Académie de Montpellier

Université Montpellier II

— Sciences et Techniques du Languedoc —

Thèse

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

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

Méthodes et algorithmes pour reconstruire les arbres de l'Évolution

par

Vincent Berry

Soutenue le 18 décembre 1997 devant le Jury composé de :

M. Jean-Paul Delahaye, Professeur, Univ. Sciences et Technologies Lille, LIFL, ... Rapporteur
M. Bruno Leclerc, Maître de Conférence, E.H.E.S.S. Paris, C.A.M.S., ... Rapporteur
Mme. Tandy Warnow, Professeur, Univ. of Pennsylvania, Philadelphie, USA, ... Rapporteur
M. Pierre Darlu, Directeur de Recherche, C.N.R.S.-I.N.S.E.R.M., Paris, ... Examinateur
M. Gilles Caraux, Professeur, E.N.S.A. de Montpellier, ... Examinateur
M. Alain Guénoche, Chargé de Recherche, C.N.R.S., L.I.M., Marseille, ... Examinateur
M. Michel Habib, Professeur, Univ. Montpellier II, L.I.R.M.M., ... Examinateur
M. Olivier Gascuel, Chargé de Recherche, C.N.R.S., L.I.R.M.M., ... Directeur de Thèse

Introduction

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

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

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

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

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

Chapitre 1

La reconstruction phylogénétique

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

L'étude des phylogénies s'insère dans le domaine de la biologie comparative, aussi appelée systématique, dont le but est de **proposer une classification des organismes vivants**, afin de comprendre les causes de leur diversité. Le classement des espèces a d'abord été pratiqué sur la base d'observations morphologiques, comme la présence ou l'absence d'ailes, le nombre de pattes, etc. L'utilisation de données moléculaires remonte aux études immunologiques du début du siècle [Nut04], mais les phylogénies moléculaires ne furent vraiment acceptées que dans les années 60, suite à la publication de phylogénies, obtenues depuis des séquences d'ADN ou protéiques, possédant de "bonnes" congruences avec les classifications morphologiques [ZP62, FM67]. La comparaison des espèces, en fonction des différences constatées dans leur génome, connut alors un engouement qui n'a

pas cessé d'augmenter depuis, et ce pour une double raison. Premièrement, les séquences étant des enchaînements de constituants chimiques simples (les nucléotides `A`,`C`,`G` et `T`), la comparaison de séquences moléculaires est dépourvue d'ambiguïté, alors que les données morphologiques sont parfois sujettes à diverses interprétations rendant les observations subjectives et donc critiquables. Deuxièmement, les millions de nucléotides, qui composent le génome des organismes, constituent un réservoir de données quasiment inépuisable en comparaison des quelques centaines de caractères morphologiques qu'on utilisait précédemment pour reconstruire des phylogénies.

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

1.1 Les données traduisant l'évolution

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

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

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

1.1.1 L'alignement de séquences moléculaires

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

dispose des trois séquences

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

un alignement possible est :

