# DE LA DIVERSIFICATION DES GALAXIES

Didier Fraix-Burnet

# DE LA DIVERSIFICATION DES GALAXIES

**Une introduction à l'astrocladistique**

Didier Fraix-Burnet

Laboratoire d'Astrophysique de Grenoble

Version 3 9 Juin 2010

*À Jean-Luc*

# Avant-propos

Dans ce livre, je souhaite raconter une belle histoire de la Science, un parfait exemple d'interdisciplinarité, une aventure humaine extraordinaire qui j'espère fascinera le lecteur autant qu'elle m'a fasciné et me fascine encore. Il fallait un lieu où je puisse expliquer la philosophie, les concepts et les méthodes pratiques de l'astrocladistique. Il fallait un lieu où certaines questions, élémentaires à première vue, nécessitaient d'être posées, afin de susciter les réflexions qui s'en dégagent. Il fallait un lieu pour que les réponses à ces questions, souvent de simples définitions, puissent être enfin clarifiées, noir sur blanc. Ce livre est le résultat de tous ces besoins. J'espère ainsi transmettre au lecteur la magie de l'astrocladistique, en montrant comment les progrès considérables de nos connaissances sur les galaxies et l'Univers impliquent de remettre à plat des notions simples, tout en envisageant le développement de nouveaux concepts. Il s'agit ni plus ni moins de revoir notre façon de raconter l'histoire de la diversité des galaxies, et donc quelque part l'histoire de l'Univers. Cette démarche est parfois déroutante, car elle se heurte notamment à la capacité somme toute limitée de l'être humain à appréhender la complexité du monde qui l'entoure.

L'astrocladistique ne laisse jamais indifférent, elle engendre l'émerveillement, la fascination, le rejet ou la peur. Les raisons en sont le plus souvent inconscientes : interdisciplinarité, méthodologie venant de la biologie évolutive, rupture culturelle, approche adaptée à l'exploitation de toutes nos connaissances, jargon incompréhensible, concepts difficiles, etc. En clair, l'adhésion sans faille ou le rejet brutal sont presque toujours associées à ce nouveau champ de recherche. Ceci fait partie de mon lot quotidien depuis que je me suis lancé dans cette aventure en 2001, et a sans aucun doute été un puissant moteur pour le projet de ce livre. Parti de la mise en forme de toutes mes notes, puis de quelques cours généraux sur la cosmologie et les galaxies, il s'est énormément enrichi des nombreuses réactions de mes collègues. Cependant, je ne pense pas qu'il réponde à toutes les questions que le lecteur pourra légitimement se poser. D'ailleurs, je n'ai pas toutes les réponses, le sujet est hautement évolutif, et je sais l'ampleur de la tâche qu'il nous reste à accomplir.

Je destine ce livre à tous mes collègues astrophysiciens, en particulier les plus jeunes qui ont peut-être découvert les bases de la cladistique durant leurs études secondaires. L'objectif de ce livre se veut pédagogique, pratique. L'astrocladistique considère les galaxies en tant que population et fait donc appel à des approches peu connues des physiciens en général. La cladistique est en elle-même une méthode conceptuellement difficile et mathématiquement encore hardue. Tout ceci rend nécessaire la description détaillée de la méthodologie dans son ensemble. J'espère que ce livre comblera un manque que les articles ne sauraient satisfaire.

Mais ce livre devrait intéresser également les systématiciens, les cladistes, les ma-

thématiciens, les statisticiens, les bioinformaticiens et bien sûr les biologistes évolutionnistes. Les astronomes amateurs y trouveront certainement leur compte, car en apprenant les bases de la cladistique sur un sujet qui les passionne, ils comprendront du même coup un peu mieux l'organisation du monde vivant et la notion finalement très générale d'évolution. Enfin, il me semble que les historiens des sciences, les épistémologues, et les sociologues des sciences peuvent y trouver matière à réflexion.

# Remerciements

S'engager dans une recherche théorique innovante et multidisciplinaire n'est pas chose facile. C'est une véritable aventure, faites de joies, de déceptions et d'incertitudes, mais surtout remplies de multiples rencontres et découvertes toutes plus merveilleuses les unes que les autres.

J'ai tout d'abord eu la chance, énorme, de trouver Philippe Choler, biologiste, sur le même campus. Mon idée lui parut immédiatement évidente, et nos discussions furent éclairantes. C'est lui, je crois, qui a inventé le mot "astrocladistique". C'est grâce à lui que ma conviction profonde que la cladistique était une piste intéressante a pris véritablement corps, et que la motivation d'aller jusqu'au bout pour le démontrer s'est forgée. Combien lui dois-je ?

Car c'est lui encore qui m'a mis rapidement en contact avec Emmanuel Douzery, biologiste spécialiste de cladistique. Il m'a beaucoup guidé, il m'a énormément conseillé, appris, et ses calculs m'ont grandement rassuré.

Je ne peux que les remercier tous les deux, car notre collaboration multidisciplinaire a parfaitement fonctionné, malgré leur surcharge monumentale de travail. Nous avons publié trois articles ensemble, les trois premiers articles fondamentaux de l'astrocladistique.

Que dire d'Emmanuel Davoust que j'ai côtoyé au tout début de ma carrière, et qui s'est pris de passion pour l'astrocladistique, lui qui a travaillé quelque temps auprès de Gérard de Vaucouleurs, un chantre de la classification de Hubble ? Son amour des catalogues, ses larges connaissances des galaxies et des amas globulaires, ont fait de lui le collaborateur idéal pour les premières exploitations astrophysiques des cladogrammes. Notre amitié de longue date me rend particulièrement heureux de pouvoir l'associer à ces premières étapes difficiles, mais stimulantes, de l'astrocladistique. Il aura joué un rôle considérable dans l'obtention des deux premiers résultats spectaculaires et décisifs, après plusieurs années d'intenses discussions et de très nombreuses analyses exploratoires souvent infructueuses.

Il y a quatre personnes qui ont permis à l'astrocladistique d'atteindre sa maturité en 2008. Tout d'abord, je ne remercierai jamais assez Pierre Pontarotti, biologiste passionné d'évolution, de m'avoir invité à sa conférence et ainsi permis de découvrir une dimension insoupçonnée de mon travail. Il s'agit de ce qu'il appelle la phylomathématique, c'est-à-dire la formalisation des mécanismes de diversification. Mais il m'a aussi amené à réfléchir sur la notion d'évolution en général, replaçant ainsi les galaxies dans le contexte plus large des populations. Enfin, et ce n'est pas rien, il m'a permis de préciser le rôle exact de l'environnement. Au retour de sa conférence, j'ai pris conscience que l'astrocladistique venait de prendre une autre dimension. Merci Pierre.

Cette conférence m'a fait rencontrer Marc Thuillard, bioinformaticien spécialiste

des réseaux, dont le travail m'a immédiatement interpelé comme pouvant être fort utile à l'astrocladistique. Notre article en commun, mais surtout nos longues discussions très mathématiques, absolument passionnantes, m'ont énormément appris sur le formalisme des méthodes de distances et de la cladistique en variables continues.

Enfin, grâce à Emmanuel, j'ai établi une collaboration humainement et scientifiquement riche avec Asis Kumar et Tanuka Chattopadhyay, respectivement statisticien et astrophysicienne. Ils sont essentiellement les premiers à appliquer des méthodes statistiques multivariées de haut vol en astrophysique avec beaucoup de succès. Je les remercie pour leur gentillesse, leur disponibilité, et surtout pour leurs compétences qui forment le maillon indispensable pour effectuer des analyses robustes.

Je dois également remercier les très nombreux collègues dont l'enthousiasme, frôlant parfois la fascination, m'a communiqué un encouragement spontané qui a joué un grand rôle pour contrer la solitude dans laquelle je me retrouvais parfois face à la communauté astrophysique.

Je remercie aussi mon équipe, notamment Guy Pelletier et Gilles Henri, ainsi que mon laboratoire, pour m'avoir fait confiance et accorder le soutien financier certes modeste mais néanmoins vital pour un tel projet.

J'ai enfin une pensée tout particulière pour Francesca et Robin, qui m'ont accompagné de si près lors des nombreuses vicissitudes du développement de l'astrocladistique.

# Table des matières

# Chapitre 1

# Introduction

Alors que les galaxies n'étaient encore pas nées dans l'esprit humain, Einstein réussissait en 1905 à décrire l'Univers en reliant l'espace et le temps tout deux régis par la gravitation. Depuis lors, la Relativité Générale s'est imposée comme l'indispensable outil pour connaître la géométrie de notre Univers, et alliée au principe cosmologique qui suppose que nous n'occupons pas de place privilégiée, elle offre un cadre prédictif et falsifiable de compréhension de l'histoire de l'Univers et de ses composantes.

Cependant, beaucoup de chemin théorique et conceptuel restait à parcourir lorsqu'en 1922 notre environnement grandit d'un facteur gigantesque avec la découverte par HUBBLE des Univers-Îles, c'est-à-dire d'entités composées d'innombrables étoiles et situées à des distances inouïes. D'autres objets semblables à notre Voie Lactée existaient donc indépendamment. Ils furent appelés galaxies. On imagine à peine le choc culturel que cela put produire. Mais que dire de la découverte par ce même HUBBLE, en 1929, que ces galaxies s'éloignaient toutes les unes des autres, et donc que l'Univers était en expansion ? En moins de dix ans, l'Univers petit et statique devenait grand et en évolution !

Cette expansion de l'Univers était prédite par les équations de la Relativité Générale, mais le blocage culturel a dérouté Einstein d'un (autre !) résultat sensationnel. Peu importe, un siècle de débats et d'exploits instrumentaux auront fourni à l'Humanité un cadre plutôt précis à notre environnement, sur des échelles spatiales et temporelles qui dépassent l'entendement. Les observations toujours plus incroyables prennent place dans un schéma étonnant de cohérence. Peut-être qu'un jour, cette image de notre Monde se révélera erronée, mais elle est actuellement trop belle pour ne pas être adoptée.

Alors quid des galaxies dans toute cette histoire ? Elles ont bénéficié elles aussi des progrès fous de la technologie. Les multiples détails observationnels, jusqu'à des distances de plus en plus grandes, se sont accompagnés de développements théoriques parfois très sophistiqués. Depuis dix ou vingt ans particulièrement, nos connaissances explosent avec le nombre de galaxies accessibles, le raffinement des observations, la compréhension des nombreux mécanismes physico-chimiques œuvrant au sein d'une même galaxie, et la puissance des ordinateurs qui tentent de simuler ces Univers-Îles. Leur statut évolutif ne fait plus de doute aujourd'hui, aussi bien du point de vue observationnel que théorique. Grâce à notre beau schéma cosmologique, nous pouvons espérer comprendre comment elles sont apparues pour la première fois, il y a envi-

ron 13 milliards d'années, au cœur des fluctuations primordiales de densité de matière noire, fluctuations bien observées aujourd'hui et issues de cette mystérieuse phase du Big Bang. Nous savons également comment elles grossissent, comment elles se déforment, comment elles se transforment, étant soumises aux lois impitoyables de la gravité et de la présence pesante de la matière noire.

Il devient ainsi de plus en plus évident que les galaxies sont des objets complexes, complexes à décrire, complexes à modéliser, complexes à appréhender dans leur diversité, et complexe à prédire dans leur évolution. Le décalage est désormais flagrant entre nos connaissances et la façon dont on les synthétise, le plus souvent à travers des classifications directement héritées de HUBBLE il y a quatre-vingts ans. Nous avons découvert que l'Univers évolue, que les galaxies évoluent, il serait temps de faire évoluer notre classification, notre façon de décrire ces gigantesques objets et notre façon de dire leur histoire. Il faut probablement renoncer à une certaine simplicité, bien confortable il est vrai. Ne doutons pas qu'il s'agit là d'un saut culturel, car décidément, les galaxies ne sont pas des étoiles.

Oui, mais comment faire ? Notre monde a largement abordé les problèmes de complexité dans le courant du XXe siècle, et de nombreux travaux, aussi bien mathématiques que philosophiques, ont été et sont encore menés. L'Humanité accumule tellement de connaissances qu'elle découvre un monde de plus en plus complexe. L'esprit humain, lui, garde une propension à la simplification : aptitude à la synthèse ou capacités limitées ? Les biologistes ont depuis longtemps été confrontés aux organismes vivants, objets complexes s'il en est. De plus, les preuves de leur évolution et de leur diversification au cours du temps ont été apportées dès le XIXe par DARWIN, et cet aspect ne fait désormais plus aucun doute. Aucun biologiste n'est capable de décrire et modéliser en détail un organisme vivant, encore moins de prédire son évolution. Pourtant, la biologie est capable de construire une histoire intelligible des êtres vivants sur la Terre, de décrire les liens de parenté, donc les différences et similitudes entre plusieurs organismes, tout en adaptant ces visions synthétiques aux progrès de la Science.

L'astrophysicien extragalacticien n'est donc pas seul dans sa problématique. Des pistes existent certainement pour exploiter au mieux les quantités astronomiques d'informations rassemblées par les grands télescopes, au sol et dans l'espace, ainsi que pour utiliser pleinement les simulations numériques traitant de la formation des galaxies et pouvoir les comparer à la réalité. Il est intéressant de constater que les biologistes ont d'emblée été confrontés à la complexité et à la diversité de leurs objets d'étude. Récemment, les progrès technologiques leur permettent de sonder en détail les mécanismes microscopiques, comme les membranes des cellules. Ils font désormais appel aux physiciens et chimistes pour décrire et comprendre les processus par lesquels les cellules absorbent ou non certaines protéines. À l'inverse, les astrophysiciens ont été confrontés d'abord aux composants élémentaires de l'Univers, les étoiles, les planètes et le gaz, alors que les progrès technologiques leur ont montré la complexité et la diversité du monde des galaxies peuplant l'Univers. Empruntant deux chemins inverses, n'est-il pas naturel que ces deux communautés se retrouvent pour partager des outils spécifiques largement éprouvés ?

La cladistique est une formidable méthodologie d'analyse des données qui fournit une représentation perfectionnée de la diversité d'objets complexes en évolution. L'objet de ce livre est de montrer comment cette approche initialement inventée par

les linguistes et largement développée pour la biologie évolutive, peut s'appliquer aux galaxies. Des questions fondamentales vont être abordées. En particulier les notions de formation, évolution, diversification, classification, et même la définition d'une galaxie. Le titre du livre résume l'ensemble de la problématique de l'astrocladistique. Dans le mot diversification, il y a bien sûr la diversité immense des objets observés à l'aide des détecteurs performants associés aux télescopes dans tous les domaines de longueur d'onde. Dans ce mot, la notion de classification se trouve évidemment associée car comment envisager d'appréhender la grande variété de ces objets et modéliser leurs comportements si nous ne simplifions pas quelque peu nos énormes catalogues en catégories pertinentes moins nombreuses ? Le mot diversification laisse entendre une action, cette action ne pouvant qu'être l'évolution, c'est-à-dire une modification au cours du temps. Ainsi la (trans)formation des galaxies est le concept qui mène à la diversification. Enfin, si les galaxies évoluent en se transformant et en se diversifiant, il en résulte nécessairement des groupes rassemblant des objets plus ou moins proches au sens évolutif, c'est-à-dire des objets partageant une histoire plus ou moins semblable. Ce sont les "liens de parenté" en biologie, terme qui laisse sous-entendre des relations peu compatibles avec la nature inanimée des objets que sont les galaxies. Quoi qu'il en soit, la représentation sous forme de ramifications, donc d'un arbre, dans l'esprit de la théorie des graphes, découle assez naturellement. C'est probablement ici que se situe la grande nouveauté culturelle de l'astrocladistique. Les clarifications que nous allons apporter dans ce livre, via une mise à plat d'un siècle de progrès sur les galaxies et sur l'Univers, et grâce aux enseignements de deux mille ans de classification des espèces vivantes, ont pour but de servir de bases solides afin d'ouvrir de nouveaux horizons.

Ce livre vise à une introduction pratique de l'astrocladistique. Il n'a nullement pour objectif d'établir une revue détaillée de nos connaissances en astrophysique extragalactique. Il s'appuie plutôt sur l'évolution de ces connaissances qui ont révélé la complexité des galaxies, impliquant de devoir les appréhender comme une population en évolution dans un environnement évolutif. L'astrocladistique s'inscrit donc logiquement dans l'évolution des progrès de l'astrophysique extragalactique. Ce livre ne présentera pas une vision complète et "phylogénétique" des galaxies peuplant notre Univers, mais s'attache davantage aux méthodes qui peuvent fournir un tel panorama.

Les deux premiers chapitres sont consacrés à des concepts indispensables à la compréhension de la méthode et de ses résultats. Les notions de diversité, de complexité et de classification sont abordées dans le Chapitre 2 sous un aspect historique afin de montrer le parallèle entre l'évolution de nos connaissances en biologie et celles en astrophysiques extragalactiques. Il montre en particulier l'intérêt des analyses multivariées dans l'étude d'objets complexes et comment dans les deux cas les progrès des connaissances les ont rendues nécessaires.

Le Chapitre 3 introduit la notion fondamentale d'évolution et ses implications en terme de classification. Nous y verrons également le rôle de l'environnement. Ces notions, devenues assez évidentes en biologie, sont ensuite clarifiées pour ce qui concerne les galaxies. Cela montrera qu'indépendamment de connaissances pointues toujours susceptibles de remises en cause majeure, nous en savons suffisamment pour que s'impose la nécessité de classifications évolutives. Nous conclurons par un résumé des principes généraux de classifications et de leur utilité en astrophysique afin de situer clairement la cladistique.

Le Chapitre 4 présente la cladistique en tant que méthode, indépendamment de tout

objet réel. Le lecteur est invité à s'exercer sur les exemples donnés car seul la pratique sur des cas simples permet de comprendre ce que les programmes feront sur des échantillons concrets. Ce chapitre est également utile pour détacher la cladistique de toute référence à la biologie et ainsi se l'approprier pleinement en tant que méthode. Nous n'aborderons pas l'aspect mathématique, sujet très difficile encore en développement, et qui n'est pas indispensable pour mettre en œuvre des analyses.

L'astrocladistique, c'est-à-dire l'utilisation de la cladistique pour les galaxies, est détaillée dans le Chapitre 5. Les descriptions des galaxies données dans les Chapitres 2 et 3, sont ici placées dans un schéma conforme à l'esprit de la cladistique. Nous verrons ainsi les concepts à la base de l'astrocladistique, en particulier comment la diversification des galaxies répond bien au processus de transmission avec modification. La fin de ce chapitre ouvrira le champ d'application de l'astrocladistique à d'autres objets, comme les amas globulaires, qui ont des liens de "parenté" provenant de leur environnement de formation.

Nous passerons à la mise en pratique de l'astrocladistique dans le Chapitre 6. En suivant pas à pas ce chapitre, qui inclut la description de quelques programmes, le lecteur devrait être capable d'effectuer lui-même ses premières analyses. La manière de construire des arbres plus complexes, les superarbres, qui permettront peut-être un jour de représenter la diversité des galaxies sur un même arbre, à l'instar de l'Arbre de la Vie, est décrite dans le Chapitre 7.

Obtenir un arbre n'est pas suffisant, il faut ensuite pouvoir le lire, c'est-à-dire le comprendre, estimer sa pertinence et l'utiliser. C'est l'objet du Chapitre 8. Les classifications multivariées sont encore assez difficiles à interpréter pour un physicien, peu formé à ce genre d'exercice. L'introduction de l'évolution est à la fois une aide et une difficulté supplémentaire. Ce chapitre, avec des résultats concrets, aiderons, je l'espère, l'astrocladiste débutant à exploiter au mieux ses analyses.

Dans tout ce livre, nous présentons un seul aspect de la cladistique, la parcimonie maximum, utilisant des variables discrètes. Il existe d'autres méthodes de reconstruction phylogénétique, pour l'essentiel encore non exploitées à ce jour en astrophysique, mais qui pourraient se révéler intéressantes. Nous présentons dans le Chapitre 9 un aperçu des différentes classes de méthodes ayant été développées. Nous y verrons notamment le cas des variables continues, qui représentent la quasi-totalité des descripteurs pour les galaxies, pour lesquelles les outils de cladistique sont encore peu développés. C'est un domaine où l'astrophysique pourrait peut-être apporter une impulsion inattendue à la bioinformatique.

Enfin, nous terminons ce livre avec des notions de taxonomie (Chapitre 10), car il semble évident qu'une nouvelle classification devra émerger de toutes les études multivariées évolutives qui se feront sur des millions de galaxies, et donc une nomenclature adaptées devra être mise au point. Cela représente un vaste programme pour le futur, nous n'esquisserons que quelques principes utiles à garder en mémoire.

# Chapitre 2

# Diversité et complexité

La classification est le reflet de notre vision de la diversité. Elle ne sert qu'à ranger les objets dans des boites qui rendent plus aisées leur désignation et leur comparaison : il est plus commode par exemple de parler des galaxies elliptiques que de nommer toutes les galaxies de cette forme. Il pourrait sembler que les galaxies sont infiniment plus simples que les organismes vivants. Est-ce bien vrai ? Pas si sûr, car la complexité d'un objet dépend beaucoup du niveau de détail des observations et du nombre de descripteurs à disposition. En réalité, les galaxies ne paraissent éventuellement simples que si on se limite à un nombre très faible d'observables. Et ce nombre tend naturellement à croître avec les progrès technologiques.

Nous allons constater dans ce chapitre que nos connaissances contemporaines des galaxies nous montrent clairement que ces objets sont complexes. Il semble alors assez légitime de regarder comment les biologistes appréhendent l'extrême diversité du monde vivant. Il est donc intéressant de se pencher sur les concepts de la classification, regroupés dans une branche de la biologie évolutive qui s'appelle la systématique. Les astrophysiciens ont cette chance de pouvoir bénéficier de plus de deux mille ans d'expérience de l'Humanité devant la complexité. Il serait dommage de ne pas en profiter.

Ce chapitre et le suivant, consacré spécifiquement à l'évolution et l'environnement, s'organisent autour du parallèle entre les évolutions de nos connaissances sur les organismes vivants d'une part et les galaxies d'autre part. Il ne s'agit en aucun cas d'analogie entre les deux types d'objets, mais bien du cheminement identique de l'évolution des observations dans les deux disciplines. C'est ce rapprochement conceptuel qui a guidé les débuts de l'astrocladistique et a grandement contribué à la formalisation qui sera développée au Chapitre 5.

## 2.1 Méthode traditionnelle de classification

### 2.1.1 Biodiversité au temps des grecs

Les êtres humains ont toujours été confrontés à la diversité du monde vivant, et leur préoccupation essentielle procédait de considérations utilitaires : les plantes pouvaient être comestibles, toxiques, médicinales ou encore cultivables, certains animaux pouvaient être domestiqués. Sans doute que, jusqu'au temps des grecs, une classification générale n'était pas indispensable, la diversité utile étant probablement peu importante.

Mais les premiers philosophes se sont intéressés à la compréhension du Monde, aux choses générales. Ainsi, du temps de ARISTOTE, est née une classification pragmatique, basée à la fois sur des critères utilitaires très anthropocentriques (propriétés alimentaires, possibilité d'être cultivé ou domestiqué), et sur des observations basiques d'apparences comme la taille des plantes, les moyens de déplacement et la température du sang des animaux.

Cette classification introduisait déjà un énorme saut culturel, car ainsi la diversité du monde vivant connu était presque entièrement couverte, avec un nombre de caractéristiques logiques et simples. Cette façon de ranger les plantes et les animaux a probablement autorisé l'émergence des premières théories unificatrices cherchant à relier les différentes classes en un schéma conceptuellement parlant. En effet, combien aurait-il pu être simple de reconnaître une plante comestible ou de déduire ses propriétés médicinales uniquement d'après la taille de ses feuilles et à son environnement immédiat ! Cette approche eut visiblement un beau succès puisque cette classification grecque perdura jusqu'au Moyen Âge.

### 2.1.2 HUBBLE découvre et classe les galaxies

C'est en 1922 que HUBBLE a découvert les galaxies, c'est-à-dire qu'il réussit à mesurer les distances considérables à laquelle se trouvaient ces "nébuleuses" (HUBBLE, 1922). Cette découverte fut certainement bouleversante car l'Univers imaginé à l'époque devenait brutalement immensément plus grand. Certes les débats rageaient entre les partisans d'un petit Univers et ceux partisans des nébuleuses "extragalactiques" dans un Univers gigantesque. Le travail de HUBBLE trancha donc, mais il est important de ne pas oublier que cette découverte est extrêmement récente par rapport à l'histoire de l'Humanité !

HUBBLE, par son travail minutieux et énorme, accumula une connaissance de la diversité des galaxies très complète à l'époque. Cette diversité s'exprimait en fonction des observables à sa disposition. Et pour cela il ne disposait que de l'imagerie, donc il n'avait accès qu'à un seul descripteur : la morphologie, dans le domaine du visible. Ainsi, pour HUBBLE, la diversité des galaxies se résumait en la diversité des morphologies.

En tout logique, comme tout être humain devant un ensemble d'objets, HUBBLE rangea les galaxies dans des classes afin d'en faciliter la dénomination et d'obtenir une vue d'ensemble plus simple pour l'esprit. Il identifia les elliptiques, les spirales et les irrégulières. De plus, il distingua parmi les spirales celles qui présentaient une barre. Ce qui lui fournissait quatre classes en tout. Cette classification, simple et visuelle, a eu un succès certain puisqu'elle est encore largement utilisée de nos jours.

## 2.2 Limites de la méthode traditionnelle

### 2.2.1 La floraison des classifications au Moyen-Âge

Avec le développement des voyages au Moyen-Âge, de très nombreux organismes vivants nouveaux étaient découverts. Le même concept de classification restait. Ainsi, pendant quelques siècles, chaque naturaliste développait sa propre classification à partir de son carnet d'observation. Chacun prenait des critères qu'il jugeait importants,

donnant à sa guise des noms le plus souvent liés à ces critères. Pour l'un, la couleur des fleurs primait, pour un autre c'était la structure des feuilles, pour un troisième les fruits constituaient un paramètre fortement discriminatoire.

C'est ainsi qu'une multitude de classifications vit le jour, toutes étaient basées sur des critères apparents et restreints, choisis très subjectivement. Parfois même, une interprétation a priori guidait le scientifique dans la définition des classes. De plus, la plupart des classifications ne comprenaient pas tous les organismes, et la plupart d'entre eux se retrouvaient appartenir à plusieurs classifications sous des noms différents. La situation devint intenable au XVIIIième siècle. En 1763, ADANSON a rédigé un merveilleux livre de revue sur l'ensemble des classifications ayant existé ou existant à l'époque (ADANSON, 1763).

Les noms des organismes étaient assez souvent multiples, voire consistaient en de petites phrases, permettant de décrire un organisme vivant selon les quelques critères apparents choisis. Cela constituait une sorte de classification multivariée, mais pas du tout objective ni commode : à partir du moment où une classe rassemble les plantes à fleurs rouges, il devient absolument impossible d'inclure une plante à fleurs jaunes même si elle est identique par ailleurs. Il est donc impossible de déceler éventuellement d'autres paramètres plus fondamentaux. Ainsi, cette manière de nommer les choses selon des descripteurs en nombre restreint est sclérosante pour obtenir une vision globale de la diversité qui puisse s'adapter à l'évolution des connaissances.

Nous allons voir dans la Sect. 2.3 suivante que le même problème se pose aujourd'hui pour les galaxies.

### 2.2.2 Les autres classifications des galaxies

Plusieurs ouvrages récents présentent un tour d'horizon de nos connaissances sur les galaxies, mais bien peu abordent le problème de la classification avec une perspective historique. Le livre de VAN DEN BERGH (1998) est à la fois une exception et également un révélateur on ne peut plus clair de la situation actuelle : la classification des galaxies n'est jamais envisagée autrement que sous l'aspect morphologique. Pourtant il existe bien d'autres classifications, et même une multitude d'autres. Cependant, elles n'atteignent pas le statut généraliste de la classification de HUBBLE, bien que fondées sur le même principe.

En effet, celles-ci rangent les galaxies dans des boites correspondants à un ou deux descripteurs. Ainsi, selon la longueur d'onde, selon le moyen observationnel utilisé, le critère de classement sera différent. On établit également des catégories en fonction du processus physico-chimique étudié en utilisant des paramètres issus de modèles eux-mêmes contraints par les observables. Concrètement, chaque échantillon de galaxies observées d'une manière donnée se verra donc catalogué et éventuellement rangé dans des boites afin d'en faciliter l'interprétation et la modélisation.

D'une manière plus poussée, des corrélations entre paramètres sont recherchées. Ces corrélations sont en effet plus contraignantes pour les modèles, et permettent dans certains cas de connaître les influences réciproques entre différents phénomènes à l'œuvre dans les galaxies. Ainsi a été découvert le plan fondamental (TREU ET AL., 2001) qui regroupe les galaxies elliptiques géantes dans un espace à trois observables (le rayon effectif, la dispersion de vitesse et la brillance de surface à ce rayon) en un plan. Ce n'est rien d'autre qu'une corrélation à trois variables, mais c'est déjà un

progrès par rapport à la classification de HUBBLE à un seul paramètre. Cependant les corrélations ne fournissent de classification qu'à travers un découpage plus ou moins arbitraire du diagramme correspondant. De toutes manières, il s'agit encore d'une classification portant sur un très petit nombre de descripteurs.

Nous verrons dans la section suivante (Sect. 2.3) comment ce genre d'approche conjuguée à la nature même des galaxies produit nécessairement une profusion de classifications. Devant la multiplication des moyens d'observations, les astrophysiciens ont donc inventé plusieurs centaines de noms, souvent sous forme d'acronyme. Une véritable zoologie est apparue en plus des types morphologiques classiques : galaxie naine, de champ, isolée, Cd, starburst, radio, ULIRG, elliptique disky ou boxy, irrégulière, merger, quasar, active (SEYFERT, BLLac, OVV, blazar, FARANOFF-RILEY, X, TeV,...), bleue, compacte, LYMAN $\alpha$, etc. Ils ont été amené à mettre au point plusieurs dizaines de classifications dont aucune ne prétend rendre compte de la diversité complète des galaxies dans toutes les observables. En conséquence, une même galaxie, appartiendra non pas à une classe, mais à plusieurs, voire beaucoup, impliquant d'inévitables incompatibilités. Ceci n'est pas sans rappeler la situation de la biologie au Moyen-Âge (Sect. 2.2.1).

## 2.3 Nature et diversité des galaxies

### 2.3.1 Constituants fondamentaux des galaxies

La classification de HUBBLE voit les galaxies comme une simple morphologie, c'est-à-dire comme la forme d'une chose non précisée. Pourtant, déjà pour HUBBLE, les galaxies étaient constituées d'étoiles puisque c'est grâce à certaines d'entre elles qu'il a pu mesurer leurs distances. La morphologie reflète donc la distribution de l'ensemble des constituants fondamentaux d'une galaxie dont les étoiles ne sont qu'une partie.

