cette procédure, il est possible de calculer, de proche en proche, la vraisemblance totale de l'arbre non enraciné *T*. On recherche ensuite l'arbre *T* et les valeurs des autres paramètres, ici l'ensemble E constitué des longueurs de branches exprimées en terme de quantité d'évolution *e,* qui maximise la

vraisemblance. On peut également obtenir une variance de ces estimations en calculant la courbure de la surface de vraisemblance, donnée par les dérivées secondes, et effectuer certains tests d'hypothèses.

2.1.1. Tests d'hypothèses

Sur un arbre *T* donné, il est possible d'effectuer des tests sur les longueurs de branches (Felsenstein, 1981b). Par exemple on peut tester que l'une des branches, e_{ij}, est de longueur nulle, simplement en comparant la vraisemblance $L(E|D,T)$, obtenue lorsque cette longueur e_{ij} est estimée, à la vraisemblance $L(E|D,T,e_{ij} = 0)$ obtenue quand on maintient cette longueur égale à 0.

Le double du logarithme du rapport de ces deux vraisemblances se distribue asymptotiquement, c'est-à-dire quand les données sont très abondantes, comme un χ^2 à 1 degré de liberté :

$$-2\ln \frac{L(E|D,T,e_{ij} = 0)}{L(E|D,T)} = \chi^2_{ddl\ =1}$$

On peut également imposer davantage de contraintes sur les longueurs de branches. Dans l'hypothèse où l'ancêtre est connu, il est possible de tester si les quantités d'évolution obéissent aux propriétés d'ultra-métricité (chapitre VII), et donc de tester l'hypothèse d'horloge moléculaire (Felsenstein, 1983a; 1985a ; 1986). Il suffit pour cela de calculer la vraisemblance de l'arbre en contraignant les quantités d'évolution entre deux UE filles et leur ancêtre commun à être égales puis à comparer cette vraisemblance à la vraisemblance obtenue sans de telles contraintes. Le double de la différence entre ces deux vraisemblances se distribue en effet asymptotiquement (quand la quantité de données augmente) comme un χ^2. Son degré de liberté est la différence entre le nombre de branches estimées sans l'hypothèse d'horloge ($2n - 3$) et le nombre de branches estimées avec l'hypothèse d'horloge ($n - 1$), soit ($n - 2$) degrés de liberté. Ce test permet donc de rejeter ou non l'hypothèse d'horloge moléculaire.

2.1.2. Exemples

Deux exemples seront donnés (figures VIII.6 et VIII.7). Dans le premier, les relations entre UE (ici des populations humaines) sont construites à partir d'une matrice de distances calculée à partir de 35 différentes fréquences alléliques. Le modèle suppose des fluctuations aléatoires de ces fréquences (dérive génique), l'absence de sélection et de mélange. Certaines longueurs ne sont pas significativement différentes de 0, sur la base d'une estimation des variances des longueurs de branches par les dérivées secondes de la surface de vraisemblance. Cet arbre montre une évidente absence de structure significative due probablement à ce que les mélanges inter-populations ne peuvent être négligés.

Le deuxième exemple montre l'application de la même méthode sur des données constituées des degrés d'hybridation d'ADN entre espèces prises deux à deux. Les méthodes phénétiques s'imposent en l'occurrence. Dans ce cas, l'hypothèse d'horloge moléculaire peut être testée. Elle n'est pas ici rejetée.

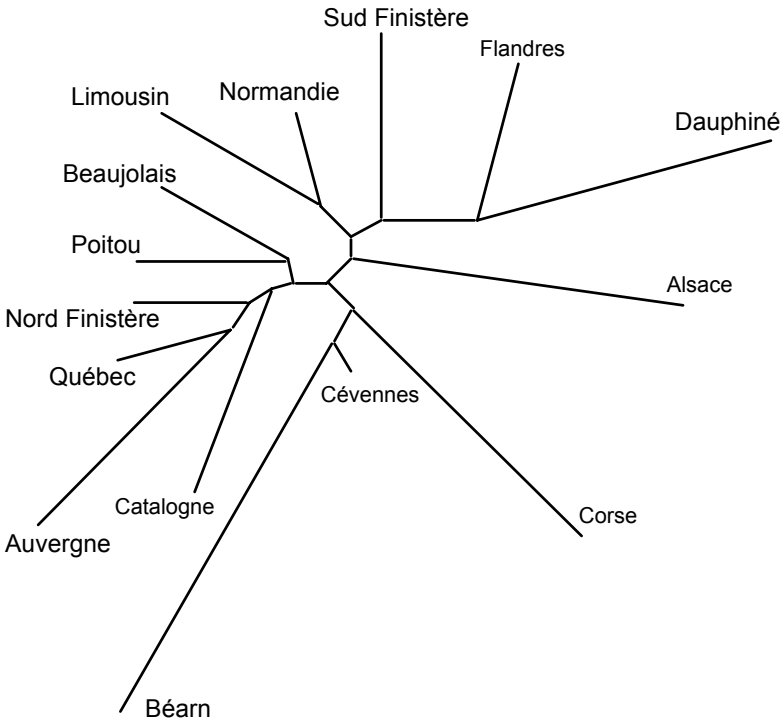


FIGURE VIII.6. *Représentation des inter-relations génétiques entre provinces françaises (Ohayon et Cambon-Thomsen, 1986) en partant d'un modèle où l'on suppose une évolution aléatoire des fréquences alléliques (dérive génique) et l'absence de sélection et de mélange. Les segments fins ne sont pas significativement différents de zéro.*

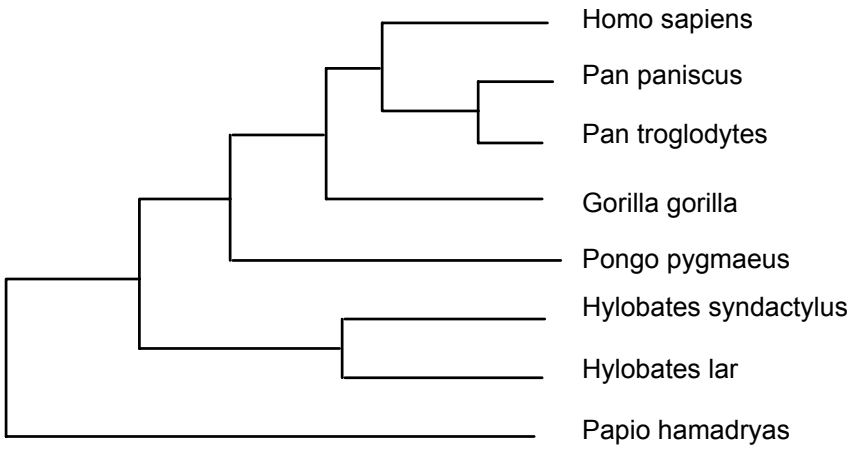


FIGURE VIII.7. *Structure de l'arbre non enraciné et longueurs de branches obtenues par maximum de vraisemblance. La matrice de distance est constituée des taux moyens d'hybridation d'ADN entre espèces prises deux à deux (d'après Felsenstein, 1987 et les données de Sibley et Ahlquist, 1987). Les logarithmes de la vraisemblance sont de 359.5 et de 357.7, respectivement sans et avec l'hypothèse d'horloge moléculaire. La différence n'est pas significative ($\chi^2 = 3.56$, d.d.l. = 7) .*

2. 2. La méthode du Treeness (Cavalli-Sforza et Piazza, 1975)

Cette méthode consiste à retrouver les relations phylogénétiques entre *n* UE à partir d'une matrice de covariance Σ ayant une dimension égale au nombre *n* d'UE. L'élément diagonal, σ_{ii}, représente la variance de l'UE i et est égal à la somme des quantités d'évolution menant de i à la racine de l'arbre (paragraphe VIII.2). L'élément σ_{ij} représente la covariance entre les taxons i et j et est égal à la somme des quantités d'évolution depuis l'ancêtre commun de i et j jusqu'à la racine de l'arbre. La matrice $\Sigma = \{\sigma_{ij}\} = f(\{e_i\})$ est fonction des paramètres *e* représentant les quantités d'évolution le long des branches. Si les *c* variables *X* qui définissent cette matrice sont multinormalement distribuées, la vraisemblance de cette distribution s'écrit :

$$L = \frac{1}{2\pi^{\frac{nc}{2}}} \left|\Sigma\right|^{\frac{-c}{2}} \exp\{-\frac{1}{2}\sum_{c} [(X - X_O)\Sigma^{-1}(X - X_O)']\}$$

Le logarithme de cette vraisemblance est donc, remplaçant les valeurs de *X* par leur réalisation x, c'est-à-dire les valeurs observées dans les différentes UE, *S* étant alors la matrice de covariance observée :

$$\ln L = -\frac{1}{2} C \left[\ln \left| \Sigma \right| + tr(\Sigma^{-1} S)\right] + Constante$$

La qualité de l'ajustement de la matrice de covariance observée *S* à la matrice théorique Σ qui correspond à l'arbre *T* peut être évaluée par le coefficient de « *treeness* » T_r :

$$T_r = \frac{|S|}{|\Sigma|}$$

S et Σ étant positives définies. T_r se distribue comme un $\chi^2 = -2c\mathrm{Ln}(T_r)$ où *c* est le nombre de caractères. Le degré de liberté est égal à $n(n+1)/2 - (2n - 1)$ (Cavalli-Sforza et Piazza, 1975 ; Astolfi *et al.,* 1978).

Le problème reste l'impossibilité d'estimer à la fois les X_O, valeurs des variables chez l'ancêtre, et les quantités d'évolution *e*. La suggestion de Cavalli-Sforza et Piazza revient à prendre comme valeur des variables chez l'ancêtre la moyenne du caractère estimée sur l'ensemble des taxons considérant que l'espérance des variables *X* est bien toujours égale à X_O (paragraphe VIII.2). Cette transformation présente l'inconvénient d'être sensible au choix des UE étudiées : la présence d'un ensemble de UE monophylétiques proches les unes des autres et largement échantillonnées fera naturellement pencher le centre de gravité des caractères vers ce groupe monophylétique, déplaçant ainsi la racine de l'arbre. Un biais dans l'échantillonnage des UE entraîne donc une distorsion dans la localisation de la racine.

Enfin, en explorant l'ensemble des différentes topologies possibles (ou un sous-ensemble raisonnable choisi selon une stratégie donnée), il est possible d'obtenir une estimation du paramètre « structure de l'arbre », *T*, qui est celle qui maximise la vraisemblance totale.

Il est important de noter que cette méthode produit un arbre enraciné, à la différence de la méthode précédente.

Exemple :

La matrice des covariances entre 5 populations européennes estimée à partir de 49 fréquences géniques différentes (24 systèmes différents : ABO, Rhésus, MNS, HLA, Gc ...) a été calculée en utilisant comme valeur moyenne des fréquences alléliques la valeur moyenne estimée sur les 5 populations. La figure VIII.8 donne l'arbre obtenu. La qualité de l'ajustement, donnée par le rapport T_r ou « *treeness* », est ici égale à 0.69, correspondant à un $\chi^2 = 36.4$, d.d.l. = 6. On voit que l'ajustement est médiocre. Le modèle évolutif choisi pour rendre compte de la matrice de covariance entre ces 5 populations est donc probablement inadéquat. Il est possible d'améliorer cet ajustement, en supposant, par exemple, que des mélanges se sont produits entre des populations ancestrales. De tels mélanges peuvent expliquer les courtes branches de la France et de l'Angleterre. Plusieurs modèles ont été développés pour rendre compte de tels mélanges dans les phylogénies (Cavalli-Sforza et Piazza, 1975 ; Lathrop, 1982 ; Darlu et Lathrop, 1993).

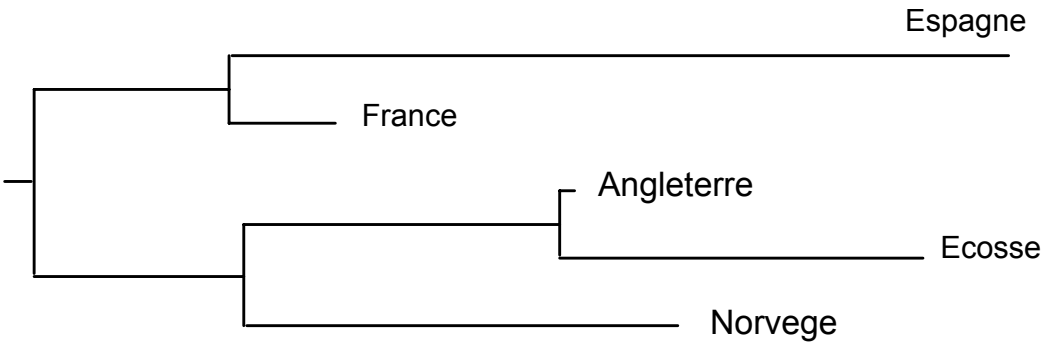


FIGURE VIII.8. *Arbre enraciné obtenu par la méthode du Treeness (Cavalli-Sforza et Piazza, 1975). Les 5 populations sont définies par 49 fréquences géniques. La méthode admet une vitesse d'évolution variable sur chaque branche.*

3. Modèle d'évolution de caractères discrets

L'une des ambitions de la reconstruction phylogénétique est de rechercher l'arbre le plus vraisemblable et d'estimer les états des caractères aux nœuds, en se fondant sur l'observation des caractères dans les UE terminaux. Ce que l'on peut écrire :

$$L(T,Y|X) = P(X|T,Y)$$

où T représente la structure d'un arbre et Y les états des caractères aux différents nœuds que l'on cherche à inférer et où X est l'ensemble des états de caractères observés sur les UE. En fait, dans le cadre de la méthode du maximum de vraisemblances, les différentes valeurs prises par Y aux nœuds sont des réalisations d'une variable aléatoire et ne sont donc pas *a priori* des paramètres que l'on peut estimer. Si l'on souhaite cependant les estimer, il faut alors les considérer comme des paramètres. Ceux-ci sont nécessairement des paramètres d'incidence comme on l'a vu plus haut. Dans ces conditions leur estimation, ainsi que celle de T, peut être inconsistante. Pour contourner cette difficulté, on peut abandonner l'idée d'estimer Y, et se contenter d'obtenir sa distribution en probabilité à partir de paramètres structuraux qui sont ceux définissant la réalisation des X.