		0	1	2	3	4	5	6	7	8	9
S_1	:	A	G	A	A	T	A	G	C	C	A
S_2	:	A	G	G	A	T	A	G	G	.	A
S_3	:	A	G	T	A	T	.	G	G	.	A

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

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

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

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

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

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

1.1.2 Le concept de similitude

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

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

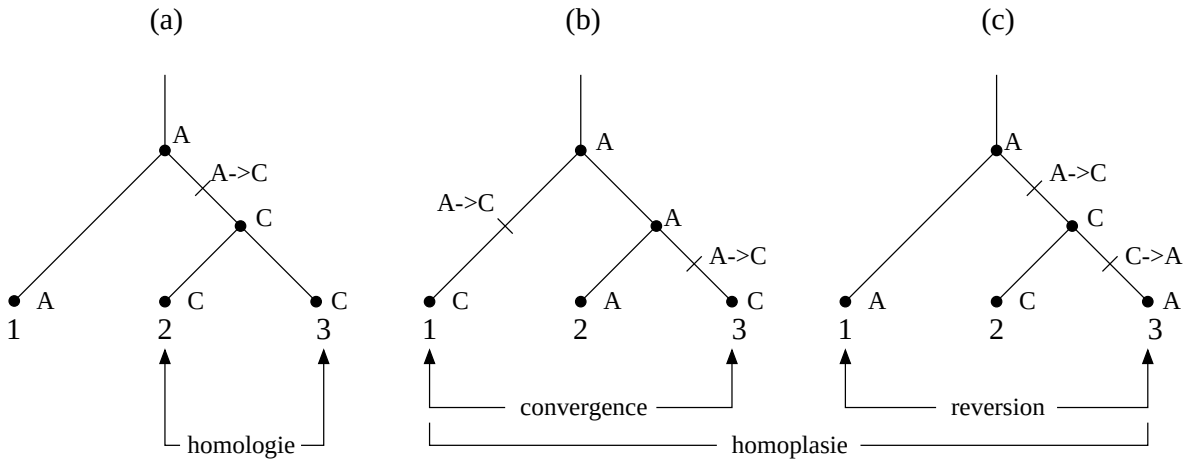

FIG. 1.1 – *Les différentes catégories de ressemblances suite à l'évolution d'un caractère : (a) ressemblance due à l'homologie, (b-c) ressemblance due à l'homoplasie suite à une convergence ou à une réversion (Fig. d'après [DT93]).*

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

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

1.2 Petite histoire des méthodes de R.P.

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

duite.

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

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

1.2.1 La phénétique

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

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

Construisant des classifications sur la base des similitudes globales, les phénéticiens avaient essentiellement recours à des **méthodes phénétiques**, *i.e.*, utilisant en entrée une matrice de similarités, ou de dissimilarités, entre les espèces

deux à deux. Des algorithmes, dits de *cluster analysis*, regroupaient en un **taxon** (une classe) les espèces les plus similaires, puis itéraient le processus jusqu'à avoir obtenue une classification hiérarchique complète des espèces étudiées. Parmi les méthodes les plus célèbres, citons la méthode du lien moyen (U.P.G.M.A.) et ses variantes.

1.2.2 Le cladisme

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

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

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

1.2.3 La place d'un modèle de l'évolution

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

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

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

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

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

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

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

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

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

La démarche ci-dessus peut sembler singulière : une méthode est posée (le maximum de parcimonie) et on étudie ensuite les hypothèses qu'elle sous-tend sur l'évolution ; toutefois, cette méthode donnant des résultats satisfaisants en pratique, elle est toujours utilisée. La démarche opposée semble plus justifiable : on pose explicitement un modèle de l'évolution, et si ce modèle est juste, les méthodes de reconstruction raisonnant dans le cadre de ce modèle sont asymptotiquement

consistantes, *i.e.*, ont une probabilité de retrouver la phylogénie correcte qui tend vers 1 quand le nombre de caractères tend vers l'infini. Les modèles de l'évolution reposent avant tout sur un mécanisme décrivant la façon dont les substitutions se produisent dans les séquences ; le mécanisme adopté est généralement Markovien (*i.e.*, sans mémoire) et suppose souvent que les substitutions sont "*indépendantes et identiquement distribuées*" (i.i.d.). Autrement dit, les changements d'états se produisant à différents endroits de la séquence et n'importe où dans la phylogénie sont indépendants et tous les sites ont la même probabilité de changer d'état n'importe où dans la phylogénie. Les modèles de l'évolution sont parfois critiqués en raison des hypotèses simplificatrices qu'ils font généralement, car il existe des jeux de données pour lesquels certaines de ces hypotèses ne sont pas raisonnables. Toutefois, ces modèles servent maintenant de base à la plupart des méthodes de distances, ainsi qu'à certaines méthodes probabilistes (reposant en même temps sur la distribution des états observés pour les espèces sur les différents caractères [Hen91]).

1.2.4 La hiérarchie actuelle des méthodes

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

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

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

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

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

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

1.3 Modèles mathématiques de l'évolution

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

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

– *Les sites évoluent selon un processus de Markov homogène :* l'évolution est vue comme un processus sans mémoire, qui peut être caractérisé en décrivant pour un site quelconque seulement le passage de l'état courant à l'état suivant dans le temps. L'hypothèse d'homogénéité indique que les taux de substitutions sont constants dans le temps, et donc identiques sur toutes les

arêtes de la phylogénie. Ceci implique que la probabilité d'un certain changement d'état (p. ex. A→C) entre deux séquences ne dépende que de la durée de l'intervalle de temps qui les sépare, et non de la position de cet intervalle dans le temps. Il s'agit d'une hypothèse forte, signifiant que les différentes lignées de la phylogénie évoluent à la même vitesse.