#### Les étoiles

Les étoiles sont le constituant le plus visible, le plus évident, des galaxies. Ce sont des objets dont on comprend relativement bien la diversité, ainsi que les processus de formation et d'évolution. Nous sommes capables de synthétiser des populations stellaires dans des galaxies, cela permet même d'estimer leur distance grâce au redshift photométrique. Mais cette connaissance théorique nous montre aussi la complexité de leur histoire collective au sein d'une même galaxie, au point que la détermination de la Fonction Initiale de Masse des étoiles est une thématique majeure de l'astrophysique stellaire actuelle (KROUPA, 2002). Est-ce que cette fonction est universelle ? Ou est-elle différente d'une galaxie à l'autre, voire d'un endroit à l'autre dans une même galaxie ?

Les observations, quant à elles, ont apporté récemment la preuve que les étoiles au sein d'une même galaxie pouvaient avoir des histoires différentes, des provenances différentes à la suite de fusions entre galaxies ou l'accrétion de galaxies plus petites. À l'évidence il n'est pas raisonnable de dater une galaxie en se basant simplement sur la population stellaire moyenne, globale, car il est maintenant établi que plusieurs épisodes de formation stellaire ont eu lieu au sein d'un même objet. Il est nécessaire

de recenser finement les différentes populations, leurs âges, leurs métallicités, leurs mouvements au sein de la galaxie, afin d'en préciser éventuellement leurs origines, et déterminer les évènements qui ont amené à la formation de la galaxie telle que nous l'observons.

Pour décrire correctement une galaxie, il faut donc prendre tout cela en compte. Les observations modernes vont indéniablement dans ce sens, complexifiant grandement les catalogues, les bases de données, et en conséquence l'utilisation synthétique de toute cette masse d'information.

### Le gaz

Le gaz est le constituant le plus primordial qu'on puisse imaginer dans l'histoire des galaxies, car c'est à partir de lui que les étoiles et la poussière se forment. Nous savons qu'il en existe beaucoup en dehors des galaxies telles que nous les voyons. Le gaz est plus malléable que les étoiles. Plus volatile, il passe facilement d'une galaxie à une autre, se comprime, se réchauffe ou se refroidit, rayonne, absorbe, se transforme. Il dépend beaucoup de son environnement immédiat, en particulier la présence d'étoiles chaudes. Sa composition chimique est intimement lié à l'histoire des étoiles. Nous savons la complexité structurelle des nuages de gaz, qu'il soit sous forme ionisé ou atomique.

Le gaz, tout comme la poussière, est moins facile à observer, car moins lumineux et moins localisé, que les étoiles. Mais est-ce que tous les détails sont importants pour décrire les particularités d'une galaxie ? Il n'est pas évident de répondre à cette question aujourd'hui. Le dénombrement, les caractéristiques physiques et la répartition des différentes régions de gaz sont sans aucun doute importantes pour raconter l'histoire de la galaxie. En revanche, il est probable que la structure fine de chacune de ces régions n'est pas pertinente quant à la spécificité de la galaxie hôte.

### La poussière

Certainement un peu moins malléable que le gaz, il n'en demeure pas moins que la poussière, composée de molécules et de grains, est plus complexe à décrire étant donnée la richesse quasi infinie de la chimie moléculaire. Sa distribution observée raconte visiblement l'histoire mouvementée des galaxies. On a souvent l'impression de voir les résidus de fusions de galaxies, dont l'une semblerait très poussiéreuse. Cette vision est sans doute trop simpliste, mais les processus de formation de la poussière peuvent être assez sophistiqués et bénéficient en tout cas de la grande variété des conditions physiques régnant dans le milieu interstellaire.

Longtemps gênante car absorbant la lumière visible, la poussière est devenue un ingrédient fondamental dans la description des galaxies et plus particulièrement pour les processus de formation des étoiles. Ces dernières se forment en effet au sein de nuages moléculaires dont la structure est très complexe. Ces zones de turbulences ne sont pas encore bien modélisées, de sorte qu'on ne sait pas si la Fonction Initiale de Masse est identique dans toutes les galaxies, et même au sein d'une galaxie. Bien que la description de la structure fine dans les zones de formation d'étoiles puisse paraître superflue car se déroulant sur des échelles spatiales et temporelles trop petites par rapport à l'échelle d'évolution d'une galaxie, les spécificités des processus de formation

d'étoiles ne sont peut-être pas universelles, et pourraient dépendre de chaque galaxie et de son histoire.

Les réactions chimiques peuvent être complexes et nécessitent des conditions très spécifiques d'environnement. La composition de la poussière est donc une sonde précise de l'état du milieu interstellaire et de son évolution. L'observation de molécules dans des galaxies très lointaines, même si faute de résolution spatiale elle ne fournit que des données globales, est précieuse non seulement pour cartographier la diversité des galaxies dans ce domaine, mais également pour aider à tracer différentes histoires à l'origine de cette diversité.

#### Les trous noirs

Ce n'est que récemment que les trous noirs ont été découverts et caractérisés au centre de certaines galaxies, au point qu'on pense aujourd'hui qu'ils sont présents dans toutes les galaxies, à l'exception sans doute des galaxies naines. Je ne parle pas ici des micro trous noirs, dont la masse est de l'ordre d'une masse solaire, qui sont la fin de vie normale des grosses étoiles et sont décrits au travers de la composante stellaire des galaxies. Il s'agit ici de trous noirs massifs, de quelques millions à quelques milliards de masses solaires. Ils "pèsent" donc tellement que leur rôle dans la galaxie est nécessairement important.

Que les galaxies se soient formées autour des trous noirs issues des premières grosses étoiles de l'Univers, ou que ces trous noirs ce soient formés plus tard au cœur des galaxies, prouve que les trous noirs supermassifs constituent très probablement une composante fondamentale de l'évolution des galaxies. Plus le trou noir central est massif, plus l'ensemble de la galaxie est autogravitant et résiste mieux aux perturbations extérieurs. Si le trou noir a grossi, c'est que de la matière lui est parvenue : comment, en quelle quantité, à quel rythme ? Les réponses à ces questions concernent l'histoire des galaxies au plus haut point. On mesure désormais des masses de trous noirs supermassifs au sein de plus en plus de galaxies. Bientôt, il faudra introduire cette composante dans la description d'une galaxie, ce qui n'est pas fait dans ce livre.

#### Le champ magnétique

Nous n'aborderons volontairement pas le champ magnétique dans ce livre. Il est indéniablement présent dans les galaxies, mais son observation reste difficile. Le rôle de cette composante dans le comportement des constituants fondamentaux des galaxies est certainement important, par exemple au cours de la formation des étoiles ou la répartition du gaz à petite ou grande échelle. Cependant, à ce jour, il n'est pas du tout évident que le champ magnétique soit un constituant à part des galaxies au même titre que les étoiles ou la poussière. L'inclure dans la description d'une galaxie pourrait peut-être présenter une certaine redondance avec les propriétés détaillées des populations stellaires, du gaz et de la poussière. Cela reste une question ouverte que nous ne discuterons pas davantage dans le présent ouvrage.

### 2.3.2 Les observables des galaxies

Nous venons de voir les constituants fondamentaux des galaxies, et on imagine aisément les descripteurs qu'il faudrait pour caractériser entièrement une galaxie. Concrè-

tement, de quelles observables correspondantes disposons-nous aujourd'hui ?

### Imagerie

Secteur traditionnel de l'astrophysique extragalactique, l'imagerie fournit d'abord une description visuelle de la répartition projetée sur le plan du ciel (en deux dimensions donc) de la lumière correspondant au domaine de longueur d'onde utilisé. Parce qu'une image touche davantage l'imaginaire qu'une courbe, ce secteur est l'outil idéal de communication. Mais une image contient aussi une grande quantité d'information et l'oeil est un merveilleux outil pour détecter des gradients ou des structures à faible contraste. Cependant, l'imaginaire étant omniprésent, la description d'une image contient inévitablement une part indéfinie de subjectivité. Il suffit pour cela de se souvenir de l'histoire des canaux de Mars. L'informatique nous aide aujourd'hui dans la reconnaissance de forme, d'importants efforts sont fournis dans ce sens, mais il faut bien avouer qu'elle déçoit parfois parce qu'elle n'atteint pas les performances qu'on espère, à savoir remplacer l'œil humain, et surtout sa subjectivité !

Alors la classification morphologique des galaxies est encore faite à l'œil, enfin pour ceux qui utilisent la classification de HUBBLE et ses perfectionnements. Cette classification est ainsi rarement univoque, car les observateurs ne sont pas toujours d'accord entre eux. Ceci démontre les limites de la classification morphologique, limites finalement révélées par les techniques informatiques et apparemment difficile à accepter par les observateurs. Pourtant, il serait aujourd'hui largement préférable de se contenter de l'objectivité de l'ordinateur et d'intégrer d'autres informations tellement précieuses.

Car ce qui importe dans la description d'une galaxie, dans ce qui relève de son histoire, ce n'est pas tant le nombre et la structure fine des bras spiraux que la répartition des étoiles, du gaz, de la poussière, des gradients de composition chimiques, de densité, de température, de taux d'ionisation. C'est donc à toutes les longueurs d'onde qu'il faut décrire la morphologie d'une galaxie, des rayons X (voire $\gamma$) aux ondes kilométriques, en passant par l'ultraviolet, le visible, les infrarouges proche et lointain, le submillimétrique et la radio millimétrique, centimétrique et métrique, que ce soit en bande spectrale large ou étroite. Seuls des indicateurs de distribution et de gradients peuvent fournir des descripteurs objectifs et surtout quantitatifs.

La richesse d'information apportée par l'imagerie seule dépasse donc plus que largement les types morphologiques de la classification de HUBBLE. Mais l'imagerie n'est pas la seule source d'information, et n'est finalement peut-être pas la plus importante.

### Photométrie

De nos jours, il n'y a plus d'imagerie effectuée sans une calibration en intensité. L'imagerie est donc systématiquement de la photométrie en deux dimensions. Il s'agit de mesurer la quantité de lumière reçue à travers le filtre d'observation, qui peut être large ou étroit. Dans le premier cas, on s'intéresse au rayonnement continu, thermique ou non, et l'information la plus pertinente pour les galaxies se trouve dans les couleurs, c'est-à-dire la pente du spectre. Ces couleurs traduisent directement la qualité thermique ou non du rayonnement. S'il est thermique, elles donnent une estimation

de la température, ce qui dans le cas des étoiles fournit grosso modo leur âge et/ou leur métallicité. Dans le domaine infrarouge, on peut en déduire la température de la poussière. S'il s'agit d'un rayonnement non thermique, les couleurs aident à déterminer la composition et les conditions physiques du plasma, avec parfois une indication de l'intensité du champ magnétique.

La photométrie intégrée n'existe presque plus, donc la photométrie peut être toujours considérée comme de l'imagerie calibrée en flux. Ce qui signifie que la description des structures, quelque soit la longueur d'onde d'observations, est nécessaire. Il semble ainsi évident que la complexité de la répartition des gradients de couleurs et des variations des paramètres physiques dans une galaxie impose l'usage d'outils mathématiques et informatiques assez sophistiqués.

### Spectroscopie

Avec des filtres à bande très étroite, on caractérise des éléments chimiques (atomes ou molécules) grâce aux raies d'émission spécifiques de chaque élément, dont les longueurs d'onde précises sont déterminées en laboratoire, ou parfois calculées théoriquement. Ces raies peuvent être vues en émission ou en absorption selon la répartition des sources lumineuses le long de la ligne de visée. Outre la présence de ces éléments, la détection et la mesure de l'intensité de plusieurs raies d'un même élément peut fournir sa densité, sa température et les conditions d'environnement dans lequel il se trouve. Les rapports de raies de différents éléments sont également utiles pour contraindre plus fortement les conditions physiques de l'environnement, mais surtout ils permettent de déterminer la composition chimique du milieu interstellaire et de comparer différentes régions ou différentes galaxies.

La spectroscopie fournit une information précieuse : la vitesse, qui se mesure par l'élargissement et le décalage Doppler des raies. L'élargissement donne l'état d'agitation interne à la région, alors que le décalage donne le mouvement de la région par rapport à l'observateur. Il est ainsi possible de reconstruire partiellement la troisième dimension en utilisant les lois de la gravitation au sein de la galaxie. Cette information cinématique est cruciale, car l'étude des différents mouvements au sein d'une même galaxie peut révéler l'origine des régions en question. C'est ainsi qu'une origine externe, c'est-à-dire par accrétion ou fusion d'une autre galaxie, peut être établie. Il s'agit donc à l'évidence d'un excellent traceur de l'histoire passée.

Il faut noter que de nouveaux détecteurs sont maintenant capable de spectro-imagerie, ce qui revient à obtenir une image dans laquelle chaque élément est un spectre. La quantité d'information est colossale, le traitement des données délicat et très gourmand en temps de calcul, mais on possède là sans doute l'outil idéal pour décrire une galaxie de manière très complète.

### Paramètres indirects

En dehors des observables directs dont nous venons de parler, il existe des descripteurs intéressants qui nécessitent non seulement une combinaison d'observables, mais également quelques hypothèses supplémentaires comme la configuration à trois dimensions, la distribution spatiale (le facteur de remplissage) ou encore l'inclinaison par rapport à la ligne de visée. Parfois de véritables modèles sont utilisés. Parmi ces pa-

ramètres, on trouve par exemple les masses (totale, du gaz, du trou noir central). Dans ces cas là, il convient d'être prudent dans l'utilisation de ces descripteurs et de bien connaître les éléments subjectifs et hypothétiques susceptibles d'influer grandement sur leur signification et leur impact.

#### Les paramètres globaux

Nous sommes habitués à décrire une galaxie avec des paramètres globaux, c'est-à-dire intégrés sur toute la galaxie, tels que la morphologie, la luminosité dans différentes bandes larges, les couleurs, le diamètre ou même la vitesse de rotation maximale. Ceci était incontournable il y a encore quelques dizaines d'années, et reste bien pratique pour des galaxies lointaines qui apparaissent comme de petites taches sur nos détecteurs modernes. Ceci est également bien pratique pour mettre dans le même sac des objets qui se ressemblent grossièrement.

Compte tenu de ce que nous savons aujourd'hui des galaxies, il devient évident que ces observables globales ne décrivent qu'un résumé des propriétés des constituants fondamentaux. Les couleurs, par exemple, donnent l'âge et la métallicité moyenne de l'ensemble des étoiles d'une galaxie, lissant ainsi les spécificités des différentes populations stellaires présentes dans une galaxie. En aucun cas avons-nous accès, avec ce genre de paramètres, à toute la complexité, et donc toute la richesse, des différents constituants des galaxies et encore moins toute leur histoire. La diversité des galaxies ne saurait se décrire en terme uniquement de paramètres globaux.

### 2.3.3 Diversité des galaxies

Avec l'imagerie seule, dans le domaine visible et une région de l'Univers limitée, HUBBLE eut tôt fait de classer les galaxies en quatre catégories basées sur des éléments évidents à l'œil : elliptiques, spirales, spirales barrées et irrégulières. Ce découpage, fondé sur une description qualitative, perdure grâce à la magie de l'image, à l'extrême simplicité de cette classification, et aux quelques grandes corrélations entre cette morphologie et certaines observables quantitatives.

De ce qui précède, il est évident que cette classification est trop limitative et ne peut en aucun cas cartographier la diversité des galaxies telle qu'elle apparaît aujourd'hui. Déjà au niveau de la morphologie dans le domaine visible, la classification de HUBBLE ne rend pas compte, loin s'en faut, des diversités de formes et de structures, en particulier pour les galaxies les plus lointaines (VAN DEN BERGH, 1998). De plus, nous avons aujourd'hui de superbes images dans presque toutes les longueurs d'onde, des rayons X à la radio centimétrique au moins. Et il est le plus souvent impossible de reconnaître une galaxie lorsqu'on change de longueur d'onde.

Bien au-delà de la simple morphologie, nous venons de voir l'impressionnant arsenal d'observables dont nous disposons, ce qui nous offre la possibilité de décrire en détail les différents constituants fondamentaux des galaxies. Tous ces descripteurs sont quantitatifs, nous trouvons des continua de distributions, chacune des observables semblant pouvoir prendre presque n'importe quelle valeur. Même quand la morphologie est quantifiée sous la forme du rapport de la taille de bulbe par rapport à la taille du disque, toutes les variations semblent possibles et la limite entre elliptique et spirale s'estompe. Ainsi, pour chaque descripteur, la diversité des galaxies est importante,

rendant problématique toute classification.

### 2.3.4 Les galaxies en tant qu'objets complexes

Que se passe-t-il quand on rassemble toutes les observables décrites ci-dessus ? Le constat est qu'il n'y pas encore de méthodologie pour considérer globalement toute cette masse d'information. Alors de nombreuses combinaisons de quelques descripteurs sont effectuées, et par recherche de corrélations fortes, les astrophysiciens tentent de ranger les galaxies dans une poignée de catégories, catégories dépendant très étroitement des observables. Et c'est bien là qu'apparaît toute la difficulté actuelle de la physique des galaxies : il existe une zoologie abondante, peu synthétique, au point qu'il est devenu impossible de parler des galaxies sans faire appel à une liste de paramètres, souvent choisis arbitrairement. La classification morphologique de HUBBLE se distingue principalement par son côté historique.

Il est donc devenu évident aujourd'hui que, désormais, les galaxies sont des objets très complexes à décrire. Il est encore plus complexe de modéliser l'ensemble des phénomènes évolutifs au sein des galaxies. Nous semblons connaître la plupart des processus physiques et chimiques qui régissent le gaz, la poussière, les étoiles, les trous noirs, et leurs interactions. Nous pensons donc comprendre beaucoup de processus élémentaires d'évolution dans les galaxies. Mais aucun modèle, aucune simulation numérique n'est encore capable de tout prendre en compte pour synthétiser un objet qui ressemble à ce qu'on observe. Et cette complexité s'accroît lorsqu'on considère les environnements variés et changeants des galaxies. Nous verrons cette question au Chapitre 3.

Mais il existe encore un autre niveau de complexité. Car même en imaginant qu'il serait possible un jour de tout modéliser grâce aux ordinateurs, les processus en jeu sont tout simplement tellement nombreux et tellement aléatoires que fondamentalement il est impossible de prédire l'évolution d'une galaxie. C'est bien en cela que les galaxies sont réellement des objets complexes, un peu à l'instar de la météorologie dont la physique de base est très bien connue et comprise mais dont pourtant la prédictabilité n'est qu'à courte échéance et toujours quelque peu incertaine. Cela n'empêche pas de classer les différents objets météorologiques, de parfaitement comprendre les raisons de leurs évolutions. Cela est bien pire en biologie, où les processus élémentaires d'évolution sont mal compris et pas toujours bien identifiés, en tout cas très difficilement modélisables. Pourtant l'histoire de la diversité des espèces vivantes est intelligible, en grande partie grâce à DARWIN, mais aussi, et surtout, grâce au développement de méthodologies d'analyses adaptées.

## 2.4 Les classifications multivariées

### 2.4.1 LINNÉ propose une nouvelle nomenclature

Durant la première moitié du XVIIIième siècle, LINNÉ élabora une nomenclature binomiale latine afin de résoudre les nombreux problèmes entre les classifications de l'époque. Il introduit l'idée de donner des noms génériques, vagues, aux organismes vivants, comme par exemple le nom d'un lieu ou d'un botaniste ayant découvert la plante, ou encore le nom commun d'usage d'un animal. Cette règle de taxonomie

distingue donc clairement le nom de la description. Le latin était choisi car c'était la langue scientifique de l'époque.

La nomenclature de LINNÉ est binomiale, parce cette structure hiérarchique en genre et espèce avait déjà été proposée par quelques botanistes, et semblait bien adaptée à la description de la diversité des organismes vivants. Il stipula que le nom de genre devait être générique, sans signification particulière, et que le nom de l'espèce servait à préciser le premier et pouvait, le cas échéant, être un peu plus spécifique. En réalité, le système de LINNÉ traduisait ouvertement un constat qui émergeait à l'époque : la diversité du monde vivant était organisée sous forme hiérarchique.

Le système de LINNÉ eut à la fois beaucoup de succès mais rencontra également de virulents détracteurs. Pourtant, il est tout de même extraordinaire que cette nomenclature soit toujours utilisée et assez performante de nos jours, alors que DARWIN et la biologie moléculaire, entre autres, ont quelque peu bouleversé nos connaissances entre temps !

### 2.4.2 Vers une classification naturelle et objective

Au-delà de la simple nomenclature, ce sont les concepts de classification qui sont revus au XVIIIième siècle. ADANSON (1763) définit les méthodes de classification en deux types : les naturelles et les artificielles. Les méthodes naturelles sont celles qui débouchent sur la découverte de l'ordre naturel des choses. Les méthodes artificielles sont celles qui facilitent la tâche des scientifiques. Ces dernières sont plus faciles à mettre en œuvre parce que c'est l'auteur qui dicte l'agencement et non pas la Nature. En conséquence, la méthode naturelle doit nécessairement prendre en compte tous les descripteurs disponibles, sans choix arbitraire et a priori de la part du scientifique. Il est clair pour lui que les classifications qui reposent sur un ou deux paramètres ne peuvent qu'être artificielles puisque les objets ainsi regroupés sont la plupart du temps hétéroclites. Afin d'illustrer ces deux concepts de classification, il cite le système de Copernic qui, de méthode artificielle, (on parlerait aujourd'hui de modèle), est devenu système naturel, voire la chose elle-même, la réalité à l'état "pur".

Ces notions nouvelles ont révolutionné la biologie et toute la cartographie de la biodiversité. Ce sont probablement les prémices des analyses de distance multivariée qu'ADANSON et ses contemporains ont inventés. Ils ont en tout cas engendré une nouvelle discipline de la biologie, la systématique, c'est-à-dire la science de la classification, qui occupe encore beaucoup les scientifiques du XXIième siècle. Nous présenterons un aperçu de ces méthodes au Chapitre 9.

### 2.4.3 Analyses multivariées en astrophysique

Habitués maintenant aux grandes masses de données, les astrophysiciens ont utilisé des méthodes multivariées pour analyser les catalogues et produire des classifications (ROBERTS AND HAYNES, 1994). L'objectif est généralement d'automatiser le rangement des objets dans des classes prédéfinies. Par exemple il peut s'agir de la reconnaissance de forme sur des images grands champs permettant un tri automatisé des étoiles, des quasars et même des différents types morphologiques de galaxies.

Très peu de tentatives ont cependant été effectuées dans un esprit de classification générale des galaxies. Des pistes fort intéressantes ont été explorées. L'une

d'entre elles consiste en la caractérisation automatique de spectres haute résolution (CONNOLLY ET AL., 1995; CABANAC ET AL., 2002; LU ET AL., 2006). Cette approche présente dans le principe deux avantages principaux : automatisation et complétude de la description. Ces deux points assurent une objectivité maximale, l'automatisation permettant en plus de couvrir de très larges échantillons. La complétude dépend certes de la qualité des données et des logiciels, mais toutes les structures du spectre sont enregistrées sans tri a priori. Par rapport à la classification de HUBBLE et toutes les classifications monoparamètres décrites précédemment, le progrès est énorme car une comparaison détaillée sur l'ensemble du spectre de chacun des objets est possible. Il ne faut pas oublier que toutes les informations dont nous pouvons disposer sur une galaxies proviennent de la lumière et se trouvent donc nécessairement inscrites dans le spectre, à la condition de disposer des spectres sur toute la surface apparente de la galaxie. Ainsi même la morphologie est essentiellement contenue dans l'information cinématique.

Une fois que l'on dispose de l'ensemble des observables imaginables avec nos moyens observationnels actuels, il est possible de comparer les objets sur l'ensemble de ces critères. Pour cela, une analyse de distance multivariée semble bien adaptée, en définissant une distance qui en général sera une simple distance quadratique sur tous les paramètres (ELLIS ET AL., 2005). Il est clair que ce type d'approche est désormais le minimum indispensable pour couvrir la diversité observée des galaxies.

Quelques tentatives ont été faites dans ce sens (en particulier WHITMORE, 1984; WATANABE ET AL., 1985). En utilisant une décomposition en composantes principales, elles aboutissent à un certain nombre de boites, caractérisées par des combinaisons d'observables, dans lesquelles sont rangées les galaxies d'une manière assez univoque. Par exemple, ils constatent que la taille et morphologie semblent plus discriminants que les autres paramètres.

Toutefois, aucune nouvelle nomenclature et/ou classification n'a émergée de ces travaux. Pourquoi ? La classification à partir d'un ou deux paramètres permet de ranger les objets dans des boites parallèles, sans lien entre elles. On peut faire une classification plus objective et large en prenant tous les paramètres, par une analyse de distance multivariée. Ainsi on trouve d'autres boites, toujours sans lien entre elles. Le résultat reste assez stérile physiquement : quand on prend un paramètre, on peut chercher à comprendre son origine, son évolution, grâce à des modèles. Mais quand on prend un ensemble de paramètres indépendants, on ne sait plus faire grand chose car ces paramètres ne sont pas nécessairement liés causalement par des processus physico-chimiques. Et lorsque ces boites sont définies par des combinaisons linéaires d'observables, on perd même toute signification physique de ce qui caractérise les objets regroupés ensemble !

Le Chapitre 9 montrera à quel point la tâche est en réalité très complexe. Les analyses multivariées sont suffisamment sophistiquées pour requérir le concours des statisticiens. Des analyses par regroupement très pertinentes ont par exemple été effectuées sur les amas globulaires (CHATTOPADHYAY AND CHATTOPADHYAY, 2007; CHATTOPADHYAY ET AL., 2008, 2009a,b), les gamma-ray bursts (CHATTOPADHYAY ET AL., 2007) et les galaxies spirales (CHATTOPADHYAY AND CHATTOPADHYAY, 2006).

Les analyses multivariées ont pour le moment échoué à fournir une nouvelle classification globale des galaxies car elles ne proposent pas de structures d'organisation

entre les boites. Il semble bien que le seul moyen de lier ces boites soit le paramètre temps, donc l'évolution. Mais cela est difficile, car il y a une multitude de paramètres et de processus physiques et chimiques dans une galaxie, et la combinaison des caractéristiques fournie par l'analyse multivariée ne prend pas du tout cela en compte. La comparaison n'est que globale, les différences étant moyennées sur toutes les observables, indépendamment de toute considération évolutive. Il faut ensuite comprendre pourquoi ces boites existent, c'est-à-dire pourquoi une galaxie se retrouve dans une boite, si elle peut changer de boite, et comment. Seule l'analyse fine des évolutions de chaque paramètre via les différents processus physico-chimiques de déroulant au sein d'une galaxie peut nous renseigner. La modélisation de tous les paramètres à la fois est bien souvent trop complexe et insuffisamment contrainte. Nous verrons que le couplage de ces méthodes multivariée avec une approche évolutive du type cladistique offre un éclairage fécond dans l'interprétation des groupes mis en évidence (Sect. 9.5). Mais il nous faut tout d'abord analyser en détail l'influence considérable de l'évolution sur la notion même de classification.

# Chapitre 3

# Évolution et environnement

Dans ce chapitre, nous allons voir que l'évolution est le lien fondamental derrière la diversité et la complexité constatées au Chapitre 2 précédent. C'est bien évidemment DARWIN qui a compris cela pour les organismes vivants, et William HENNIG un siècle plus tard qui a ouvert la voie des méthodes phylogénétiques visant à construire des classifications représentant ces liens. Pour les galaxies, nous allons ici mettre en évidence le fait que la diversité des galaxies est en grande partie due à l'évolution.

Le rôle de l'environnement dans le monde vivant est relativement bien identifié. Il est lui-même en évolution qui à la fois influe sur et dépend de l'évolution de l'organisme considéré. L'Univers, qui constitue l'environnement des galaxies, commence seulement à être suffisamment compris pour pouvoir, là encore, apercevoir un parallèle instructif entre les microcosmes biologiques et les évolutions comparées d'une galaxie et de son environnement gravitationnel.

L'objectif de ce chapitre est donc de placer ces deux ingrédients nécessaires à la logique, espérons-le rendue naturelle, du développement de l'astrocladistique. Afin de synthétiser la problématique de la représentation de la diversification des galaxies, avant de plonger plus en détails dans la méthode dans le reste de ce livre, nous résumerons en fin de chapitre les enseignements de la systématique à propos de la classification d'objets complexes en évolution.

## 3.1 Classification évolutive

### 3.1.1 DARWIN et l'organisation hiérarchique de la diversité

Il faudra attendre 1850 et DARWIN pour comprendre l'origine de l'organisation hiérarchique de la diversité des organismes vivants, si bien décrite par la nomenclature linnénne et la classification sous-jacente (Sect. 2.4.1). Car c'est bien l'évolution qui explique cette structure en genre, espèce et sous-espèce, ainsi que d'autres niveaux plus élevés (famille, ordre). C'est plus particulièrement le mécanisme darwinien de transmission avec modification qui en est la cause.

Un organisme, quel qu'il soit, se reproduit (la réplication), d'une manière ou d'une autre, par sexuation ou duplication, en donnant naissance à un autre individu. Mais cette réplication n'est pas un clonage parfait et lors de la transmission, il y a modification ou innovation. Ceci reste vrai, aussi bien pour la sélection naturelle de DAR-

WIN (adaptation), ou l'influence de l'environnement de LAMARCK (acclimatation) ou encore les équilibres ponctués de GOULD(voir quelques références dans VIGNAIS, 2001). La transmission assure la ressemblance, donc le lien de parenté, et génère l'ossature de la structure hiérarchique. Les modifications engendrent la diversité en créant les embranchements engendrés par les innovations.

Chaque nouvel individu va être à la fois très ressemblant et un peu différent de ses parents. D'un côté, il va partager avec eux une innovation héritée d'un ancêtre plus éloignés, et d'un autre côté, il va posséder un ou plusieurs traits innovants qui pourront peut-être se propager dans toute la lignée de ses descendants. Graduellement, de génération en génération, la ressemblance deviendra moins évidente avec les parents plus ancestraux, et il sera possible, sur des critères arbitraires, de décider que le nouvel individu appartient à une nouvelle espèce. Cette continuité ne nous est pas accessible autrement que par certaines expérimentations de laboratoires grâce à des organismes qui se renouvellent très vite à l'échelle humainie, car nous ne pouvons pas identifier la totalité des individus qui constituent une suite généalogique complète sur des millions d'années.

La reconstitution de l'histoire évolutive de la diversité des organismes vivants range donc les espèces selon leurs liens de parenté et leur degré de divergence (voir Sect. 3.3.3), sachant que de nombreux chaînons sont certainement manquants. Ceci nous donne la fausse illusion que les espèces apparaissent brutalement, et que les frontières sont bien définies et hermétiques. La réalité de l'évolution des espèces est toute autre, ainsi qu'il peut en être vérifié en laboratoire.