Après quelques généralités, deux modèles différents pour résoudre cette question seront développés : le modèle d'évolution de type Poisson (Felsenstein, 1981), et le modèle d'évolution indépendant du temps proposé par Sober (1985 ; 1988), modèle que ce dernier considère comme étant identique au modèle de parcimonie, ce que conteste avec arguments Goldman (1990). La discussion sur ce modèle nous conduira à préciser les rapports entre vraisemblance et parcimonie.

3 .1. Généralités

La question est d'exprimer, en terme de probabilité, les différentes valeurs que peuvent prendre les variables Y aux nœuds (ou UEH), sachant les valeurs observées, X, sur les UE.

Soit un caractère X pouvant se présenter sous s états différents X_a, X_b, ...X_s. Lorsque le caractère est codé « présent » ou « absent », $s = 2$. Lorsqu'il s'agit d'un site nucléotidique quatre états sont possibles, $s = 4$: Adénine, Guanine, Cytosine et Thymine ou Uracile.

La notation $X_{d,D}$ indique que l'UEH D possède le caractère X dans l'état d (figureVIII.9). La probabilité d'observer l'état d du caractère X chez D, sachant que l'on observe l'état a dans l'UE A qui en est dérivé et sachant que la quantité d'évolution séparant D de A est égale à e_{AD} peut s'écrire de la façon suivante :

$$p(X_{d,D}|X_{a,A},e_{AD})$$

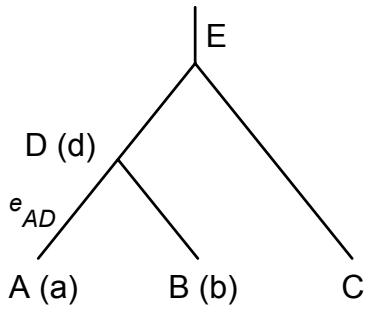


FIGURE VIII.9. *Les UE A et B possèdent le caractère X sous l'état a et b respectivement.* e_{AD} *est la quantité d'évolution séparant D de A.*

Cette probabilité est fonction de la quantité d'évolution entre *A* et *D* (voir paragraphe VIII.2) et de la probabilité de transformation de l'état *d* vers l'état *a* caractère *X*. Différents modèles peuvent être proposés pour évaluer cette probabilité. Certains d'entre eux seront exposés plus loin.

La vraisemblance de $X_{d,D}$ l'état *d* du caractère *X* chez *D*, est donnée par le produit de deux expressions comparables : l'une concernant l'UE fille *A*, l'autre l'UE fille *B*. La justification de ce produit vient de ce que les évolutions de *D* vers *A* et de *D* vers *B* sont indépendantes. Chacune de ces expressions représente la probabilité que *D* soit X_d sachant que *A* (et B) est dans l'un des *s* états possibles, pondéré par la vraisemblance que *A* (et B) soit effectivement dans cet état :

$$L(X_{d,D}|X_{a,A},X_{b,B}) = \sum_{h=1}^{s} [p(X_{d,D}|X_{h,A},e_{AD}).L(X_{h,A}). \sum_{h=1}^{s} [p(X_{d,D}|X_{h,B},e_{BD}).L(X_{h,B})$$

Lorsque *A* (ou *B*) est une UE sur laquelle l'état du caractère X peut être observé, la vraisemblance d'un état du caractère *X* dans ce taxon *A* (ou *B*) est :

$L(X_{h,A})$ = 1 si *A* est *h*, 0 autrement.
$L(X_{h,B})$ = 1 si *B* est *h*, 0 autrement.

Dans ce cas la formule donnant la vraisemblance de X_d se simplifie grandement. En revanche, si *A* (ou *B*) est une UEH, la vraisemblance de chacun de ses états devra être estimée par la même formule.

Les vraisemblances de chacun des *s* états possibles du caractère *X* dans l'UEH D, peuvent donc être estimées à partir des observations effectuées sur les taxons *A* et *B*. De la même façon, il est possible de calculer la vraisemblance de tous les états possibles du caractère *X* dans l'UEH E, simplement en appliquant la formule précédente, les UE filles étant cette fois *D* et *C*. Le calcul de la vraisemblance de l'arbre *T* peut s'étendre ainsi jusqu'à la racine *O* de l'arbre. Elle sera alors égale, pour le caractère *X* à la racine :

$$L_X = \sum_{h=1}^{s} [\pi_{h,O}.L(X_{h,O})]$$

où $\pi_{h,O}$ est la probabilité *a priori* que *X* soit dans l'état *h* à la racine de l'arbre. La somme est étendue sur l'ensemble des *s* états du caractère *X*.

La vraisemblance totale de l'arbre *T* est obtenue en multipliant les vraisemblances calculées pour chacun des *c* caractères *X* évoluant indépendamment :

$$L(T,E) = \prod_c L_X$$

Dans le calcul de cette vraisemblance, les paramètres du modèle sont la structure *T* de l'arbre et *E*, l'ensemble des quantités d'évolution *e* attachées à chacune des branches. Les états des caractères aux nœuds ne sont pas ici des paramètres à estimer. Ce sont des états conditionnés, selon une loi de probabilité à définir, par les réalisations des variables au niveau des UE observées.

Maintenant que l'on sait calculer la vraisemblance totale d'un arbre ayant une structure donnée *T*, on peut rechercher la structure de l'arbre *T* et les longueurs de branches, exprimées en terme de quantité d'évolution, qui maximisent cette vraisemblance selon des procédures décrites antérieurement.

La formulation d'hypothèses plausibles sur les probabilités des événements (changements d'état, substitutions de bases ou d'acides aminés) telles qu'elles sont décrites dans la première formule du paragraphe VIII.3.1 constitue la difficulté essentielle d'une telle approche. Plusieurs modèles ont été proposés dans ce but. Nous n'en décrirons ici que deux : l'un dérive du modèle d'évolution de type Poisson, le second en est une version simplifiée proposée par Sober pour son analogie avec la méthode de parcimonie (Sober, 1988). Ce point d'ailleurs sera discuté dans une autre partie (VIII.4).

3.2. Modèle d'évolution de type Poisson, fonction du temps

Supposons que les probabilités de changements d'état par unité de temps, λ pour un caractère *X*, ne dépendent ni du sens ni de la nature de la transformation, ni de sa position dans l'arbre. Dans ce cas la probabilité d'observer *x* changements d'état pendant un temps *t* est donnée par :

$$p(x,\lambda) = \frac{(\lambda t)^x}{x\,!} e^{-\lambda t}$$

Lorsqu'il n'y a que deux états possibles, 0 et 1 par exemple, la probabilité que l'ancêtre *D* (figure VIII.9) ait le même état de caractère que l'UE fille *A* est égale à la somme de la probabilité qu'il n'y ait aucun changement entre *A* et *D*, de la probabilité qu'il y ait deux changements dont une réversion (0 vers 1 et 1 vers 0), de la probabilité qu'il y ait 4 changements, etc.

$$p(X_{i=0,D} | X_{i=0,A}, \lambda t) = p(x=0,\lambda t) + p(x=2,\lambda t) + p(x=4,\lambda t) + ... = \frac{1}{2}(1+e^{-2\lambda t})$$

De même la probabilité pour que *A* et *D* ne soient pas dans le même état du caractère *X* s'écrit :

$$p(X_{i=0,D} | X_{i=1,A}, \lambda t) = p(x=1,\lambda t) + p(x=3,\lambda t) + p(x=5,\lambda t) + ... = \frac{1}{2}(1-e^{-2\lambda t})$$

La valeur λ*t* correspond ici à la quantité d'évolution e_{AD} entre *A* et *D*. A moins de pouvoir évaluer λ de manière indépendante, il n'est pas possible de distinguer entre λ et *t,* le temps séparant *A* de *D* (voir paragraphe VIII.2).

L'exemple donné ici pour estimer les probabilités de changements lorsque les caractères se présentent sous deux états distincts peut se généraliser aux situations où ils se présentent sous plusieurs états. C'est le cas par exemple pour un site nucléotidique où 4 états sont possibles (A,T,C,G pour l'ADN). Dans ce cas on peut écrire, g et h étant deux états distincts du caractère X (Bishop et Friday, 1985) :

$$p(X_{g,D}|X_{h,A},\lambda t) = \frac{1}{4}(1 - e^{-2\lambda t}) + \delta\, e^{-2\lambda t}$$

avec $\delta = 0$ lorsque $g = h$, $\delta = 1$ lorsque $g \neq h$.

Il est possible de préciser davantage les probabilités de changements, lorsqu'il s'agit d'un site nucléotidique par exemple. Ainsi peut-on supposer que les probabilités de changements d'état du caractère sont différentes selon la nature du changement : c'est le cas du modèle à deux paramètres de Kimura (1980) où les probabilités de transition sont différentes des probabilités de transversion. Des modèles à 4 paramètres (Takahata et Kimura, 1981), 6 paramètres (Kimura (1981b), 12 paramètres (Gojobori *et al.,* 1982) ont également été proposés. Un modèle général a été développé par Tajima et Nei (1984).

Un exemple

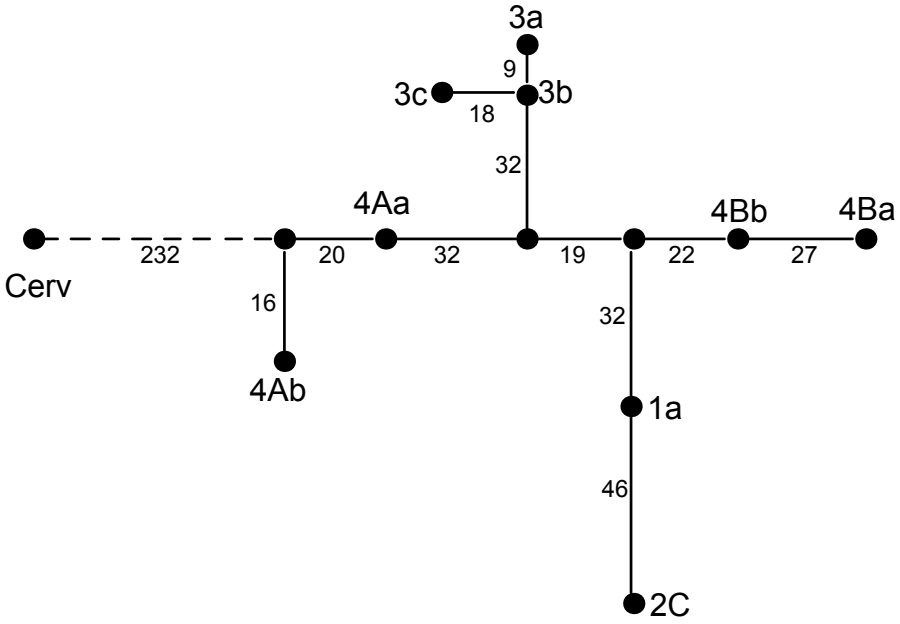


FIGURE VIII.10 *: Arbre non enraciné obtenu par maximum de vraisemblance à partir des séquences alignées de 114 sites de l'extrémité 5' de l'ARN ribosomal 16S dans le genre* Mus *(Fort, 1982 ; Fort* et al.*, 1984). Cerv :* Mus Cervicolor*; 4Aa et 4Ab , 4Ba et 4Bb :* Mus spicilegus *(Bulgarie) ; 1a :* Mus domesticus *; 2C :* Mus musculus castaneus *; 3a, 3b, 3c :* Mus spretus. *Les fréquences des nucléotides utilisées* a priori *sont les fréquences observées sur l'ensemble des sites. Le rapport transition/transversion donne une vraisemblance optimale quand il est proche de 2 (programme DNAML de Felsenstein, 1990). Les nombres sont proportionnels à la longueur des branches.*

3.3. Modèle d'évolution indépendant du temps

Dans ce modèle, la probabilité qu'aucun changement d'état du caractère *X* ne soit observé entre *D* et *A* est donnée par :

$$P(X_{h,D}|X_{h,A},e_{AD}) = \pi$$

La probabilité d'observer des états différents en *A* et *D* est :

$$P(X_{g,D}|X_{h,A},e_{AD}) = (1- \pi)$$

π peut prendre une valeur quelconque entre 0 et 1. Il est important de remarquer que ces probabilités ne sont pas fonctions de la quantité d'évolution entre *D* et *A*, e_{AD}, et sont donc indépendantes du temps *t*, à la différence du modèle précédent. Autrement dit on considère dans ce modèle qu'un événement a autant de chance de se produire durant un intervalle de temps court que durant une longue période de temps. En conséquence, une *différence* d'état entre deux taxons peut correspondre à un nombre élevé de *changements* d'état, sans que la probabilité de l'évènement soit différente : dans le cas d'un caractère présentant deux états, 0 et 1, on peut en effet passer de 0 à 1 selon le schéma 0➞1 (1 pas) ou bien selon le schéma suivant : 0➞1➞0➞1 (3 « événements ») avec, dans ce modèle, la même probabilité. De même l'absence de différence peut dissimuler plusieurs changements d'état : 0➞1➞0 par exemple ou davantage (2, 4, 6, ...pas).