- *Les substitutions sont indépendantes et identiquement distribuées sur les sites* (*i.i.d.*) : les substitutions affectant un site ne dépendent pas de sa position dans la séquence, ni des substitutions affectant les autres sites. Ceci permet de poser que la probabilité qu'une séquence se transforme en une autre séquence au cours d'un certain intervalle de temps, est le produit de la probabilité du changement d'état correspondant pour chacun des sites. De plus, tous les sites suivent un processus d'évolution identique, *i.e.*, ont la même probabilité de changer d'état au cours d'un intervalle de temps t.

- *Le processus évolutif est réversible* : le sens de l'évolution est indifférent, la probabilité pour un site de passer d'un état i à un état j dans un intervalle de temps t est identique à sa probabilité de passer d'un état j à un état i dans le même laps de temps.

- *La probabilité d'une substitution durant un intervalle de temps infinitésimal, dt, est proportionnelle à la durée.* La constante de proportionnalité est le taux de substitution déjà mentionné précédemment. On peut ainsi écrire $dP = Rdt$. On pose aussi que la probabilité que deux substitutions aient lieu durant l'intervalle dt, est négligeable vis-à-vis de la probabilité qu'il y en ait au plus une.

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

1.3.1 Les premiers modèles de l'évolution

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

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

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

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

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

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

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

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

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

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

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

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

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

et pour le modèle de Kimura à deux paramètres:

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

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

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

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

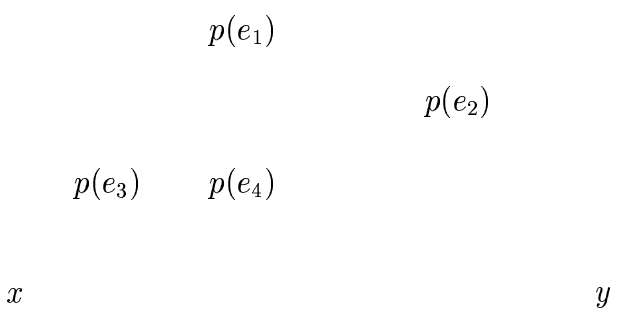

FIG. 1.2 – *Phylogénie dont les séquences évoluent suivant le modèle NCF. À chaque arête e_i est associée une probabilité $p(e_i)$ qu'une substitution affecte tout caractère entre les deux extrémités de l'arête.*