### 3.1.2 William HENNIG invente la cladistique

La découverte de DARWIN (DARWIN, 1859) a permis d'expliquer un constat observationnel concernant la répartition hiérarchique de la diversité particulièrement bien décrite par la nomenclature de LINNÉ. Mais la classification était fondée sur une sorte d'analyse de distance multivariée, totalement déconnectée de la notion d'évolution. Un siècle après DARWIN, William HENNIG, en 1950, mis au point la cladistique (HENNIG, 1965), méthodologie de classification qui compare les objets non pas sur leur ressemblance, même globale, mais sur leur rang dans l'évolutiion, rang défini par l'ensemble des états évolutifs des descripteurs des organismes vivants (Chapitreřefchapintroclad).

Puisque l'évolution se fait par transmission avec modification, ce sont donc les caractéristiques héritées en commun qui vont rapprocher les organismes vivants. En quelque sorte, il ne s'agit plus vraiment de savoir si les espèces se ressemblent, mais plutôt de savoir pourquoi elles se ressemblent. Les différences sont ici remplacées par la notion de divergences des chemins évolutifs. Les objets d'étude seront décrits avec tous les descripteurs disponibles, leurs valeurs étant considérées comme des états évolutifs. En plaçant ainsi les objets de proche en proche, le schéma organisationnel qui en découle est un arbre illustrant les divergences des chemins évolutifs, représentation issue de la théorie des graphes. Cette méthodologie s'est largement imposée depuis les années 1980 et produit la plupart des arbres de la vie aujourd'hui.

La cladistique a quelque peu révolutionné la classification des organismes vivants en faisant éclater certains groupes pourtant bien connus (exemple des poissons) et en rapprochant des espèces a priori très différentes (oiseaux et dinosaures). Elle a cer-

tainement changé profondément la vision de la classification, qui était très liée à une image un peu poussiéreuse de la zoologie descriptive. Désormais la classification est un outil dynamique et évolutif, s'adaptant parfaitement à l'évolution des connaissances, devenu encore davantage un instrument indispensable à la compréhension de la diversité.

Mais la cladistique est avant tout un outil d'analyse de données, fondée sur le concept de regroupement d'objets pouvant déboucher sur une classification évolutive. Elle peut s'appliquer à tout groupe d'objets complexes en évolution, et à tout type de descripteurs à condition qu'ils caractérisent l'état évolutif de ces objets. Ainsi, elle s'accommode très bien des données moléculaires dont nous parlons ci-dessous.

### 3.1.3 Les classifications génétiques

La biologie moléculaire a pris son essor vers le milieu du XXième siècle. Elle a permis de fournir non plus des descripteurs morphométriques, le plus souvent décrits à l'œil, mais des descripteurs concernant les constituants fondamentaux des organismes vivants, à savoir les gènes (VIGNAIS, 2001). En décrivant les séquences de nucléotides, il est possible de comparer objectivement deux organismes vivants. Les gènes, portés par l'ADN, sont le vecteur de la transmission, donc des liens de parenté. En conséquence, les différences sont le résultat d'une modification apparue lors de cette transmission. Ainsi, l'analyse cladistique s'applique bien à la biologie moléculaire, et ce d'autant plus qu'il est aisé d'inclure des modèles de probabilités de mutation et de transfert de gènes. Des techniques spécifiques, mathématiques et informatiques, ont été développées et constituent de nos jours un formidable champ de recherche dans le domaine de la bioinformatique.

En réalité, la cladistique, appelée aujourd'hui systématique phylogénétique, n'est qu'une méthodologie parmi d'autres approches phylogénétiques que nous décrirons brièvement au Chapitre 9. Cependant, elle garde toujours sa spécificité d'être étroitement liée aux innovations transmises lors des réplications d'organismes.

La comparaison entre les arbres génétiques et les arbres biologiques (dérivés des données morphométriques) montre généralement un bon accord global, mais il est encore impossible de dire objectivement lesquels sont les plus proches de la réalité (BROWER AND VOGLER, 1996; NICHOLS, 2001; DEGNAN AND ROSENBERG, 2006; NAKHLEH ET AL., 2006; POLLARD ET AL., 2006). L'approche génétique présente l'énorme avantage de pouvoir décrire de la même manière à peu près n'importe quel organisme, qu'il soit fossile ou vivant. C'est également le seul moyen raisonnable d'étudier les bactéries et les protozoaires pour lesquels les données morphométriques sont bien difficiles à définir, surtout dans un contexte évolutif. Néanmoins, les phénomènes de mutations et de transferts de gènes sont cruciaux. L'analyse cladistique apporte ici un puissant moyen d'investigation notamment avec ses extensions récentes pour prendre en compte ces phénomènes de réticulation qui concrètement tendent à joindre deux branches parallèles d'un arbre (MAKARENKOV ET AL., 2006).

### 3.1.4 Évolution par embranchement et évolution réticulée

Le concept darwinien d'évolution par transmission avec modification produit une structure hiérarchique qui peut se représenter sous la forme d'un arbre. Une espèce

donnée va finir par engendrer une nouvelle espèce qui constitue une divergence par rapport à la lignée initiale. Cette divergence est un embranchement sur un arbre, à la manière des arbres généalogiques ou ceux issus de la théorie des graphes. De tels arbres, appelés dendogrammes ou phénogrammes, peuvent être construits à partir d'une matrice de distance issue d'une analyse de distance multivariée, auquel cas deux objets sont proches parce que leur similitude globale est grande (Sect. 2.4.2). Ils sont a priori dépourvus de toute signification évolutive, celle-ci ne pouvant qu'être ajoutée a posteriori, grâce à l'hypothèse ad hoc que la ressemblance globale traduit le lien de parenté. Au contraire, dans le cadre darwinien, ou plus généralement évolutif, on souhaite visualiser l'histoire de la diversification, c'est-à-dire situer les divergences évolutives les unes par rapport aux autres. Pour cela, il semble préférable d'utiliser nos connaissances sur les mécanismes élémentaires d'évolution. Dans le contexte darwinien, il s'agit de l'évolution par embranchement créée par la transmission avec modification. Comme nous le verrons au Chapitre 4 la cladistique est intrinsèquement conçue pour construire des arbres basés sur ce mécanisme d'évolution. Ces arbres sont appelés cladogrammes.

Il existe cependant un autre type d'évolution qui est issue du transfert de gènes entre deux espèces. C'est le cas par exemple des espèces hybrides obtenues par accouplement. Mais cela concerne particulièrement les premières bactéries (archébactéries, eubactéries et eucaryotes) qui semblent visiblement avoir beaucoup évolué par échanges de gènes. Ce mécanisme d'évolution complique le schéma hiérarchique car il relie deux espèces situées sur deux branches différentes d'un arbre évolutif, créant ainsi un schéma réticulé. Ce type d'évolution s'appelle donc évolution réticulée, ou aussi horizontale car les arbres ont longtemps été présentés de bas en haut, dans le sens opposé des arbres généalogiques. Ce type d'évolution introduit du bruit dans l'analyse cladistique et peut même la rendre inutilisable si elle domine trop l'évolution par embranchement. Grâce aux multiples analyses cladistiques, les biologistes réalisent aujourd'hui que des améliorations de l'approche cladistique doivent être développées, afin de pouvoir construire des réticulogrammes. Les débats restent largement ouvert. Par exemple, d'après WOESE (2000), les deux types d'évolution seraient à la fois distinctes et fondamentales dans l'évolution vivante : l'évolution réticulée serait à l'origine des innovations ("mutations"), et l'évolution par embranchement serait responsable de la propagation de ces innovations et de la diversité. On peut quand-même craindre de perdre la simplicité de l'organisation hiérarchique d'un cladogramme, rendant un peu plus difficile la détermination de l'histoire des différentes sortes de cellules primitives à l'origine de tous les organismes vivants. Les réseaux sont l'outil adapté, comme généralisation des arbres hiérarchiques, même si la lecture du schéma de diversification est plus difficile (Sect. 9.4).

### 3.1.5 Adaptation et acclimation : le rôle de l'environnement

DARWIN a, le premier et bien avant la découverte des gènes, compris que la diversité du monde vivant était causée par le processus de transmission avec modification. Il faut cependant ajouter que selon DARWIN ces modifications apparaissent spontanément lors de la réplication, et que seuls les individus les mieux adaptés sont sélectionnés car la probabilité qu'ils puissent transmettre leur innovation est plus grande. C'est la fameuse sélection naturelle dans laquelle l'environnement, comprenant les autres organismes vivants et le monde inanimé, intervient a posteriori.

LAMARCK a cependant proposé que l'environnement peut agir avant la réplication, en ayant une influence telle que certaines caractéristiques de l'organisme s'infléchissent et se transmettent ainsi modifiées. C'est un processus d'acclimatation qui a pu être observé dans de nombreux cas, et même en laboratoire.

Il apparaît aujourd'hui que les deux processus sont certainement à l'œuvre, la différence fondamentale étant que l'acclimatation est bien plus souvent réversible que ne peut l'être l'adaptation. Quel que soit le processus, le mécanisme, souvent appelé darwinien, de transmission avec modification est toujours valable et produit l'organisation essentiellement hiérarchique des organismes vivants.

Évolution des espèces et évolution de l'environnement sont donc étroitement associées. Il suffit de penser à un système clos comme une île, peuplée de différents animaux et plantes, pour comprendre qu'il existe un équilibre complexe de relations d'influence, chaque individu ayant nécessairement un impact sur son environnement dont il dépend à son tour.

## 3.2 Formation, évolution et environnement des galaxies

### 3.2.1 Le diagramme de HUBBLE en diapason

Les liens entre des classes d'objets naturels reposent nécessairement sur des modèles ou des théories, et sortent du champ de l'observation proprement dit. Il est toujours possible d'affirmer que les classes ont toujours été ainsi. Cependant, cette approche est stérile et insatisfaisante. Finalement la question pertinente n'est pas tant de savoir comment sont composées ces classes, mais plutôt pourquoi et comment ces classes sont apparues. La notion d'évolution est donc inévitable puisque dans la formation d'un objet, il y a inévitablement transformation de quelque chose en autre chose.

HUBBLE utilisa des arguments physiques pour proposer que les galaxies elliptiques devaient finir par s'aplatir en un disque au cours du temps et former des galaxies spirales, barrées ou non. Ainsi, il relia les classes morphologiques en invoquant l'évolution comme facteur de diversification. C'est ainsi qu'est né le très fameux diagramme en diapason, dit diagramme de HUBBLE (HUBBLE, 1936). Il reste encore la seule classification évolutive jamais proposée pour les galaxies dans leur ensemble.

Seulement, ce schéma extrêmement simple, fondé sur des arguments physiques élégants et irréfutables, avait un défaut important : les galaxies irrégulières n'y avaient pas leur place. Ce point met en évidence une erreur conceptuelle qui n'est certainement pas du fait de HUBBLE mais qui s'est installée sournoisement au cours des décennies. Il s'agit de la confusion entre la classification de HUBBLE et le diagramme en diapason. La classification de HUBBLE comprend quatre classes et inclut toutes les galaxies. Le diagramme en diapason n'est qu'une proposition d'un lien entre ces classes, sous la forme d'un schéma évolutif les reliant (ressemblant étrangement à un "arbre", voir Section 6.1), mais ne les incluant pas toutes parce que le modèle expliquant ces liens est incomplet.

Autrement dit, la classification morphologique de HUBBLE, en quatre classes, est une synthétisation des *observations*, le diagramme de HUBBLE est un *modèle* appliquée à cette synthétisation. La meilleure preuve en est que les types morphologiques des galaxies n'ont pas changé, ils se sont raffinés avec les progrès techniques. Par contre le schéma évolutif n'est plus du tout celui de HUBBLE puisque nous savons

maintenant que d'une part le temps mis par une galaxie elliptique pour s'aplatir est beaucoup plus long que l'âge de l'Univers, et d'autre part, deux galaxies spirales qui fusionnent donnent presque toujours une galaxie elliptique. La persistance de cette erreur conceptuelle est dommageable car nombre d'astrophysiciens cherchent à expliquer le diagramme de HUBBLE en le considérant comme une donnée observationnelle (opérant ainsi une modélisation d'un modèle !), en amalgamant de plus évolution des galaxies et évolution de leur morphologie.

Depuis les premiers travaux de HUBBLE, avec les progrès techniques, les gros télescopes notamment, la structure des galaxies est apparue de plus en plus détaillée aux yeux des astronomes. La diversité des formes augmentait tout en restant incluse dans les quatre classes de HUBBLE. Les bras spiraux devenaient plus ou moins nombreux, plus ou moins nets d'une galaxie à l'autre. Même les rondeurs des galaxies elliptiques allaient de la sphère presque parfaite à des formes oblongues très marquées. Le livre de VAN DEN BERGH (1998) fait une revue assez complète des différentes classifications morphologiques ayant été mises au point. On peut citer les systèmes de DE VAUCOULEURS, SANDAGE, VAN DEN BERGH, ELMEGREEN, MORGAN, etc. Tous se basent sur le diagramme en diapason de HUBBLE, mais apportent nombre de sous-classes précisant la complexité des formes. Une nomenclature propre à chaque système est développée, à laquelle on associe le type morphologique qui prend parfois une allure de paramètre quantitatif comme chez de Vaucouleurs. Le schéma retenu est toujours un rangement dans le sens de la complexification, surajouté à l'ossature du diagramme en diapason de HUBBLE. Inévitablement, peut-être inconsciemment, le mot "évolution" y est attaché, principalement dans l'acception de "progressivité". C'est sans doute un effet du rasoir d'OCKHAM qui stipule la simplification des hypothèses interprétatives, c'est-à-dire ici la parcimonie du schéma de classification auquel on donne un sens évolutif. Nous reviendrons sur cette notion dans le Chapitre 4.

Le système de classification de MORGAN a apporté une approche plus objective et s'est distingué de la classification de HUBBLE tout en utilisant un critère apparenté à la morphologie. Il s'agit de la concentration de lumière qui est un paramètre assez facilement mesurable, et ce même automatiquement sur des galaxies lointaines. Le système DDO (David Dunlap observatory) est peut-être l'un des premiers à utiliser une observable quantitative, à savoir la luminosité, mais l'a simplement associée sous forme de sous-classes des types morphologique usuels.

Comment peut-on expliquer le succès évident auprès de la communauté astrophysique de la classification de HUBBLE et surtout du diagramme en diapason ? Outre sa simplicité qui fait rêver, la raison se trouve probablement dans des corrélations générales trouvées entre le type morphologique et des observables quantitatives et objectives. Par exemple, il existe une évolution indiscutable de la couleur ou de la luminosité des galaxies le long du diagramme de HUBBLE. Comme ces deux paramètres sont reliés à l'âge des étoiles, la tentation d'attribuer un sens temporel à la classification morphologique est si grande qu'elle perdure de nos jours encore. D'autres corrélations existent, peut-être un peu plus grossières, comme par exemple le fait que les galaxies elliptiques sont beaucoup moins pourvues en hydrogène neutre que les galaxies spirales. Cependant, les corrélations observées ne sont jamais serrées au point de pouvoir utiliser une observable quantitative objective en remplacement des types morphologiques. Inversement, ces corrélations ne justifient en aucun cas l'utilisation de la seule morphologie comme traceur de l'état évolutif des galaxies.

### 3.2.2 Trajets évolutifs des galaxies

Chacune de ces classifications basées sur celle de HUBBLE, cherche, par la modélisation, à fournir un schéma évolutif intelligible. Mais la structure reste fortement contrainte par le diagramme en diapason. Il existe quelques autres représentations de possibles schémas évolutifs qui s'appuient exclusivement sur des diagrammes binaires, donc, d'une manière ou d'une autre, sur des corrélations à deux ou trois paramètres. Ces corrélations ne sont pas toujours bien expliquées quoique très souvent considérées comme étant dues à un lien physique. Comme une dispersion plus ou moins importante est presque toujours présente, il est naturel de penser que celle-ci est due à un paramètre supplémentaire. Si ce paramètre peut être relié à l'évolution, alors des modèles permettent de tracer un chemin évolutif sur ce diagramme binaire.

La recherche d'une corrélation entre deux paramètres revient à porter l'un d'entre eux en fonction de l'autre, donc à accorder un poids infini à leurs valeurs précises. Ceci implique que les valeurs relatives ont un sens, excluant d'emblée toute variation possible au sein d'une même classe d'objets. Pour illustrer ce phénomène, il suffit de considérer la corrélation entre une observable (par exemple la taille d'une galaxie) et sa métallicité moyenne. Nous savons que toutes deux augmentent avec le temps au cours de l'évolution de l'Univers, ce qui se traduit par une corrélation positive entre ces deux paramètres. Il est néanmoins assez évident que les galaxies grossissent à des rythmes différents, s'enrichissent en métaux lourds à des rythmes différentes, de sorte que pour un redshift donné, des galaxies de tailles variées et de métallicités variées sont présentes. Des galaxies peuvent évoluer de la même manière, mais à des époques différentes de l'Univers. La corrélation ne porte donc en réalité que sur la moyenne des tailles ou des métallicités à un instant donné, et une bonne partie de la dispersion est due au mélange des objets à des stades évolutifs différents. Si on imagine les nombreux paramètres caractérisant les galaxies, pouvant amener le même genre de dispersion, on imagine aisément que des corrélations peuvent passer complètement inaperçues à cause de la dispersion, et que les autres ne sont pas forcément très significatives physiquement.

De plus, il est important de bien distinguer une corrélation physique d'une corrélation évolutive ou historique, c'est-à-dire fortuite entre deux processus indépendants : leurs évolutions respectives sont corrélées uniquement parce que le temps se déroule dans le même sens. Comme pour la taille et la métallicité évoquées ci-dessus ou encore la vitesse de rotation et le taux de formation d'étoiles : est-ce qu'une faible vitesse de rotation et un faible taux de formation d'étoiles (notamment trouvées dans les elliptiques) sont interdépendants (par exemple une faible vitesse empêcherait la formation d'étoiles) ou deux conséquences indépendantes de l'histoire complexe de la galaxie ?

Une illustration assez frappante d'un tel amalgame entre corrélation physique et corrélation historique est donnée par l'étude de (DISNEY ET AL., 2008) qui effectue une analyse en composantes principales de galaxies décrites par six paramètres. Les observables ne sont à l'évidence pas toutes indépendantes (comme le rayon à deux niveaux d'isophotes), mais elles évoluent toutes avec le temps. Bien que la conclusion des auteurs semble soulever un mystère, à savoir que la diversité des galaxies serait gouvernée par un seul paramètre (non indentifié) en apparente contradiction avec les modèles actuels, n'est-il pas aisé de voir que ce paramètre est tout simplement le temps ? En effet celui-ci explique à lui tout seul que toutes les variables se retrouvent

dans la première composante principale. La preuve d'une corrélation historique démontrée dans le cas d'une analyse astrocladistique effectuée sur le plan fondamental des galaxies. Il apparaît très clairement que la corrélation apparente entre l'indice $Mg_2$ et la dispersion centrale de vitesse n'est en réalité que la succession des états évolutifs des galaxies (FRAIX-BURNET ET AL., 2010).

Toujours dans l'espoir que les galaxies soient des objets simples, autant que les étoiles, ces diagrammes binaires avec trajet évolutif ont été l'objet d'une tentative d'analogie avec le diagramme de Hertzsprung-Rüssel (DUTIL, 2001). L'idée est certainement intéressante car elle introduit la notion de trajet évolutif non linéaire dans un espace de paramètres à quelques dimensions. Malheureusement, cette approche n'a visiblement pas pu être étendue à l'ensemble de la diversité des galaxies. Les galaxies ne seraient-elles quand même pas des objets beaucoup plus complexes que les étoiles ? Et ne faut-il pas les classer d'une manière appropriée avant toute chose ?

### 3.2.3 Environnement des galaxies : un Univers en évolution

Notre vision des galaxies en ce début de XXIème siècle met en évidence la forte intrication entre la structure de l'Univers et l'évolution des galaxies. Pendant longtemps, ces objets ont été les traceurs visibles de notre Monde à grande échelle, sondeurs de l'immensité de l'espace. Désormais, l'Univers invisible devient prépondérant et son exploration est devenue nécessaire pour appréhender l'histoire de la diversité des galaxies. Comprendre l'un pour mieux connaître l'autre, et réciproquement, voilà comment il nous faut aujourd'hui regarder les galaxies : non plus uniquement depuis notre bonne vieille Terre, avec nos gigantesques télescopes, mais également en nous plaçant depuis les fins fonds de l'Univers, grâce à un arsenal théorique rassurant face à un espace-temps déroutant. Évolution des galaxies et cosmologie sont donc plus que jamais intimement liées. Le concept même de galaxie devient plus flou à mesure qu'on s'approche de la période de l'apparition des premiers objets de l'Univers. Il devient nécessaire de définir ce qu'on entend par galaxie dans un contexte hautement évolutif (Chapitre 5).

Notre vision du Monde se place nécessairement dans un cadre, contexte culturel et scientifique, nourri des connaissances acquises au cours des siècles, et enrichi des découvertes du moment. L'Univers n'échappe pas aux paradigmes qui permettent aux scientifiques de forger de nouvelles hypothèses, d'imaginer de nouvelles observations, dans le but toujours renouvelé de pousser les frontières du paradigme dominant, quitte à le faire exploser. Le Big Bang désigne en réalité un ensemble divers de paradigmes entrant tous dans le cadre de l'expansion observée de l'Univers, et évoluant au gré des découvertes observationnelles et théoriques. Le paradigme dominant actuellement est bien différent de celui d'il y a dix ou vingt ans, et sera certainement tout autre dans un futur proche. Néanmoins, quelques grandes lignes sont largement confirmées, et les progrès de nos connaissances dans ce domaine sont remarquables à l'échelle de l'Humanité. Voici donc brièvement comment nous imaginons notre Univers aujourd'hui, en nous limitant à ce qui nous semble pertinent pour les galaxies et leur histoire.

Un certain nombre de faits observationnels marquants, depuis la découverte de la récession des galaxies par HUBBLE en 1929 et celle du rayonnement de fond cosmologique à 3K par Penzias et Wilson en 1963, jusqu'aux observations récentes et détaillées de la composition chimique de l'Univers, des fluctuations primordiales de

densité ainsi que des Supernovae très lointaines, entrent dans un cadre théorique relativement simple et compréhensible, qui constitue le paradigme du Big Bang. Plus exactement, on parle même du modèle de concordance, modèle servant aujourd'hui de standard et fixant de nombreuses variables à des valeurs plutôt précises (SPERGEL ET AL., 2006). Le cadre de la Relativité Générale et la Mécanique Quantique suffisent pour raconter une histoire de l'Univers depuis 13,7 milliards d'années jusqu'à nos jours. Luxe suprème, de nombreuses simulations numériques permettent de mettre l'Univers en boite et de regarder en particulier la matière visible évoluer depuis la recombinaison (redshift z=1000) jusqu'à nos jours (z=0) alors que les galaxies les plus lointaines observées à ce jour ne sont qu'à un redshift de 10 au maximum. Cependant, d'ici peu, nous aurons accès à la totalité de l'Univers visible.

La cosmologie n'a certainement pas dit son dernier mot, la physique non plus, mais voici l'essentiel de l'histoire de l'Univers qui nous intéresse pour appréhender l'histoire des galaxies.

L'Univers s'est donc retrouvé un jour dans un état d'expansion. Pourquoi ? Nous ne savons encore pas vraiment car nous avons atteint la limite de notre physique bien établie. C'est le fameux temps de Planck ($10^{-43}$s) en deçà duquel une théorie quantique de la Gravitation est nécessaire, une théorie reliant l'infiniment petit à l'infiniment grand en quelque sorte. Mais l'intérêt d'une limite est de pouvoir être franchie, c'est ce qui fait avancer la connaissance, et les théoriciens ne manquent pas d'idées. Ne doutons pas que ces recherches aboutiront, laissons-nous simplement le temps de digérer la découverte par HUBBLE de l'expansion de l'Univers, découverte très récente à l'échelle de l'Humanité, puisque datant d'il y a moins d'un siècle.

Le terme Big Bang ne doit pas tromper. Il ne s'agit pas d'une explosion, mais bien d'une expansion, qui n'a ni centre ni origine. Puis ce mouvement s'emballe en un épisode appelé Inflation. Mystérieuse Inflation qui pourtant semble pouvoir être décrite avec notre physique et expliquée par un état de la matière très particulier en ces temps là (GUTH, 2001). L'important est que cette période d'expansion folle est nécessaire, imposée par l'homogénéité quasi parfaite de l'Univers à grande échelle et sa structure à l'âge de 380 000 ans environ (z=1000) qu'on observe aujourd'hui sous forme de fond diffus à 3K. À cette époque, le cosmos avait suffisamment refroidi (T$\simeq$ 3000K) pour que les électrons s'associent aux protons en formant les premiers atomes, permettant à l'Univers de devenir ainsi transparent donc observable (période de la recombinaison). L'Univers était 1000 fois plus petit qu'aujourd'hui, mais tellement beaucoup plus grand qu'avant l'Inflation, que le principe de causalité aurait rendu impossible une telle homogénéité en si peu de temps. L'Inflation a dû avoir lieu. Elle prédit également que l'Univers doit avoir une géométrie plate, et c'est exactement ce qu'on trouve.

La matière dite baryonique (électrons-protons-neutrons), celle que nous connaissons le mieux et qui compose les galaxies, s'est forgée quelque temps après l'Inflation. Les noyaux d'atomes, conglomérats de protons et neutrons, se sont ensuite formés selon des compositions relatives faciles à calculer, et très dépendantes des conditions physiques. Des contraintes très fortes proviennent là encore de l'observation du rayonnement de fond diffus à 3K, pourtant émis un peu plus tard dans l'histoire de l'Univers. Les proportions des différents noyaux d'atomes sont donc a priori bien connues. Lors de la recombinaison, les électrons se sont associés à ces noyaux pour former les premiers atomes neutres de l'Univers. Ce sont alors principalement les étoiles qui ont

transformé ces éléments chimiques initiaux en éléments plus lourds. Ces prédictions théoriques et les observations effectuées dans des étoiles très vieilles, donc de composition quasi primordiale, sont d'un accord tout simplement remarquable

L'observation de l'Univers à la recombinaison est donc cruciale. Tout d'abord il nous fixe la composition chimique : beaucoup d'hydrogène, de l'hélium, et un peu de Deutérium, de Lithium et de Béryllium, qui sont tous des éléments dits légers. Cette composition chimique joue un rôle majeur dans la formation des premières étoiles, et d'une manière un peu plus générale dans la formation des premiers objets. De même, leur évolution future est grandement influencée par la composition chimique de l'Univers à tout instant et en tout lieu.

Ensuite, le fond diffus observé est très homogène comme il a été dit plus haut, mais pas parfaitement, à $10^{-5}$ près "seulement". Ces toutes petites fluctuations ont pourtant de grandes conséquences, car le spectre de puissance (c'est-à-dire leur distribution aux différentes échelles spatiales) traduit précisément la structure géométrique de l'Univers et donc en même temps la quantité de matière gravitationnelle, comprenant la matière baryonique et la matière noire (non-baryonique).

Combinée à la valeur du taux d'expansion à z=0 (constante de HUBBLE $H_0$), mesurée indépendamment grâce à l'observation des supernovae lointaines, nous pouvons décrire aujourd'hui notre Univers ainsi :

— âge 13,7 milliards d'années ($H_0 = 72$ km/s/Mpc)
— géométrie plate
— énergie du vide 70% (constante cosmologique = 0.7)
— matière noire 26%
— matière baryonique 4%

La matière visible, celle que nous connaissons, celle qui compose les galaxies, ne représenterait donc que 4% de l'Univers, tout le reste étant à peu près totalement inconnu. Pourtant, cet inconnu joue un rôle primordial pour l'histoire des galaxies.

Premièrement, l'énergie du vide se comporte comme une force répulsive. C'est elle qui est responsable de l'expansion, expansion qui va en s'accélérant. Pour former une structure autogravitante, comme nous le verrons plus loin pour les galaxies, il faut que la gravitation vainque l'énergie du vide. Cela ne peut se faire qu'avec une certaine quantité de matière, et à une certaine échelle car les deux forces n'ont pas la même dépendance vis à vis de la distance (TRIAY, 2005). C'est une contrainte notable pour comprendre en particulier la formation des premières structures de l'Univers, donc des premiers objets.

Deuxièmement, la matière noire est six fois plus abondante que la matière baryonique. C'est donc elle qui va dicter sa loi et structurer le champ gravitationnel que la matière visible va subir. L'environnement extérieur dans lequel les galaxies se diversifient, en dehors de leurs congénères, est donc un champ de matière noire dont on commence seulement à cartographier la distribution. Il faut garder à l'esprit que d'autres théories, comme la gravité modifiée (MOND), pourraient mettre à mal l'existence de la matière noire et même de l'énergie noire. Cependant, cela ne modifiera pas l'image de galaxies évoluant dans un environnement gravitationnel structuré.

Nous avons aujourd'hui accès à presque toute la période observable de l'Univers, qui correspond de plus à toute l'histoire des galaxies. Bien sûr, il y a un grand trou entre z=10 et z=1000, mais cela ne représente qu'une assez faible portion temporelle de l'âge de l'Univers (moins de 1%). De plus, nos ordinateurs ont aujourd'hui suffi-

samment progressé pour qu'on puisse calculer le devenir de toute premières fluctuations observées de densité dans un Univers en expansion, avec des lois physiques qui semblent encore valables à des milliards d'années-lumière de nous !

C'est ainsi qu'on prédit numériquement que de petits halos de matière noire, issus des fluctuations primordiales, plus ou moins autogravitants, vont apparaître, puis fusionner pour former de plus grandes structures, dans une évolution hiérarchique que nous évoquerons dans la Sect. 5.7.1. La répartition de la matière noire finit par devenir filamentaire, occupant les bords de cellules pratiquement vides qui sont peut-être générées par la force répulsive de l'énergie du vide (TRIAY, 2005).

Quant à la matière baryonique, donc initialement sous forme de gaz d'éléments légers, puis plus tard sous forme de galaxies, on imagine en première approximation qu'elle se concentre au cœur des halos de matière noire et se répartit ensuite le long des filaments. Cette distribution spatiale reproduit presque parfaitement la distribution des galaxies observée jusqu'à un redshift de 2 environ (BAUGH ET AL., 2004). Au-delà, nous ne connaissons que trop peu de galaxies pour conclure. Cette image a donné naissance au scénario hiérarchique de formation des galaxies dans lequel ces dernières grossissent par fusion à l'instar des halos de matière noire (voir par exemple références dans DISNEY ET AL., 2008). Cependant, nous savons aujourd'hui que des galaxies très massives existaient déjà tôt dans l'histoire de l'Univers, redonnant quelque vitalité au scénario de formation par effondrement monolithique. La distinction claire que fait l'astrocladistique entre des objets évolutifs et leurs environnement montre que les galaxies se diversifient selon des processus physiques bien identifiés indépendamment de l'environnement dont la structure évolue par un processus hiérarchique (Chapitre 5). Les galaxies ne constituent pas le cœur des halos de matière noire, leur évolutions et leur trajectoires sont seulement influencées par la présence des halos, avec une probabilité plus grande d'orbiter et de se rapprocher de leurs cœurs.