Ces possibilités de « changements multiples » ne sont donc pas prises en compte, et la probabilité de réalisation de l'événement ne dépend pas du temps. Ce modèle reprend celui donné en exemple dans l'introduction de ce chapitre.

La vraisemblance s'exprime alors en fonction de la structure de l'arbre *T*, de π et des probabilités *a priori* des différents états à la racine de l'arbre (voir les formules du paragraphe VIII.1.2) :

$$L(T,\pi) = \prod_c L_X$$

La structure *T* et la valeur π estimées seront celles maximisant cette vraisemblance.

4. Parcimonie et vraisemblance

Dans l'analyse de parcimonie, chaque différence d'état (et non pas chaque changement d'état) d'un caractère *X* entre l'origine et la fin d'une branche (0 et 1 ou 1 et 0 par exemple) compte pour une longueur unité et contribue donc à ajouter, quel que soit le prix de cette « différence », une unité à la longueur totale de l'arbre *T*. L'arbre *T* le plus économique (le plus parcimonieux) sera celui nécessitant, pour un ensemble donné de caractères, le nombre minimal de différences d'état entre les deux extrémités de toutes les branches Pour obtenir ce nombre, il est évident qu'il est nécessaire de faire des inférences sur l'état des caractères aux nœuds de l'arbre. En d'autres termes, cette méthode suppose que l'on estime à la fois des paramètres de structure (ici *T)* et les valeurs, *Y*, prises par

les variables aux nœuds ; comme on l'a vu, ces variables sont alors des paramètres d'incidence. En conséquence, il n'est pas assuré qu'une telle approche possède la qualité de consistance que l'on souhaite, à la différence de la méthode précédente où de tels paramètres étaient traités comme des paramètres de nuisances et supprimés de l'inférence en leur affectant des probabilités conditionnelles calculées à partir d'une loi de distribution, la loi de Poisson en l'occurrence, ou à partir d'une probabilité constante.

La méthode de parcimonie revient à maximiser la vraisemblance suivante, où Y représente ici les états des caractères aux nœuds que l'on veut inférer et X les états des caractères sur les UE observées :

$$L(T,Y|X) = P(X|T,Y)$$

Le modèle d'évolution est celui décrit au paragraphe VIII.1.2 : la probabilité qu'aucune différence d'état ne soit observée sur une branche donnée dont l'origine est I et l'extrémité est J s'écrit :

$$P(I=1|J=1) = P(I=0|J=0) = \pi$$
$$P(I=1|J=0) = P(I=0|J=1) = (1-\pi)$$

Soit v^i_{00} le nombre de branches commençant et finissant par l'état 0 du caractère i,

v^i_{01} le nombre de branches commençant par l'état 0 et finissant par l'état 1,

v^i_{10} le nombre de branches commençant par l'état 1 et finissant par l'état 0,

v^i_{11} le nombre de branches commençant et finissant par l'état 1.

On peut écrire la vraisemblance de la distribution des états des c caractères dans l'arbre T de la façon suivante :

$$L = \prod_{i=1}^{c} \pi^{(v^i_{00} + v^i_{11})} . (1-\pi)^{(v^i_{01} + v^i_{10})}$$

$$L = \pi^{\sum_i (v^i_{00} + v^i_{11})} . (1-\pi)^{\sum_i (v^i_{01} + v^i_{10})}$$

Pour maximiser cette vraisemblance, il suffit de déterminer les états des variables Y telles que l'exposant du premier terme soit maximal ou, de manière équivalente telle que l'exposant du second terme soit minimal, c'est-à-dire minimiser le nombre de pas. La condition est, de plus, que π soit supérieur à 1/2. La vraisemblance est donc ici conditionnée à l'attribution des états aux nœuds.

L'exemple suivant, inspiré de Goldman (1991) permet de bien illustrer cette différence entre vraisemblance et parcimonie. Considérons la matrice suivante de caractères, codés 0 ou 1 (Tableau VIII.1). Pour simplifier, supposons que l'état ancestral soit l'état 0 et que le choix ne se porte que sur deux arbres différents T_1 et T_2 de la figure VIII.11. On relève l caractères distribués comme le caractère X, m comme le caractère Y, n comme Z et w comme W.

UE	A	B	C	D	Nombre
Caractères					
X	1	1	0	0	l
Y	0	0	1	1	m
Z	0	1	1	0	n
W	1	1	1	0	w

TABLEAU VIII.1. *Quatre types de distributions de caractères (X, Y, Z, W) dans quatre UE (A, B, C et D). l, m, n et w représentent les nombres d'occurrence de chacun de ces types.*

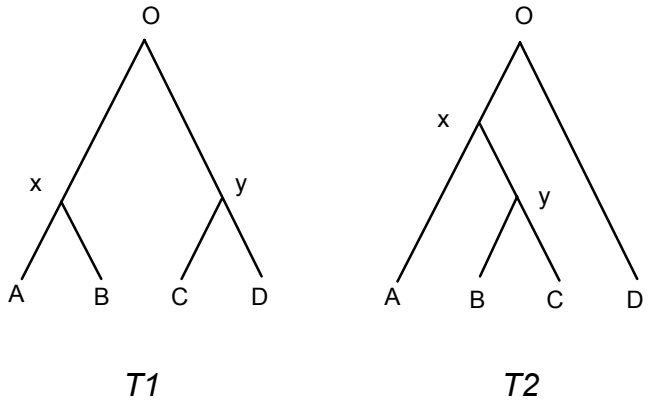


FIGURE VIII.11 *: Exemple de deux arbres T_1 et T_2 ayant 4 UE (A, B, C, D) et deux UEH (x, y).*

L'arbre T_1 rend compte de la distribution du caractère de type X (et Y) en supposant un seul changement, puisque l'ancêtre est O. En revanche deux changements sont nécessaires pour rendre compte de la distribution des caractères de type Z et W. Le nombre total de changements, C_1, requis pour rendre compte des observations de la matrice de données dans le cas de l'arbre T_1 est donc :

$$C_1 = (l + m) + 2(n + w)$$

Un raisonnement identique montre que le nombre de changements nécessaires dans le cas où l'arbre serait T_2 est :

$$C_2 = 2(l + m) + (n + w)$$

Appliquons le critère de parcimonie : on dira que l'arbre T_1 est plus parcimonieux que l'arbre T_2 si $C_1 < C_2$, donc si $n + w < l + m$

Qu'en est-il d'une approche par la méthode de vraisemblance ? Choisissons un modèle identique à celui décrit dans l'exemple donné au début de ce chapitre, plus simple cependant en ce sens que l'état ancestral est cette fois connu. Comme on l'a vu, il faut envisager *toutes* les combinaisons possibles des différents états aux

différents nœuds et en calculer la probabilité. Le tableau VIII.2 résume ces probabilités.

La vraisemblance de l'arbre T_1 est donnée par :

$$L(T_1|D) = p(X)^l p(Y)^m p(Z)^n p(W)^w$$

où $p(X)$, $p(Y)$, $p(Z)$ *et* $p(W)$ sont les probabilités figurant dans la colonne « somme » du tableau VIII.2 correspondant à l'arbre T_1. De la même façon on peut calculer la vraisemblance de l'arbre T_2. L'arbre T_1 est choisi lorsque $L(T_1|D) > L(T_2|D)$; l'arbre T_2 dans le cas contraire.

On peut trouver facilement des cas où le critère de parcimonie ne donne pas le même résultat que le critère de vraisemblance. Par exemple, posant $\pi = 0.75$ et $l = 2$, $m = n = w = 1$. Dans ce cas $l + m > n + w$. La parcimonie nous conseille l'arbre T_1. En revanche on a : $L(T_1|D) = 6.82.10^{-7} < L(T_2|D) = 7.13.10^{-7}$. La vraisemblance nous conseille donc l'arbre T_2.

De la même façon, on peut trouver des exemples où la parcimonie ne nous permet pas de choisir entre T_1 et T_2 ($l + m = n + w$) tandis que la vraisemblance le permet.

	Etat	x=0 y=0	x=0 y=1	x=1 y=0	x=1 y=1	Somme
T1	X	$\pi^4\rho^2$	$\pi\rho^5$	$\pi^5\rho^*$	$\pi^2\rho^4$	$\pi\rho(1-3\pi+3\pi^2)$
	Y	$\pi^4\rho^2$	$\pi^5\rho^*$	$\pi\rho^5$	$\pi^2\rho^4$	$\pi\rho(1-3\pi+3\pi^2)$
	Z	$\pi^4\rho^{2*}$	$\pi^3\rho^3$	$\pi^3\rho^3$	$\pi^2\rho^4$	$\pi^2\rho^2$
	W	$\pi^3\rho^3$	$\pi^2\rho^4$	$\pi^4\rho^{2*}$	$\pi^3\rho^3$	$\pi^2\rho^2$
T2	X	$\pi^4\rho^{2*}$	$\pi^3\rho^3$	$\pi^3\rho^3$	$\pi^4\rho^{2*}$	$2\pi^3\rho^2$
	Y	$\pi^4\rho^{2*}$	$\pi^3\rho^3$	$\pi\rho^5$	$\pi^2\rho^4$	$\pi\rho^2(1-2\pi+2\pi^2)$
	Z	$\pi^4\rho^2$	$\pi^5\rho^*$	$\pi\rho^5$	$\pi^4\rho^2$	$\pi\rho(1-4\pi+6\pi^2-2\pi^3)$
	W	$\pi^3\rho^3$	$\pi^4\rho^2$	$\pi^2\rho^4$	$\pi^5\rho^*$	$\pi^2\rho(1-2\pi+2\pi^2)$

TABLEAU VIII.2. *Probabilités des différentes distributions des caractères* X, Y, Z et W *en fonction de l'arbre choisi (T_1 ou T_2) et selon les états des caractères aux nœuds* x *et* y.

Cet apparent paradoxe où la vraisemblance est en désaccord avec la parcimonie s'explique aisément. En effet, dans le tableau VIII.2, des astérisques figurent, pour chaque caractère, la combinaison d'états aux nœuds x et y qui présente la probabilité maximale : ainsi, pour l'arbre T_1 et le caractère X, la probabilité la plus élevée, sachant les états aux nœuds, est $\pi^5\rho$ correspondant à x = 0 et y = 1, étant entendu que $\pi > 0.5$, c'est-à-dire que le changement est plus rare que le non

changement. Du point de vue de la parcimonie formulée dans le cadre de la vraisemblance, les critères à comparer pour choisir l'arbre T_1 ou T_2 utilisent les probabilités conditionnelles les plus élevées (connaissant x et y), celles marquées « * » dans le tableau VIII.2 :

$$L(T_1,x,y\ /D)\ = (\pi^5\ \rho)^{l+m}(\pi^4\rho^2)^{n+w} \text{ et } L(T_2,x,y\ /D)\ = (\pi^5\ \rho)^{n+w}(\pi^4\rho^2)^{l+m}$$

ce qui revient bien à choisir T_1 si $l + m > n + w$, toujours à la condition que $\pi > 0.5$. La conclusion est donc que l'arbre le plus parcimonieux n'est pas nécessairement l'arbre le plus vraisemblable.

5. Parcimonie, vraisemblance et consistance

Dans le paragraphe VIII.4, on a montré que, d'un point de vue probabiliste, la méthode de parcimonie revenait à attribuer une valeur constante aux probabilités d'observer un état différent entre les deux extrémités d'une branche. Que peut-il se passer lorsque cette hypothèse n'est pas exacte et donc lorsque ces probabilités sont différentes d'une branche à l'autre ? Cette question revient également à poser celle de la consistance de la méthode de parcimonie. Rappelons que la consistance est une propriété statistique qui fait que l'estimation d'un paramètre converge vers la vraie valeur de ce paramètre au fur et à mesure que les données s'accumulent.