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

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

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

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

1.3.3 Correction des distances observées

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

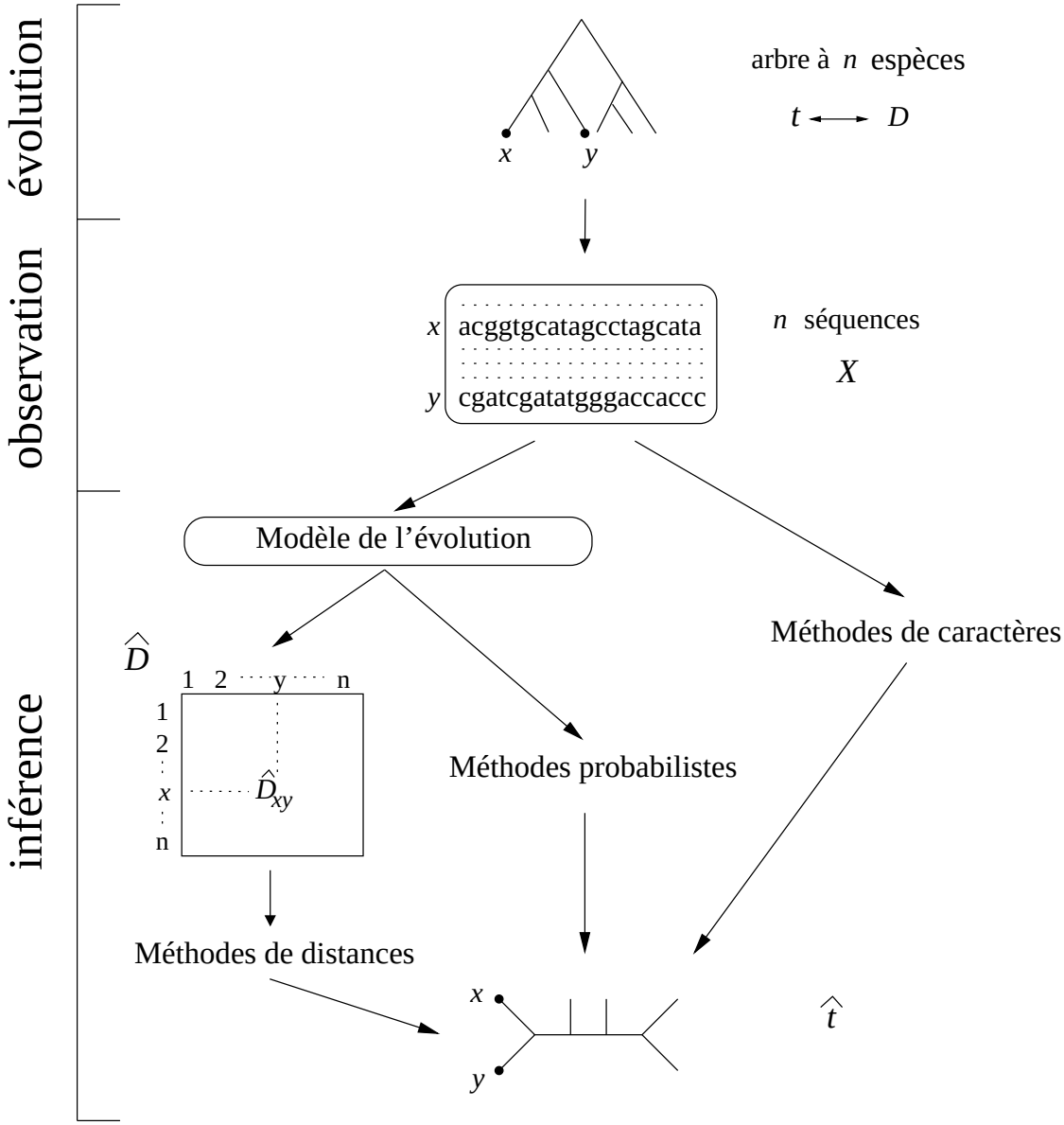

FIG. 1.3 – *La reconstruction phylogénétique vue comme une procédure d'estimation statistique.*

La Fig. 1.3 présente la reconstruction phylogénétique selon un point de vue lié à l'estimation statistique. L'histoire évolutive d'un ensemble E de n espèces y

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

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

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

Notre préoccupation principale dans cette thèse est de retrouver les différents groupes selon lesquels se répartissent les espèces étudiées, sans précision des dates auxquelles ces différents groupes se sont séparés, ni des quantités d'évolution les séparant. Autrement dit, nous nous intéressons avant tout à la *structure* de t, *i.e.*, au schéma de branchement suivant lequel les espèces se sont séparées. En ce sens,

nous adoptons une *approche discrète de la R.P.*, cherchant simplement à savoir quelles sont les arêtes de la phylogénie inconnue, sans souci de leurs longueurs respectives. Nous délaissons l'évaluation des longueurs d'arêtes, non seulement parce qu'une fois la structure de la phylogénie fixée, des méthodes efficaces existent pour fixer les longueurs de ses arêtes en fonction des distances entre espèces, mais aussi parce que la mesure exacte de ces distances dépend du contexte (p. ex. du type ou de la provenance des caractères initiaux) ou encore du modèle de l'évolution utilisé pour corriger les distances observées.

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

1.5 Définitions préliminaires

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

1.5.1 Phylogénies

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

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

(a) (b) (c)

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

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

1.5.2 Le problème de la racine

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

étudiées peut aussi rendre la reconstruction bien plus difficile (p. ex. augmentation de la distance évolutive maximum, ...).