### 3.2.4 Les premiers objets de l'Univers

Le gaz d'éléments légers se concentre donc initialement grâce aux petites surdensités que sont les halos de matière noire. Il va alors se condenser et, étant encore dépourvu d'éléments atomiques lourds, s'effondrer très rapidement pour former de très grosses étoiles dont le rayonnement va terminer la période dite des Âges Sombres depuis la recombinaison. Ces étoiles vont vivre peu de temps et exploser violemment en éjectant les premiers éléments lourds synthétisés qui se retrouveront dans les nouvelles étoiles qui vont se former.

Des ensembles complexes, constitués d'étoiles et de gaz plus ou moins enrichis en éléments lourds, vont ainsi nécessairement apparaître. Ces structures de matière baryonique sont plus ou moins autogravitantes, subissant en permanence la compétition entre gravitation et expansion, ainsi que l'influence des différents halos de matière noire et autres étoiles alentours.

Comment définir les premiers objets de l'Univers ? Si on pose comme condition qu'ils soient autogravitants, jusqu'à quel degré peuvent-ils l'être à cet instant de l'Univers qui est encore relativement dense ? Pour les étoiles prises individuellement, il n'y pas de doute possible. Mais les nuages de gaz qui les ont formées, ne sont-ils pas apparus avant ? Et même si on admet que les étoiles sont les véritables premiers objets autogravitants de l'Univers, nous sommes encore loin d'une galaxie telle que nous les

connaissons autour de nous. La question des premiers objets de l'Univers n'est encore pas résolue, et ce problème n'est pas aussi évident qu'il n'y paraît dans le contexte des galaxies. Nous voyons ici la nécessité de définir proprement ce qu'on appelle une galaxie si on veut pouvoir retracer l'histoire complète de ces objets (Sect. 5.1).

### 3.2.5 Évolution et diversification des galaxies

En toute logique, l'évolution d'une galaxie n'est rien d'autre que l'évolution de ces constituants fondamentaux que sont les étoiles, le gaz et la poussière (Sect. 2.3 et 5.1). Indépendamment de la galaxie qu'on considère, chacun des constituants a des raisons propres d'évoluer.

Les étoiles vieillissent toutes seules selon leur masse et leur métallicité. Elles changent en nombre, en distribution d'âge et de masse lorsque de nouvelles étoiles se forment au sein de la galaxie, ou lorsque d'autres étoiles sont accrétées par la galaxie, par exemple lors de fusions de deux galaxies. Les étoiles les plus jeunes sont plus métalliques car formées avec du gaz transformé au cœur des étoiles massives des générations précédentes. Enfin, les étoiles changent leurs cinématiques et leurs répartitions en fonction des influences gravitationnelles internes à la galaxie, ou externes (autres galaxies, matière noire ; voir Sect. 5.4).

Le gaz modifie sa distribution, sa concentration, sa cinématique, sa température, son taux de ionisation en fonction de l'environnement gravitationnel et radiatif en perpétuel changement. Il peut se condenser pour former des étoiles, et sa composition s'enrichit en éléments lourds grâce aux étoiles qui explosent en supernovae et relâchent le produit de leur combustion thermonucléaire.

Poussières et molécules modifient leurs distributions, densités, cinématiques, températures suite à de nombreux évènements, comme les perturbations gravitationnelles ou les ondes de chocs générées par des explosions d'étoiles. Elles sont chauffées et même détruites par les rayonnements des étoiles chaudes ou même en provenance du noyau central de la galaxie. Les rayons cosmiques également peuvent les détruire. Les poussières et molécules voient leur composition chimique, et donc leur taille et leur complexité, se modifier selon les conditions locales et suite aux perturbations de tous ordres. Elles peuvent même aller jusqu'à se condenser en des nuages de plus en plus denses pour former des étoiles.

Les structures morphologiques des galaxies, tels que les bulbes, disques, halos stellaires, barres et bras spiraux, sont simplement les traces visibles des propriétés cinématiques principalement des étoiles. Elles traduisent la distribution des orbites stellaires dans la galaxie, qui sont modifiées selon les aléas de l'environnement gravitationnel local.

Enfin, l'activité du noyau des galaxies, avec ou sans la présence d'un trou noir central, est alimentée très certainement à la suite de perturbations gravitationnelles impliquant des modifications drastiques des orbites du gaz et des étoiles de la galaxie.

In fine, ce sont les descripteurs tirés des observations qui vont caractériser leurs états évolutifs, donc l'état évolutif de la galaxie.

Quand on parle de l'évolution d'une galaxie, c'est donc bien de tout cela dont on parle, dont on devrait parler. Car la formation et l'évolution des galaxies sont une problématique majeure de l'astrophysique contemporaine. Cependant, qu'entend-on exactement par formation, par évolution ? Le Chapitre 5 traitera ces notions plus pro-

fondément, mais dans le contexte de ce chapitre nous pouvons dores et déjà en préciser les contours.

Par formation, on entend en principe création ou apparition. Mais la formation des galaxies concerne à la fois la formation des premiers objets de l'Univers qui peuvent être considérés comme galaxies (encore faut-il définir cet objet), et la formation d'une galaxie dans l'état où nous l'observons à un redshift quelconque. Ce deuxième cas est généralement ignoré, ou largement amalgamé avec l'évolution au travers d'un concept flou de transformation. Pourtant, chaque fois qu'une galaxie se modifie significativement, elle forme bel et bien un nouvel objet (Sect. 5.4). La formation des galaxies est donc généralement une expression assez vague que l'astrocladistique précise en identifiant clairement les processus d'évolution qui sont des processus de transformation des galaxies (Sect. 5.3).

Le terme "évolution" sous-entend à la fois modification et temporalité : la modification se déroule au cours du temps. Par évolution des galaxies on confond généralement l'évolution des galaxies en tant que population (les toutes premières galaxies n'ont rien à voir avec celles à z=0) et l'évolution de l'individu (une galaxie toute seule évolue parce que ces constituants évoluent). La modification au cours du temps peut être diversifiante si le nombre et la complexité des constituants évolutifs, la variété des environnements, et la nature des processus d'évolution rendent peu probables des changements identiques pour tous les objets. Par exemple, dans le modèle hiérarchique de formation des galaxies, la diversité des masses des galaxies serait apparue au cours du temps, progressivement, les galaxies massives étant le résultat du mélange de galaxies plus petites. Peut-ont imaginer qu'elles grossissent ainsi toutes au même rythme ? Nous reviendrons sur les processus d'évolution et de formation des galaxies au Chapitre 5, mais la diversité observée des galaxies plaide largement en faveur d'une évolution avec diversification.

Il semble de plus que le terme "évolution" s'applique bien à une unité, que cela soit la population des galaxies, une galaxie individuelle, ou un paramètre particulier (comme la masse), associant une quantité en tant que fonction du temps. Il est alors facile de comprendre la notion relative de "plus ou de moins évolué". Par contre, lorsqu'un objet est intrinsèquement multivarié, c'est-à-dire lorsque plusieurs paramètres sont nécessaires pour le caractériser fidèlement, les évolutions différentes de ceux-ci rendent généralement les comparaisons peu pertinentes : quel est le plus évolué entre un éléphant et une fourmi, ou entre une galaxie naine possédant des étoiles très vieilles et une grosse galaxie spirale possédant surtout des étoiles jeunes ?

La diversification, naturellement multivariée, est l'action menant à la diversité. Par exemple, dans le modèle initial de formation des galaxies par effondrement monolithique, la diversité (notamment en masse) serait apparue au moment de la formation de toutes les galaxies au tout début de l'Univers. Elle aurait donc été générée initialement, en un laps de temps très "court", sans modification ultérieure. Dans ce cas extrême, la diversité ne serait pas liée au temps, donc à l'évolution, mais serait primitive. Cependant, comme les interactions et les fusions ont nécessairement modifié et donc fait évoluer les galaxies (Sect. 3.2.3, 5.4), ce modèle paraît trop radical. D'une manière plus générale, la diversification sous-entend une augmentation de la variété, les modifications étant donc multiples et diverses.

En conclusion, il nous semble que le terme de diversification est plus précis et moins ambigu que "évolution", il inclut les notions de formation, d'évolution et même

de classification. C'est en cherchant à définir proprement ces notions et l'objet "galaxie" que nous conclurons au Chapitre 5 que le terme de diversification recouvre complètement les véritables préoccupations des astrophysiciens lorsqu'ils parlent de "formation et évolution des galaxies".

## 3.3 Classer des objets complexes en évolution

Dans ce chapitre et le précédent, nous avons constaté que la biologie évolutive et l'astrophysique des galaxies sont confrontées à la même problématique. Tant du côté des organismes vivants que des galaxies, il est nécessaire de synthétiser les observations des objets complexes en évolution et d'en comprendre les relations. La systématique, qui est la science de la classification des organismes vivants, a, au cours des Âges, clairement identifié les objectifs, les possibilités et les limitations de la classification de tels objets, et a permis le développement de méthodes adaptées. Plus de deux mille ans d'histoire de la classification des espèces vivantes ont fourni à l'Humanité une expérience précieuse sur les meilleures manières d'appréhender la complexité du monde qui nous entoure. Nous allons maintenant nous détacher des objets eux-mêmes pour esquisser les concepts.

### 3.3.1 Classer, pour quoi faire ?

Classer relève de l'observation pure et fait partie des usages les plus courants de l'être humain. Ranger, classer, organiser, sont des activités nécessaires dans la vie de tous les jours, une sorte de pulsion qui fait écho à un besoin de l'esprit humain de simplifier les choses.

Comme le dit si bien ADANSON en 1763, classer sert à soulager la mémoire. Il est plus simple de ranger de nombreux objets dans quelques boites que de devoir les décrire tous un par un. Cependant, un deuxième intérêt réside dans la possibilité ainsi offerte de comprendre les différences entre ces boites, de comprendre leurs liens, leurs relations. Il est en effet plus simple d'appréhender la diversité du monde lorsque qu'une logique l'explique. Les physiciens connaissent bien cette approche avec les théories d'Unification, puis de Grande Unification, des interactions fondamentales. Mais pour classer, il faut comparer, c'est-à-dire choisir des descripteurs et une méthode objective de mesure d'une "distance".

### 3.3.2 Les trois façons de comparer les objets

La systématique nous apprend qu'il existe trois manières de comparer les objets. Nous les détaillons en nous servant de la table 3.1.

| | c1 | c2 | c3 |
|---|---|---|---|
| A | 0 | 0 | 1 |
| B | 0 | 1 | 0 |
| C | 0 | 1 | 1 |

TABLE 3.1 – Un petit échantillon avec 3 objets (A, B, C) décrits par trois paramètres (c1, c2, c3) qui ne possèdent que deux valeurs, ou deux états, possibles (0 et 1).

### Apparence

Dans cette approche traditionnelle (Sect.2.1), la plus spontanée, on sélectionne un ou deux critères apparents (par exemple les ailes ou les pattes pour les animaux, ou les bras spiraux pour les galaxies) et on mesure ensuite une distance, ce qui est facile avec si peu de descripteurs. Bien entendu, le choix est nécessairement subjectif lorsque les objets d'étude sont un tant soit peu complexes. Cela peut suffire selon l'utilisation voulue, et c'était visiblement le cas du temps des grecs jusqu'au Moyen Âge. Par exemple, pour l'échantillon de la table 3.1, en choisissant chacun des paramètres un par un, on obtient trois classifications plausibles (ABC ou "A et BC" ou "AC et B"), mais incompatibles entre elles. Aucun argument objectif ne permet de décider laquelle est la plus vraisemblable. Il est donc vain d'espérer décrire l'ensemble de la diversité d'un grand ensemble d'objets par cette approche.

### Similitude globale

Basée sur tous les descripteurs, elle s'appelle également phénétique ou analyse de distance multivariée selon les disciplines (Sect.2.4). L'objectivité de la méthode est garantie par la prise en compte de tous les paramètres. C'est ce qu'a proposé ADANSON en 1763, et cette approche a apparemment résolu tous les problèmes antérieurs de la classification en biologie. La mesure de distance la plus courante est l'analyse en moindres carrés qui calcule une distance quadratique moyenne sur tous les paramètres. Il s'agit donc bien de similitude globale puisqu'elle moyenne les écarts d'un paramètre à un autre. Sur l'échantillon de la table 3.1, on obtient ainsi que $distance(A,C) = distance(B,C) = 1$ et $distance(A,B) = 2$, donc C se trouve entre A et B. Il y a nécessairement trois classes, déterminées objectivement, dont chacun des objets ici est un représentant. Qu'est-ce qui relie ces trois classes ? Si on suppose qu'il s'agit de l'évolution, deux séquences sont possibles : A $\rightarrow$ C $\rightarrow$ B ou son inverse B $\rightarrow$ C $\rightarrow$ A. On peut aussi envisager que C est le plus ancestral et deux branches divergent vers A et B. Nous n'avons aucun moyen objectif de déterminer la plus vraisemblable de ces trois possibilités. De plus, l'hypothèse que l'évolution relie ces trois classes doit être justifiée. Comment le faire si ce n'est en examinant les comportements des trois paramètres et en admettant qu'ils puissent caractériser un certain état évolutif ? Cette analyse se fait a posteriori, et si des conflits apparaissent, il est bien difficile de les interpréter et les corriger. Car l'hypothèse fondamentale faite ici est que la ressemblance globale traduit le lien de parenté. Elle semble invérifiable avec cette seule analyse, non évolutive. Elle ne prend pas en compte les innovations apparues dans la diversification, ni les convergences qui font que deux objets peuvent se ressembler mais en empruntant des chemins évolutifs différentes.

En dehors d'une plus grande objectivité, cette approche n'est pas donc a priori pas beaucoup plus satisfaisante que l'approche traditionnelle, car, au bout du compte, on trouve d'autres boites, également parallèles, mais pour lesquelles il est encore plus difficile de découvrir les liens. En effet, autant il peut être aisé de modéliser quelques paramètres, de comprendre leur origine et leur évolution, autant cela devient vite très complexe et aléatoire pour un ensemble important de variables indépendantes.

Cependant, cette approche peut être utilisée comme méthode phylogénétique lorsque les paramètres sont sélectionnés pour leurs propriétés évolutives (Chapitre 9).

**Histoire commune**

Deux objets peuvent être dits similaires s'ils ont une histoire commune. Cette approche est basée sur l'évolution de tous les paramètres et non plus seulement sur leurs valeurs comme pour la comparaison à partir de la similitude globale. D'emblée, il est affirmé qu'une classe évolue en une autre classe par évolution des descripteurs. Cette approche a donné naissance à la cladistique. Pour un même paramètre, les différences sont considérées comme reflétant un chemin évolutif, et plus l'écart est grand, plus les objets sont éloignés dans l'évolution. En considérant un par un l'ensemble des paramètres, on construit un arbre, appelé cladogramme, qui cherche a produire un schéma évolutif le plus simple possible. Ainsi, la cladistique prend en compte l'évolution dans sa stratégie même de regroupement des objets et de représentation de la diversité. En elle-même, elle n'est pas une méthode de classification, elle fournit un schéma évolutif pour l'ensemble des objets étudiés, à partir duquel une classification peut être proposée.

Reprenons l'exemple de la table 3.1. Cette fois-ci, les valeurs des paramètres sont considérés comme des états évolutifs, par exemple le "0" représentant l'état initial, ou ancestral, et le "1" représentant l'état final, ou évolué, ou encore dérivé. On suppose que ces trois objets dérivent d'un "ancêtre commun", c'est-à-dire en réalité d'une espèce ancestrale commune, hypothèse qui peut être vérifiée par la cohérence du résultat. L'objet A possède un état évolué, tout comme B alors que C en a deux. En conséquence, A et B sont plus proches (moins diversifiés) que C de l'ancêtre commun. Les données disponibles ne permettent pas de déterminer si C dérive de A ou de B. Contrairement à la similitude globale, le sens de l'évolution est directement contenu dans les données de la matrice Table 3.1. Quoi qu'il en soit, le résultat est différent. En terme de classes, A et B définissent deux lignées différentes, C appartenant à l'une d'entre elles. Nous analyserons les arbres obtenues avec cette matrice au Chapitre 4 (Sect. 4.2.3).

La cladistique est décrite en détail dans le reste de ce livre. Sa philosophie est très différente des méthodes d'apparence et de similitude globale, mais mathématiquement elle peut se comparer aux analyses de distances multivariées (Chapitre 9), la distance ici étant une distance évolutive. Malheureusement, cette distance évolutive est le plus généralement très difficile à connaître et même définir. En ce sens, la cladistique est une méthode générale pour des objets complexes en évolution car fondée sur la notion de transmission avec modification. Par nature elle construit les relations évolutives pas à pas. Sa mise en œuvre ne fait donc appel qu'à de l'algorithmique, praticable manuellement pour un petit nombre d'objets. Néanmoins, cette simplicité n'est que toute relative, car outre la grande sophistication des algorithmes nécessaires dans la pratique, la cladistique reste peu intuitive et il est assez difficile d'en saisir toutes les subtilités aussi bien d'un point de vue mathématique, statistique, que phylogénétique. Elle requiert un effort certain, et ce livre espère guider le lecteur.

### 3.3.3 Sur la notion d'espèce

Classer, c'est définir des groupes. Nous avons dit plus haut que la cladistique, méthode comparant les objets sur leur histoire qu'ils partagent, n'est pas une méthode de classification à proprement parler. Elle construit les liens de parenté, au chercheur

de définir les frontières des groupes. De même, aucune des deux autres méthodes de classification ne fournit aisément des groupements. Les objets naturels sont intrinsèquement variables au sein d'une même classe. Les méthodes de regroupement présupposent toujours un ou plusieurs critères séparant les groupes. Deux exemples l'illustrent.

Premièrement, dans la méthode traditionnelle, on peut classer facilement les objets selon leur couleur : bleu, rouge, vert. Malheureusement, ces couleurs ne sont pas tranchées, puisqu'elles sont définies par la longueur d'onde, variable quantitative et continue. Existe-t-il une limite précise entre le bleu et le vert ? Si oui, elle est nécessairement arbitraire. Sinon, comment classer ? Mais de part de d'autres de cette limite, deux objets peuvent être bien plus proches qu'avec un quelconque de leur congénère. Il y a des nuances de bleu et de vert. En déplaçant la limite, on change les classes.

Deuxièmement, comment définir la distance entre deux classes ? Doit-on prendre le centre (moyenne, médiane, centre de gravité...) ou les bords ? Que se passe-t-il en cas de chevauchement ? Il n'y a aucune réponse absolue à ce problème, pourtant important pour positionner les classes relativement les unes aux autres.

En biologie, une classe bien connue est l'espèce. Il est communément admis que l'espèce a une définition simple basée sur l'interfécondité. La réalité est tout autre, car cette notion a beaucoup évolué au cours du temps et pose toujours autant de problèmes. L'espèce semble d'abord étroitement liée à la nomenclature de LINNÉ qui attribue le premier nom au genre et le deuxième à l'espèce. Cela ne définit absolument pas l'espèce, et il semble que cette notion a toujours présenté avant tout un avantage pratique dépendant des connaissances du moment, sans jamais trouver de consensus sur une définition unique et objective. De nombreuses propositions ont été formulées. L'utilisation de l'interfécondité pour caractériser les espèces est née relativement récemment, et cette définition s'est déjà modifiée depuis sa première formulation par Ernst Mayr en 1942. Des découvertes assez récentes montrent de toutes manières que les frontières entre espèces ainsi définies ne sont pas très rigides.

La cladistique a mis en lumière des contradictions dans les définitions de la notion d'espèce et pose finalement la question de l'utilité d'un tel concept. Un certain nombre d'auteurs ont entrepris une révision complète de la nomenclature, donc celle de LINNÉ, en cherchant à la fonder sur l'histoire des organismes, c'est-à-dire sur les cladogrammes et en lui permettant l'évoluer avec l'évolution des connaissances. Ainsi, la notion d'espèce pourrait disparaître (Chapitre 10). Mais le débat est loin d'être clos.

La classification possède donc deux aspects qu'il faut bien distinguer. Il y a d'abord l'organisation relative des objets, ensuite il y a le découpage en groupes. Il peut y avoir plusieurs manières d'organiser les objets, et à partir d'une même organisation il peut y avoir plusieurs découpages possibles. Dans ce livre, nous parlons essentiellement du premier point.

# Chapitre 4

# Introduction à la cladistique

Dans ce chapitre, nous allons donner quelques bases de la cladistique. Il est hors de question d'être complet, plusieurs ouvrages traitant d'une manière pédagogique différentes facettes de cette approche (DARLU AND TASSY, 1993; WILEY ET AL., 1991; LIPSCOMB, 1998, en particulier). Nous nous contenterons de présenter succinctement les différents concepts et les différentes définitions nécessaires à une analyse cladistique, et utiles dans le contexte des galaxies.

## 4.1 Principes et définitions pour la cladistique

### 4.1.1 Principes généraux

La cladistique est née parce que William HENNIG (HENNIG, 1965) a compris qu'on ne pouvait regrouper des organismes simplement sur leur ressemblance, alors que leurs liens de parenté sont le résultat de toute une histoire évolutive. En effet, deux espèces peuvent posséder certaines caractéristiques similaires après avoir suivi des chemins évolutifs indépendants à partir de deux espèces ancestrales (en général appelées simplement "ancêtres", voir plus loin) différents. De plus, le partage d'un ancêtre commun par un groupe doit pouvoir être déterminé d'une manière non ambiguë, ce qui revient à pouvoir identifier des caractéristiques propres à un tel ancêtre et qui ont été propagées aux descendants. Ce qui caractérise une espèce, quelqu'en soit sa définition, c'est le partage de caractéristiques issues d'une innovation apparues chez un ancêtre commun.

C'est bien entendu grâce aux descripteurs qu'on peut retrouver la trace des espèces ancestrales qui ont donné en héritage une modification génétique ou morphométrique. Ces descripteurs sont appelés caractères et sont définis par un certain nombre d'états considérés comme évolutifs. Par définition, l'évolution d'un individu est l'évolution de ces caractéristiques. Ainsi chacune d'entre elles est divisée en un certain nombre d'états évolutifs qui servent en cladistique à mesurer le degré de diversification.

Pour se faire, on distingue les états "dérivés", qui proviennent d'une innovation, des états "ancestraux". La cladistique regroupe les organismes qui partagent les états dérivés apparus chez un ancêtre commun. Le regroupement se fait de proche en proche selon un schéma hiérarchique représenté sous forme d'arbre. En conséquence, une analyse cladistique présuppose nécessairement que les taxons considérés sont tous issus

d'une espèce ancestrale commune. Si ce n'est pas le cas, ou si l'ancêtre commun est trop éloigné, il est possible de diviser l'échantillon. L'évolution par embranchement est aussi un ingrédient nécessaire à la cladistique. L'évolution réticulée (Sect. 3.1.4), si elle est présente, amène du bruit dans l'analyse et la reconstruction de la phylogénie devient plus problématique.

L'arbre issu d'une analyse cladistique s'appelle un cladogramme. Il représente en quelque sorte un arbre généalogique des espèces. Chaque nœud (point d'embranchement) représente un ancêtre hypothétique, individu ou espèce. Ce point est essentiel, parce que premièrement, il est impossible de connaître tous les individus ou toutes les espèces, rendant donc illusoire l'espoir d'identifier formellement un ancêtre, et deuxièmement, les découvertes nous apportent régulièrement de nouvelles espèces qu'il est ainsi plus facile d'insérer dans un arbre.

### 4.1.2 Quelques définitions

Comme tous les domaines scientifiques, la cladistique a développé son propre jargon, parfois emprunté à la théorie des graphes, et établi des définitions précises indispensables. Les termes spécialisés rendent rapidement les textes rebutants et l'apprentissage un peu long pour un non-initié. Sans être exhaustif, le glossaire à la fin du livre regroupe les définitions principales, et nous présentons ici les différentes catégories de termes employés en expliquant leur utilité.

Tout d'abord, le cadre dans lequel on se situe s'appelle la systématique phylogénétique, qui prend également le nom de cladisme, dont l'objectif est une classification basée sur les relations phylogénétiques. La phylogenèse est, par construction, l'histoire évolutive des espèces.

Les objets d'étude doivent être désignés. Dans l'analyse cladistique, ils peuvent être des individus, des spécimens, des groupes, des espèces. Le terme "taxon" a donc été attribué aux objets qui prennent place au bout des branches (les feuilles) de l'arbre phylogénétique, afin de faciliter grandement la description des données initiales et des résultats. À chaque nœud des arbres correspond un taxon hypothétique, complètement fictif et imaginaire, qui est l'ancêtre commun à tous les taxons en aval de ce nœud. Ce terme "ancêtre" doit toujours être pris dans le sens de taxon-ancêtre, la cladistique posant comme principe qu'il est impossible d'identifier un individu-ancêtre, et qu'il est interdit d'identifier même une espèce-ancêtre car nous ne pouvons jamais être certain de connaître toutes les espèces et leur phylogénie véritable. En conséquence, les taxons identifiables sont toujours aux extrémités des branches et ne sont jamais des ancêtres. Par contre, ils peuvent ainsi être proche (au sens de la diversification) d'un tel ancêtre. En réalité, ces nœuds peuvent être vu comme des évènements au cours desquels une innovation est apparue.

Deux groupements formels sont définis par la cladistique. Le premier, qui s'appelle un clade, correspond à un groupe évolutif qu'on cherche à découvrir et à caractériser. Sa définition, très précise, est étroitement liée au cladogramme. Il comprend l'ancêtre et tous les taxons ayant hérités des innovations apparues chez celui-ci. Il est donc caractérisé d'une manière non-ambiguë par un certain nombre d'états dérivés apparus chez l'ancêtre et transmis à tous les autres taxons du clade. Il doit être à la base de la classification des objets d'étude car il représente un véritable groupe dans le processus de diversification et d'évolution.

Le deuxième groupement formel s'appelle un groupe de comparaison (ou outgroup ou extragroupe) supposé posséder un ancêtre commun avec le groupe d'étude (ingroup ou intragroupe) puisqu'il est inclus dans l'analyse. Son intérêt réside dans l'enracinement de l'arbre, c'est-à-dire qu'il permet d'orienter le sens de l'évolution et de situer les groupements évolutifs les uns par rapport aux autres. En effet, la proximité évolutive de deux taxons dépend de l'échelle de diversité considérée et l'ampleur de l'échantillon étudié. Ainsi, ce groupe de comparaison s'avère essentiel dans une analyse cladistique.

Différents groupements peuvent être effectués plus ou moins indépendamment du cladogramme. Ils sont alors séparés en trois classes selon qu'ils incluent ou non l'ancêtre commun et tous les descendants. Dans le premier cas, il s'agit de clades et sont dit monophylétiques. Dans le deuxième cas ils sont dits paraphylétiques. Parfois, des groupements rassemblent des objets ayant plusieurs ancêtres et sont dits polyphylétiques, comme par exemple le groupe des animaux ayant des ailes. Les groupements paraphylétiques et polyphylétiques ne sont pas très informatifs pour la compréhension de l'évolution car ils correspondent à des groupements "artificiels" car mélangeant des lignées partielles ou différentes.

Enfin, les descripteurs sont appelés caractères dès lors qu'on peut leur attribuer plusieurs états évolutifs. Les caractères peuvent être des observables ou d'autres paramètres. La notion d'états évolutifs laisse penser que les caractères sont nécessairement discrets, mais ceci n'est pas indispensable (voir Sect. 6.5, 9.4). Ce sont les états de caractères qui d'une part vont permettre la construction de l'arbre, et d'autre part définir les clades. Le comportement des caractères est donc décrit par plusieurs particularités.

Les synapomorphies, sur lesquelles la cladistique s'appuie, sont des états de caractères dérivés partagés par tous les membres du clade dont fait partie l'ancêtre commun à l'origine de ces caractéristiques nouvelles. Au contraire, les homoplasies (incluant notamment les analogies comme l'endothermie des oiseaux et des mammifères) amènent du bruit dans l'analyse, car elles perturbent l'identification des lignées engendrées par la transmission d'une innovation. Elles sont le résultat soit d'une évolution parallèle (même caractéristique apparue indépendamment dans plusieurs lignées, exemple des ailes des oiseaux et des chauves-souris, ou de la masse d'une galaxie acquise par accrétion lente ou par fusion violente), d'une convergence (même caractéristique mais issue d'évolutions différentes du caractère dans plusieurs lignées différentes, comme les galaxies naines comparées aux galaxies naines de marée) ou d'une régression (caractéristique ancestrale retrouvée après la perte d'une caractéristique acquise, exemple de l'absence de poils chez les baleines qui sont pourtant des mammifères). Les homoplasies brouillent donc les pistes des liens de parenté, alors qu'elles sont la base des méthodes de comparaison par similitude globale comme les analyses de distances multivariées. Enfin les autapomorphies, caractéristiques d'un unique taxon, ne permettent évidemment aucun regroupement, et restent donc assez neutre dans l'analyse.

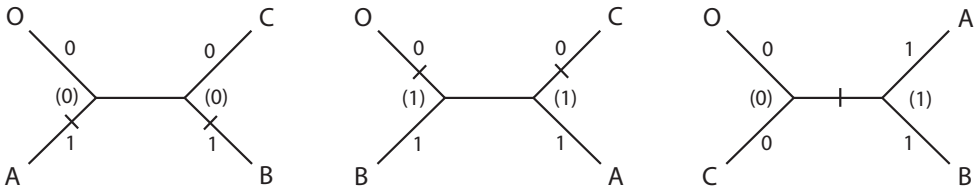
## 4.2 Construction d'un cladogramme

### 4.2.1 Un seul caractère à deux états

La méthode la plus efficace pour comprendre l'essence de la cladistique est de construire des arbres à la main. Considérons donc d'abord un exemple très simple avec quatre taxons (O, A, B, C) décrits par un seul caractère ayant deux états (0 et 1) :

| | |
|---|---|
| O | 0 |
| A | 1 |
| B | 1 |
| C | 0 |

La démarche consiste d'abord à construire tous les arbres possibles avec ces quatre taxons, c'est-à-dire à envisager toutes les combinaisons ou arrangements possibles. Dans le cas présent, il y en a trois :



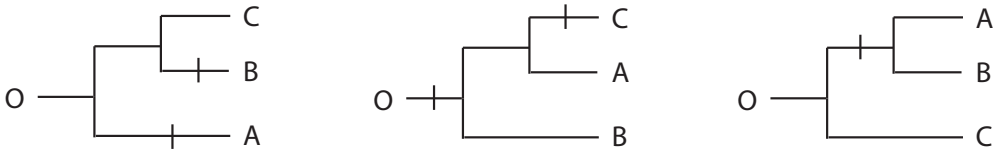

Le schéma évolutif est donné par le comportement des caractères le long de l'arbre. Pour le matérialiser, il suffit de choisir l'un quelconque des taxons et d'indiquer le changement d'état éventuellement nécessaire pour "aller" vers un autre taxon et de le marquer par une barre sur la branche adéquate comme sur la figure. Une fois ce travail effectué pour tous les taxons et tous les arbres, on peut déduire la valeur du caractère au nœud (chiffre entre parenthèses sur la figure) qui correspond à l'état reconstruit, donc prédit, du caractère pour l'ancêtre hypothétique. Il suffit ensuite de compter le nombre de barres pour connaître le nombre total de changements d'états, appelé nombre de pas de l'arbre, nécessaires dans le schéma évolutif proposé. C'est une mesure de sa complexité. Sur la figure, l'arbre de droite est le plus simple puisqu'il n'implique qu'un seul changement en tout, contre deux pour les autres.