Les méthodes de vraisemblance sont généralement consistantes, quand le nombre de paramètres à estimer n'augmente pas plus vite que ne s'accumulent les données. Cette condition est satisfaite pour les constructions phylogénétiques quand la structure de l'arbre et les longueurs de branches sont les seuls paramètres à estimer. Il est clair que si les états des caractères aux nœuds étaient également considérés comme des paramètres à estimer, ceux-ci augmenteraient en même temps que le nombre de caractères et les estimations de tels paramètres par la méthode de vraisemblance ne seraient plus nécessairement consistantes. Or, on l'a vu, la méthode de parcimonie estime la structure de l'arbre par optimisation des états des caractères aux nœuds ; on peut donc suspecter cette méthode d'être inconsistante au sens statistique du terme.

Ces interrogations ont principalement été posées par Felsenstein (1978b), à partir d'une situation simple ne comprenant que 4 UE, les caractères ne prenant que deux états différents (0 et 1), puis à partir d'une situation plus complexe où les caractères peuvent se présenter sous 4 états (A, T, C, G : les 4 acides nucléiques) (Felsenstein, 1983b). Le cas d'un nombre d'UE plus élevé a été abordé par Hendy et Penny (1988). Dans l'exemple donné ici, on présentera uniquement le cas de 4 UE et de caractères binaires.

Soient quatre UE (A, B, C, D). Supposons qu'elles puissent se connecter selon les deux seuls arbres non enracinés de la figure VIII.12 : arbres T_1 et T_2. Les paramètres a, b, c, d, e sont les *probabilités d'observer une différence d'états entre les deux extrémités de chacune des 5 branches*. Les caractères observés ne sont présents que sous deux états : 0 et 1. Supposons maintenant, pour simplifier

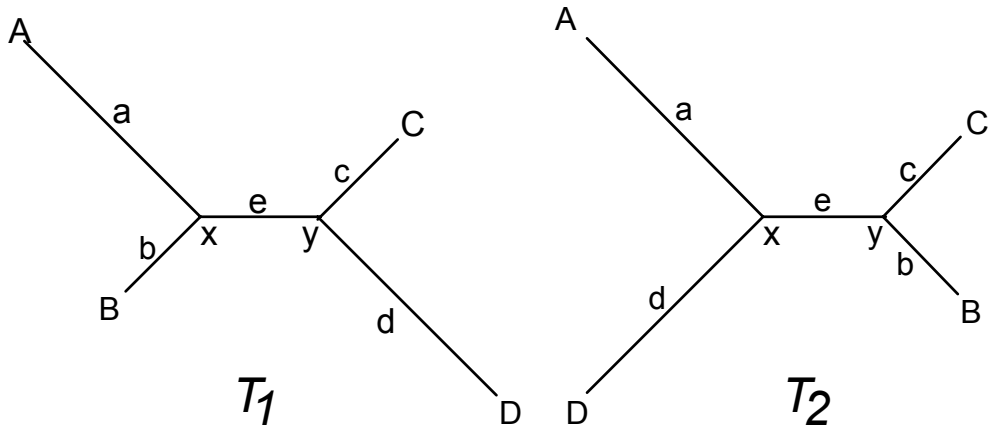


FIGURE VIII.12. *Les quatre UE (A, B, C et D) sont respectivement dans les états 1, 1, 0 et 0. Deux arbres non enracinés sont considérés : l'arbre de type T_1 et l'arbre de type T2. Les valeurs a, b, c, d et e correspondent aux probabilités d'observer un état différent entre le début et la fin de la branche.*

l'exemple, que l'on observe exclusivement les deux distributions suivantes des caractères : A=1, B=1, C=0, D=0 d'une part et A=1, B=0, C=0, D=1 d'autre part (plus schématiquement on écrira : 1100 et 1001 respectivement, dans l'ordre des UE ABCD). Il est possible de calculer la probabilité d'observer de telles distributions dans le cas de l'arbre T_1 et dans celui de l'arbre T_2. Pour cela il faut envisager toutes les quatre combinaisons d'états aux deux nœuds x et y : 00, 10, 01, 11. On a donc :

$$p(1100|T_1) = p(1100|T_1,x=0,y=0)+p(1100|T_1,x=0,y=1)+p(1100|T_1,x=1,y=0)+ p(1100|T_1,x=1,y=1)$$

$$p(1100|T_1) = ab(1-c)(1-d)(1-e)+(1-a)(1-b)(1-c)(1-d)e+abcde+(1-a)(1-b)cd(1-e)$$

De la même façon peut-on calculer :

$$p(1001|T_1) = a(1-b)(1-c)d(1-e)+(1-a)b(1-c)de+a(1-b)c(1-d)e+(1-a)bc(1-d)(1-e)$$

$p(1100|T_2) = p(1001|T_1)$, remplaçant b par d et d par b,

$p(1001|T_2) = p(1100|T_1)$, remplaçant b par d et d par b.

Soit n' et n" les nombres de caractères respectivement distribués selon 1100 et 1001. Supposons que l'arbre véritable soit l'arbre T_1. Le critère de parcimonie nous conduit à choisir correctement l'arbre T_1 dès lors que la proportion n'/N des caractères de type 1100 est supérieure à la proportion n"/N des caractères de type 1001 :

$$n'/N > n''/N$$

N est ici le nombre total de caractères observés. Lorsque N augmente, ces proportions convergent vers leurs probabilités, c'est-à-dire vers $p(1100|T_1)$ et $p(1001|T_1)$ respectivement. Pour que la méthode de parcimonie soit consistante, il

faut donc que l'inégalité suivante, correspondant au cas où N est grand, soit également vérifiée :

$$p(1100|T_1) > p(1001|T_1)$$

Or on peut démontrer qu'il n'en est pas toujours ainsi. La figure VIII.13 en effet représente les variations du rapport :

$$L = \frac{p(1100|T_1)}{p(1001|T_1)}$$

en fonction de la valeur de la probabilité d'observer une différence entre les deux extrémités d'une branche et en fonction des variations de cette probabilité d'une branche à l'autre. Ce rapport n'est pas toujours supérieur à 1.

Pour simplifier, posons que a = d, b = e = c et que le rapport b/a = r. Comme la méthode de parcimonie n'est consistante que dans les cas où *L* est supérieur à 1, la figure VIII.13 montre que cela est vrai à la condition que a soit petit et que le rapport r soit assez grand, donc à la condition que les changements soient « rares » et que les « longueurs » de branches (au sens de probabilités d'observer une différence d'état entre les extrémités des branches) ne soient pas trop différentes.

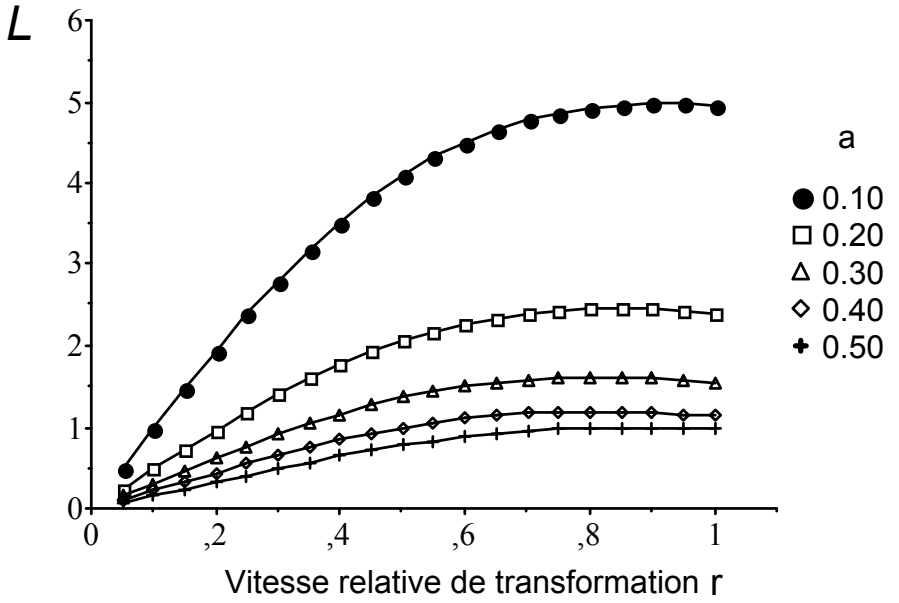


FIGURE VIII.13. *Variation du rapport de vraisemblance L de l'arbre T_1 en fonction de la probabilité de transformations le long des branches (ρ= a) et de sa variation relative r selon les branches (voir figure* VIII.12*). La méthode de parcimonie n'est consistante que dans les cas où L est supérieur à 1.*

Ces problèmes de consistance ont été soulevés par Felsenstein (1978b) qui démontre leur existence aussi bien dans le cas des méthodes de parcimonie appliquées à des caractères orientés ou non orientés que dans le cas des méthodes de compatibilité. Par ailleurs Hendy et Penny (1988) ont souligné que la propriété

de consistance n'est pas plus démontrée quand le nombre d'UE augmente, même lorsque les longueurs de branches (au sens de probabilités d'observer une différence d'état entre les extrémités des branches) sont identiques pour toutes les branches.

Par ailleurs, pour illustrer des situations où l'on observe une contradiction entre l'arbre choisi par la méthode de parcimonie et celui choisi par la méthode de vraisemblance, calculons le rapport de vraisemblance suivant à partir de l'exemple précédent :

$$L = \frac{p(1100|T_1)^{n'} p(1001|T_1)^{n''}}{p(1100|T_2)^{n'} p(1001|T_2)^{n''}}$$

Si l'on suppose que les deux arbres T_1 et T_2 ont une probabilité *a priori* identique, c'est-à-dire que l'on part de l'idée initiale que chacun des arbres a autant de chance d'être le bon, alors la méthode de vraisemblance conduit à choisir l'arbre T_1 quand ce rapport est supérieur à 1, et l'arbre T_2 quand il est inférieur.

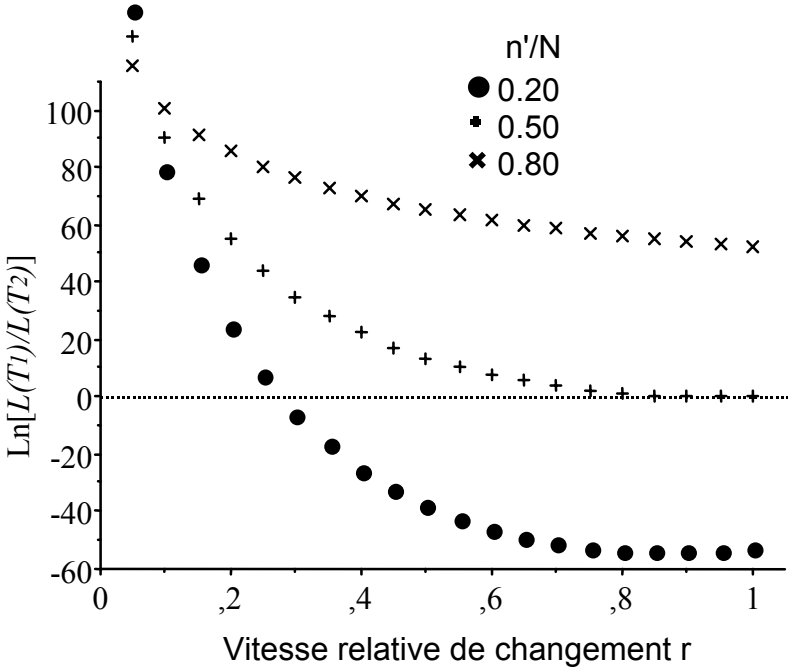


FIGURE VIII.14. *Variations du rapport entre les probabilités d'observer les arbres non enracinés T_1 et T_2, en fonction du rapport r entre les probabilités de changements le long des branches « rapides » et « lentes » et en fonction de la proportion de caractères 1100 parmi N caractères. La courbe correspondant à n'/N = 0.5 délimite une région supérieure où la parcimonie choisit l'arbre T_1 et une région inférieure où elle choisit l'arbre T_2. La méthode de vraisemblance choisit l'arbre T_1 au-dessus de la ligne d'ordonnée zéro et l'arbre T_2 en dessous. On voit donc que la méthode de parcimonie conduit à choisir l'arbre T_2, et non T_1, dans toute la région comprise entre la ligne d'ordonnée nulle et la courbe en croix.*

En revanche la méthode de parcimonie conduit à choisir l'arbre T_1 quand n' (nombre de caractères de type 1100) est supérieur à n" (nombre de caractères de type 1001), donc quand n'/N > 0.5, et T_2 dans le cas inverse.

Les courbes de la figure VIII.14 ont été calculées avec a = 0.20, a = d, b = c = e ; b/a = r. La zone de choix de T_1 par la méthode de vraisemblance est celle située au-dessus de l'ordonnée $L = 0$. La zone de choix de l'arbre T_1 par la méthode de parcimonie est celle située au-dessus de la courbe pour laquelle n' = n". Il existe donc une zone où les deux méthodes conduisent à un choix contradictoire. Cette contradiction ne disparaît que lorsque r = 1.

6. Conclusions

Les différents résultats et développements présentés dans les paragraphes 4 et 5 conduisent aux remarques générales suivantes.

Rappelons d'abord que chaque méthode phylogénétique fait appel à un *modèle d'évolution*. explicite ou non. Toute méthode se propose d'estimer et comparer, *à l'intérieur d'un même modèle*, les valeurs prises par les paramètres du modèle, valeurs qui constituent les hypothèses du modèle. L'objet de la méthode n'est donc, en aucun cas, de chercher à réfuter un modèle.