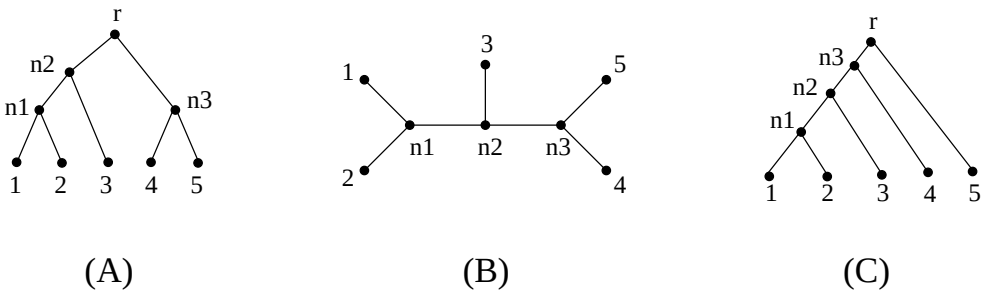

FIG. 1.5 – *Enracinement d'une phylogénie. La phylogénie enracinée (A), resp. (C), est obtenue en insérant une racine sur l'arête* (n_2, n_3)*, resp.* $(n_3, 5)$ *de la phylogénie (B).*

1.5.3 Degré de résolution

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

3. $d(x)$ dénote le degré du noeud x.

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

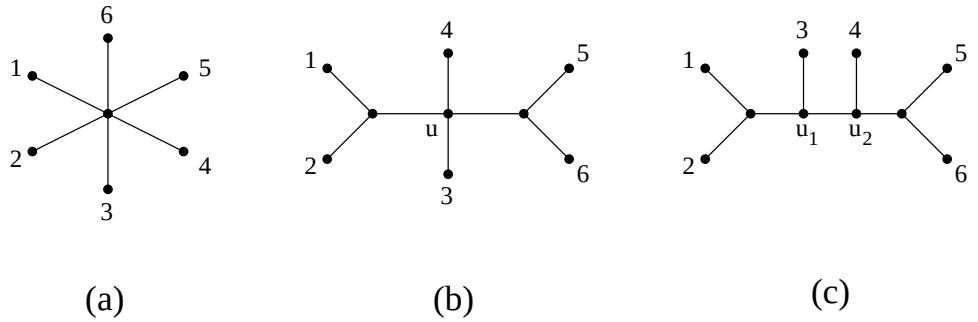

FIG. 1.6 – *Trois exemples de* phylogénies *sur les espèces* $1, \ldots, 6$ *: (a) la phylogénie étoile, (b) une autre phylogénie* partiellement *résolue, (c) une phylogénie* totalement *résolue correspondant à une résolution possible du noeud u de la phylogénie précédente.*

1.5.4 Bipartitions

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

les phylogénies sur le même ensemble d'espèces induisent les mêmes bipartitions triviales. Pour cette raison, on s'intéresse uniquement aux bipartitions non-triviales induites par une phylogénie, et on compare les méthodes de reconstruction sur leur capacité à retrouver les arêtes internes d'une phylogénie.

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

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

Théorème 1 *[LeQ69, Bun71, EJM76]*

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

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

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

1.5.5 Quadruplets et r4-arbres

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

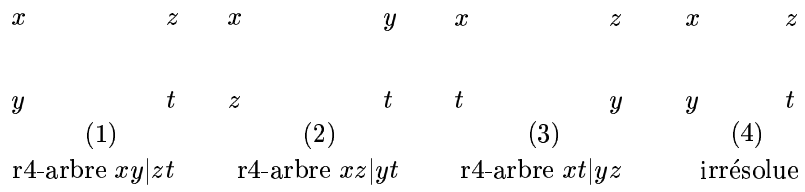

FIG. 1.7 – *Phylogénies possibles pour un quadruplet d'espèces.*

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

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

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

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

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

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

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

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

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

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

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

1.5.6 Valuation d'une phylogénie

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

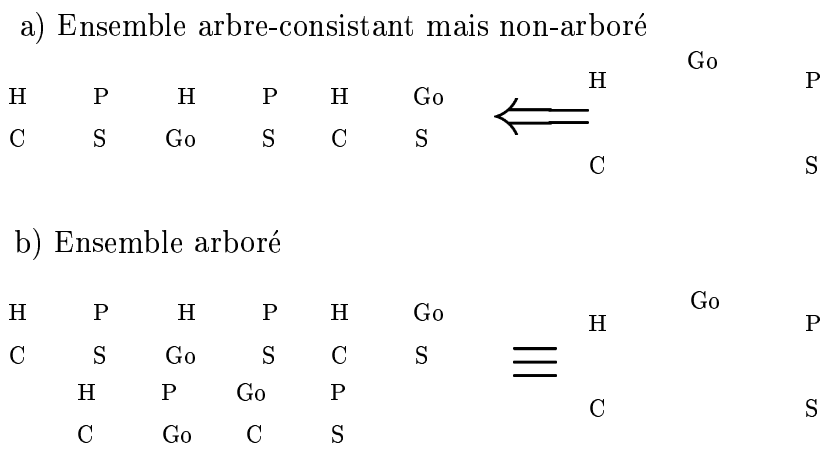

FIG. 1.8 – *Propriétés d'ensembles de r4-arbres .*

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

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

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

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

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

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

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

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

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

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

1.5.7 Un aperçu de la combinatoire des phylogénies

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

donc :