Dans cet exemple, le caractère décrivant ces taxons est dit informatif vis à vis du nombre de pas, parce qu'il permet de séparer des arbres selon ce critère. S'il avait été par exemple 0 pour A, alors tous les arbres auraient eu le même nombre de changements d'états, il n'aurait pas été possible de les distinguer par ce biais.

### 4.2.2 Enracinement des arbres

Ces arbres ne sont pas enracinés, c'est-à-dire qu'aucune indication du sens de l'évolution ou de la diversification n'est donnée. Pour cela, il faut identifier au moins un état ancestral et un état dérivé. En pratique, il paraît inimaginable de pouvoir le faire pour chacun des caractères, sans introduire beaucoup d'arbitraire. L'enracinement de l'arbre se fait le plus souvent grâce à un groupe de comparaison (extragroupe) qui est supposé proche de l'ancêtre commun qu'il partage avec l'échantillon d'étude (intragroupe). Ce sont plus concrètement ses descripteurs qui indiquent les états ancestraux

des caractères. Le choix de ce groupe de comparaison est en général difficile, plusieurs peuvent être choisis afin de comparer les phylogénies obtenues. Cet objet peut tout-à-fait être fictif. Bien souvent, l'extragroupe est un membre de l'échantillon. Dans notre exemple, nous allons décréter que l'état ancestral du caractère est "0", et c'est donc le taxon O qui joue le rôle de l'extragroupe. On représente alors les trois cladogrammes précédents sous la forme enracinée suivante :




Il n'y a pas véritablement d'information nouvelle, seul le sens de la lecture est imposé, ce qui est tout de même important pour raconter l'histoire évolutive de ces taxons. On s'aperçoit clairement sur l'arbre de droite que A et B forment un groupe dit monophylétique, c'est-à-dire partageant un ancêtre commun (qui serait au nœud de bifurcation vers A et B) ayant acquis l'état 1 du caractère. Le taxon C ne s'est pas différencié de l'ancêtre. C'est l'interprétation la plus simple de l'histoire évolutive de ce groupe, traduite par le nombre de pas le plus faible (1 au lieu de 2).

L'arbre de gauche est plus compliqué, parce qu'il montre que B et C formerait un groupe, mais que le caractère aurait évolué de l'état 0 vers l'état 1 indépendamment pour A et pour B. Ceci s'appelle une convergence ou une évolution parallèle. Dans le cas présent, cette explication est moins satisfaisante car elle nécessite d'expliquer pourquoi cet état est apparu indépendamment dans deux lignées.

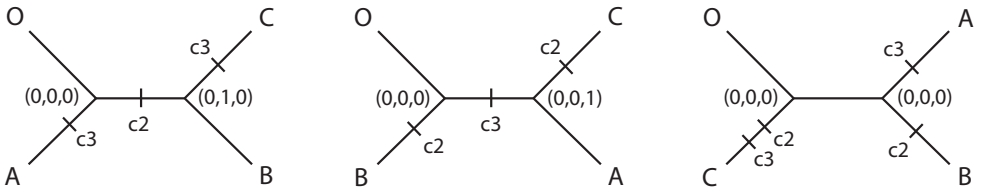
L'arbre du milieu regrouperait les taxons A et C au prix d'un premier changement d'état du caractère (0 vers 1) sur la branche venant de 0 suivi d'un changement inverse (de 1 vers 0) sur la branche menant à C. Ceci s'appelle une régression qui constitue, avec les convergences et les évolutions parallèles, des homoplasies créant du bruit dans l'analyse cladistique. En effet, elles ne représentent en rien un partage de caractéristiques issus d'un ancêtre commun mais ressemblent plus à une simple similitude due plus ou moins au hasard de l'évolution.

### 4.2.3 Plusieurs caractères à deux états

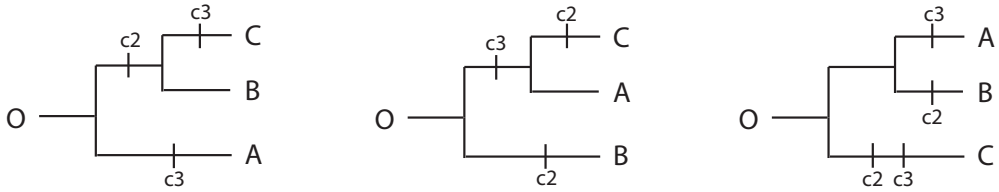
Reprenons la matrice (voir tab. 3.1) du Chapitre 3 en y ajoutant un groupe de comparaison O.

| | c1 | c2 | c3 |
|---|---|---|---|
| O | 0 | 0 | 0 |
| A | 0 | 0 | 1 |
| B | 0 | 1 | 0 |
| C | 0 | 1 | 1 |

Le même travail que pour un seul caractère peut être effectué, le nombre d'arbres possibles ne dépendant que du nombre de taxons.




Le nom du caractère dont l'état change est indiqué, et les valeurs des trois caractères sont données pour chaque nœud. À l'évidence, le caractère c1 n'est pas informatif puisqu'il est constant. Les arbres enracinés sont :

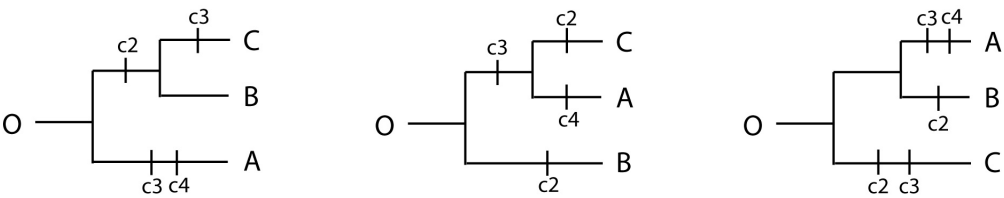



L'arbre de droite est le moins simple avec quatre pas au lieu de trois pour les deux autres arbres qu'on ne peut pas départager sur ce critère. Dans ces deux cas, nous devons conclure à la présence d'une convergence ou d'une évolution parallèle d'un des caractères c2 ou c3 selon l'arbre choisi. C'est déjà une information évolutive fondamentale, qui peut être acceptable ou au contraire inciter à réexaminer les données initiales, à recommencer l'analyse avec un autre groupe de comparaison, à considérer la possibilité qu'il n'y ait pas qu'un seul ancêtre commun pour cet échantillon, ou enfin à attendre que d'autres descripteurs deviennent disponibles pour effectuer une nouvelle analyse.

Ajoutons maintenant de l'information supplémentaire sous la forme d'un caractère c4 :

| | c1 | c2 | c3 | c4 |
|---|---|---|---|---|
| O | 0 | 0 | 0 | 0 |
| A | 0 | 0 | 1 | 1 |
| B | 0 | 1 | 0 | 0 |
| C | 0 | 1 | 1 | 0 |

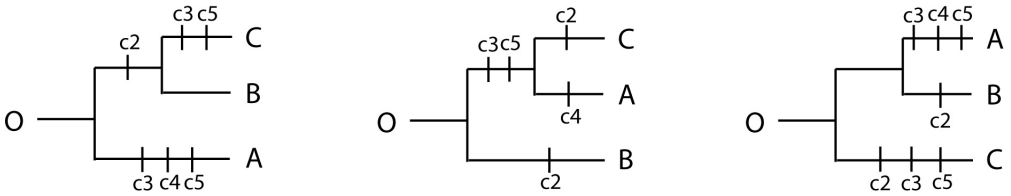
Ce caractère est non informatif puisqu'il ne caractérise qu'un seul des taxons. Il s'agit d'une autapomorphie, qui dans notre cas ne fait qu'ajouter un pas aux trois arbres. Ce caractère supplémentaire ne permet donc pas de départager les arbres de gauche et du milieu :




Enfin, ajoutons un cinquième caractère :

| | c1 | c2 | c3 | c4 | c5 |
|---|---|---|---|---|---|
| O | 0 | 0 | 0 | 0 | 0 |
| A | 0 | 0 | 1 | 1 | 1 |
| B | 0 | 1 | 0 | 0 | 0 |
| C | 0 | 1 | 1 | 0 | 1 |

Les taxons A et C partagent des états dérivés pour deux caractères (c3 et c5). Il apparaît maintenant que l'arbre du milieu est le plus simple avec 5 pas, contre 6 pour celui de gauche et 7 pour celui de droite.




Concrètement, les caractères c3 et c5 sont corrélés, ce qui ne veut pas dire qu'ils indiquent la même chose. Ils indiquent seulement que leurs évolutions sont compatibles, donnant un poids supplémentaire au signal phylogénétique qu'ils représentent. L'arbre du milieu a le nombre de pas le plus faible et minimise le nombre d'homoplasies en donnant plus de poids aux synapomorphies. En multipliant les caractères, on espère ainsi dégager le signal phylogénétique qui est donc compatible avec le maximum d'évolutions de caractères (appelées également transformations de caractères).

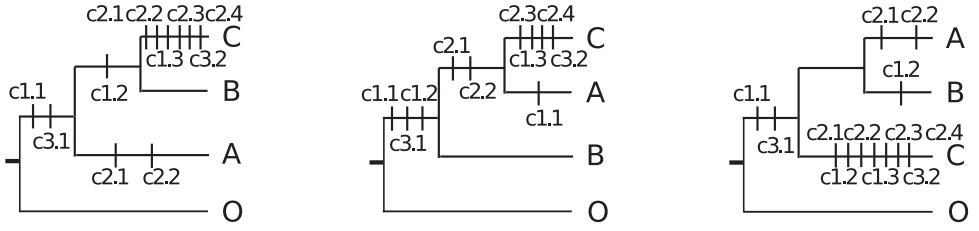
Nous voyons ainsi comment de l'information nouvelle, pouvant provenir de découvertes ou de technologies nouvelles, peut faire évoluer une phylogénie même déjà établie. Un cladogramme est conçu pour être évolutif. Cela doit être vrai également pour toute classification qui dépend nécessairement du niveau de détail dans la description des objets étudiés et de la méthode utilisée.

### 4.2.4 Caractères à plusieurs états

Les caractères peuvent souvent être représentés par plusieurs états évolutifs, ce qui permet de décrire plus finement la diversification et d'utiliser des descripteur quantitatifs et continus. Chacun des caractères peut même posséder son propre schéma évolutif, c'est-à-dire un schéma plus ou moins complexe de la transformation de ce caractère. Il est tout à fait possible d'inclure cette information dans une analyse cladistique. Dans le présent ouvrage, nous nous limiterons à une évolution simple du passage d'un état à un autre, sans embranchement, et nous illustrons ici quelques possibilités avec la matrice suivante :

| | c1 | c2 | c3 |
|---|---|---|---|
| O | 0 | 0 | 0 |
| A | 1 | 2 | 1 |
| B | 2 | 0 | 1 |
| C | 3 | 4 | 2 |

Les arbres correspondants sont :




Le chiffre situé après le nom du caractère correspond à l'état dans lequel celui-ci passe. Par exemple, c2.3 signifie que le caractère c2 passe à l'état 3. En supposant que le nombre de pas est égal à la valeur absolue de la différence entre les états (hypothèse de WAGNER, voir Sect. 4.3) nous trouvons que le nombre de pas total est, de gauche à droite, 11, 10 et 12. L'arbre du milieu est donc le plus simple. Il présente une régression du caractère c1, revenant à l'état 1 sur la branche de A, après avoir évolué de 0 vers 1 puis 2 depuis le groupe de comparaison O. Il est à noter que le cadre de cette hypothèse de WAGNER, le nombre de pas resterait inchangé si on supposait que ce caractère passe à l'état 2 (qui serait c1.2 sur l'arbre) en deux endroits, sur la branche de B d'un côté et sur la branche de C de l'autre. Il y aurait alors une évolution parallèle mais plus de régression. Seule une bonne connaissance de ce caractère pourrait ici indiquer la solution la plus plausible.

Tous ces exercices deviennent vite fastidieux avec un nombre un peu plus important de taxons et de caractères mais il est essentiel de pratiquer et de bien comprendre les exemples précédents. Concrètement, la recherche des arbres se fait par ordinateur. Cependant, mais l'exploration tous les arbres est un problème "NP-hard" c'est-à-dire qu'elle est impossible à effectuer en un temps raisonnable pour un nombre de taxons important. Les algorithmes déploient des astuces numériques, les méthodes heuristiques, qui permettent de trouver les arbres les plus simples en des temps de calculs raisonnables, sans toujours garantir qu'elles soient réellement les meilleures (voir Sect. 6.8).

## 4.3 Principe de parcimonie et critère d'optimisation.

L'échantillon de taxons et les caractères correspondants fournissent une matrice regroupant une bonne partie de l'information. La construction d'un arbre en lui-même ne dépend que du nombre de taxons, et le nombre d'arrangements possibles est très grand. La sélection du meilleur arbre s'effectue selon un critère d'optimisation basé sur les caractères et leurs états évolutifs, c'est-à-dire sur le comportement des caractères le long de l'arbre. Nous en avons vu une illustration dans les exemples précédents.

Il existe plusieurs critères (parcimonie, maximum de vraisemblance, moindre carrés, ...), mais le plus utilisé en cladistique est le principe de parcimonie : de deux solutions, la meilleure est la plus simple. C'est un principe général en Sciences, aussi connu sous le nom de rasoir d'OCKHAM. Autrement dit, le meilleur arbre est celui qui présente la phylogénie la plus simple, ou encore le schéma évolutif le plus simple, mesuré par le nombre total de changements des états des caractères. C'est ce nombre, appelé nombre de pas, qui est optimisé. On peut montrer que le principe de parcimonie minimise les homoplasies, c'est-à-dire les états finaux de caractères apparus indépen-

damment dans des lignées différentes. Il revient aussi à optimiser les évolutions de tous les caractères en les rendant toutes le plus simple possible compte tenu de l'ensemble.

Cependant, si simplifier le schéma évolutif revient à minimiser le nombre de changements d'états total, compter ce nombre de pas nécessite d'attribuer un coût évolutif pour passer d'un état à un autre. Il existe principalement quatre modèles d'évolution des caractères :

— WAGNER : les états des caractères sont réversibles et additifs. Le nombre de pas est égal à la valeur absolue de la différence entre les états. Le sens n'importe donc pas. Par exemple, passer de l'état 3 à l'état 6 coûte 3 pas, soit 3 fois plus que pour passer de 0 à 1 ou de 4 à 3. Cette hypothèse de WAGNER est également appelée celle des caractères ordonnés.
— FITCH : les états des caractères sont réversibles et non additifs. De sorte que passer de l'état 0 à l'état 3, cela coûte un seul pas, comme de 4 vers 2. On dit que les caractères sont désordonnés.
— DOLLO : les états dérivés des caractères ne peuvent apparaître qu'une fois. Les évolutions parallèles sont donc interdites. C'est une hypothèse très forte, sans doute valable dans quelques cas bien particuliers.
— CAMIN-SOKAL : les états des caractères sont irréversibles. C'est une contrainte extrêmement forte sur l'évolution du caractère, et une hypothèse qui est toujours difficile à connaître a priori et à prouver. Ce critère est très rarement utilisé.

Il apparaît maintenant évident que le principe de parcimonie ne donnera pas le même résultat selon le modèle envisagé. Ce choix doit s'effectuer pour chacun des caractères individuellement, car ils n'ont pas nécessairement tous le même comportement. Dans tout cet ouvrage, nous ne considérerons que le critère de WAGNER qui est à la fois souple et plutôt bien adapté pour les galaxies. Nous verrons au Chapitre 6 quand le critère de Fitch peut être également utile. Ces modèles d'évolution des caractères ne sont pas arbitraires, ils permettent d'injecter des connaissances supplémentaires afin de contraindre davantage l'analyse. Le choix du modèle peut facilement être modifier pour être testé et comparé.

Des schémas évolutifs plus complexes peuvent être également intégrés, lorsque des embranchements ou des sauts sont attendus pour un caractère donné. Cela peut être précisé dans l'analyse, à l'aide d'une matrice d'évolution pour ce caractère ou même d'un arbre évolutif pour le caractère.

## 4.4 Estimation de la solidité d'un arbre

Une fois le meilleur arbre obtenu selon le critère d'optimisation choisi, il reste à déterminer à quel point l'information contenue dans les données impose cette organisation hiérarchique particulière. Cette information s'appelle la force du signal phylogénétique. Il s'agit donc de mesurer ce signal par des méthodes statistiques. Nous présentons d'abord les deux méthodes les plus utilisées en cladistique pour quantifier la solidité d'un arbre, puis les indicateurs liés aux caractères.

### 4.4.1 Bootstrap

C'est la méthode la plus largement utilisée pour tester la fiabilité des branches internes (HOLMES, 2003). Le bootstrap consiste à effectuer de nombreuses réplications (1000 pour que la méthode soit statistiquement significative) de la matrice initiale par un tirage aléatoire des caractères avec remise. Ainsi, pour chaque réplique, certains caractères peuvent apparaître plusieurs fois et d'autres peuvent être absents. Ceci revient à imposer à chaque caractère un poids aléatoire compris entre 0 et le nombre total de caractères. Une analyse effectuée pour chacune des matrices répliquées produit les arbres les plus parcimonieux. Il suffit ensuite de calculer la fréquence à laquelle un embranchement donné apparaît. Les nœuds ayant une valeur de bootstrap supérieure à 95% sont considérés comme étant extrêmement fiables, alors que ceux dont le bootstrap est inférieur à 50% sont considérés comme tout à fait non significatifs.

La méthode de bootstrap a relativement peu de sens avec un petit nombre de caractères (quelques unités), mais son plus gros défaut est qu'elle est très gourmande en temps de calcul. Idéalement elle devrait effectuer autant de recherche par parcimonie que de matrices répliquées (1000 le plus souvent). En pratique, la recherche de l'arbre le plus parcimonieux est donc moins poussée que lors d'une analyse directe.

### 4.4.2 Decay index

L'indice de decay (ou de dégénérescence ou de BREMER) est défini comme le nombre de pas supplémentaires nécessaire pour faire disparaître un nœud. On regarde donc parmi les arbres moins parcimonieux la présence ou l'absence de ce nœud. Ce test a l'avantage de relativiser légèrement le critère de parcimonie qui ne retient pas les arbres un tout petit peu moins parcimonieux. Pourtant, leurs phylogénies ne sont pas vraiment plus compliquées et l'indice de decay mesure en quelque sorte la sensibilité d'une branche à ce critère de parcimonie.

### 4.4.3 Indices liés aux caractères

Le comportement des caractères le long de l'arbre peut être examiné afin de déterminer comment ils contribuent à cette phylogénie. Plusieurs indices permettent de mesurer un peu plus finement que le bootstrap ou le decay, la force et la nature du signal phylogénétique dans les données. Mais comme nous allons le voir, ils n'ont malheureusement pas de signification absolue contrairement au nombre de pas.

On définit d'abord les quantités suivantes, qu'on calcule pour chaque caractère individuellement :

— $s$ : nombre de pas sur l'arbre considéré.

— $m$ : nombre minimum possible de pas. Il se calcule pour chaque caractère en ne considérant pas les changements homoplasiques (convergences, évolutions parallèles et régressions).

— $g$ : nombre maximum possible de pas. Il correspond également au nombre minimum possible de pas dans la plus mauvaise configuration de l'arbre, à savoir quand il est totalement non résolu (aucun signal phylogénétique).

Chacune de ces quantités peut être ensuite calculée pour l'ensemble des caractères par simple addition.

On calcule ensuite les indices suivant :

— Consistency Index ou indice de cohérence $CI = m/s$.
— Homoplasy Index $HI = 1 - CI$.
— Retention Index $RI = (g-s)/(g-m)$.
— Rescaled Consistency Index $RCI = RI \times CI$.

L'indice *RI* mesure le rapport entre le nombre réel de synapomorphies et le nombre apparent. L'indice *RCI* est un meilleur estimateur que *CI*, étant plus fiable, et éliminant les autapomorphies.

Dans l'idéal, tous ces indicateurs doivent approcher 1, sauf *HI* qui doit être proche de 0, avec *RI* supérieur au *CI* dans les bons cas. Cependant, ces indices dépendent beaucoup du nombre de taxons. En particulier, l'indice de cohérence *CI* diminue quand celui-ci augmente. Ils n'ont donc pas de valeurs en tant qu'estimateurs absolus et permettent surtout de comparer plusieurs arbres issus de la même matrice. Calculés pour chaque caractère indépendamment, ils permettent de comparer les caractères entre eux en quantifiant leur comportement évolutif le long de l'arbre. Ceux qui ont les meilleurs indices représentent les caractères qui soutiennent le mieux la phylogénie, ou, dit autrement, qui sont les plus compatibles avec elle. Les moins bons d'entre eux ont probablement un comportement plutôt chaotique dans le schéma évolutif proposé par le cladogramme. Lorsque ces indices sont calculés globalement pour l'arbre, ils permettent de comparer rapidement des arbres obtenus avec la même matrice, même si le nombre de pas est prépondérant avec le critère de parcimonie.

## 4.5 Autres applications de la cladistique

Bien que développée pour la biologie, la cladistique n'est en rien liée à la nature des organismes vivants comme le montre le présent chapitre. Tout groupe d'objet pouvant être décrits par des caractères dont les valeurs peuvent représenter des états évolutifs est susceptible de bénéficier d'une analyse cladistique. La seule condition est une diversification suffisante par le biais d'une évolution par embranchement, c'est-à-dire d'une transmission avec modification. Néanmoins, l'analyse cladistique elle-même teste la validité de l'hypothèse sur le type d'évolution. Il existe au moins trois autres applications qu'il est instructif de considérer.

La premier exemple d'application concerne la propagation des copies d'ouvrages anciens (ROBINSON AND ROBERT, 1996, voir par exemple). Les caractères sont ici les mots et leur orthographe. L'idée est que pour multiplier les copies d'un ouvrage donné, les scribes recopiaient plusieurs fois un exemplaire, chacune des nouvelles copies pouvant devenir à nouveau l'exemplaire de référence d'un scribe suivant. Chaque erreur des scribes se propageait ainsi, engendrant une diversité des copies par embranchement, chaque lignée étant caractérisée par une erreur propagée. Une analyse cladistique permet alors de déterminer la copie la plus originelle.

Ce sont en réalité les linguistes qui ont inventé la méthode cladistique, même si elle a été surtout formalisée et développée par les biologistes évolutifs. Les caractères concernent à la fois l'orthographe des mots mais également leur prononciation. Le mécanisme de diversification est le plus souvent comparable à de la transmission avec modification, plusieurs langues finissant par naître à partir d'une seule. Ainsi une cartographie des origines des différentes langues à travers le monde a pu être établie (voir par exemple WELLS, 1987).

Enfin, un développement plus récent et un peu plus insolite concerne l'organisation des systèmes de production dans les entreprises dans le domaine de l'économie évolutive (KASTELLE, 2005). En effet, ces systèmes sont tous créés à partir d'une expérience transmise et qui est ensuite améliorée. Il y a donc à l'évidence transmission avec modification, sachant que de nombreuses possibilités d'amélioration sont toujours possibles, dont plusieurs probablement mises en œuvre quelque part. Un des aspects intéressants dans cette démarche est l'identification des caractères, qui concerne par exemple le partage du travail, la formation continue, les productions parallèles, les sous-traitances multiples, etc. Un des intérêts direct serait de fournir des aides dans les stratégies de développement des entreprises.

# Chapitre 5

# Astrocladistique : concepts, définitions

Nous allons maintenant montré comment il est possible d'utiliser la cladistique pour cartographier la diversité des galaxies avec une approche résolument évolutive. Dans ce chapitre, nous précisons les notions de formation, d'évolution et de diversification des galaxies (FRAIX-BURNET ET AL., 2006b,c). Ces précisions sont indispensables pour comprendre comment les définitions et méthodes exposées au Chapitre 4 peuvent se transposer aux galaxies, et permettre une application concrète détaillée dans le Chapitre 6. Nous verrons qu'aucune astrophysique nouvelle n'est nécessaire, il s'agit seulement de cadrer nos connaissances dans une perspective "galactogénétique".

## 5.1 Définition de l'objet "galaxie"

Depuis notre Terre, HUBBLE découvrit des ensembles d'étoiles très éloignés, beaucoup plus éloignés que les étoiles de notre Voie Lactée. Il montra donc l'existence de ce que Kant avait nommé les "Univers-Îles". Cette expression traduit bien la notion d'objet autogravitant indépendant. Ceci nous paraît aujourd'hui tellement évident que l'idée de définir ce qu'on entend par galaxie semble absurde. Pourtant, le développement de nos connaissances sur l'Univers, et incidemment celui de l'astrocladistique, exige une définition claire.

Notre objectif est finalement de retracer l'histoire des galaxies dans toute leur diversité. Si on pense à la formation des toutes premières galaxies, la question que l'on se pose est : "quand sont-elles apparues ?", question étroitement liéée à la nature des premiers objets de l'Univers (Sect. 3.2.4). Si on regarde l'a diversité l'ensemble' des galaxies que nous connaissons, nous constatons que peu de galaxies finalement sont réellement isolées, et que très vraisemblablement les plus isolées ne l'ont pas toujours été. L'histoire des galaxies, telle que nous la comprenons aujourd'hui, montre qu'elles sont nagent dans un environnement gravitationnel très structuré et en perpétuel changement (Sect. 3.2.3). Le caractère autogravitant des "Univers-Îles" est clairement insuffisant pour définir une galaxie.

Au fond, une galaxie, c'est quoi ? Il est certain que ces objets sont formés de trois constituants fondamentaux : les étoiles, le gaz, la poussière. Mais est-ce qu'ils sont toujours présents tous ensemble ? N'a-t-on pas des galaxies très pauvres en gaz, en

particulier chez les galaxies naines où l'autogravitation est faible ? Et n'existerait-il pas des galaxies, certainement très jeunes, formées de gaz et de quelques étoiles, sans poussière ou presque ? Même si nous ne connaissons pas nécessairement d'exemple flagrant, qui nous dit que nous n'en découvrirons pas, surtout en remontant à des redshifts de plus en plus grands ? Peu après la recombinaison, des nuages de gaz, piégés dans les halos de matière noire, ont fini par former des étoiles, qui au bout d'un certain temps ont permis la formation de poussières (Sect. 3.2.4). N'a-t-on pas ici, avec ces nuages de gaz très primordiaux, des embryons de galaxies, puisque si les petits halos de matière noire sont quasiment autogravitants, il en va un peu de même pour la matière baryonique piégée en son sein ? Mais sans poussière, et encore sans étoile. Comment désigner ces objets ?

Étant donné que la construction de ce qu'on voudrait appeler "galaxie" résulte d'une évolution continue et que leur complexification se fait de manière progressive, il serait très limitatif et sans doute stérile de s'en tenir à une définition des galaxies correspondant aux Univers-Îles que, depuis HUBBLE, nous nous efforçons de connaître dans tous les détails. Il semble plus judicieux d'adopter une définition très ouverte, qui a l'avantage énorme de rendre accessible la genèse complète des galaxies. Ces objets nous paraissent très particuliers aujourd'hui, mais étaient-ils aussi aisément identifiables il a douze ou treize milliards d'années ? L'histoire complète des galaxies doit raconter comment elles ont pu devenir ce que nous voyons, à partir d'un simple nuage de gaz, que cela soit peu après la recombinaison ou à n'importe quel autre moment dans l'histoire de l'Univers.

Nous proposons la définition suivante :

**Définition**

Une galaxie est un ensemble autogravitant d'étoiles, de gaz *et/ou* de poussières

Cette définition est très ouverte, et n'exclut aucune composante supplémentaire que nos connaissances jugeront utiles d'ajouter. Elle est volontairement ouverte pour évoluer et s'adapter aux progrès des descripteurs que nous pourront inventer, observer et cataloguer. Par exemple, on pourrait ajouter des éléments comme les trous noirs qui semblent exister au centre de toutes les galaxies. Après tout, nous savons que les premières étoiles étaient grosses et ont très probablement formé des trous noirs. La question non résolue actuellement est de savoir si les galaxies se sont formées autour de ces trous noirs.

En nous plaçant au niveau des tout premiers objets de l'Univers et en essayant de définir ce qu'on peut appeler "galaxie", il apparaît assez évident que les galaxies ne peuvent qu'évoluer. Celles que nous observons aujourd'hui n'ont certainement rien à voir avec ces objets primordiaux. Collectivement, il y a bien eu évolution de l'objet en tant que population. D'après notre définition, une galaxie change également individuellement, parce que les constituants fondamentaux servant à la définir évoluent, les étoiles vieillissant et le gaz s'enrichissant en éléments lourds. Au-delà de la définition qui précise le genre d'objets constituant la population "galaxies", la description des individus ou des différentes variétés observés correspond donc toujours à un état donné dans un schéma évolutif plus large.

## 5.2 Galactogenèse et astrocladistique

Nous avons vu aux Chapitres 2 et 3 que les galaxies sont des objets complexes en évolution, objets qu'on peut décrire par un nombre important de paramètres caractérisant les constituants fondamentaux qui sont à l'origine même de l'évolution des galaxies. De plus, les observables sont associées à des processus physico-chimiques, parfois bien compris, de sorte que l'évolution de chaque caractère peut souvent être modélisée ne serait-ce que dans ces grandes lignes. Le concept d'états évolutifs d'un caractère a donc ici un sens évident, même si la quasi-totalité des variables sont quantitatives et continues.

En conséquence, les ingrédients sont réunis pour entreprendre une analyse phylogénétique des galaxies. L'objectif en est la compréhension de la galactogenèse, définie comme l'étude de la diversification des galaxies. Pour toute analyse phylogénétique, indépendamment de la méthode utilisée, il est indispensable de détailler la vie d'une galaxie et des galaxies dans leur ensemble, et de définir proprement les termes, afin de déterminer précisément les processus de diversification. C'est ce que nous allons faire dans la suite de ce chapitre.

Différentes approches sont possibles. La cladistique en est une, idéalement adaptée au cas d'une évolution par embranchement dont un ingrédient nécessaire est le processus de transmission avec modification. Des considérations théoriques donnent des indications a priori sur ce point (ce chapitre), et la mise en œuvre de la méthode confirmera ou infirmera l'intérêt de l'approche (chapitres suivants). Si l'évolution par embranchement n'est pas vérifiée, ou seulement partiellement, il faudra alors envisager des modifications à l'analyse cladistique, en prenant en compte une éventuelle évolution réticulée (Sect. 3.1.4).