1) La solution de parcimonie peut s'intégrer dans le cadre d'un modèle probabiliste d'évolution dont les paramètres peuvent s'estimer par la méthode du maximum de vraisemblance. Il s'agit d'un modèle d'évolution qui attribue une probabilité fixe, non dépendante du temps, à l'existence d'une *différence* entre le début et la fin d'une branche. Pour chaque caractère, le nombre de changements par branche est toujours minimal, compte tenu des états aux extrémités des branches. Les changements supplémentaires qui pourraient survenir sur une même branche sans altérer les états aux extrémités des branches ne sont pas pris en considération. La probabilité d'existence d'une différence entre début et fin d'une branche doit être inférieure à la probabilité d'absence de différence. Cette condition revient à assumer que le changement est rare par rapport au non changement.

2) La méthode de parcimonie effectue des inférences sur les états des caractères aux nœuds. Elle revient donc à ne prendre en considération que la vraisemblance *conditionnée* par les états des caractères aux nœuds. *Elle choisit la combinaison d'états de caractère la plus probable, même si ce choix peut conduire à une vraisemblance de l'arbre plus faible*. Une telle méthode peut ne pas être consistante (au sens statistique du terme), c'est-à-dire que l'augmentation du nombre de caractères peut conduire à conforter un choix d'arbre erroné.

3) L'arbre le plus parcimonieux n'est pas nécessairement l'arbre le plus vraisemblable et inversement.

Les difficultés techniques de mise en place des méthodes de vraisemblance expliquent pourquoi elles sont aussi rarement utilisées actuellement. En effet, dès que le nombre d'UE est élevé, il devient vite impossible d'avoir quelques chances d'obtenir l'arbre le plus vraisemblable dans des temps raisonnables, même avec le plus performant des ordinateurs. Des progrès sont donc à espérer pour bientôt dans cette voie.

CONCLUSION

Au terme de cette revue des méthodes d'analyse phylogénétique, il est difficile d'esquiver la question de savoir si certaines méthodes sont meilleures que d'autres. Toutefois, ce livre n'a pas d'autre finalité que d'expliciter les performances et les limites de chaque type d'approche, les buts poursuivis et les options choisies en fonction de ces buts. En ce sens, un classement des méthodes, de la plus mauvaise à la meilleure, ne se conçoit pas. Les commentaires ultimes que nous ferons à propos des approches cladistique, phénétique et probabiliste, ne seront qu'un bref rappel de ce qui les différencie fondamentalement.

Le succès grandissant des *méthodes cladistiques* vient en partie de ce qu'elles reposent sur les notions de caractère et d'homologie, notions qui sont au cœur de la pensée évolutionniste. En cherchant à construire un schéma de parenté en reconstituant des traits ancestraux à chaque nœud de l'arbre, l'approche cladistique répond à l'attente des phylogénéticiens qui considèrent les caractères individuels comme les éléments ultimes sur lesquels s'appliquent les phénomènes évolutifs.

Les limites du cladisme peuvent être perçues dans les discussions autour du lien entre le principe de parcimonie - principe de base du cladisme - et le processus évolutif. Ces discussions ont été abordées à la fin des chapitres V et VIII. Nous n'en retiendrons ici que deux aspects.

— Pour que les résultats obtenus par l'application du principe de parcimonie puissent être interprétés en termes de relation de parenté entre taxons, avec le moins d'erreurs d'inférence possible, il faut que l'apparition, la transformation ou la disparition des caractères, soient des événements qui surviennent *rarement* au cours du processus évolutif.

Mais que signifie cette rareté des transformations évolutives, vis-à-vis de la diversité taxinomique et du temps de l'évolution (de plus de trois milliards d'années à quelques générations, selon le matériel étudié) ? En réalité, cette rareté du changement ne peut être évaluée correctement que par des analyses phylogénétiques concrètes et par des calibrations géologiques ou historiques. En fonction de ce qui est connu – même à grands traits – de l'évolution biologique, il est difficile de quantifier d'une façon universelle cette rareté des changements.

— Le deuxième aspect concerne une réflexion sur la nature des caractères et de leurs changements. Dans la perspective cladistique, seuls les caractères qui changent (présents sous différents états) sont pertinents pour l'analyse. L'hypothèse est faite que tous ces caractères se valent *a priori* et que leur changement ont *a priori* le même poids. Cette utilisation des caractères d'abord,

leur gestion ensuite sont une particularité, souvent discutée, de la méthode cladistique.

En ce qui concerne les *méthodes probabilistes*, dont le domaine d'application encore marginal reste celui de la biologie moléculaire, leur spécificité tient à ce que les processus évolutifs, en particulier les transformations des caractères, y sont exprimés en termes probabilistes, dans le cadre d'un certain modèle défini *a priori.* L'arbre recherché est alors l'arbre le plus probable compte tenu à la fois des données observées et du modèle probabiliste d'évolution qui est retenu.

L'utilisation de méthodes probabilistes oblige à émettre de façon explicite des hypothèses *a priori* sur les probabilités de transformations des caractères. A ce niveau, il faut insister sur le fait qu'il n'est pas possible de formuler de telles hypothèses en dehors de toute considération phylogénétique préalable. Il est clair que ces méthodes probabilistes ne peuvent s'appliquer à n'importe quel caractère. En effet, si on peut raisonnablement affecter une probabilité à la mutation d'une base en une autre au niveau de l'ADN, il devient en revanche peu justifié ou très acrobatique d'affecter une probabilité à l'apparition de la bipédie par exemple. Il semble donc bien que la multitude des caractères morphologiques, à la différence de la plupart des caractères moléculaires, se révèlera longtemps impropre à toute hypothèse sur leur probabilité de changement. De plus, l'information morphologique n'est pas stéréotypée. Il peut se faire qu'au cours de l'histoire, un caractère morphologique se transforme dix fois sans passer deux fois par le même état. Un tel cas est impossible pour ce qui est des changements de nucléotides.

Par ailleurs, cette méthode prend en considération tous les caractères, même ceux qui ne changent pas, contrairement à la méthode cladistique; elle ne permet de décrire les états des caractères à chaque nœud de l'arbre qu'en terme de probabilité. En cela elle diffère donc de l'approche cladistique qui attribue effectivement des états de caractères aux nœuds.

Insistons de nouveau sur le fait que la validité des méthodes probabilistes est essentiellement dépendante du degré de « réalisme » du modèle choisi pour rendre compte des transformations des caractères. En revanche, une fois ce modèle choisi et à l'intérieur de celui-ci, il devient possible de tester plusieurs hypothèses évolutives.

On a montré par ailleurs les difficultés propres à l'usage des *méthodes phénétiques* à des fins de construction phylogénétique. Parce qu'elles se fondent sur le concept de similitude globale, ces méthodes s'opposent clairement aux méthodes cladistiques et probabilistes. On a vu que les constructions phénétiques sont intelligibles en terme d'arbre phylogénétique à la condition de tenir, *a priori,* l'homoplasie pour négligeable. Elles ne permettent pas de localiser sur l'arbre les états de caractères homoplasiques ni les homologies, que ce soit de manière certaine ou en probabilité. Les méthodes phénétiques considèrent que le taxon, en tant qu'ensemble indissociable de caractères, est l'unité de l'évolution. Elles ne s'intéressent donc pas aux caractères en tant que tels, à l'inverse des approches cladistique et probabiliste.

Toutes les méthodes de reconstruction souffrent d'un sérieux handicap : il est en effet impossible de refaire l'histoire et de confronter la réalité historique aux résultats obtenus par une méthode quelconque. Cependant il est possible de contourner ce problème de deux façons :

— La première revient à effectuer des simulations d'évolution d'espèces sous différents modèles, afin de bien préciser les conditions d'évolution dans lesquelles telle ou telle méthode redonne bien, et avec quelles incertitudes ou quelle robustesse, l'arbre véritable simulé. Les travaux sur ce thème sont extrêmement nombreux (Astolfi *et al.,* 1981 ; Tateno *et al.,* 1982 ; Nei *et al.*, 1983 ; Saitou et Imanishi, 1989).

— La deuxième consiste à effectuer une véritable expérimentation phylogénétique « en laboratoire » C'est ce qu'ont tenté Hillis *et al.* (1992). L'objet de leur étude est l'évolution du bactériophage T7 en présence d'un agent mutagène. La destinée du virus en laboratoire a été dirigée ; les lignées étaient divisées à des intervalles préétablis : les ancêtres et les dichotomies étaient connus. Une phylogénie de huit taxons terminaux a ainsi été créée à partir d'un ancêtre commun. Les cartes de sites de restriction de l'ADN des taxons ont été analysées afin d'inférer une hypothèse phylogénétique et la comparer à l'histoire connue.

Le résultat de ces analyses montre que les méthodes de reconstruction phylogénétique ne sont pas de pures spéculations et ont quelque lien avec la réalité. Toutes les méthodes testées (4 méthodes de distances et une méthode de parcimonie) ont donné l'arbre correct. Aucune n'a donné les véritables longueurs de branches mais la corrélation entre les longueurs véritables et estimées va de 0,91 (parcimonie) à 0,82 (UPGMA). Comme on le sait, la méthode de parcimonie a la particularité d'inférer les états des caractères aux nœuds, c'est-à-dire les caractères des ancêtres. Or l'analyse de parcimonie a estimé correctement 97,3% des états, 1,4% des estimations étant ambiguës et 1,3% fausses.

Selon Hillis *et al.*, cette étude légitime l'emploi des méthodes de reconstruction phylogénétique et illustre la puissance de résolution de l'approche de parcimonie. Il convient cependant d'être prudent. En effet il est probable que le mode d'évolution du phage T7 dans cette expérience est justement celui qui est requis pour que toutes les méthodes donnent le même résultat et, qui plus est, le bon. En outre, il faut se garder de généraliser à l'ensemble des êtres vivants les résultats obtenus à partir d'un seul exemple. En particulier les vitesses d'évolution peuvent être très variables selon les branches et l'homoplasie répartie de manière non aléatoire.

Cet exemple grandeur nature, si l'on ose dire, est néanmoins un moyen de tempérer un certain pessimisme qui a pu naître à la suite de la comparaison des méthodes présentées dans ce livre. En effet, la figure V.26 conçue comme la phylogénie vraie, a servi de test aux différentes approches et les résultats ont été discordants. La structure de l'arbre vrai est en effet telle qu'elle met en échec toutes les méthodes. Seule la méthode cladistique a fourni le bon arbre mais cet arbre était l'un des deux obtenus par parcimonie (figure V.27). On a vu que le choix entre plusieurs arbres parcimonieux repose sur différents critères. L'un de ces critères est la pondération successive (paragraphe V.4.3.) : l'application de ce critère amène à retenir le mauvais arbre (tableau V.11). Les méthodes phénétiques

ont fourni des résultats erronés (figures VII.5 et 11) ainsi que l'analyse de compatibilité (figure VI.4). Les raisons de ces échecs éclairent les principes inhérents aux présupposés méthodologiques des différentes approches. L'homoplasie, la longueur dissymétrique des branches des groupes frères (les vitesses d'évolution inégales) sont responsables des erreurs. Quant aux méthodes probabilistes, elles sont inopérantes dans l'état actuel des possibilités informatiques, pour résoudre un problème phylogénétique à 14 taxons.

La construction d'un arbre phylogénétique n'est pas autre chose qu'une recherche de la meilleure interprétation possible de la matrice de caractères. Mais l'arbre phylogénétique lui-même ne peut pas être meilleur que la matrice des caractères qu'il est censé interpréter. C'est pourquoi il faut insister sur l'importance de l'étape initiale de la recherche phylogénétique : l'identification des caractères, qu'ils soient morphologiques ou bien moléculaires.

La recherche phylogénétique fondée sur les caractères morphologiques a déjà une longue histoire. Près de deux siècles de connaissance anatomique à finalité phylogénétique ont donné aux observations une signification qui dépasse la simple distribution taxinomique : dimensions fonctionnelle et adaptative, ontogénique et chronologique. Ce savoir biologique n'a pourtant pas éliminé l'écueil que constitue le phénomène d'homoplasie. En outre, il reste beaucoup à faire en matière de compréhension de la morphologie. Même des groupes aussi familiers que les mammifères, recèlent toujours des questions morphologiques non résolues qui posent un défi aux phylogénéticiens.

Par ailleurs, les recherches en biologie moléculaire connaissent un essor remarquable. Mais l'étude des caractères moléculaires est encore dans l'enfance. On sait que la structure tridimensionnelle des molécules n'est pas sans influence sur la mutabilité des sites. Par ailleurs, la structure et l'organisation du génome offre un champ de recherches à venir qui affecteront certainement notre compréhension de ce que sont les caractères moléculaires et en conséquence les constructions d'arbres. De plus, les analyses phylogénétiques sont tributaires de méthodes d'alignement de séquences, qui sont sans aucun doute encore perfectibles.