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

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

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

Chapitre 2

Critères et méthodes de R.P.

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

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

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

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

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

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

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

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

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

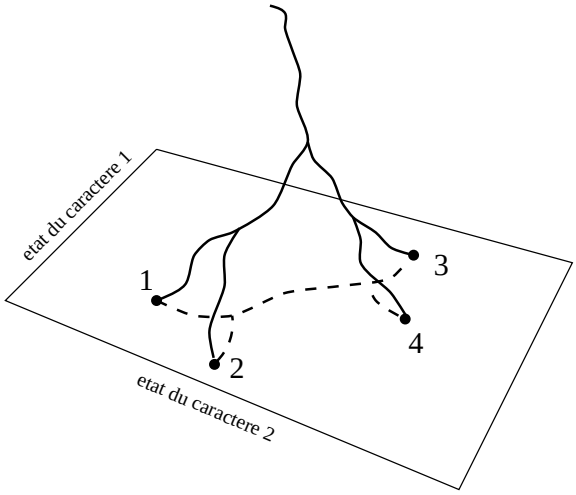

FIG. 2.1 – *(empruntée à Cavalli-Sforza et Edwards [CSE67]). Une phylogénie et sa projection sur le plan "actuel". Par souci de simplification, seules deux dimensions associées à des caractères sont représentées (la dimension temporelle est représentée verticalement). .*

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

2.1 Les méthodes de caractères

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

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

2.1.1 Le Maximum de Parcimonie (MP)

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

Définition du critère

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

$$S_0 = AAAAAA$$

$$S_1 = GAAAAA \qquad\qquad S_2 = AAAGAA$$

$$S_{11} = GGAAAA \qquad S_{12} = GAGAAA \qquad S_{21} = AAAGGA \qquad S_{22} = AAAGAG$$

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

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

Dans le cas général, nous ne disposons initialement que des séquences associées aux feuilles de la phylogénie (les espèces contemporaines). Il existe théoriquement

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

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

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

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

Méthodes d'optimisation de MP

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

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

Les nombreux algorithmes heuristiques proposés [Swo90, Fel93] restent les plus

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

2.1.2 La méthode de Compatibilité

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

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

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

Tous les caractères rejetés par cette procédure nécessitent plusieurs substitutions pour expliquer leurs états aux espèces sur la phylogénie proposée, ils sont

donc jugés homoplasiques. Selon Darlu et Tassy ([DT93], p.148), la méthode de compatibilité peut être considérée comme une méthode utilisant le principe de parcimonie, en ce sens qu'elle retient comme phylogénie celle qui *minimise le nombre* de caractères jugés homoplasiques (les plus coûteux à expliquer).

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

2.1.3 La méthode du Maximum de Vraisemblance (MV)

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

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

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

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

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

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

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

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

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

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

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

2.2 Les méthodes de distances

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

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

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

2.2.1 La méthode de représentation multidimensionnelle (MDS)

Edwards et Cavalli-Sforza [ECS63, CSE67] furent parmi les premiers à appliquer une approche de distances en R.P. Ils disposaient de données consistant en des fréquences de gènes observées pour différentes espèces, qu'ils proposaient de convertir en une matrice de dissimilarités $\hat{D}$. Leur idée était de représenter, sur la base de $\hat{D}$, l'ensemble des n espèces étudiées par un ensemble de n points dans

un espace euclidien à $n-1$ dimensions. La longueur d'une phylogénie reliant les espèces est simplement obtenue en sommant la longueur de ses différentes arêtes, chacune étant définie comme la distance euclidienne séparant ses deux extrémités. Cavalli-Sforza et Edwards proposaient de rechercher l'arbre de Steiner de longueur minimum pour relier les points associés aux espèces. Autrement dit, on cherche ici la phylogénie dont la quantité d'évolution est minimale, puisque la distance euclidienne entre deux points représente leur distance évolutive.