Nous avons désigné par "astrocladistique" l'application de la cladistique au cas des galaxies. Il s'agit à la fois d'une méthodologie et d'un outil pour l'étude de la galactogenèse. Les concepts et définitions posés dans ce chapitre ne sont pas spécifiques à la cladistique, ils traduisent l'approche phylogénétique en astrophysique dont l'astrocladistique est la toute première tentative. À notre sens, ils répondent plus largement à un besoin de clarification de nos connaissances modernes sur les galaxies dans le paradigme cosmologique actuel (voir Chapitre 3).

## 5.3 La formation des galaxies

Comment les galaxies se sont-elles formées ? Grande question qui en contient réellement deux distinctes : d'un côté l'apparition des toutes premières galaxies de l'Univers, et d'un autre côté l'origine des autres galaxies que nous observons à tout redshift. La première question relève entièrement de la cosmologie, de l'étude de l'évolution des conditions physiques de l'Univers jusqu'à la première possibilité de former des objets qu'on pourrait appeler galaxies. La seconde question relève à la fois de la cosmologie (évolution ultérieure aux premiers objets des conditions physiques de l'Univers) et de la physique extragalactique, c'est-à-dire de l'étude de la vie des galaxies : comment peuvent-elles se former, évoluer, disparaître ? Cela concerne davantage l'histoire évolutive des individus.

Quand les galaxies sont-elles apparues ? Là encore, deux aspects doivent être dis-

tingués. Premièrement, l'apparition des toutes premières galaxies détermine le début de l'histoire des galaxies en tant que population, par rapport à l'histoire de l'Univers. Elle encadre la recherche de l'espèce ou des espèces ancêtrales communes à toutes les galaxies de l'Univers. Deuxièmement, l'évolution établie des galaxies implique qu'elles se transforment en produisant de nouvelles variétés. La question ici est de savoir quand une espèce de galaxies particulière est apparue pour la première fois dans l'histoire de l'Univers. Par exemple, quand sont apparues les premières galaxies elliptiques, ou massives ?

Il y a une ambiguïté à parler de formation des galaxies si on ne distingue pas le "comment" du "quand", et ensuite si on ne précise pas "dans l'état observé", qu'il s'agisse des toutes premières galaxies ou non. En somme, lorsque nous observons une galaxie particulière, nous devons nous poser la question de savoir quand et comment ce genre d'objet, avec toutes ses propriétés spécifiques, est apparue pour la première fois dans l'Univers. Cette ambiguïté se retrouve dans deux modèles de formation des galaxies bien connus. Le premier, dit monolithique, suppose, pour simplifier, qu'elles se sont toutes formées presque en même temps par effondrement d'un nuage de gaz dans un état proche de ce que nous observons aujourd'hui. Ici, le terme de formation semblerait signifier à la fois la formation initiale (la première apparition d'une galaxie donnée en tant qu'individu) et primordiale (la première apparition d'un objet appelé galaxie). Le deuxième modèle, dit hiérarchique, suppose que des galaxies, petites et à peu près identiques initialement, se sont formées en même temps et ont fourni les graines pour former par fusions et accrétions les galaxies actuelles, en particulier les plus massives d'entre elles. Le terme de formation est ici ambigu, car est-ce que les petites galaxies ont évolué pour devenir grosses, auquel cas formation et évolution sont le même phénomène, ou est-ce que les grosses galaxies se sont formées à partir des petites, auquel cas l'évolution n'a plus de place ? Autrement dit, chaque modèle précise le "comment", mais implique plus ou moins des idées du "quand" : dans le modèle monolithique il n'y a qu'une seule période de formation, ne laissant plus beaucoup de place à l'évolution, alors que dans le modèle hiérarchique, les galaxies se forment à peu près continuellement durant l'âge dans l'Univers, mais il tend à mélanger les concepts de formation (initiale) et d'évolution.

L'ambiguïté précédente s'explique parce qu'aucune distinction claire n'est faite entre individu et population, entre évolution d'un individu et évolution de la population, entre formation d'un individu particulier avec ses caractéristiques spécifiques et formation d'une espèce de galaxies. Bien sûr, la notion d'espèce n'existe pas en astrophysique extragalactique parce qu'aucune classification multivariée et évolutive n'a encore été mise au point. L'astrocladistique veut répondre à ce manque qui empêche une vision claire de la formation des galaxies en tant que population, c'est-à-dire une véritable galactogénèse.

La réalité fait certainement appel à un processus de formation graduel, continu, même pour les galaxies primordiales, une sorte de panachage entre ces deux modèles un peu extrêmes. Nous en avons discuté dans le Chapitre 3. Il est largement reconnu aujourd'hui que les galaxies changent, et donc que leur formation se déroule tout au long de l'âge de l'Univers. Le modèle hiérarchique est préféré aujourd'hui, mais pourtant la sémantique n'en est pas totalement clarifiée pour autant. De plus, ce modèle est basé sur un critère principal, la masse, qui est assez bien corrélé à la taille et la forme des galaxies. Bref ce modèle est étrangement lié à la classification de HUBBLE

et surtout à l'évolution hiérarchique des halos de matière noire, et peu corrélé à la variété des observables disponibles aujourd'hui. Il faut noter également que le modèle d'effondrement monolithique regagne du terrain grâce à des observations de grosses galaxies à des redshifts très élevés, laissant ainsi trop peu de temps pour que le modèle hiérarchique s'applique à ces objets.

Nous définissons donc la formation comme étant la formation d'une galaxie ou d'une classe de galaxies telle que nous pouvons l'observer à un instant donné. La formation correspond à la première apparition d'un objet ou d'une classe. Comme elle est le résultat de toute une longue histoire, un héritage d'une succesion de processus d'évolution, de transformation (Sect. 5.4) la notion de formation inclut nécessairement la notion d'évolution, c'est-à-dire de changement, de modification au cours du temps. Lorsque les modifications deviennent importantes, le nouvel objet ainsi formé devient suffisamment distinct de son progéniteur, ou ancêtre, plus ou moins lointain, pour qu'on les range dans deux classes différentes. C'est exactement la même chose avec les espèces vivantes pour lesquelles on parle plus volontiers de l'apparition d'une espèce que de sa formation. Si on reprend l'exemple du modèle hiérarchique de formation des galaxies, de nombreuses fusions et accrétions, à partir de galaxies naines initialement, ont eu lieu avant que les grosses galaxies actuelles apparaissent comme nous les connaissons. Est-ce que ces galaxies naines ressemblent à celles que nous observons aujourd'hui ? Des galaxies massives ont aussi pu se former bien plus vite, peut-être selon le scénario monolithique, cohabitant donc dans nos observations avec des objets de tailles semblables mais d'origines très différentes. Cependant, tout ceci ne concerne qu'un seul aspect très apparent des galaxies, et laisse de côté la plus grosse partie de la physique et de la chimie des galaxies. L'histoire ainsi racontée est très incomplète. La notion de formation doit donc intégrer tous les évènements évolutifs de tous les ingrédients qui la composent et la caractérisent, évènements qui ont pu apparaître à des moments différents, avec des intensités différentes pour chacune des lignées menant aux galaxies que nous observons. C'est cette histoire complexe de la formation des galaxies qu'il nous faut restituer.

En résumé, la formation des galaxies est la suite des processus d'évolution qui ont généré les propriétés spécifiques d'une galaxie ou d'une espèce de galaxies donnée. La diversification est l'ensemble des évènements qui ont créé, au cours de l'histoire de l'Univers, la diversité des objets, existants ou ayant existé, qui constituent la population des galaxies.

## 5.4 Les processus de transformation des galaxies

Au Chapitre 3, nous avons finalement montré que l'évolution, au-delà du sens générique de changement au cours du temps, des galaxies concerne les individus. Nous allons préciser maintenant le mécanisme de cette évolution qui s'apparente davantage à une transformation. En effet, elle est intimement liée à la physique d'une galaxie et à ce qu'on pourrait appeler la vie d'une galaxie. Nous pouvons identifier cinq processus élémentaires d'évolution, c'est-à-dire cinq processus par lesquelles une galaxie se modifie, change ses caractéristiques, se transforme donc : l'assemblage, qui correspond à la création d'un nouvel individu, l'évolution séculaire qui frappe toutes les galaxies, les interactions qui perturbent et stimulent la précédente, l'acquisition et la perte de

matière, pouvant toutes les deux se dérouler d'une manière plus ou moins violente.

Ces processus sont ici individualisés car ils correspondent chacun à une physique et un évènement bien particuliers. Cette façon de présenter à le double avantage de détailler les mécanismes concrets par lesquels une galaxie peut se transformer, et de démontrer qu'un phénomène de transmission avec modification existe bel et bien chez les galaxies. Ces évènements peuvent tout-à-fait être plus ou moins simultanés, cela ne change rien à cette conclusion.

### 5.4.1 Assemblage

Ce terme d'assemblage montre bien l'apparition d'un objet qu'on décide d'appeler galaxie et adhère parfaitement à la définition donnée en Sect. 5.1. Il s'agit de l'assemblage d'une certaine quantité de matériau, des étoiles, du gaz et/ou de la poussière. L'assemblage correspond à une formation initiale, mais pas nécessairement primordiale au début de l'Univers, d'une galaxie qui n'a pas été engendrée par une autre galaxie. Il s'agit d'un ensemble auto-gravitant d'étoiles, de gaz et/ou de poussières qui ne résulte a priori pas de la transformation d'une galaxie parente ou progénitrice par les processus listés ci-dessous. On imagine aisément que peu après la recombinaison, de nombreuses galaxies se sont ainsi formées pour la première fois. Cependant, il semble tout-à-fait possible que des galaxies se créent en d'autres instants de l'Univers, peut-être même encore actuellement, à partir de nuages intergalactiques, de fragments arrachés à plusieurs galaxies, voire plus hypothétiquement de trous noirs solitaires. Chaque composante peut avoir déjà une histoire (par exemple du gaz éjecté de supernovae ou des étoiles déjà formées), mais l'objet constitué par l'ensemble des composantes n'est pas le descendant direct d'un processus de transformation d'une galaxie.

### 5.4.2 Évolution séculaire

Aucune galaxie n'échappe à une évolution interne indépendante de tout stimulus externe. Au minimum, les étoiles vieillissent, explosent, se forment et interagissent avec leur environnement pour modifier la chimie interstellaire, la composition du gaz, de la poussière, les mouvements orbitaux, tout ceci change de manière et à un rythme globalement imprévisibles. La complexité intrinsèque de ces objets fait que plusieurs galaxies identiques, isolées, vont inévitablement se transformer différemment, engendrant l'apparition de nouvelles espèces de galaxies. Le processus peut être plutôt lent par rapport à l'âge de l'Univers, de sorte que la diversification sera sans doute modeste mais bien réelle.

### 5.4.3 Interaction

Les galaxies n'étant que rarement isolées du fait principalement de la portée infinie de la Gravitation et de la grande quantité de matière noire, les interactions sont maintenant connues pour être un moteur essentiel de la modification des propriétés des galaxies, donc de leur évolution. Ce processus est nécessairement diversifiant pour deux raisons. La première est que les paramètres d'impacts sont aléatoires rendant les conditions de la rencontre (vitesses relatives, angles relatifs, distances, champ gravitationnel dû à la matière noire ou à d'autres galaxies) toujours différentes. La deuxième

raison vient de la nature complexe des objets qui font que les conséquences seront très variables.

Les interactions génèrent une grande diversité. Tout d'abord les conséquences d'une interaction seront toujours beaucoup plus importantes que pour l'évolution séculaire, de sorte que l'objet après l'interaction sera certainement assez différent de celui avant pour pouvoir le ranger dans une classe distincte. Ensuite, dans le cas où deux galaxies suffisamment différentes pour appartenir à deux classes différentes interagissent, il est très probable que les deux galaxies après l'interaction soient également très différentes d'avec les progénitrices et entre elles, montant à quatre le nombre total de classes en jeu. Ce processus est donc hautement diversifiant et probablement très fréquent.

### 5.4.4 Fusion – accrétion

Le processus de fusion est en quelque sorte une interaction catastrophique. Lorsque les masses des galaxies sont comparables, on parle de fusion majeure, sinon on parle de fusion mineure. Dans les deux cas, un phénomène essentiel se produit, et il semble souvent oublié : de deux galaxies, on obtient un unique nouvel objet. Il y a donc disparition d'au moins l'un des deux protagonistes. Mais n'y a-t-il pas plus simplement disparition de deux objets et apparition d'un nouveau ? Ceci est évident dans le cas des fusions majeures, et en particulier dans le scénario bien connu de fusion de deux galaxies de forme spirale qui le plus souvent résulte en une galaxie de forme elliptique. Les deux galaxies spirales ont donc bel et bien disparu et ont transmis leur matériau pour former un tout nouvel objet. Pour les fusions mineures, on aurait tendance à penser que seule la plus petite disparaît. Injustice, car on doit nécessairement retrouver des traces de sa présence, au sein de la plus grosse. Cette dernière n'est d'ailleurs pas nécessairement très ressemblante après l'évènement si on prend le soin de regarder toutes les propriétés en détail.

Il est intéressant de noter que les conséquences des fusions majeures sont souvent qualifiées de catastrophiques. Cela est essentiellement définie à partir de le morphologie, ce qui constitue un critère très subjectif. De plus, il est limitatif en terme de diversification car il n'est pas impossible que les autres formes de fusion ou accrétion ne soient pas aussi perturbantes pour un ensemble d'autres descripteurs comme la cinématique, les compositions chimiques ou les populations stellaires. De ce point de vue, les interactions sont sans doute souvent assez catastrophiques.

L'accrétion s'assimile beaucoup à une fusion mineure, sauf qu'initialement une seule galaxie est présente. Le matériau accrété n'est pas désigné par le terme de galaxie mais faudrait-il encore que la définition d'un tel objet fût bien posée. Selon notre définition (Sect. 5.1), beaucoup d'accrétions pourrait finalement être des fusions mineures ! Quoi qu'il en soit, la galaxie accrétante ne sera plus tout-à-fait la même, donc là encore, un individu disparaît pour engendrer par transmission un nouvel objet nécessairement différent.

L'accrétion et la fusion font grossir les galaxies, elles sont les processus privilégiés par le scénario hiérarchique de formation des galaxies. Cela n'implique cependant pas que toutes les galaxies grossissent. Les fusions majeures sont certainement assez rares, les fusions mineures et l'accrétion dépendent beaucoup de l'environnement et de l'époque de l'Univers. Il doit donc exister des galaxies qui n'ont pas grossi beaucoup.

Parmi elles, il n'est pas interdit d'imaginer des galaxies assez massives et formées par assemblage. Le seul critère de taille ne suffit à l'évidence pas pour distinguer entre ces deux chemins évolutifs pourtant très différents.

Nous y reviendrons un peu plus loin, mais le processus de fusion illustre parfaitement bien comment une nouvelle galaxie se forme à partir de matériau transmis directement par ces ancêtres directs. Dans la majorité des cas cette transmission ne se fait pas sans dégâts, c'est-à-dire sans modifications. La diversification est donc forte, avec le plus souvent trois objets très différents, donc sans doute trois classes, intervenant au cours d'un tel processus.

### 5.4.5 Éjection – balayage

L'éjection se rapporte davantage aux étoiles et le balayage au gaz, mais tous les deux sont un peu l'opposé du processus de fusion – accrétion. L'éjection d'un paquet d'étoiles se fait principalement par effet de marée, lors d'une rencontre proche entre deux galaxies. Cet éjectat peut former des galaxies, petites, qui sont appelées naines de marée. C'est un cas évident de formation de galaxies à n'importe quel âge de l'Univers puisque d'une galaxie sont nés deux objets différents. Nous pouvons avoir ici trois classes en tout. Chacune des deux nouvelles galaxies est formée à partir du matériau transmis par la galaxie initiale, matériau modifié suite à la forte perturbation subie ne serait-ce qu'en ce qui concerne la dynamique. .

Le balayage du gaz ou de la poussière peut s'envisager lorsqu'une galaxie passe à travers un nuage intergalactique, mais également lors d'une rencontre. Seules deux classes (une avant, une après) peuvent donc être identifiées durant ce processus à moins que le gaz forme un ensemble auto-gravitant, ce qui est peut probable. Il faut cependant noter que le matériau éjecté, étoiles, gaz ou poussière, peut très bien rester diffus dans un premier temps et donc ne pas entrer dans la définition d'une galaxie, puis se rassembler plus tard en un tel objet. Il s'agirait alors là d'un cas d'assemblage tel que discuté plus haut.

## 5.5 La diversification des galaxies

La diversification des galaxies passe nécessairement par les cinq processus de transformation évoqués ci-dessus. Pour montrer que l'évolution des galaxies est diversifiante, nous allons envisager plusieurs scénarios.

### 5.5.1 Objets simples, sans interaction et apparus au même instant de l'Univers

Si les objets sont suffisamment simples, leur évolution intrinsèque est toute tracée, toutes les galaxies changent de la même manière, au même rythme, donc en même temps. On peut leur définir facilement un âge. Cela serait par exemple le cas d'étoiles de masse et de métallicité données apparues toutes au même instant, leur trajet évolutif est parfaitement déterminé. Puisqu'il n'y a pas d'interaction, les galaxies ne risquent pas d'être perturbées. Dans ce scénario, on pourrait penser qu'une seule classe existe, l'évolution se faisant au sein de cette classe. Néanmoins, les objets peuvent avoir suffisamment changé pour qu'on ait du mal à les reconnaitre. Dans l'exemple d'une étoile,

les astronomes distinguent plusieurs types correspondant à plusieurs stades par lesquelles une même étoile passe au cours de son évolution. Il existe donc nécessairement plusieurs classes de galaxies qui sont apparues dans l'histoire de l'Univers, même si c'est le même objet qui se transforme. C'est la notion de lignée si bien représentée sur un cladogramme. Dans notre cas où toutes les galaxies évoluent au même rythme, à chaque instant, c'est-à-dire à chaque redshift, un seule classe est présente, les représentants des autres classes ayant tous disparus ou n'étant encore pas apparu. En conséquence, il y a bien diversité engendrée par l'évolution.

La notion d'âge peut être un peu plus complexe déjà dans un cas aussi simple. On peut effectivement parler d'un âge égal au temps depuis le tout premier assemblage de tous les objets. Cependant, il peut être plus adéquat de définir aussi un âge au sein de chaque classe, correspond au temps écoulé depuis qu'un objet est entré dans cette classe.

### 5.5.2 Objets simples, sans interaction et apparus en des instants différents de l'Univers

Si les objets sont initialement apparus en des temps différents, autrement dit si le processus initial d'évolution ne commence pas au même instant pour tous les objets, alors en un instant donné il existera plusieurs classes correspondant à différents états évolutifs. La notion de population à un redshift donné perd déjà un peu de signification puisqu'il s'agit d'une même population à différents stades évolutifs. Nous avons donc bien diversification, celle-ci étant davantage visible dans les observations.

S'il est encore possible de déterminer un âge pour chaque individu ou pour chaque classe, il n'est déjà plus possible de le définir collectivement. Ainsi parler de la période de formation des étoiles géantes rouges n'a pas de sens. À la rigueur peut-on parler de l'époque d'apparition des premières géantes rouges dans l'Univers.

### 5.5.3 Objets complexes, sans interaction

Dans un objet complexe, de nombreux processus évoluent chacun à leur rythme de manière généralement non linéaire, et peuvent interagir entre eux, de sorte qu'il est impossible de prédire leurs comportements avec exactitude. Comment alors donner un âge à de tels objets ? Même en imaginant pouvoir déterminer un âge à partir de leur première apparition, cette quantité ne refléterait pas du tout leur état évolutif. Par ailleurs, il est impensable que tous les objets initiaux puissent suivre rigoureusement le même chemin évolutif. On s'attend donc inévitablement à une diversification vraie, à l'apparition de plusieurs classes à partir d'une même classe, indiquant une évolution par embranchement.

### 5.5.4 Objets complexes, avec interactions

Lorsqu'en plus de la complexité des objets, on autorise les interactionsqui sont un moteur essentiel de l'évolution des galaxies, alors la diversification sera maximale puisque les paramètres d'impact sont aléatoires. Par contre l'évolution par embranchement pourrait sembler en défaut lors des phénomènes d'accrétions et fusions puisqu'ainsi on hybride deux lignées (mais voir Sect. 5.6.3).

Si le destin des galaxies est de fusionner toutes sous l'influence de la gravitation, alors il ne restera plus qu'un seul objet à la fin. Malgré tout, il y aura bien eu diversification, même si la totalité des classes, sauf une, auront disparu. Il semblerait que l'Univers soit en expansion accélérée, rendant improbable ce scénario catastrophique.

Enfin, nous avons vu que les galaxies interagissent beaucoup avec leur environnement, et sont notamment sous l'influence de la matière noire et plus généralement de la distribution de la matière gravitationnelle. Donc les processus d'évolution des galaxies liés aux perturbations extérieures dépendent aussi du lieu dans l'Univers, ce qui ajoute un élément supplémentaire à la diversification.

## 5.6 La diversité engendrée par l'évolution

Nous avons précisé la notion de formation des galaxies, identifié les cinq processus d'évolution, et détaillé les mécanismes de diversification. Nous pouvons maintenant aborder le point de vue de la systématique et montrer pourquoi l'approche cladistique semble tout-à-fait justifiée dans le cas des galaxies.

### 5.6.1 Transmission avec modification

Dans chacun des cinq processus d'évolution, le nouvel individu est constitué de matériau transmis par son ou ses progéniteurs. Même dans le cas de l'assemblage, le matériau peut avoir été balayé d'une galaxie, un peu à la manière des étoiles qui se forment à partir de gaz enrichi au sein d'autres étoiles ayant explosé en supernovae. Il y a donc toujours transmission, de gaz, de poussière, d'étoiles, c'est-à-dire d'un matériau qui porte inévitablement son histoire avec lui, incluant toutes les modifications qu'il a subi lors des processus évolutifs précédents.

Il y a bien entendu toujours modification, par définition de l'évolution et telle que précisé à travers les cinq processus élémentaires. L'importance perçue et la pertinence de cette modification dépendent de la finesse de la description complète de la galaxie. Dans le cas des espèces vivantes, les modifications sont très très faibles à chaque génération, ne justifiant pas la distinction d'une nouvelle espèce à chaque fois. Pour les galaxies, chacun des processus d'évolution va créer un nouvel individu qu'il ne sera parfois ni facile ni utile de distinguer de son progéniteur. Cependant, les processus faisant appel à des perturbations extérieures sont probablement assez violents et peuvent impliquer une modification assez radicale des objets.

Une fois assemblée, chaque galaxie va être invariablement soumise à l'un des quatre autres processus d'évolution. En toute rigueur, son matériau pris séparément pourra être soumis aux cinq évènements possibles qui ont des probabilités d'apparaître plusieurs fois au cours du temps de HUBBLE (FRAIX-BURNET ET AL., 2006c). L'histoire de la diversité des galaxies est donc écrite dans les transmissions avec modification subies par les constituants fondamentaux qui forment les galaxies.

Le processus de diversification des galaxies passe donc par un ensemble de cinq processus élémentaires d'évolution, répétés plusieurs fois dans l'histoire d'une galaxie, avec des probabilités, donc des fréquences, différentes. Le terme adopté est "histoire" et non pas "vie" d'une galaxie, car il doit être clair maintenant que la vie d'une galaxie ne dure que le temps entre deux processus évolutifs consécutifs. Au cours de la

vie d'une galaxie, il ne se passe finalement pas grand chose. Lorsqu'elle change de classe, elle disparaît au profit d'un autre objet. L'histoire d'une galaxie, héritage de sa genèse et de tous ses ancêtres, est essentielle pour expliquer ses caractéristiques. Comprendre la diversification des galaxies consiste donc à regrouper des galaxies ayant eu des histoires similaires en retraçant au mieux les enchaînements des divers processus d'évolution. Seule la description aussi fine et complète que possible des composants fondamentaux des galaxies peut fournir des indices suffisamment précis.

### 5.6.2 Une évolution par embranchement

À chaque processus d'évolution, une galaxie va donner naissance à une galaxie, voire plusieurs, aux propriétés différentes. Ce nouvel individu appartiendra ou non à la même classe selon son degré de changement et selon la définition de cette notion. Lors de l'évolution séculaire, ou lors d'accrétions ou de fusions de faible importance, il est possible que le changement de classe ne se fasse qu'après plusieurs cycles de modifications, un peu à la manière des organismes vivants dont une espèce différente apparaît après plusieurs générations d'individus. Mais en général pour les galaxies, ce changement sera plus brutal et ne nécessitera qu'un évènement parmi les processus évolutifs.

Une galaxie va donc le plus souvent donner naissance à un objet d'une autre classe comme nous l'avons vu dans la description des cinq processus évolutifs. La complexité des galaxies et des processus va inévitablement engendrée plusieurs classes différentes à partir de la même classe. La diversification va ainsi se produire par embranchement, impliquant une organisation hiérarchique de la diversité des galaxies représentable sous la forme d'un arbre. L'application de la cladistique aux galaxies est donc dores et déjà justifiée. Les analyses astrocladistiques effectuées à ce jour ont confirmé cette déduction (FRAIX-BURNET ET AL., 2006b,c,a; FRAIX-BURNET, 2006; FRAIX-BURNET ET AL., 2009).

### 5.6.3 Les fusions sont-elles assimilables à des hybridations ?

À première vue, les fusions pourraient ressembler à des hybridations et donc s'apparenter à une évolution horizontale ou réticulée. Nous avons vu chez les espèces vivantes que les hybridations et les échanges de gènes pouvaient perturber fortement une analyse cladistique. Est-ce que la fusion de deux galaxies n'est pas une hybridation par excellence ?

Imaginons deux galaxies issues de deux lignées différentes, donc ayant chacune hérité de nombreuses caractéristiques propres à ses ancêtres. Elles peuvent partager certains états de caractères par homoplasies (convergences ou évolutions parallèles), mais leurs histoires se distinguent très nettement. Par exemple l'une a pu hériter d'un énorme trou noir central, qui a influencé toutes les générations futures, et l'autre a pu hériter d'une très faible quantité de gaz. Ces deux galaxies se rencontrent, s'attirent et fusionnent pour ne former qu'un seul objet. Dans notre exemple, cela pourrait donner un trou noir massif et peu de gaz. N'y a-t-il pas d'autres moyens d'en arriver là ? Probablement pas, car le trou noir massif retiendrait plus facilement le gaz autour de lui, et inversement, sans gaz il y a peu de chance de nourrir le trou noir pour le faire

grossir. Il semblerait donc que le nouvel objet soit un hybride, un mélange bizarre de deux objets plus “standards”.

Le raisonnement ci-dessus est cependant faux pour deux raisons. La première est que nous oublions tous les autres descripteurs et qu'il est dangereux de ne raisonner que sur des paramètres restreints. Considérer ces galaxies seulement avec le trou noir et le gaz semble bien trop simpliste et certainement irréaliste. Si on prend en compte l'ensemble des caractères et leurs états évolutifs, le côté hybride du nouvel objet serait très probablement moins apparent.

La deuxième raison est que les constituants fondamentaux des galaxies ne sont pas des gènes. Ces derniers se remplacent les uns aux autres au sein des chromosomes des organismes vivants : lors d'une hybridation, un gène chasse l'autre et l'expression du premier prend la place du second. Pour les galaxies, il s'agit toujours d'un simple mélange des constituants et donc de leurs propriétés. De plus, lors de la fusion, de nombreux caractères vont changer, et même violemment. Le nouvel objet ne sera donc pas une addition exclusive de propriétés comme pour les hybridations, mais une dérivation de propriétés nouvelles obtenues à partir d'une addition inclusive et perturbante.

Dans le cas des fusions mineures, on imagine bien que la classe à laquelle appartient la plus grosse des galaxies va engendrer ainsi une nouvelle classe, apparaissant comme une nouvelle branche divergente, à condition toutefois que la trace de la galaxie absorbée puisse être détectée. Mais dans le cas des fusions majeures de deux galaxies différentes et de masses semblables, quelle lignée considérer ? Nous ne pouvons exclure que la nouvelle classe apparaisse sur l'arbre comme la connexion entre deux branches.

L'évolution réticulée n'est donc a priori pas impossible dans le cas des fusions de galaxies. Cependant, il est probable que ces évènements soient beaucoup moins fréquents que les fusions mineures et les interactions. Ainsi, l'évolution par embranchement domine très certainement, comme semble le prouver les premiers résultats de l'astrocladistique, et en particulier une analyse effectuée sur des galaxies simulées (FRAIX-BURNET ET AL., 2006c).

Il faut tout de même insister sur le fait que l'analyse cladistique est un test de l'évolution par embranchement. Le résultat de l'analyse, sa robustesse, sa signification physique ou biologique, révèle la possible organisation de la diversité et son intérêt. Il existe nombre d'outils adaptée à la réticulation même s'ils reposent sur la notion de réseaux qui sont plus difficiles à interpréter que les arbres.

### 5.6.4 Une notion d'espèce pour les galaxies ?

Nous avons vu au Chapitre 2 que l'intérêt d'une classification réside à la fois dans la simplification de la description de la diversité et dans la possibilité ainsi offerte de comprendre les raisons de cette diversité. L'évolution des galaxies consiste donc en une sucession de transformations selon les processus décrits précédemment. L'ampleur du changement est très variable. Il peut être inobservable avec nos moyens actuels, ou au contraire apparaître flagrant sur un ou plusieurs des descripteurs accessibles à l'observateur. Il peut être complexe en influençant de nombreux paramètres, mais il peut également ne pas être très significatif du point de vue de la diversification des galaxies. Le danger ici est de mesurer cette ampleur du changement à l'aide de critères subjectifs.

Le degré de modification des propriétés d'une galaxie doit pouvoir être estimé objectivement car le processus de diversification des galaxies ne dépend pas des capacités et compétences de l'observateur. On voit ici pleinement l'intérêt d'une approche multivariée. Il est donc impératif de prendre en compte toutes les observables à notre disposition à un instant donné, sachant que cette information va évoluer avec le temps.

Il est ensuite nécessaire de synthétiser ensuite la diversité des galaxies ainsi obtenue sur un schéma évolutif, puis de classer. L'intérêt immédiat est de mettre en lumière les processus d'évolution les plus pertinents et les conditions dans lesquelles ils le deviennent. Dans quel cas l'évolution séculaire est-elle suffisante pour rendre utile la définition de deux classes distinctes et identifiables ? Dans quelle configuration une interaction produit-elle une ou deux classes nouvelles ? Mais avant toute chose, comment reconnaître ces évènements dans les descripteurs ?

Nous connaissons bien les avantages pratiques de la notion d'espèce en biologie, même si elle est quelque peu remise en question par la cladistique. Nous avons vu (Sect. 3.3.3) qu'elle a plusieurs définitions possibles. Le terme lui-même n'est d'ailleurs en aucun cas spécifique de la biologie. Nous avons évité dans ce livre de l'employer à propos des galaxies, mais son usage futur n'en est pas interdit. Les notions de classe, groupe évolutif et clade, ont des définitions précises suffisantes pour cet ouvrage. Nous reviendrons plus loin dans ce livre, en particulier dans le dernier chapitre, sur les concepts de classification et de taxonomie. Notons tout de même que le mot "type" est à éviter pour les galaxies car il est invariablement associé au type morphologique.