C'est sur leur capacité à rendre compte de l'évolution des caractères que seront jugées les améliorations futures des méthodes d'analyse. En manière de conclusion sur ce point, nous emprunterons des propos déjà anciens – et toujours d'actualité – tenus par Walter Fitch : « le futur nous apportera des méthodes dont la puissance de résolution phylogénétique sera supérieure à ce que l'on possède aujourd'hui ; cela se fera en comprenant mieux les caractères et en utilisant des méthodes visant à cette compréhension » (Fitch, 1984).

A la fin de cet ouvrage d'apparence technique, la finalité de la construction phylogénétique ne doit cependant pas être perdue de vue. Loin d'être une gymnastique spéculative de l'esprit, elle est au contraire une étape obligée dans la compréhension des mécanismes évolutifs qui ont conduit à la diversité – actuelle et fossile – du monde vivant. Il est impossible de comprendre ou de localiser des changements de vitesse d'évolution sans une phylogénie préalable, impossible également de trancher la question de la neutralité des gènes ou du rôle de la

sélection si l'on n'a pas une image aussi claire que possible de la parenté entre les molécules de différents génomes. Impossible encore de préciser l'organisation du génome sans une histoire évolutive des éléments qui le constituent.

Il faut cependant aller plus loin et reconnaître que les reconstructions phylogénétiques doivent s'enrichir et intégrer toutes les informations concernant les mécanismes évolutifs eux-mêmes. En d'autres termes – ceux de *pattern* (structure) et de *process* (processus) (Eldredge et Cracraft, 1980) – il faut admettre que la reconstruction phylogénétique, en tant que démarche heuristique d'analyse des caractères, constitue ce que l'on peut appeler une analyse de *pattern.* Elle alimente les inférences que l'on peut effectuer sur les mécanismes de l'évolution, c'est-à-dire sur l'analyse des processus. Sans analyse préalable de la structure, pas d'inférence possible sur l'arbre évolutif, pas de conclusion possible sur les processus. Mais également, pas d'analyse phylogénétique sans considérations, à un niveau ou à un autre, sur les processus. C'est bien en combinant ces deux démarches que l'on peut espérer comprendre l'histoire du monde vivant, notre Histoire.

RÉFÉRENCES BIBLIOGRAPHIQUES

Adams, E. N. I, 1972. Consensus techniques and the comparison of taxonomic trees. *Syst. Zool.,* 21: 390-397.

Agassiz, J. L. R, 1859. *An essay on classification.* Longman (London).

Alberch, P., Gould, S. J., Oster, G. F. et Wake, D. B., 1979. Size and shape in ontogeny and phylogeny. *Paleobiology,* 5: 296-317.

Allard, M. W., 1990. Further comments on Goodman's Maximum Parsimony Procedure. *Cladistics,* 6(3): 283-290.

André, H. M., 1988. Age-dependent evolution : from theory to practice. In : Humphries, C. (Ed.), *Ontogeny and systematics.* British Museum (Natural History) (Londres). pp. 137-187.

Archie, J. W., 1989a. A randomization test for phylogenetic information in systematic data. *Syst. Zool.,* 38(3): 239-252.

Archie, J. W., 1990. Homoplasy excess statistics and retention indices : a reply to Farris. *Syst. Zool.,* 39(2): 169-174.

Archie, J. W., 1989b. Homoplasy excess ratios : new indices for measuring levels of homoplasy in phylogenetic systematics and a critique of the consistency index. *Syst. Zool.,* 38(3): 253-269.

Astolfi, P., Kidd, K. K. et Cavalli-Sforza, L. L., 1981. A comparison of methods for reconstructing evolutionary trees. *Syst. Zool.,* 30(2): 156-169.

Astolfi, P., Piazza, A. et Kidd, K. K., 1978. Testing of evolutionary independence in simulated phylogenetic trees. *Syst. Zool.,* 27(4): 391-400.

Ax, P., 1984. *Das Phylogenetische System.* Gustav Fischer (Stuttgart).

Baer, K. E. von, 1828. *Uber Entwicklungsgeschichte der Thiere. Beobachtung und Reflexion.* Bornträger (Königsberg).

Balakrishnan, V. et Sanghvi, L. D., 1968. Distance between populations on the basis of attribute data. *Biometrics,* 24: 859-865.

Baroin, A., Perasso, R., Su, L. H., Burgerolle, G., Bachellerie, J. P. et Adoutte A. 1988. Partial phylogeny of the unicellular eucaryotes based on rapid sequencing of a portion of 28S ribosomal RNA. *Proc. Natl. Acad. Sci., USA,* 85 : 3474-3478.

Barriel, V., 1991. Caractères ostéologiques et odontologiques chez les Hominoidea. Essai de parcimonie. *Bull. et Mém. Soc. Anthrop. Paris,* 3(1-2): 45-72.

Barriel, V. et Darlu P., 1990. Approche moléculaire de la phylogénie des Hominoidea. L'exemple de la pseudo êta-globine. *Bull. et Mém. Soc. Anthrop. Paris.,* 2(1): 3-24.

Barthélemy, J. P .et Guénoche, A., 1988. *Les arbres et les représentations des proximités*. Masson (Paris).

Bonde, N., 1981. Problems of species concepts in paleontology. In : Martinell J. (Ed.), *Concept and method in paleontology*. Universitat de Barcelona (Barcelona). pp. 19-34.

Bonde, N., 1984. Primitive features and ontogeny in phylogenetic reconstructions. *Vidensk. Meddr. Dansk. Naturh. Foren,* 145: 219-236.

Bowcock, A. M., Kidd, J. R, Mountain, J. L., Hebert, J. N., Carotenuto, L., Kidd K. K. et Cavalli-Sforza, L. L., 1991. Drift, admixture and selection in human evolution : a study with DNA polymorphism. *Proc. Natl. Acad. Sci.,* 88(3): 839-843.

Brown, N. M., Prager, E. M., Wang, A. et Wilson, A. C., 1982. Mitochondrial DNA sequences of primates : tempo and mode of evolution. *J. Mol. Evol.,* 18: 225-239.

Buckup, P.A. et Dyer, B.S., 1991. Transformation series analysis (TSA) is dependent on initial order of character states. *Syst. Zool.,* 40(4): 500-502.

Bulmer, M., 1991. Use of the method of generalized least squares in reconstructing phylogenies from sequence data. *Mol. Biol. Evol.,* 8(6): 868-883.

Cain, A.J. et Harrison, G.A., 1960. Phyletic weighting. *Proc. zool. Soc. of London,* 135: 1-31.

Camin, J. H. et Sokal, R. R., 1965. A method for deducing branching sequences in phylogeny. *Evolution,* 19: 311-326.

Candolle, A. P. de, 1813. *Théorie élémentaire de la Botanique ou exposition du principe de classification élémentaire naturelle et de l'art de décrire et d'étudier les végétaux*. Déterville (Paris).

Cavalli-Sforza, L. L. et Edwards, A. W. F, 1966. Estimation procedures for evolutionary branching processes. *Bull. Inst. Internat. Statist.,* 41: 803-808.

Cavalli-Sforza, L. L. et Edwards, A. W. F, 1967. Phylogenetic analysis: models and estimation procedures. *Am. J. Hum. Genet.,* 19: 233-257.

Cavalli-Sforza, L. L. et Piazza, A., 1975. Analysis of evolution: Evolutionary rates, independence and treeness. *Theor. Pop. Biol.,* 8(2): 127-165.

Cavender, J. A., 1978. Taxonomy with confidence. *Math. Biosci.,* 40: 271-280.

Cavender, J. A., 1981. Tests of phylogenetic hypotheses under generalized models. *Math. Biosci.,* 54: 217-229.

Cavender, J. A., 1989. Mechanized derivation of linear invariants. *Mol. Biol. Evol.,* 6(3): 301-316.

Cavender, J. A. et Felsenstein, J., 1987. Invariants of phylogenies in a simple case with discrete states. *J. of Classif.,* 4: 57-71.

Czecanowski, J., 1909. Zur Differentialdiagnose der Neandertalgruppe. *Korrespondenzblatt Deutsch. Ges. Anthropol. Ethnol. Urgesch,* 40: 44-47.

Darlu P., 1992. Are parsimony and compatibility methods relevant to inter language evolution ?. In : Piazza, A. et Cavalli-Sforza, L. L. (Eds.), *Language change and biological evolution.* Stanford University Press (sous presse) (Stanford).

Darlu, P. et Lathrop, G. M., 1993. Estimation of admixture in evolutionary trees. *J. of Evol. Biol. (sous presse),*

Darlu, P.; Ruhlen, M. et Cavalli-Sforza, L. L., 1990. A taxonomic analysis of linguistic families. In : Wang N.S.Y. (Ed.), *Language change and linguistic evolution.* (London).

Darwin, C., 1859. *On the origin of species.* John Murray (London).

Darwin, C., 1872 (6e édition). *On the origin of species.* John Murray (London).

Dayhoff, M. O. (Ed), 1969. *Atlas of Protein Sequence and Structure.* Md. : Natl. Biomed. Res. Found. 5 (Silver Springs).

De Beer G., 1954. *Archaeopteryx lithographica.* British Museum Natural History (Londres).

De Beer, G., 1930. *Embryology and evolution.* Clarendon Press (Oxford).

De Beer, G., 1958 (3e édition). *Embryos and ancestors.* Clarendon Press (Oxford).

d'Udekem-Gevers, M., 1990. *L'analyse cladistique : problème et solutions heuristiques informatisées.* Biosystema 4, Société Française de Systématique, *(*Paris) .

Duncan, T. et Stuessy, T. F. (Eds), 1984. *Cladistics : perspectives on the reconstruction of evolutionary history.* Columbia University Press (New-York).

Dupuis, C., 1988. Le taxinomiste face aux catégories. *Cahiers des Naturalistes,* 44: 49-109.

Edwards, A. N. F, 1972. *Likelihood.* Cambridge University Press (Cambridge).

Edwards, A. W. F et Cavalli-Sforza, L. L., 1963. The reconstruction of evolution. *Ann. Hum. Genet.,* 27: 104-105.

Edwards, A. W. F et Cavalli-Sforza, L. L., 1964. Reconstruction of evolutionary trees. *Systematics Association Publication,* 6: 67-76.

Efron, B., 1979. Bootstrap methods : an other look at the jackknife. *Ann. Statist.,* 7: 1-26.

Efron B., 1982. *The Jackknife, the Bootstrap and other resampling plans.* Society for industrial and applied mathematics (Philadelphie)

Eldredge, N. et Cracraft, J., 1980. *Phylogenetic patterns and the evolutionary process.* Columbia University Press (New-York).

Estabrook, G. F., 1972. Cladistic methodology: a discussion of the theoretical basis for the induction of evolutionary history. *Ann. Rev. Ecol. Syst.,* 3: 427-456.

Estabrook, G. F., Johnson, C. S. Jr et McMorris, F. R., 1976. A mathematical foundation for the analysis of cladistic character compatibility. *Math. Biosci.,* 29: 181-187.

Estabrook, G. F., Strauch, J. G. et Fiala, J. K., 1977. An application of compatibility analysis to the Blackiths' data on orthopteroid insects. *Syst. Zool.,* 26: 269-276.

Faith, D. P., 1985. Distance methods and approximation of most-parsimonious trees. *Syst. Zool.,* 34(3): 312-325.

Farris, J. S., 1966. Estimation of conservation of characters by constancy within biological populations. *Evolution,* 20: 587-591.

Farris, J. S., 1967. The meaning of relationship and taxonomic procedure. *Syst. Zool.,* 16: 44-51.

Farris, J. S., 1969. A successive approach to character weighting. *Syst. Zool.,* 18: 374-385.

Farris, J. S., 1970. Methods for computing Wagner trees. *Syst. Zool.,* 19: 83-92.

Farris, J. S., 1972. Estimating phylogenetic trees from distance matrices. *Am. Nat.,* 106: 645-668.

Farris, J. S., 1977a. Phylogenetic analysis under Dollo's law. *Syst. Zool.,* 26: 78-88.

Farris, J. S., 1977b. On the phenetic approach to vertebrate classification. In : Hecht M.K. , Goody, P. C., Hecht B. M. (Eds.), *Major Patterns in Vertebrate Evolution.* Plenum Press (New York). pp. 823-850.

Farris, J. S., 1981. Distance data in phylogenetic analysis. In : Funk V.A. ; Brooks D.R. (Eds.), *Advances in Cladistics.* The New York Botanical Garden (Bronx, New York). pp. 3-23.

Farris, J.S., 1982. Outgroups and parsimony. *Syst. Zool.,* 31: 328-334.

Farris, J. S., 1983. The logical basis of phylogenetic analysis. In : N.I. Platnick et V.A. Funck (Eds.), *Advances in cladistics, Vol. 2.* Columbia University Press (New York). pp. 7-36.

Farris, J. S., 1985. Distance data revisited. *Cladistics,* 1(1): 67-85.

Farris, J. S., 1986. Distances and statistics. *Cladistics,* 2(2): 144-157.

Farris, J. S., 1988. Hennig86, version 1.5, user's manual. Published by the author.

Farris, J. S., 1989a. The retention index and homoplasy excess. *Syst. Zool.,* 38(4): 406-407.

Farris, J. S., 1989b. The retention index and the rescaled consistency index. *Cladistics,* 5(4): 417-419.