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

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

2.2.2 Les méthodes d'ajustement

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

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

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

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

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

2.2.3 Les méthodes de scores

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

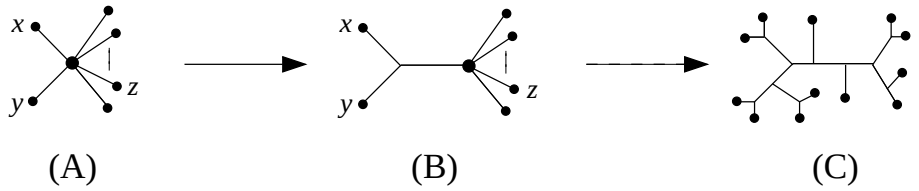

FIG. 2.3 – *Processus agglomératif employé pour construire une phylogénie.*

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

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

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

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

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

2.2.4 La méthode d'évolution minimale (ME)

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

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

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

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

Une heuristique pour ME

Examiner toutes les phylogénies possibles n'est pas raisonnable en raison de leur nombre exponentiel (cf. équation (1.7)). Les phylogénéticiens ont plutôt recours à un algorithme heuristique présenté par Saitou et Nei [SN87]. Cet algorithme est appelé *Neighbor-Joining* (NJ), puisqu'il adopte, comme *Addtree* (cf. section 2.2.3), un principe agglomératif le conduisant à *joindre*, à chaque étape, des noeuds *voisins* dans la phylogénie. L'algorithme NJ fonctionne donc selon le même principe qu'*Addtree* à la différence qu'à chaque étape, il choisit de regrouper les deux

espèces x et y, de sorte que la longueur de la phylogénie résultant de leur agglomération en un sous-arbre commun, estimée aux moindres carrés, soit minimale. En ce sens, l'algorithme optimise à chaque étape le critère ME. Toutefois, cette optimisation est approximative puisque l'algorithme est glouton, et rien n'indique qu'ajouter la plus courte arête à chaque étape conduise à la phylogénie la plus courte.

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

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

2.2.5 La méthode FPM

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

Plus précisément, la condition des quatre points assure qu'une distance $\hat{D}$ sur quatre espèces x, y, z, t correspond exactement à une phylogénie si et seulement

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

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

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

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

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

2.2.6 La méthode ordinale

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

proposé par A. Guénoche).

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

- On note $\mathcal{P}(\hat{D})$ la préordonnance associée à une dissimilarité $\hat{D}$, obtenue en ordonnant les valeurs de $\hat{D}$. Si $\hat{D}_{xy} < \hat{D}_{zt}$, on aura $xy < zt$ dans la préordonnance, ce qui peut être noté par le *couple* $(xy < zt)$.
- Une préordonnance $\mathcal{P}(\hat{D})$ est dite *arborée* ssi il existe une distance d'arbre D t.q. $\mathcal{P}(D) = \mathcal{P}(\hat{D})$.
- On dit que les couples $(xy < zt)$ et $(xz < yt)$ sont *emboîtés* si les *paires* de l'un ou de l'autre couple sont les valeurs extrêmes des quatre paires, *i.e.*, on a dans $\mathcal{P}(\hat{D})$ l'un des deux ordres $xy < xz < yt < zt$ ou $xz < xy < zt < yt$ (cf. Fig. 2.4-(a)).
- On dit que le couple $(xz < yt)$ *domine* le couple $(xy < zt)$ si dans $\mathcal{P}(\hat{D})$ on a $xy < xz$ et $zt < yt$ (cf. Fig. 2.4-(b)).
- On appelle *condition ordinale des quatre points* la condition telle que tout quadruplet d'espèces x, y, z, t ait deux couples emboîtés et que le troisième couple ne domine pas les deux autres.