Sur quels critères classer ? La systématique nous apporte la réponse : la classification ne peut se faire qu'a posteriori. Concrètement, lorsqu'on utilise la cladistique pour retracer l'histoire évolutive à partir des descripteurs, c'est l'arbre final cartographiant la diversité qui pourra servir de guide pour définir une classification pertinente vis à vis de l'évolution. Cette classification sera ainsi basée sur une analyse objective, et elle sera adaptée à la fois à la finesse des descripteurs (on pourrait dire la résolution du signal phylogénétique) et à une signification réelle en terme de diversité évolutive des galaxies.

La classification ne peut intervenir qu'après regroupement des objets d'étude. Ces regroupements nécessitent une méthode associée à des critères. Il faut donc commencer les analyses astrocladistiques, afin d'établir des cladogrammes, avec des galaxies prises comme taxons-individus. L'arbre n'est pas un arbre généalogique des individus, les nœuds de l'arbre représentent un taxon-ancêtre hypothétique mais volontairement non identifié (voir Chapitre 4). Chaque galaxie est alors considérée comme typique d'une classe, même si éventuellement elle est unique.

L'analyse cladistique est décorrélée de la classification proprement dite puisque cette dernière nécessite d'abord de regrouper les objets. La notion d'espèce ou de classe ne peut donc pas être définie trop tôt dans l'avancée de l'astrocladistique. D'ailleurs, son utilité n'apparaît que si les liens entre les différentes catégories sont établies, et ce sont les cladogrammes qui les visualisent. Il faut donc procéder méthodiquement et effectuer d'abord des analyses cladistiques d'échantillons conséquents et ensuite seulement apprendre à lire des arbres dont les interprétations aboutiront sur une ou plusieurs classifications. Ces points seront repris plus en détails au cours des prochains chapitres.

## 5.7 Extension des concepts : le lien évolutif à travers l'environnement

L'astrocladistique a été initialement pensée pour les galaxies qui présentent un mécanisme de transmission avec modification menant à la diversité. Étant donnée l'intrication entre les objets astrophysiques et leur environnement, la structure de l'Univers, il apparaît que ce mécanisme peut s'appliquer à d'autres objets astrophysiques. Nous l'illustrons ici sur deux exemples.

### 5.7.1 Évolution hiérarchique des halos de matière noire

Les halos de matière noire grossissent par fusion (Sect. 3.2.3). C'est un processus hiérarchique par excellence, impliquant des objets caractérisés uniquement par leur masse ou leur taille. Ce processus présente un mécanisme de transmission avec modification simple : les masses de deux halos qui fusionnent sont additionnées.

La figure 5.1a (FRAIX-BURNET, 2009) illustre la représentation habituelle sous forme d'arbre de fusions. Ce genre d'arbre est en réalité un arbre généalogique, tel qu'il peut être directement établi à partir des simulations numériques. Cette représentation est toutefois limitée puisqu'il est impossible de représenter l'ensemble des halos existant ou ayant existé de notre Univers. Il peut également induire en erreur, la plupart des arbres de fusions schématisés laissant penser que tous les petits halos disparaissent. Ceci est inexact comme on peut le voir sur l'arbre de la figure 5.1a.

En réalité, la masse d'un halo caractérise une "espèce", de sorte qu'une représentation sous la forme d'un cladogramme est plus appropriée. Sur la figure 5.1b nous montrons un tel cladogramme qui décrit la diversification des halos de matière noire. Il n'y a pas d'échelle de temps même si le sens de la diversification va vers le bas. La masse (ou la taille) est le seul critère de diversification, elle augmente donc vers le bas. Chaque nœud de bifurcation correspond à un évènement de fusion qui engendre un halo de masse supérieure. Le processus de formation de halos de matière noire apparaît immédiatement sur ce cladogramme, laissant la possibilité aux plus petits halos de coexister avec les plus gros ou d'avoir disparu.

La distribution de la matière noire dans l'Univers est plus complexe que ces simples halos et semble créer un réseau filamentaire avec des cellules presque vides de galaxies. En prenant en compte d'autres paramètres que la masse ou la taille, notamment en incluant des données géométriques, la structure de la matière noire pourrait être décrite sur un cladogramme très probablement plus complexe que celui de la figure 5.1b.

### 5.7.2 Amas globulaires : des galaxies très particulières

Les amas globulaires sont des ensembles d'étoiles essentiellement dépourvus de gaz et de poussières. Ils entrent donc tout-à-fait dans la définition des galaxies (Sect. 5.1). Par contre, en première approximation, ces objets ne sont que très peu modifiés par les interactions extérieures, ne se mélangent pas, et leur évolution séculaire est due au vieillissement des étoiles, aucune formation stellaire n'étant possible. La diversité des amas globulaires est donc principalement expliquable par l'assemblage, c'est-à-dire par leur formation. Il n'y a donc pas de transformation de ces objets autrement que

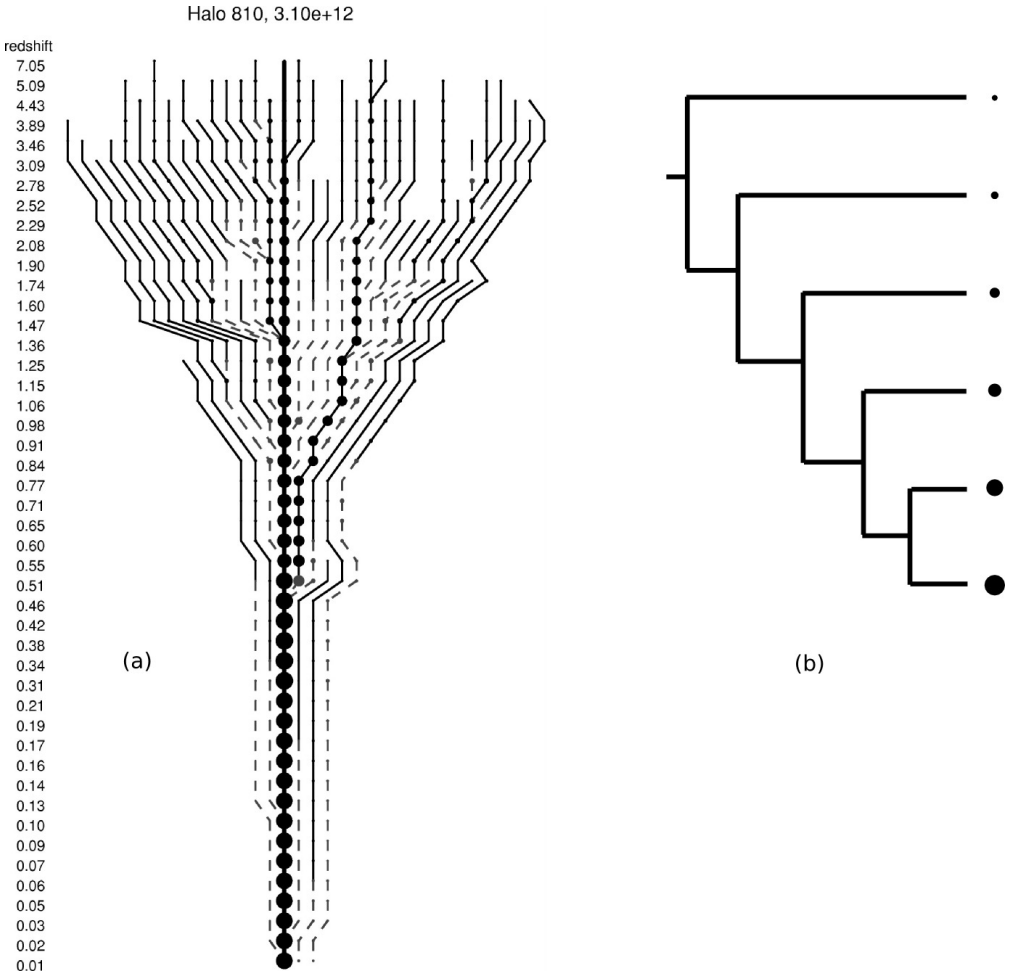



FIGURE 5.1 – (a) Un arbre de fusion de halos de matière noire tel que représenté habituellement. Le temps évolue vers le bas, le redshift étant indiqué à gauche. Le trait vertical en gras au milieu correspond au progeniteur "principal" du halo terminal à $z = 0$ (époque actuelle). La taille d'un disque est proportionnelle à la masse du halo (Stewart et al 2008, avec le permission de l'AAS) (b) Cladogramme correspondant, la masse étant le seul paramètre.

le vieillissement de leur population stellaire, ce qui permet de leur définir un âge de manière significative.

Le matériau avec lequel ils se forment, c'est-à-dire avec lequel leur constituant unique, les étoiles, se forme, est composé de gaz et de poussières en général enrichis en métaux puis éjectés par d'autres étoiles. Ce gaz et ces poussières sont donc les vecteurs de la transmission d'une histoire dont hérite l'amas globulaire. Il y a bien entendu modification puisque ce gaz et ces poussières se transforment en étoiles.

Nous voyons ainsi que le processus de transformation avec modification utilise pleinement notre définition large d'une galaxie et met en lumière l'intérêt de décomposer les processus de transformation des galaxies. Les liens évolutifs entre les amas globulaires apparaissent nettement via l'environnement. Cependant l'environnement ne doit pas être pris au sens géographique du terme. Il s'agit ici davantage du matériau utilisé pour l'assemblage ainsi que de ses propriétés chimiques et physiques qui caractérisent une certaine histoire dont hériteront les amas globulaires formés dans les mêmes conditions, et pas nécessairement au même endroit ni au même moment. Nous y reviendrons lors de l'interprétation d'une analyse cladistique des amas globulaires de notre Galaxie (Sect. 8.5.2).

# Chapitre 6

# Astrocladistique : méthodes

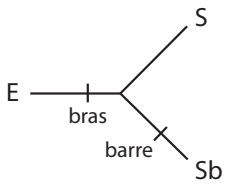
Les définitions et concepts ayant été posés au chapitre précédent, la mise en œuvre d'une analyse astrocladistique consiste à reprendre ce qui a été fait au Chapitre 4 avec des taxons qui s'appellent des galaxies. Concrètement, il s'agit d'établir les matrices qui permettront la construction des cladogrammes à l'aide de quelques programmes dont nous présenterons un rapide survol. Le présent chapitre décrit la mise en œuvre pratique d'une analyse astrocladistique, et le Chapitre 7 suivant dévoilera différentes stratégies pour aborder le cas des grands échantillons de galaxies.

## 6.1 Le diagramme de HUBBLE redémontré

Du temps de HUBBLE, les descripteurs étaient au nombre de deux : la forme plus ou moins applatie, qu'on peut traduire par la présence ou l'absence de bras spiraux (ou de manière équivalente d'un disque), et la présence ou l'absence d'une barre. En 1936, HUBBLE pensait que, selon la loi de Jeans, les galaxies elliptiques allaient s'aplatir en un disque. Nous avons là tous les ingrédients pour une analyse cladistique simple à partir de la matrice suivante :

| | bras | barre |
|---|---|---|
| E | 0 | 0 |
| S | 1 | 0 |
| Sb | 1 | 1 |

L'état codé "0" signifie absence et le "1" présence. De la même manière que pour les exercices du Chapitre 4, on trouve que l'arbre le plus parcimonieux non enraciné est celui-ci :




Les traits en travers des branches mentionnent un changement d'état, dans un sens ou dans l'autre, du caractère indiqué. Pour passer du taxon E aux taxons S ou Sb, il

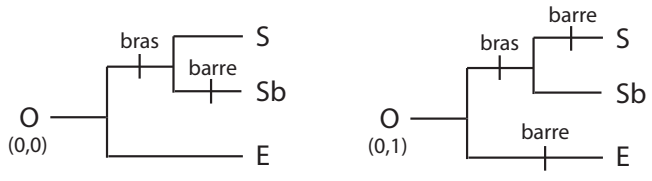
faut acquérir des bras spiraux, pour passer de S à Sb, il faut acquérir une barre, et de Sb pour E il faut d'abord perdre la barre puis perdre les bras.

Cet arbre ressemble étrangement au diagramme de HUBBLE à l'énorme différence près que ce dernier est le plus souvent supposé être enraciné, ou plus exactement le sens de l'évolution est censé aller de gauche à droite (hypothèse de HUBBLE) ou de droite à gauche (hypothèse moderne). Cette supposition est extrêmement forte et sans véritable justification. En effet, pourquoi l'évolution n'irait-elle pas verticalement ou selon un chemin plus complexe ? Quelle que soit l'évolution réelle de la morphologie, qu'est-ce qui prouve que les galaxies de type ancestral ont toute la même forme et qu'on connaît toutes les formes possibles ? De plus, dans l'hypothèse moderne, il y aurait ainsi deux types d'ancêtres, les spirales et les spirales barrées. Cela sous-entend nécessairement qu'ils n'ont pas d'ancêtres communs : est-il vraiment raisonnable de penser que les galaxies spirales et les galaxies spirales barrées seraient issues, même après une longue histoire évolutive, d'objets complètement différents ? Ce résultat semble assez difficile à concilier avec nos connaissances actuelles de l'évolution de l'Univers. Il faudrait qu'il soit basé sur une analyse plus rigoureuse non seulement de la physique complexe des galaxies, mais déjà de la manière de représenter la diversité présente dans la matrice très simple ci-dessus.

Pour enraciner l'arbre en cladistique, il est nécessaire de polariser les états de caractères, afin de contraindre le sens de l'évolution. Pour ce faire, on introduit un groupe de comparaison "O" sous la forme ici d'un ancêtre complètement fictif, objet à jamais inconnu mais qui laisse la porte grande ouverte à de nouvelles découvertes. Par exemple, selon l'hypothèse de HUBBLE, cet ancêtre devait être elliptique et s'aplatir ensuite par friction dynamique. Donc pour le caractère "bras" l'état ancestral est 0, l'état dérivé 1. Cette information de polarisation est inconnue pour le caractère "barre", il sera donc codé " ?" dans la matrice. Cela signifie qu'il faudra effectuer une recherche des arbres les plus parcimonieux pour chacune des deux valeurs possibles (0 et 1) pour ce caractère, et comparer les résultats afin de sélectionner les meilleurs scénarios. La nouvelle matrice se présente comme ceci :

| | bras | barre |
|---|---|---|
| O | 0 | ? |
| E | 0 | 0 |
| S | 1 | 0 |
| Sb | 1 | 1 |

Pour chacune des deux possibilités du caractère "barre" pour le groupe de comparaison O, nous obtenons un seul diagramme le plus parcimonieux :



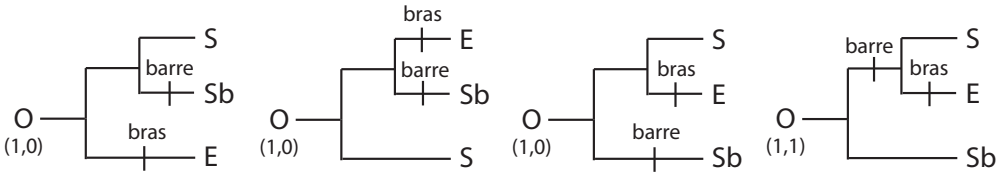

L'arbre de gauche est le plus parcimonieux de tous car totalisant deux pas seulement, stipulant que l'hypothèse la plus probable (car la plus simple) est que l'ancêtre n'avait pas de barre.

Il est utile de remarquer dès maintenant que sur les deux cladogrammes enracinés, le groupe de comparaison O peut être placé à gauche parce qu'il est fictif et sert seulement à indiquer la valeur des caractères au nœud situé à sa droite. Il est tout à fait possible de polariser la matrice avec un des membres de l'échantillon, par exemple E puisqu'il possède un caractère (bras) dans l'état ancestral. Dans ce cas E ne représente pas l'ancêtre, seulement un objet lui ressemblant le plus par rapport aux autres. Les cladogrammes sont alors les mêmes qu'à gauche, sans le taxon fictif "O". Dans les deux cas, les galaxies de type E sont clairement les plus proches, les moins diversifiées, de l'ancêtre commun. Il serait ainsi aisé d'insérer un nouveau type morphologique, qui apparaîtrait comme une nouvelle branche quelque part sur l'arbre. Ceci n'est pas possible dans le diagramme de HUBBLE.

De nos jours, l'hypothèse évolutive de HUBBLE n'est pas considérée comme correcte, car le temps d'aplatissement dynamique d'une galaxie elliptique est plus long que l'âge de l'Univers. On tend plutôt à penser que les galaxies elliptiques sont principalement formées à partir de fusion de galaxies spirales. Ce qui implique que l'état ancestral du caractère bras est 1, d'où la nouvelle matrice :

| | bras | barre |
|---|---|---|
| O | 1 | ? |
| E | 0 | 0 |
| S | 1 | 0 |
| Sb | 1 | 1 |

Quatre cladogrammes les plus parcimonieux sont possibles, ils totalisent tous deux pas :




Les trois arbres de gauche correspondent à l'absence de barre chez l'ancêtre et sont équivalents à celui du milieu (deuxième en partant de la gauche) dont la représentation est plus parlante car le taxon S n'a pas évolué par rapport à l'ancêtre commun. Il n'y a donc en réalité que deux arbres possibles, les deuxième et quatrième en partant de la gauche.

Dans l'analyse cladistique, on peut donc imposer facilement une hypothèse d'évolution de certains caractères soit en fonction des connaissances du moment, soit pour tester et comparer plusieurs hypothèses. Le choix du groupe de comparaison n'est pas du tout anodin car il oriente la diversification. Par exemple, dans l'hypothèse de HUBBLE, le critère de parcimonie n'autorisait qu'une seule solution possible et prédisait même que l'ancêtre devait ne pas posséder de barre. Dans l'hypothèse moderne, deux solutions sont possibles, impliquant que la morphologie seule ne permet pas d'établir une séquence évolutive.

L'enracinement de l'arbre a donc une influence importante sur l'interprétation en terme de définition des groupes évolutifs et de leur hiérarchie historique. Il peut quand

même être intéressant d'utiliser les cladogrammes non enracinés lorsqu'aucun indice solide d'états ancestraux n'est disponible, car l'information phylogénétique reste présente.

L'analyse cladistique effectuée à partir des seuls types morphologiques des galaxies montre à quel point il est limitatif de s'en tenir au seul diagramme de HUBBLE en diapason, et également à quel point cela est dangereux en terme d'interprétation évolutive de la diversité morphologique des galaxies. On vient même de montrer que les seules données morphologiques sont insuffisantes pour établir une séquence d'évolution morphologique. L'astrocladistique est plus proche de la physique des galaxies car elle s'impose de prendre en considération tous les descripteurs disponibles, bien au-delà de la morphologie, et sans choix a priori. L'objectif de l'astrocladistique est donc de reprendre le travail présenté dans cette section, en utilisant les données modernes sur les galaxies, donc en étendant la matrice à toutes les observables à notre disposition, et progressivement à toutes les galaxies connues.

## 6.2 Mise en œuvre pratique

L'analyse dans la section précédente considérait d'emblée la reconstruction de la phylogénie de classes de galaxies, puisque les taxons étaient représentatifs des types morphologiques déjà établis. Ces classes étaient basées sur deux caractères seulement et issues d'une classification a priori. Ceci est notablement insuffisant pour décrire la diversité des galaxies, mais aucune classification englobant l'ensemble des observables et prenant l'évolution en compte n'existe encore. Comment faire dans ce cas ?

Le plus simple est de partir d'objets individuels et de considérer qu'ils représentent chacun une classe encore inconnue. C'est l'organisation phylogénétique de ces objets qui permettra d'établir une classification basée sur la diversification et l'évolution des galaxies.

Il est de toute évidence exclu de prendre toutes les galaxies connues dès le départ, principalement parce que le nombre de descripteurs est nécessairement limité. Concrètement, l'approche consiste plutôt à choisir un échantillon restreint de galaxies afin de déterminer leur phylogénie. Ensuite, il sera possible de combiner plusieurs arbres ou définir un groupe évolutif dont un représentant (réel ou fictif) sera utilisé dans d'autres phylogénies (voir Chapitre 7). Ceci justifie encore une fois l'utilisation du terme "taxon" pour désigner les objets à analyser.

## 6.3 Choix des échantillons de galaxies

Imaginons vouloir comparer des organismes vivants sans rien connaître de la taxonomie. On a toutes les chances de chercher à découvrir des liens de parenté entre taxons d'origines très éloignées dans le processus de diversification, dont l'espèce ancestrale commune, si elle a jamais existé, est si différente que les caractéristiques qu'elle a transmise peuvent être perdues ou cachées dans la masse des mutations.

C'est à peu près la situation dans laquelle se trouve l'astrocladistique aujourd'hui. Deux solutions sont possibles et seront détaillées au Chapitre 7.

La première consiste à regrouper a priori les galaxies. Il est d'abord possible de s'inspirer des classifications actuelles, basées sur des critères observationnels res-

treints. Mais il est certainement préférable et plus objectif d'utiliser des méthodes de distance multivariée afin de rassembler les objets par ressemblance globale. Cependant, le résultat de l'analyse cladistique sera inévitablement influencé par ce tri en amont, et n'évitera pas la nécessité d'analyser toutes les galaxies pour bien couvrir la diversité de ces objets. Cette approche n'est donc valablement utilisable qu'à des fins de comparaison des résultats avec ceux de l'analyse cladistique.

La deuxième solution est d'aborder ce problème par des échantillons peu importants, susceptibles pour une raison subjective et pas nécessairement correcte, d'avoir partagé une histoire commune. Il s'agit un peu de pêche à la ligne, mais l'analyse cladistique est justement là pour vérifier l'hypothèse d'un ancêtre commun et pour déterminer les liens de parenté si le signal phylogénétique est suffisamment fort. Des premiers résultats dépendront les choix des échantillons suivants, et en particulier l'identification de groupes évolutifs facilitera grandement l'extension rapide du domaine de diversité étudié.

Concrètement, quelques milliers de taxons (galaxies ou groupes) constituent une taille d'échantillon raisonnable avec les ordinateurs d'aujourd'hui. Au-delà, il faut se tourner vers des approches du type superarbre qui seront décrites dans le chapitre suivant (Chapitre 7).

## 6.4 Sélection des caractères

Par essence même, la cladistique refuse de choisir a priori les descripteurs car ceci risque de se faire essentiellement sur des critères subjectifs. Par contre, parmi les descripteurs, certains ne peuvent certainement pas caractériser un état évolutif d'une galaxie. Ces caractères, non pertinents pour l'analyse cladistique, sont reconnaissables par le fait qu'il est impossible de leur attribuer un état ancestral et un état dérivé d'une manière non ambiguë, principalement parce qu'ils sont éphémères, et éventuellement récurrents, dans l'histoire d'une galaxie. Par exemple la présence ou l'absence de barre ne caractérise probablement pas un état évolutif, la présence d'un sursaut d'étoiles ne dit a priori pas grand chose sur l'état de diversification de la galaxie hôte.

Néanmoins, il est prudent de ne pas rejeter trop vite ces caractères et ceux pour lesquels le comportement au cours du temps est incompris. L'analyse cladistique s'occuper de faire le tri, fournissant par là même des arguments objectifs de rejet lors d'une analyse ultérieure. Utiliser tous les descripteurs disponibles permet de faire ressortir les grandes tendances évolutives, et minimise l'influence des conflits possibles entre des évolutions de caractères, et met également en évidence les caractères aberrants. L'analyse détaillée de l'arbre fournira les explications nécessaires de ces déviances (Chapitre 8).

Les redondances entre caractères peuvent fausser l'analyse, en attribuant un poids artificiel à certains descripteurs. Il ne s'agit pas de caractères corrélés, mais bien d'observables donnant exactement la même information physico-chimique d'une même composante élémentaire de la galaxie. C'est parfois le cas pour des raies d'émission, souvent multiples pour un même élément placé dans les mêmes conditions. Par exemple, les intensités lumineuses mesurées en large bande sont essentiellement redondantes, car une galaxie très lumineuse dans une bande optique le sera également dans toutes les bandes optiques, peut-être même dans l'infrarouge. Il est donc préfé-

rable de remplacer toutes les bandes larges par des couleurs, sauf probablement une par grand domaine (X, optique, infrarouge lointain, radio) pour caractériser une quantité totale comme la masse d'étoiles.

Les redondances sont gênantes car elles imposent un poids important au véritable caractère pertinent pour l'évolution. Dans l'analyse cladistique, il est tout-à-fait possible de pondérer les caractères, mais il faut beaucoup de prudence et des arguments solides pour le faire. Laisser des caractères redondants force la pondération d'une manière souvent cachée. Cependant, le résultat de l'analyse doit mener assez aisément à la mise en évidence de ce genre de redondance.

Si l'objectif de l'astrocladistique est d'appréhender la diversité des galaxies dans leur aspect évolutif, alors ce sont des millions d'objets qu'il faudra analyser, ces objets étant complexes et ayant une histoire complexe. Plus grand sera le nombre de descripteurs, plus fine sera la résolution des liens évolutifs et des différences, plus précise et contrainte sera l'histoire de la diversification ainsi reconstruite. Cependant, le nombre des observables restera toujours en nombre limité par rapport au nombre d'objets connus. En réalité, nous ne cherchons pas à établir un arbre généalogique des galaxies, ce qui est impossible, mais nous cherchons à synthétiser nos connaissances de la diversité. En conséquence, notre objectif est de ranger les galaxies dans des classes bien moins nombreuses. Les méthodes phylogénétiques telles que la cladistique permettent de regrouper les objets par affinités évolutives, et ainsi de simplifier de manière pertinente la représentation de la diversité.

Quelles sont donc les descripteurs pertinents ? Ce sont a priori toutes les observables mentionnées en Section 2.3. De préférence, il faut privilégier les caractères objectifs, c'est-à-dire qui ne sont pas issus d'un arbitrage humain, comme la morphologie. Cette dernière peut très avantageusement être remplacée par des mesures de dimensions, comme le rapport disque/bulbe, des rapports d'aplatissement, ou des données cinématiques (dispersion de vitesse notamment). Les grandeurs issues de modèles peuvent être intéressantes, comme par exemple la masse totale d'une galaxie, car elles représentent une donnée physique intrinsèque caractérisant l'objet. On peut se demander s'il n'est pas préférable d'utiliser directement les observables qui permettent de déduire cette grandeur, plutôt que de compter sur un modèle nécessairement imparfait incluant souvent une petite dose d'arbitraire ?

Un point crucial de l'analyse cladistique mérite d'être souligné. Le choix des objets de l'échantillon et des paramètres associés n'est jamais définitif dans le sens où la méthode est une méthode d'investigation, pas de classification. C'est-à-dire que l'analyse cherche à mettre en évidence une structure d'organisation des données et des paramètres, structure qui est trop complexe pour être visible directement. C'est ensuite le rôle du chercheur d'examiner cette structure et de décider son intérêt, à savoir si elle est cohérente avec ce que nous savons, si les homoplasies ne dominent pas trop, si elle éclaire de manière nouvelle notre compréhension de l'évolution à la fois des objets et des paramètres. Le résultat de l'analyse fournit lui-même les indications de la validité de cette analyse.

En conséquence, peu importe le choix de l'échantillon, peu importe le choix des paramètres, on peut toujours apprendre quelque chose d'une analyse phylogénétique. Néanmoins, pour une compréhension globale de la diversification, il est nécessaire que les analyses tendent à inclure l'ensemble des objets connus et le maximum de paramètres observables.

Nous avons vu précédemment qu'il est tout-à-fait possible d'utiliser des caractères insuffisamment renseignés, c'est-à-dire dont la valeurs ne sont pas connues pour tous les objets de l'échantillon. Comme il faut envisager toutes les possibilités pour ce genre de caractères, les temps de calculs deviennent bien plus longs, et les contraintes sur les résultats moins importantes, de sorte que le nombre de solutions possibles est plus grand. Le nombre acceptable de valeurs inconnues dépend beaucoup de la force du signal phylogénétique présent dans les autres caractères, mais il peut être jusqu'à 30% environ. Cela signifie qu'il n'est pas nécessaire de se limiter aux rares observables très bien connues car on risque de ne pas cartographier pleinement la diversité des galaxies. De plus, le résultat de l'analyse fournit des prédictions sur ces valeurs inconnues.

Il doit rester bien clair à l'esprit que la disponibilité des caractères évoluent dans le temps avec les progrès technologiques, rendant les analyses cladistiques toujours perfectibles et révisables.

## 6.5 Coder des valeurs continues

Définir des états évolutifs pour les caractères revient à coder les observables. Pour les descripteurs qualitatifs, ceci est en général facile, comme nous l'avons vu pour la présence/absence de bras et de barre. Mais la quasi totalité des observables astrophysiques sont des données quantitatives et surtout continues. Les processus à l'origine de l'évolution des composantes élémentaires des galaxies génèrent toujours un continuum de grandeur. Il n'est ainsi pas évident a priori de définir un état ancestral d'un état dérivé sachant qu'entre les deux la frontière est à la fois floue et quelque peu arbitraire.

La cladistique en variables continues a connu des développements importants récemment. Nous en présenterons un aperçu en Sect. 9.4. Dans ce livre, nous n'abordons que la cladistique en variables discrètes, les paramètres continus étant donc discrétisés ou encore codés.

Le codage s'effectue à l'aide de nombres entiers en découpant en tranches ("bins") la plage de valeurs pour un caractère donné. Trois questions se posent : quel doit être le nombre de tranches, doivent-elles toutes avoir la même taille, quelles doivent être les tailles optimales pour mieux représenter l'état évolutif des différents groupes de taxons en présence ? Les réponses à ces questions ne sont pas immédiates, elles sont soumises à un certain arbitraire dont l'influence sur le résultat de l'analyse cladistique devra être estimé au moins a posteriori.

Un exemple concret peut nous aider à appréhender le processus de codage. L'âge d'un enfant est presque toujours codé en nombre d'années. Pourtant, ce critère ne donne pas parfaitement l'état évolutif d'un enfant puisqu'entre deux enfants du même âge, il peut y avoir un bien plus grand écart en nombre de mois qu'entre deux enfants d'âges consécutifs. On peut résoudre ce problème en comptant le nombre de mois, mais la même question apparaîtra si on tient compte des jours. La solution, nécessairement imparfaite, dépend de l'utilisation qu'on souhaite faire de ce découpage, de l'échantillonnage de l'ensemble des enfants pour chaque âge, et enfin de la précision de la mesure dont on dispose (année seulement ou aussi mois de naissance ?). Pour des variables astrophysiques, la question qui se pose le plus souvent est de savoir si on garde la valeur elle-même ou son logarithme.