Farris, J. S., 1991. Excess homoplasy ratio. *Cladistics,* 7: 81-91.

Felsenstein, J., 1973. Maximum likelihood estimation of evolutionary trees from continuous characters. *Am. J. Hum. Genet.,* 25: 471-492.

Felsenstein, J., 1978a. The number of evolutionary trees. *Syst. Zool.,* 27: 27-33.

Felsenstein, J., 1978b. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.,* 27: 401-410.

Felsenstein, J., 1981a. A likelihood approach to character weighting and what it tells us about parsimony and compatibility. *Biol. J. of the Linn. Soc.,* 16: 183-196.

Felsenstein, J., 1981b. Evolutionary trees from gene frequencies and quantitative characters: finding maximum likelihood estimates. *Evolution,* 35(6): 1229-1242.

Felsenstein, J., 1983a. Statistical inference of phylogenies. *J. R. Statist. Soc. A,* 146(3): 246-272.

Felsenstein, J., 1983b. Inferring evolutionary trees from DNA sequences. In : Weir B.S. (Ed.), *Statistical analysis of DNA sequence data.* Dekker (New York). pp. 133-150.

Felsenstein, J., 1984a. Distance methods for inferring phylogenies : a justification. *Evolution,* 38: 16-24.

Felsenstein, J., 1984b. The statistical approach to inferring evolutionary trees and what it tells us about parsimony and compatibility. In : Duncan, T. et Stuessy, T. F. (Eds.), *Cladistics : perspectives on the reconstruction of evolutionary history.* Columbia University Press (New York). pp. 169-191.

Felsenstein, J., 1985a. Confidence limits on phylogenies with a molecular clock. *Syst. Zool.,* 34(2): 152-161.

Felsenstein, J., 1985b. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution,* 39: 783-791.

Felsenstein, J., 1986. Distance methods : a reply to Farris. *Cladistics,* 2(2): 130-143.

Felsenstein, J., 1987. Estimation of hominoid phylogeny from a DNA hybridization data set. *J. Mol. Evol.,* 26: 123-131.

Felsenstein, J., 1988. Phylogenies from molecular sequences : inference and reliability. *Ann. Rev. of Genet.,* 22: 521-565.

Felsenstein, J., 1990. *Phylogeny inference package. Version 3.3,*. Department of Genetics, University of Washington (Seattle).

Fitch, W.M., 1970. Toward defining the course of evolution : minimum change for a specific tree topology. *Syst. Zool.,* 20(406-416)

Fitch, W.M., 1971. Distinguishing homologous from analogous proteins. *Syst. Zool.,* 20: 406-416.

Fitch, W. M., 1975. Toward finding the tree of maximum parsimony. In : Estabrook G.F. (Ed.), *Proc. Eighth Int. Conf. on Numerical Taxonomy.* Freeman (San Francisco). pp. 189-230.

Fitch, W. M., 1976. Molecular evolutionary clocks. In : Ayala F.J. (Ed.), *Molecular evolution.* Sinauer Ass. Inc. Publishers, (Sunderland, Massachusetts).pp. 160-178.

Fitch, W. M., 1984. Cladistic and other methods : problems, pitfalls, and potentials. In : Duncan, T. et Stuessy, T. F. (Eds.), *Cladistics : perspectives on the reconstruction of evolutionary history.* Columbia University Press (New York). pp. 221-252.

Fitch, W. M. et Margoliash, E., 1967. Construction of phylogenetic trees. *Science,* 155: 279-284.

Forster, M. R., 1986. Statistical covariance as a measure of phylogenetic relationship. *Cladistics,* 2(4): 297-317.

Fort, P., 1982. *Variabilité de l'extrémité 5' du RNA ribosomal mitochondrial 16S dans le genre* Mus. *Modes d'évolution différents des génomes nucléaire et mitochondrial.* Thèse de 3° cycle, U.S.T.L. (Montpellier, France).

Fort, P., Bonhomme, F., Darlu, P., Piachaczyk, M., Jeanteur, P. et Thaler, L., 1984. Clonal divergence of mitochondrial DNA versus populational evolution of nuclear genome. *Evolutionary Theory.*

Gaudry, A., 1866. *Considérations générales sur les animaux fossiles de Pikermi.* F. Savy (Paris).

Gingerich, P. D., 1979. The stratophenetic approach to phylogeny reconstruction in vertebrale paleontology. In : Cracraft J. et Eldredge N.(Eds.), *Phylogenetic analysis and paleontology.* Columbia University Press (New York). pp. 41-77.

Gojobori, T., Ishii, K. et Nei, M., 1982. Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotide. *J. Mol. Evol.,* 18: 414-423.

Goldman, N., 1990. Maximum likelihood inference of phylogenetic trees, with special reference to a Poisson process model of DNA substitution and to parsimony analyses. *Syst. Zool.,* 39(4): 345-361.

Goodman, M. M., 1969. Measuring evolutionary divergence. *Jap. J. Gen.,* 44(1)310-316

Goodman, M. M., 1989. Emerging alliance of phylogenetic systematics and molecular biology : a new age of exploration. In : Feinholm B., Bremer K.et Jörnvall H. (Eds.), *The hierarchy of life.* Nobel symposium 70, Excerpta Medica (Amsterdam). pp. 43-61.

Goodman, M. M., 1990. Response to remarks by Allard (1989). Concerning Kimura's "Damning" criticism of Goodman. *Cladistics,* 6(2): 195-196.

Goodman, M. M., Czelusniak, J., Moore, G. W., Romero-Herrera, A. E. et Matsuda, G., 1978. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst. Zool.,* 28: 132-163.

Gould, S. J., 1977. *Ontogeny and phylogeny.* Belknap Press of Harvard University Press, Cambridge

Gregorius, H. R., 1978. The concept of genetic diversity and its formal relationship to heterozygosity and genetic distances. *Math. Biosci.,* 41: 253-271.

Haeckel, E., 1866. *Generelle Morphologie der Organismen.* Georg Reimer (Berlin).

Haeckel, E., 1877. *Anthropogénie.* Reinwald C. et Cie. (Paris).

Hendy, M. D. et Penny, D., 1982. Branch and bound algorithms to determine minimal evolutionary trees. *Math. Biosci.,* 59: 277-290.

Hendy, M. D. et Penny, D., 1989. A framework for the quantitative study of evolutionary trees. *Syst. Zool.,* 38(4): 297-309.

Hennig, W., 1950. *Grundzüge einer Theorie der Phylogenetischen Systematik.* Deutscher Zentralverlag (Berlin).

Hennig, W., 1966. *Phylogenetic Systematics.* University of Illinois Press (Urbana).

Hennig, W., 1969. *Die Stammesgeschichte der Insekten.* Kramer (Frankfurt).

Hennig, W., 1981. *Insect Phylogeny.* John Wiley and Son (New York).

Holmquist, R., Miyamoto, M. M. et Goodman, M., 1988. Analysis of higher-primate phylogeny from transversion differences in nuclear and mitochondrial DNA by Lake's methods of evolutionary parsimony and operator metrics. *Mol. Biol. Evol.,* 5(3): 217-236.

Jaccard, P., 1908. Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Mat.,* 44: 223-270.

Jacquard, A., 1973. Distances généalogiques et distances génétiques. *Cah. Anthrop. Ecol. hum.,* 1: 11-85.

Jorde, L. B., 1985. Human genetic distance studies: present status and future prospects. *Ann. Rev. Anthropol.,* 14: 343-373.

Jukes, T. H.et Cantor, C. R., 1969. Evolution of protein molecules. In : Munro H.N. (Ed.), *Mammalian Protein Metabolism.* Academic Press (New-York). pp. 21-132.

Kidd, K. K. et Sgaramella-Zonta, L. A., 1971. Phylogenetic analysis : concepts and methods. *Am. J. of Hum. Gen.,* 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., 1981a. Was globin evolution very rapid in its early stages ? A dubious case against the rate-constancy hypothesis. *J. Mol. Evol.,* 17: 110-113.

Kimura, M., 1981b. Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Acad. Sci. USA,* 78(1): 454-458.

Kimura, M. et Ohta, T., 1972. On the stochastic model for estimation of mutational distance between homologous proteins. *J. Mol. Evol.,* 2: 87-90.

Kluge, A. G., 1985. Ontogeny and phylogenetic systematics. *Cladistics,* 1: 13-27.

Kluge, A. G. et Farris, J. S., 1969. Quantitative phyletics and the evolution of anurans. *Syst. Zool.,* 18(1): 1-32.

Kraus, F., 1988. An empirical evaluation of the use of the ontogeny polarization criterion in phylogenetic inference. *Syst. Zool.,* 37: 106-141.

Lake, J. A., 1987a. A rate-independent technique for analysis of nucleic acid sequences : evolutionary parsimony. *Mol. Biol. Evol.,* 4(2): 167-191.

Lake, J. A., 1987b. Determining evolutionary distances from highly diverged nucleic acid Sequences : operator netrics. *J. Mol. Evol.,* 26(1-2): 59-73.

Lalouel, J. M., 1980. Distance analysis and multidimensional scaling. In: : Mielke J. H. et Crawford M. H. (Eds.), *Current developments in Anthropological genetics. Vol. 1 : Theory and methods.* Plenum Press (New York). pp. 209-250.

Lam, H. J., 1950. Proposal to indicate a taxonomic group of any rank with the term taxon (plural taxa). In : J. Lanjouw (Ed.), *Botanical nomenclature and taxonomy.* Union inter. Sc. biol., Colloquia, Ser. B., vol. 2 (Paris). pp. 1-88.

Lamarck, J. B. Monet de, 1809. *Philosophie zoologique.* Dentu (Paris).

Lance, G. N. et Williams, W. T., 1967. A general theory of classificatory sorting strategies; I. Hierarchical systems. *Computer J.,* 9: 373-380.

Lankester, E. R., 1870. On the use of the term homology in modern zoology, and the distinction between homogenetic and homoplastic agreements. *Ann. Mag. Nat. Hist.,* 4(6): 34-43.

Lanyon, S. M., 1985. Detecting internal inconsistencies in distance data. *Syst. Zool.,* 34(4): 397-403.

Lathrop, G. M., 1982. Evolutionary trees and admixture: phylogenetic inference when some populations are hybridized. *Ann. Hum. Genet.,* 46: 245-255.

Lee, Y. M., Friedman, D. J. et Ayala, F. J., 1985. Superoxyde dismutase: an evolutionary puzzle. *Proc. Natl. Acad. Sci. USA,* 82: 824-828.

LeQuesne, W. J., 1969. A method of selection of characters in numerical taxonomy. *Syst. Zool.,* 18(2): 201-205.

LeQuesne, W. J., 1972. Further studies based on the uniquely derived character concept. *Syst. Zool.,* 21(3): 281-288.

Lewin, R., 1987. When does homology mean something else? *Science,* 237: 1570.

Li, W. H., 1989. A statistical test of phylogenies estimated from sequence data. *Mol. Biol. Evol.,* 6(4): 424-435.

Lipscomb, DL, 1989. Relationships among the eukaryotes. In : Fernholm B. , Bremer K.et Jörnvall H. (Eds.), *The hierarchy of life.* Nobel Symposium 70, Excerpta Medica, (Amsterdam). pp. 161-178.

Lipscomb, D. L., 1990. Two methods for calculating characters : Transformation Series Analysis and the iterative FIG/FOG method. *Syst. Zool.,* 39: 277-288.

Mabee, P. M., 1989. An empirical rejection of the ontogenetic polarity criterion. *Cladistics,* 5: 409-416.

Maddison, W. P., Donoghue, M. H. et Maddison, D. R., 1984. Outgroup analysis and parsimony. *Syst. Zool.,* 33: 83-103.

Maeda, N., Wu, C. I., Bliska, J. et Reneke, J., 1988. Molecular evolution of intergenic DNA in higher primates : pattern of DNA changes, molecular clock and evolution of repetitive sequences. *Mol. Biol. Evol.,* 1: 1-20.

Mahalanobis, P. C., 1936. On the generalized distance in statistics. *Proc. Natl. Inst. Sci. India.,* 2: 49-55.

Maslin, P. T., 1952. Morphological criteria of phyletic relationships. *Syst. Zool.,* 1: 49-70.

Matile, L., Tassy, P. et Goujet, D., 1987. *Introduction à la systématique zoologique*. Biosystema 1*,* Société française de systématique (Paris).

Mayr, E., 1965. Classification and phylogeny. *Amer. zool.,* 5: 165-174.

Mayr, E., 1969. *Principles of systematic zoology.* McGraw Hill (New York).

Mayr, E., 1981. Biological classification : toward a synthesis of opposing methodologies. *Science,* 214: 510-516.

Mayr, E., 1986. La systématique évolutionniste et les quatres étapes du processus de classification. In : Tassy, P. (Ed.), *L'ordre et la diversité du vivant.* Fayard, Fondation Diderot (Paris). pp. 143-160.

Mayr, E., 1988. The limits of reductionism. *Nature,* 331: 475.

Mayr, E., Linsley, E. G. et Usinger, R., 1953. *Methods and principles of systematic zoology.* McGraw-Hill (New York).