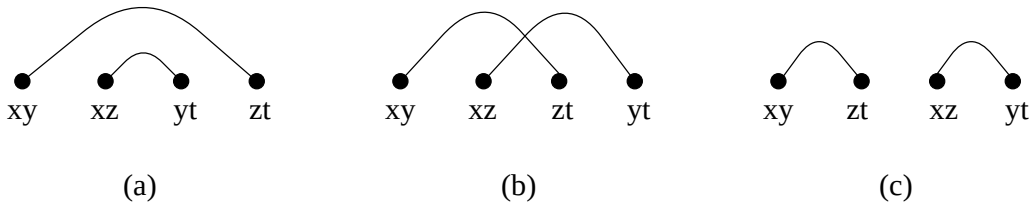

FIG. 2.4 – *Différentes configurations pour les couples* $(xy < zt)$ *et* $(xz < yt)$. *(a) Les deux couples sont emboîtés. (b) et (c) le couple* $(xz < yt)$ *domine le couple* $(xy < zt)$.

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

une phylogénie sur la base des quadruplets vérifiant cette condition par la règle suivante [Gué97b, KHM97, Kea97] :

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

2.3 Propriétés souhaitables pour une méthode de reconstruction

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

2.3.1 Complexité polynomiale

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

2.3.2 Convergence

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

méthode est dite convergente si elle infère une certaine phylogénie avec une probabilité de plus en plus élevée lorsque le nombre de caractères disponibles augmente.

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

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

Conditions sur les longueurs relatives des arêtes

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

inégaux). Ceci a été montré sous le modèle NCF [PHH87, HP89] et sous celui de Jukes et Cantor [TN94a]. Signalons que plusieurs jeux de données réels, pour lesquels la phylogénie supposée est soumise au problème d'attraction des longues branches, ont été exhibés [Lak86, PHS91].

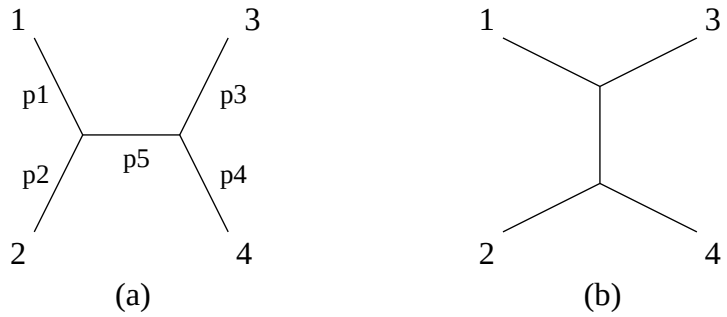

FIG. 2.5 – *(a) Phylogénie sur 4 espèces, où à chaque arête est affectée une probabilité p_i que tout site subisse une substitution le long de cette arête. (b) Phylogénie incorrecte inférée par MP quand $p_1 = p_3 >> p_2 = p_4 = p_5$.*

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

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

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

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

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

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

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

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

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

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

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

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

On sait aussi que pour toute distance d'arbre $D=t$, il existe une matrice quelconque $\hat{D}$ et une distance d'arbre $D'=t'$ tels que $L_\infty(\hat{D}, D)=\frac{l(e)}{2}$ et $L_\infty(\hat{D}, D')=\frac{l(e)}{2}$ [ESSW97b, Att97c]. Ceci caractérise donc le rayon le plus large autour de t, et

au sens de L_∞, dans lequel on peut garantir qu'une méthode raisonnable retrouve la phylogénie corecte avec une probabilité de 1. En effet, pour toute matrice de données $\hat{D}$, dès que $L_\infty(\hat{D}, D) > \frac{l(e)}{2}$, t n'est plus la phylogénie (unique) la plus proche de $\hat{D}$ au sens de L_∞. Ce dernier résultat implique que les méthodes de distances basées sur FPM ainsi que les méthodes *Addtree* et NJ ont des garanties de retrouver la phylogénie corecte de façon sûre, aussi larges que possible au sens de L_∞ !

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

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

2.3.4 Puissance statistique

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

Résultats empiriques

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

soumises au problème d'attraction des longues branches sont difficiles à reconstuire même pour les méthodes de distances : celles-ci demandent un nombre irréaliste de caractères pour avoir ne serait-ce qu'une chance sur deux de retrouver la phylogénie correcte [HH93, HHS94].

Résultats théoriques

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

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

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

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

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

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

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

1. nombre maximum d'arêtes entre deux feuilles.

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

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

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

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

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

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

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

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

2.3.5 Robustesse

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

2.3.6 Facteurs d'approximation

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

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

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

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

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

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

Agarwala *et al.*

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