Meacham, C. A., 1984. The role of hypothesized direction of characters in the estimation of evolutionnary history. *Taxon,* 33(1): 26-38.

Michener, C. D. et Sokal, R. R., 1957. A quantitative approach to a problem in classification. *Evolution,* 11: 130-162.

Mickevich, M. F., 1982. Transformation Series Analysis. *Syst. Zool.,* 31(4): 461-478.

Mickevich, M. F. et Johnson, M. S., 1976. Congruence between morphological and allozyme data in evolutionary inference and character evolution. *Syst. Zool.,* 25: 260-270.

Mickevich, M. F. et Lipscomb, D. L., 1991. Parsimony and the choice between different transformations for the same character set. *Cladistics,* 7: 111-139.

Mickevich, M. F. et Mitter, C., 1981. Treating polymorphic characters in systematics: a phylogenetic treatment of electrophoretic data. In : Funk V.A. et Brooks D.R. (Eds.), *Advances in Cladistics.* The New York Botanical Garden (Bronx, New York). pp. 45-58.

Mickevich, M. F. et Mitter, C., 1983. Evolutionary patterns in allozyme data : a systematic approach. In : Platnick N. I. et Funk V. A. (Eds.), *Advances in cladistics, vol. 2.* Columbia University Press (New-York). pp. 169-189.

Mickevich, M. F. et Weller, S. J., 1990. Evolutionary Character analysis : tracing character change on a cladogram. *Cladistics,* 6: 137-170.

Mitchell, P. C., 1901. On the intestinal tract of birds with remarks on the valuation and nomenclature of zoological characters. *Trans. Linnean Soc. London, Zool.,* 2(8): 173-275.

Miyamoto, M.M., et Goodman M., 1986. Biomolecular systematics of eutherian mammals : phylogenetic patterns and classification. *Syst. Zool.*, 35:230-240.

Miyamoto, M. M. et Slightom, J. L. et Goodman, M., 1987. Phylogenetic relationships of humans and African apes as ascertained from DNA sequences (7.1 kbp) of the ??-Globin region. *Science.,* 238: 369-373.

Moore, G. W., 1976. Proof for the maximum parsimony ("Red King") algorithm. In : Goodman M. et Tashian R.E. (Eds.), *Molecular Anthropology.* Plenum Press (New-York). pp. 117-137.

Moore, G. W., Barnabas, J. et Goodman, M., 1973. A method for constructing maximum parsimony ancestral amino acid sequences on a given network. *J. Theor. Biol,* 38: 459-485.

Mueller, L. D. et Ayala, F. J., 1982. Estimation and interpretation of genetic distance in empirical studies. *Genet. Res.,* 40: 127-137.

Nei, M., 1972. Genetic distance between populations. *Amer. Nat.,* 106: 283-292.

Nei, M., 1987. *Molecular evolutionary genetics.* Columbia University Press (New-York).

Nei, M., Stephens, C. et Saitou, N., 1985. Methods for computing the standard errors of branching points in an evolutionary tree and their application to molecular data from humans and apes. *Mol. Biol. Evol.,* 2(1): 66-85.

Nei, M., Tajima, F. et Tateno, Y., 1983. Accuracy of estimated phylogenetic trees from molecular data. II Gene frequency data. *J. Mol. Evol.,* 19: 153-170.

Nelson, G., 1973 a. The higher-level phylogeny of vertebrates. *Syst. Zool.,* 22: 87-91.

Nelson, G., 1973 b. Negative gains and positive losses: a reply to J.G. Lundberg. *Syst. Zool.,* 22: 330.

Nelson, G., 1978. Ontogeny, phylogeny, paleontology and the biogenetic law. *Syst. Zool.,* 27: 324-345.

Nelson, G., 1979. Cladistics analysis and synthesis: principles and definitions with a historical note on Adanson's "Famille des Plantes" (1763-1764). *Syst. Zool.,* 28: 1-21.

Nelson, G., 1985. Outgroup and ontogeny. *Cladistics,* 1: 29-45.

Nelson, G. et Platnick, N., 1981. *Systematics and biogeography: cladistics and vicariance.* Columbia University Press (New York).

Ohayon, E. et Cambon-Thomsen, A., 1986. *Génétique des populations humaines.* Editions INSERM (Paris).

Owen, R., 1845. *Lectures on the comparative anatomy.* Longman (Londres).

Panchen, A. L., 1992. *Classification, evolution, and the nature of biology.* Cambridge University Press (Cambridge).

Patterson, C., 1982. Morphological characters and homology. In: : Joysey, K. A. et Friday A. F (Eds.), *Problems of phylogenetic reconstruction.* Academic Press (Londres). pp. 21-74.

Patterson, C., 1983. How does phylogeny differ from ontogeny ? In : Goodwin B.C. , Holder N. et Wylie C. C. (Eds.), *Development and evolution.* Cambridge University Press (Cambridge) pp. 1-31.

Patterson, C., 1987. Introduction. In : Patterson C. (Ed.), *Molecules and morphology in evolution : conflict or compromise ?* Cambridge University Press (Cambridge). pp. 1-22.

Patterson, C., 1988. Homology and molecular biology. *Mol. Biol. Evol.,* 5: 603-625.

Patton, J. C. et Avise, J. C., 1983. An empirical evaluation of qualitative hennigian analyses of protein electrophoretic data. *J. Mol. Evol.,* 19: 244-254.

Penny, D., 1982. Towards a basis for classification: the incompleteness of distance measures, incompatibility analysis and phenetic classification. *J. Theor. Biol.,* 96: 129-142.

Pogue, M. C. et Mickevich, M. F. , 1990. Character definitions and character state delineation : the bête noire of phylogenetic inference. *Cladistics,* 6: 319-361.

Rao, C. R., 1980. Diversity and dissimilarity coefficients: a unified approach. *Technical report 80-10.* Dpt. of Mathematic and Statistics, University of Pittsburgh (Pittsburgh).

Reeck, G. R., De Haën C., Teller, D. C., Doolittle, R. F., Fitch, W. M., Dickerson, R. E., Chambon, P., McLachlan, A. D., Margoliash, E. et Jukes, T. H. et Zuckerkandl, E., 1987. "Homology" in proteins and nucleir acids : a terminology muddle and a way out of it. *Cell,* 50: 667.

Renyi, A., 1966. *Calcul des probabilités.* Dunod (Paris).

Reynolds, J., Weir B.S. et Cockerham, C. C., 1983. Estimation of the coancestry coefficient : Basis for a short-term genetic distance. *Genetics,* 105: 767-779.

Ruhlen, M., 1975. *A guide to the languages of the world.* Ruhlen, M., publ. (Stanford).

Saitou, N. et Imanishi, T., 1989. Relative efficiencies of the Fitch-Margoliash, Maximum-Parsimony, Maximum-Likelihood, Minimum-Evolution and Neighbor-joining methods of Phylogenetic tree construction in obtaining the correct tree. *Mol. Biol. Evol.,* 6(5): 514-525.

Saitou, N. et Nei, M., 1987. The Neighbor-joining Method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.,* 4(4): 406-425.

Sanchez-Mazas, A. et Langaney, A., 1986. Measure and representation of the genetic similarity between populations by the percentage of isoactive genes. *Theoria,* 4: 143-154.

Sankoff, D., 1990. Designer invariants for large phylogenies. *Mol. Biol. Evol.,* 7(3): 255-269.

Sankoff, D.et Cedergren,.R. J., 1983. Simultaneous comparison of three or more sequences related by a tree. In : Sankoff D. et Kruskal B. (Eds.), *Time warps, string edits, and macromolecules : the theory and pratice of sequence comparison.* Addison-Wesley (Reading, Massachusset). pp. 253-263.

Sarich, V. M. et Wilson A.C., 1973. Generation time and genomic evolution in primates. *Science,* 179: 1144-1147.

Scherer, S., 1989. The relative-rate test of the molecular clock hypothesis : a note of caution. *Mol. Biol. Evol.,* 6(4): 436-441.

Schoch, R.M., 1986. *Phylogeny reconstruction in paleontology.* Van Nostrand Reinhold Company (New York).

Sibley, C. G. et Ahlquist, J. E., 1987. DNA hybridization evidence of hominoid phylogeny : results from expanded data set. *J. Mol. Evol.,* 20: 2-15.

Simpson, G. G., 1961. *Principles of animal taxonomy.* Columbia University Press (New York).

Smith, C. A. B, 1977. A note on genetic distance. *Ann. Hum. Genet.,* 40: 463-479.

Sneath, P. H. A et Sokal, R. R., 1973. *Numerical taxonomy.* Freeman (San Francisco)

Sober, E., 1988. *Reconstructing the past. Parsimony, evolution, and inference.* A Bradford Book, Massachusetts Institut of Technology (Cambridge).

Sober, E., 1985. A likelihood justification of parsimony. *Cladistics,* 1: 209-233.

Sokal, R. R., 1983. A phylogenetic analysis of the caminalcules. I the data base. *Syst. Zool.,* 32(2): 159-184.

Sokal, R. R. et Michener, C. D., 1958. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin,* 38: 1409-1438.

Sokal, R. R. et Rohlf, F. J., 1962. The comparison of dendrograms by objective methods. *Taxon,* 11: 33-40.

Sokal, R. R. et Sneath, P. H. A, 1963. *Principles of numerical taxonomy.* Freeman, San Francisco

Steel, M. A., Hendy, M. D. et Penny, D., 1988. Loss of information in genetic distances. *Nature,* 336: 118.

Swofford, D. L., 1985. *PAUP, Version 2.4. User's manual.* Illinois Natural History Survey (Champaign).

Swofford, D. L., 1990. *PAUP, version 3.0. User's manual.* Illinois Natural History Survey (Champaign).

Swofford, D. L., Olsen G.J., 1990. Phylogeny reconstruction. In : Hillis D.M.et Moritz C. (Eds.), *Molecular Systematics.* Sinauer Ass. (Sunderland, Massachusetts).pp.411-501.

Szalay, F. S., 1981a. Functional analysis and the practice of the phylogenetic method as refected by some mammalian studies. *Am. Zool.,* 21: 37-45.

Szalay, F. S. , 1981b. Phylogeny and the problem of adaptive significance : the case of the earliest primates. *Folia Primatol.,* 36: 157-182.

Tajima, F. et Nei, M., 1984. Estimation of evolutionary distance between nucleotide sequences. *Mol. Biol. Evol.,* 1(3): 269-285.

Takahata, N., 1989. Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics,* 122: 957-966.

Takahata, N. et Kimura, M., 1981. A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes. *Genetics,* 98: 641-657.

Tassy, P., 1982. Les principales lichotomies dans l'histoire des Proboscidea (Mammalia) : une approche phylogénétique. *Géobios, Mém. sp. 6,* : 225-245.

Tassy, P., 1986 (coord.). *L'ordre et la diversité du vivant.* Fayard-Fondation Diderot (Paris).

Tassy, P., 1988. The classification of Proboscidea: how many cladistic classifications? *Cladistics,* 4: 43-57.

Tassy, P., 1991. *L'arbre à remonter le temps.* Christian Bourgois Editeur (Paris).

Tassy, P. et Darlu, P., 1987. Les Elephantidae : nouveau regard sur les analyses de parcimonie. *Géobios,* 20: 487-494.

Tateno, Y., Nei, M. et Tajima, F., 1982. Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. *J. Mol. Evol.,* 18: 387-404.

Thompson, E. A., 1973. The method of minimum evolution. *Ann. Hum. Genet.,* 36: 333-340.

Thompson, E. A., 1975. *Human evolutionary trees.* Cambridge University Press (Cambridge).

Voorzanger, B. et Van der Steen, W. J., 1982. New perspectives on the biogenetic law ? *Syst. Zool.,* 31: 202-205.

Wagner, W. H. Jr, 1961. Problems in the classification of ferns. In: : *Recent Advances in Botany.* University of Toronto Press (Montreal). pp. 841-844.

Wagner, W. H. Jr, 1984. Applications of the concepts of groundplan divergence. In : Duncant T. et Stuessy T. F. (Eds.), *Cladistics : perspectives on the reconstruction of evolutionary history.* Columbia University. Press (New-York). pp. 95-118.

Wallace, A. G., 1856. Attempts at a natural arrangement of birds. *Ann. Mag. Nat. Hits.,* 18(2): 193-216.

Ward, S. C. et Kimbel, W. H., 1983. Subnasal alveolar morphology and the systematic position of Sivapithecus. *Amer. J. Phys. Anthr.,* 61: 157-171.

Wheeler, Q. D., 1990. Ontogeny and character phylogeny. *Cladistics,* 6: 225-268.

Wiley, E. O., 1976. The phylogeny and biogeography of fossil and recent gars (Actinopterygii : Lepisosteidae). *Misc. Publ. Mus. Natur. Hist. Univ. Kansas,* 64: 1-111.

Wiley, E. O., 1981. *Phylogenetics: the theory and practice of phylogenetic systematics.* John Wiley and sons (New York).

Wiley, E. O., Siegel-Causey, D., Brooks, D. R.et Fink, V. A., 1991. *The compleat cladist.* Sp. Publ. Lawrence, The University of Kansas Museum of Natural history.

INDEX