Mais le problème se complique parce que les enfants n'évoluent pas tous de la même manière selon le critère choisi. Par exemple, la taille est extrêmement variable pour un âge donné ou pour une maturité donnée. Comment, dans ces conditions, coder la taille d'un enfant ? Cela dépend beaucoup de l'utilisation faite de ce codage, mais il est impensable d'utiliser ce critère pour déterminer l'état évolutif de l'enfant globalement. De même que pour l'âge, il s'agit ici d'un descripteur, parmi de très nombreux autres, et il est vain d'espérer retracer l'histoire évolutive avec un seul d'entre eux. On entrevoit ici le fait que la manière de coder un caractère n'est pas sans influence sur son poids vis à vis des autres caractères. Il est donc a priori préférable de coder les caractères de manière cohérente, sauf pour des cas particuliers bien justifiés. Le choix de coder la valeur quantitative ou son logarithme n'est donc pas anodin.

Pour un état évolutif global donné, la disparité des tailles est grande, de sorte que les plages de tailles se recoupent d'un état à un autre voisin. Coder simplement par tranche induit donc un découpage artificiel qu'il est presque impossible d'éviter. Il existe des outils statistiques sophistiqués permettant de détecter des tranches à partir d'une valeur moyenne et un écart type pour chacune d'entre elles, mais ces méthodes ne peuvent pas être totalement infaillibles.

La solution simple du découpage de la valeur quantitative par tranches reste quand-même satisfaisante, si on garde à l'esprit que les caractères spécifiques aux groupes évolutifs débordent très probablement sur plusieurs des tranches prédéfinies. Il s'agira donc de choisir un compromis entre la taille des tranches, qui doit être idéalement supérieure à la précision de la mesure, et leur nombre, qui doit répartir au mieux l'échantillon dans ces tranches.

Le découpage ne doit pas être nécessairement régulier, mais il doit représenter le chemin évolutif du caractère. Prendre des tranches de tailles différentes, c'est donner de l'importance aux plus petites, car en terme évolutif dans l'analyse, il coûtera aussi cher de franchir une petite tranche qu'une grande. Les justifications doivent être solides pour autoriser un tel schéma.

L'expérience nous montre que le critère principal pour déterminer le nombre de tranches est la stabilité du résultat vis à vis de ce choix. En pratique, nous avons constaté qu'en deçà de 10 ou 15 tranches, le résultat varie significativement avec ce nombre. Bien sûr, cela peut dépendre des échantillons, mais il nous semble préférable de prendre systématiquement 30 tranches lorsque le logiciel le permet.

La question des incertitudes de mesure est liée à ce qu'on pourrait appeler la variance cosmique, sauf bien entendu lorsque que les incertitudes sont très grandes (50% par exemple). En effet, il ne faut pas oublier que nous cherchons à grouper des objets dont les processus de transformations sont intrinsèquement continus. Chaque groupe est donc caractérisé par une certaine valeur moyenne ou médiane associée à chacun des paramètres complétée d'une dispersion. Il n'est donc pas évident de distinguer cette dispersion des incertitudes de mesures. Cela revient à dire qu'une certaine incertitude règne dans les contours exacts des groupements. Il ne faut pas perdre de vue que nous faisons de la statistique. Si cela s'avère utile, il est tout de même possible d'être plus rigoureux en attribuant plusieurs valeurs à un même caractère pour un même objet pour introduire explicitement une incertitude de mesure comme pour préciser l'évolution possible du caractère.

## 6.6 Évolutions des caractères

Les états évolutifs des caractères étant identifiés par le codage, il reste à préciser comment chaque caractère évolue en contraignant au mieux la séquence de transformation des caractères, c'est-à-dire la suite des différents états codés possibles. Cette information supplémentaire aidera à diminuer le nombre de phylogénies possibles dans l'échantillon en rendant le résultat plus pertinent.

Nous avons vu en Sect. 4.3 les quatre contraintes générales, dont deux seulement sont a priori réellement utiles : l'hypothèse de WAGNER (caractères ordonnés) et l'hypothèse de FITCH (caractères désordonnés). Pour les galaxies, il apparaît que la première est la plus adaptée dans la majorité de cas. Nous avons affaire à des variables continues, régies par des processus physico-chimiques qui rendent moins probables les sauts importants des valeurs dans tous les sens. Néanmoins, ce genre d'hypothèse doit être appliquée à chaque caractère, donc l'hypothèse de Fitch pourrait certainement être utile dans certains cas, ne serait-ce que pour évaluer l'influence de ces hypothèses.

D'une manière plus générale, les contraintes sur les évolutions des caractères sont introduites sous la forme d'une matrice dite matrice de pas (stepmatrix), qui donne le coût du passage d'un caractère à un autre. Par exemple il est possible d'imposer l'irréversibilité en interdisant par un coût infini (valeur "i") le passage dans un sens donné :

```
USERTYPE StepBmV (Stepmatrix) =
 10
 0 1 2 3 4 5 6 7 8 9
 0 1 2 3 4 5 6 7 8 9
 i 0 1 2 3 4 5 6 7 8
 i i 0 1 2 3 4 5 6 7
 i i i 0 1 2 3 4 5 6
 i i i i 0 1 2 3 4 5
 i i i i i 0 1 2 3 4
 i i i i i i 0 1 2 3
 i i i i i i i 0 1 2
 i i i i i i i i 0 1
 i i i i i i i i i 0
 ;
```

Par l'intermédiaire de cette matrice de pas, il est possible d'introduire des schémas évolutifs plus complexes lorsqu'il peut y avoir des embranchements dans la séquence de transformation des caractères. Ces cas sont sans doute rares en astrophysique, mais l'analyse cladistique ne préjuge de rien, et laisse beaucoup de souplesse pour introduire des contraintes très précises.

## 6.7 Choix du groupe de comparaison

Les contraintes précédentes sur l'évolution des caractères ne donnent pas automatiquement la direction de l'évolution, c'est-à-dire l'état ancestral pour chacun des caractères. Ceci est rarement connu pour l'ensemble des descripteurs. Ce n'est pas véritablement obligatoire, car l'arbre non-enraciné est également très informatif.

Cependant, la détermination de l'histoire évolutive des galaxies passe nécessairement, pour chaque échantillon, par la détermination d'espèces ancestrales à partir desquelles les propriétés se sont propagées et modifiées. De plus, il est beaucoup plus aisé de comparer plusieurs arbres si leurs racines sont les mêmes.

La détermination d'un groupe de comparaison doit être effectuée pour chaque échantillon. Tant que l'astrocladistique n'aura pas produit de nombreuses phylogénies, il sera sans doute nécessaire de considérer un des membres de l'échantillon comme étant le plus proche de l'ancêtre. Mais il est impératif de considérer plusieurs choix et de justifier celui retenu par les conséquences sur l'interprétation de l'arbre enraciné, et plus particulièrement sur les évolutions de tous les caractères impliquées par celui-ci (Chapitre 8). Bien qu'idéalement ce choix ne peut pas être basé sur un seul des descripteurs, il est souvent pertinent de considérer la métallicité comme un bon indicateur d'ancestralité lorsqu'elle est faible grâce à la définition d'une galaxie (Sect. 5.1) que nous donnons en astrocladistique.

## 6.8 Programmes et calculs

### 6.8.1 Principes généraux

La matrice codée, complétée par l'ordonnancement des caractères et l'éventuel groupe de comparaison, peut maintenant être utilisée pour construire des arbres. L'emploi d'un ordinateur est indispensable pour rechercher tous les arbres possibles et sélectionner les plus parcimonieux. Cependant, ce problème est connu pour être NP-hard, c'est-à-dire non soluble en un temps polynomial. Il est donc impossible d'explorer tous les arbres au-delà de quelques dizaines de taxons. Plusieurs stratégies différentes existent, elles sont dites heuristiques. Elles sont souvent rapides, mais ne peuvent jamais garantir totalement detrouver l'arbre le plus parcimonieux. De nombreux algorithmes ont été développés et la recherche reste très active dans ce domaine. Nous n'aborderons pas du tout le fonctionnement de ces algorithmes dans ce livre. Nous ne décrivons ici que deux approches très simples à comprendre et à mettre en œuvre, assez rapides et efficaces.

La méthode la plus utilisée est une méthode assez simple qui explore l'espace formé par tous les arbres possibles en recherchant le minimum absolu sans se faire piéger dans les minima locaux. Pour se faire, un arbre est construit petit à petit, à partir d'un taxon initial, en ajoutant des taxons un par un. À chaque étape, des combinaisons aléatoires des branches permettent de trouver les arbres partiels les plus parcimonieux. Une fois tous les taxons intégrés, on recommence en changeant les taxons initiaux. Plus on itère le processus, plus les chances de trouver l'arbre ou les arbres les plus parcimonieux sont grandes, mais le temps de calcul peut devenir vite prohibitif. Il n'est donc jamais certain que le résultat est le meilleur dans l'absolu. Néanmoins, cette approche reste relativement rapide et efficace.

Il existe également une méthode un peu dérivée, dite "ratchet", qui consiste, à partir de chaque minimum local détecté, de déformer l'espace des phases afin d'échapper plus sûrement à ce minimum. La déformation s'effectue en attribuant des poids aléatoires au caractères. Une nouvelle recherche est lancée, et lorsqu'un nouveau minimum est trouvé, l'espace des phases est remis dans sa forme originale (la pondération aléatoire des caractères est supprimée) et un nouveau calcul vérifie si ce point est un minimum plus important que le premier. Cette approche est bien plus rapide et efficace que la méthode heuristique simple.

### 6.8.2 Programmes

Nous présentons ici quelques programmes développés pour la biologie évolutive et déjà utilisés dans le cadre de l'astrocladistique. Il ne s'agit certainement pas d'une liste exhaustive de ce qui existe, de ce qui est utilisable, et surtout de ce qu'il serait souhaitable de développer spécifiquement dans le cadre des galaxies, mais ce petit bestiaire permet de faire déjà beaucoup de choses. Nous ne donnons pas d'aide détaillée pour chacun d'eux, seulement quelques pistes pour démarrer et les fonctions qu'ils remplissent particulièrement bien pour l'astrocladistique. Notons que la biologie moléculaire, dont les caractères et leurs états sont spécifiques, amène de très nombreux développements de méthodes mathématiques et de logiciels qui semblent à première vue difficiles à utiliser directement dans le cas des galaxies. Nous n'en parlerons pas ici, mais ces pistes vaudraient certainement la peine d'être explorées.

#### PAUP

Le logiciel PAUP* (SWOFFORD, 2003) est très répandu et très utilisé, il est bien ciblé pour l'analyse cladistique de données morphométriques, donc il est idéalement adapté au cas des galaxies. C'est un très riche ensemble de programmes, particulièrement utile en astrocladistique pour la recherche des arbres les plus parcimonieux, par plusieurs méthodes, ainsi que pour le calcul des estimateurs de solidité (bootstraps et decay). C'est donc le logiciel central, dont la possibilité d'écrire des scripts de commande est particulièrement précieuse pour les calculs sur des fermes d'ordinateurs. C'est le seul logiciel payant de la liste donné ici.

#### Méthodes de ratchet : Pauprat, PRAP et MBPR

PAUPrat (http ://users.iab.uaf.edu/∼derek_sikes/software2.htm) est un programme gratuit générant un script exécutable par PAUP et contenant toutes les instructions permettant d'effectuer une recherche heuristique selon la méthode ratchet.

PRAP2 (http ://bioinfweb.info/Software/PRAP2) est une implémentation du ratchet plus récente et plus conviviale utilisant également PAUP.

MBPR (Multi-Batch Paup Ratchet, http ://mathbio.sas.upenn.edu/mbpr/) est une version très puissante et efficace puisqu'il effectue ses calculs en 10 batches en parallèle, augmentant de beaucoup le temps et l'efficacité de convergence vers une solution parcimonieuse.

#### TNT (Tree analysis using New Technology)

Cet ensemble de programmes (http ://www.zmuc.dk/public/phylogeny/tnt/) implémente des méthodes algorithmiques poussées, et permettent tout particulièrement d'effectuer des analyses avec des variables continues (GOLOBOFF ET AL., 2006). Il prend intérgralement en compte la particularité de ce type de données quantitatives, c'est-à-dire qui incluent à la fois une incertitude de mesure et une variance naturelle au sein d'un même groupe. Il est donc particulièrement adapté pour les galaxies, mais n'a pas été encore exploité en astrocladistique parce que trop récent.

**autodecay**

AutoDecay (http ://www.bergianska.se/index_forskning_soft.html) est un petit programme gratuit qui aide au calcul des indices de decay (ou de BREMER) en générant un script de commandes pour PAUP.

**Clann**

Ce logiciel Clann (CREEVEY AND MCINERNEY, 2005) est destiné à la construction des superarbres dont il sera question au chapitre suivant (Chapitre 7). À partir d'un ensemble d'arbres, il génère une matrice dont les caractères, purement artificiels, servent à décrire la structure des arbres et la place de chaque taxon sur ceux-ci. Cette matrice est ensuite utilisé dans un calcul de parcimonie avec PAUP.

**TreeView**

Ce petit programme gratuit (http ://taxonomy.zoology.gla.ac.uk/rod/treeview.html ou http ://darwin.zoology.gla.ac.uk/ rpage/treeviewx) développé par Rod Page est très pratique pour visualiser rapidement un arbre, changer sa racine, et imprimer un arbre calculé dans PAUP par exemple. Il possède également quelques fonctions d'éditions de l'arbre, comme pour modifier et supprimer des branches, ce qui peut être utile pour l'impression d'arbres complexes. La description des arbres est standard et se réalise dans un fichier ASCII à l'aide de parenthèses imbriquées.

**FigTree**

FigTree (http ://tree.bio.ed.ac.uk/software/figtree/) est un programme java de visualisation un peu plus puissant que TreeView.

**TreeGraph**

Treedyn (http ://treegraph.bioinfweb.info/) est un logiciel de visualisation encore un peu plus complet.

**Treedyn**

TreeGraph (http ://www.treedyn.org/) est un autre logiciel de visualisation, en java, plus complet que les précédents. Il permet de faire intéragir des données complémentaires sur des graphes en parallèle. Ceci est extrêmement utile pour l'analyse et l'interprétation de l'arbre. Il est également bien adapté à de très grands arbres (plus de 1000 taxons).

**Mesquite**

Mesquite (MADDISON AND MADDISON, 2004) est un logiciel gratuit sophistiqué tournant à l'aide de Java, donc indépendant de la plateforme. Il est modulaire et incorporant des programmes de calculs élaborés, pouvant faire appel directement à PAUP. Un de ses intérêts majeurs réside dans ces fonctionnalités très puissantes de visualisation et d'analyse des arbres en connexion directe avec la matrice. Ainsi, il est possible

de projeter les états des caractères sur l'arbre en code de couleur, ce qui permet immédiatement de comprendre comment se comportent les caractères le long de l'arbre. Il est également très facile de comparer plusieurs arbres car la sélection par taxon ou par branche se reporte immédiatement dans les fenêtres des fichiers associés, permettant par exemple de visualiser facilement si un groupe évolutif donné se retrouve sur d'autres arbres.

## 6.9 Arbre consensus

Il est fréquent que plusieurs arbres également parcimonieux soient trouvés. À ce niveau, aucun autre critère objectif ne permet de sélectionner l'un ou l'autre. Ces arbres sont-ils différents et où se situent les écarts ? La comparaison consiste principalement à effectuer un examen synthétique mettant en évidence les structures identiques, les structures compatibles et les structures incompatibles.

Un excellent moyen de faire cette synthèse est de calculer un arbre consensus. Il y a plusieurs sortes de consensus, dont deux principales. On peut calculer un consensus majoritaire qui garde intacts les embranchements trouvés dans une majorité des arbres, les autres nœuds devenant des polytomies (dits également non résolus car plus de deux branches sont issues d'un même nœud). La "majorité" est par défaut 50% mais peut être n'importe quel nombre décidé à l'avance. On peut également opter pour un consensus strict qui ne gardera que les embranchements retrouvés dans tous les arbres, les autres devenant non résolus. Cela revient à choisir 100% pour la "majorité".

Il n'y pas de règle pour choisir un consensus plutôt qu'un autre. Le consensus strict est certainement le plus conservateur, mais il n'est peut-être pas toujours très informatif selon l'utilisation. De nombreux débats sont toujours possibles sur ce sujet. S'il est presqu'entièrement résolu, il est préférable de garder le consensus strict. Dans les autres cas, il est bon de regarder les valeurs de chaque nœud sur le consensus majoritaire.

## 6.10 Estimateurs statistiques

Il existe deux types d'estimateurs de solidité, ceux pour les nœuds et ceux pour les caractères (Sect. 4.4). Ces indicateurs peuvent être facilement calculés à l'aide des programmes présentés plus haut, notamment PAUP, et doivent si possible être donnés avec l'arbre.

Les premiers (bootstraps et decay) sont les plus importants car ils sont conçus pour apporter un regard objectif sur la signification réelle de la structure de l'arbre vis à vis des données. En quelque sorte, il servent à évaluer si cet arbre est vraiment représentatif du signal phylogénétique, en quantifiant la variabilité de chaque nœud face à des données légèrement perturbées et dans les arbres à peine moins parcimonieux.

Les indicateurs de caractères sont utiles pour déterminer quantitativement le comportement global de chacun des caractères dans le schéma évolutif représenté par le cladogramme. Il servent également à comparer plusieurs arbres obtenus avec la même matrice.

Il existe d'autres indicateurs statistiques que nous ne présentons pas ici, l'estimation de la fiabilité d'un arbre via à vis de la phylogénie réelle restant difficile et un

peu controversée. D'une manière générale, la manière dont s'est déroulée l'analyse donne déjà une indication qualitative du résultat. D'autres analyses complémentaires peuvent également être effectuées pour tester la robustesse du résultat vis à vis des données et des hypothèses introduites. Enfin, le résultat ne prendra toute sa signification qu'à l'aune de l'interprétation des groupements de taxons et des comportements des caractères le long de l'arbre, sujet du Chapitre 8.

## 6.11 Exemple de calcul

Nous donnons ici un exemple concret de procédure d'analyse avec les logiciels MBPR et PAUP. De nombreuses options peuvent êtrre modifiées. Celles qui sont présentées permettent d'obtenir de bons résultats en astrophysique, en un temps de calcul raisonnable.

La matrice codée adopte généralement le format NEXUS qui est un format standard lu par la quasi-totalité des logiciels. Voici un exemple de matrice sous format NEXUS :

```
   #NEXUS
BEGIN TAXA;
        TITLE exemple;
        DIMENSIONS NTAX=5;
        TAXLABELS
A
B
C
D
E
        ;
END;
BEGIN CHARACTERS;
    TITLE  'Matrix in file "exemple.nex"';
    DIMENSIONS  NCHAR=3;
    FORMAT DATATYPE = STANDARD GAP = - MISSING = ?
                SYMBOLS = " 0 1 2 3 4 5 6 7 8 9 A B C D E F G H I J K L M N O P Q R S T " ;
    CHARSTATELABELS
1       'car1',
2       'car2',
3       'car3';
        MATRIX
A       5 G D
B       O 6 0
C       E 7 1
D       M 9 2
E       8 A T
;
END;
BEGIN PAUP;
 ctype ord: 1 - 3;
 set outroot = monophyl torder = right maxtree = 10000 tcompress = yes;
 weight 1 : 1-3 ;
 outgroup A ;
END;
```

Elle est composée de différents blocs encadrés par un `BEGIN bloc;` et un `END;`. Nous avons ici un premier bloc décrivant les taxons et leurs noms, suivi du bloc des caractères précisant le type de valeurs (continues, codées ou génétiques) ainsi que les symbols utilisés pour le codage. Dans le cas présent, les 30 bins sont codés de 0 à T. Dans ce bloc, les noms des caractères sont donnés dans l'ordre des colonnes

de la matrice, suivis la matrice elle-même dont l'ordre des taxons doit correspondre exactement à celui du bloc des taxons. Enfin, un bloc de commande PAUP permet de préciser le type de contrainte sur l'évolution des caractères, différent paramètres jouant sur la présentation de l'arbre ou le nombre maximale d'arbres à conserver à chaque itération. Ces paramètres sont permanents tant que la session de PAUP n'est pas fermée.

Les arbres sont représentés par le format standard NEWICK sous forme de parenthèses imbriquées : (A(B(C,D))) dans lequel le niveau le plus interne regroupe les éléments les plus proches sur l'arbre. Il est possible d'ajouter également une longueur de branche : (A :1.0(B :0.24(C :9.18,D :0.01))). Cette représentation peut s'intégrer à un fichier au format NEXUS dans un bloc `BEGIN tree;`.

La recherche de l'arbre le plus parcimonieux s'effectue d'abord avec la méthode du ratchet avec MBPR. Ce logiciel consiste en un script perl qui va appeler PAUP plusieurs fois pour effectuer de nombreuses itérations.

```
perl mbpr.pl --log=paup.log --batch=batch.nex --store=tree --weight=15 --iters=50 --save=10
                                              --echo=paup --limit=1000 --data=exemple.nex
```

Ce calcul va produire 10 fichiers contenant 50 arbres chacun qui sont les meilleurs arbres obtenus à chaque itération. Ces arbres sont ensuite utilisé comme point de départ pour une recherche plus approfondie autour des arbres les plus parcimonieux par la suite de commande PAUP suivante :

```
log start file=fichier.log;
set tcompress=yes torder=right outroot=monophyl root=outgroup maxtrees=1000 increase=no;
execute exemple.nex ;
gettrees mode=3 duptrees=eliminate file=tree0.tre;
gettrees mode=7 duptrees=eliminate file=tree1.tre;
gettrees mode=7 duptrees=eliminate file=tree2.tre;
gettrees mode=7 duptrees=eliminate file=tree3.tre;
gettrees mode=7 duptrees=eliminate file=tree4.tre;
gettrees mode=7 duptrees=eliminate file=tree5.tre;
gettrees mode=7 duptrees=eliminate file=tree6.tre;
gettrees mode=7 duptrees=eliminate file=tree7.tre;
gettrees mode=7 duptrees=eliminate file=tree8.tre;
gettrees mode=7 duptrees=eliminate file=tree9.tre;
pscores ;
filter best;
hsearch start=current swap=TBR steepest=no multrees=yes nreps=1000 nchuck=10
                                                                   chuckscore=PLAFOND ;
pscores ;
filter best;
condense collapse=no deldupes=yes ;
contree all /majrule=yes strict=yes grpfreq=no indices=no showtree=yes replace=yes
                                                append=no treefile=paupoptimcons.tre;
log stop;
```

La commande `hsearch` est celle qui effectue la recherche heuristique des arbres les plus parcimonieux. L'option `chuckscore` limite la recherche exhaustive aux arbres ayant un nombre de pas (score) plus petit que PLAFOND. Cette valeur est en principe choisie égale au meilleur score augmenté de 5 ou 10%. La commande `pscores` permet d'afficher le nombre de pas de chaque arbre en mémoire, ainsi que tous les indicateurs avec les options correspondantes :

```
pscores /single=all CI=yes HI=yes RI=yes RC=yes ;
```

Ces indications ne sont valables que pour un jeu de contraintes donné, notamment le type d'évolution des caractères (WAGNER ordonné ou autre), mais ne dépendent pas de l'enracinement choisi.

En générale, cette deuxième recherche plus approfondie amène à des arbres encore plus parcimonieux. Il faut réitéré jusqu'à convergence. Comme il s'agit d'une exploration heuristique et non exhaustive de l'espace des arbres, il peut être utile de recommencer toute la procédure plusieurs fois (MBPR+hsearch).

Enfin une estimation du bootstrap se fait sous PAUP avec la commande :

```
bootstrap nreps=1000 grpfreq=no search=heuristic/addseq=random nchuck=10 chuckscore=1 nreps=5;
```

Il faut avoir conscience qu'un bootstrap crée 1000 matrices en tirant les caractères au sort avec remise (toutes les matrices ont donc la même dimension). Il est donc peu pertinent d'effectuer un bootstrap avec quelques caractères seulement. De plus, la recherche de l'arbre le plus parcimonieux est nécessairement moins poussée du fait de temps de calcul prohibitifs. Le résultat du bootstrap sera donc un arbre légèrement différent du plus parcimonieux trouvé précédemment, mais l'indication de solidité des nœuds est tout de même intéressante.

# Chapitre 7

# Vers l'arbre des galaxies : les superarbres

Les procédures précédentes nous permettent d'effectuer une analyse cladistique sur quelques milliers d'objets au maximum. Or, nous sommes confrontés à la réalité de l'Univers avec le nombre immense de galaxies qui le peuplent. Nous ne les connaissons pas toutes, loin s'en faut, mais les catalogues sont déjà remplis de millions d'objets extragalactiques. Heureusement, nous ne sommes pas intéressés par leur généalogie, car le nombre de descripteurs reste et restera sans doute pour longtemps extrêmement faible face à ce gigantisme. L'important est de pouvoir cartographier la diversité des galaxies à travers leur évolution. L'objectif de l'astrocladistique est donc de rassembler les galaxies dans des groupes évolutifs dont les liens de parenté devront être déterminés. Oui, mais par où commencer ? Ce chapitre présente quelques stratégies destinées à englober un grand nombre d'objets connus au sein d'un scénario évolutif basé sur l'analyse cladistique.

## 7.1 Le problème des grands échantillons

La comparaison de plusieurs objets entre eux nécessite un nombre plus ou moins grand de descripteurs caractérisés par plusieurs états évolutifs. Plus les descripteurs seront nombreux, plus le nombre de groupes évolutifs différents qu'on pourra identifier sera grand, plus cette comparaison sera donc fiable et fine. À l'inverse, s'ils sont trop peu nombreux, elle deviendra grossière et peu informative. Il est difficile d'établir une règle générale car cela dépend considérablement du problème, c'est-à-dire de la nature des objets, des descripteurs correspondants et du signal phylogénétique associé.

L'analyse cladistique ne cherche pas à établir une généalogie d'individus mais plutôt une généalogie d'espèces, de sorte que les objets d'études sont en réalité des taxons, c'est-à-dire soit des groupes, soit des individus qui représentent généralement des groupes. La notion de taxon présente donc un avantage considérable puisqu'il est tout à fait possible de mélanger dans une même analyse cladistique des individus et des groupes. Si nous sommes capables de regrouper les objets, alors il est aisé d'étendre une analyse cladistique à un nombre formidable de galaxies. Seul le nombre final de groupes différents peut être limité par le nombre de descripteurs. L'analyse des grands échantillons sembleraient ainsi résolue, sauf que l'objectif de l'analyse cladistique est

justement de regrouper les taxons, sur des critères évolutifs. Par où commencer : regrouper ou analyser ?

Nous allons répondre à cette question au cours de ce chapitre, et nous détaillerons en particulier les stratégies utilisées pour aborder un grand catalogue de galaxies et pour insérer un échantillon sur un arbre déjà établi. Il ne faut cependant pas oublier que ces approches ne compensent pas le manque de descripteurs. De nouvelles observations, plus complètes, plus fines, seront toujours nécessaires. Elles viendront alimenter de nouvelles analyses astrocladistiques qui s'appuieront sur les résultats déjà obtenus, et pourront donc utiliser les méthodes présentées dans ce chapitre.

## 7.2 Regrouper les galaxies

Face à un trop grand nombre de galaxies, il est naturel de tenter de regrouper les objets. Oui mais sur quel critère ? On rejoint ici la problématique générale de la classification (voir Chapitre 2) et il serait dangereux de regrouper les objets a priori sur des critères peu compatibles avec les principes de la cladistique. Tant qu'une ébauche de classification évolutive n'existe pas, il est préférable d'éviter une telle approche.

Imaginons un important catalogue de galaxies, décrites au mieux par une centaine de descripteurs. Aucune analyse cladistique globale n'est envisageable, un regroupement des objets est nécessaire. Nous n'avons encore pas le recul des biologistes et avons donc peu d'idées sur l'organisation évolutive de la diversité des galaxies. Les seuls regroupements a priori actuellement utilisés reposent sur quelques critères observationnels. Faire un tel regroupement risque de fausser l'analyse astrocladistique car ces quelques caractères seront implicitement affectés d'un poids très important par rapport aux autres. L'objectivité de l'analyse n'est pas respectée. Il semblerait préférable d'utiliser une mesure de distance multivariée afin de dégager quelques groupes. Mais l'information évolutive des caractères est perdue et en aucun cas l'analyse cladistique qui s'en suivra ne devra être considérée comme complète. En réalité, l'analyse de distance multivariée, en particulier lorsqu'elle construit un arbre de distance (phénétique), peut servir de guide à la définition de sous-échantillons qui devront être analysés séparément et éventuellement recombinées selon les techniques décrites plus loin. Ce n'est donc qu'un processus itératif qui permettra d'établir la monophylie, c'est-à-dire la robustesse vis à vis de la diversification, des groupes ainsi préalablement définis.

Il n'y a donc guère que dans le cas où des groupes évolutifs ont déjà pu être établis par l'astrocladistique qu'un regroupement a priori peut se justifier. En conséquence, le traitement des grands échantillons ne peut se concevoir que progressivement, en analysant d'abord de petits échantillons, en définissant quelques probables groupes évolutifs, puis en essayant d'insérer de nouveaux objets à l'intérieur ou entre les groupes.

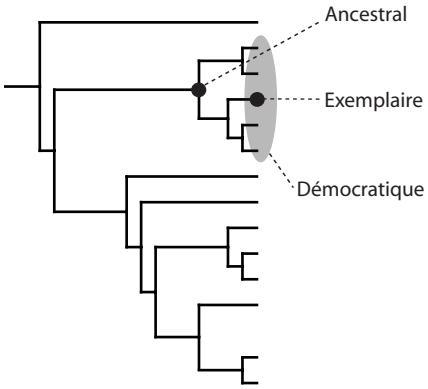
## 7.3 Taxons supraspécifiques

Le seul regroupement logique dans le cadre d'une analyse astrocladistique consiste naturellement à se baser sur un arbre évolutif. Les regroupements peuvent alors s'effectuer par branche en introduisant la notion de taxon supraspécifique, entité représentant les objets de la branche et les remplaçant dans les analyses ultérieures. Ainsi, de proche en proche, l'ensemble de l'échantillon initial pourra être placé sur un arbre, soit

sur une branche particulière, soit à l'intérieur d'un groupe évolutif représenté par un taxon supraspécifique.

Comment caractériser un groupe évolutif afin de le remplacer par un taxon pour une analyse cladistique plus large ? Le problème concret est ici de donner une valeur à chacun des caractères afin d'entrer l'élément dans la matrice. Or nous avons à faire à des variables continues, donc présentant une certaine disparité au sein d'un groupe donné (voir Sect. 6.5). De plus, le groupe est décrit par un certain nombre de propriétés bien spécifiques (les synapomorphies ou caractères dérivés d'un ancêtre commun) qui n'incluent pas tous les descripteurs.

On peut définir trois types de taxon supraspécifique à partir de l'arbre :




— *ancestral* : ancêtre commun hypothétique et fictif dont les caractéristiques sont déduites de l'analyse par parcimonie.
— *démocratique* : objet fictif dont chaque caractère prend l'état le plus fréquent dans le groupe. Il est également appelé "Common Equals Primitive".
— *exemplaire* : un des membres du groupe considéré comme étant représentatif du groupe, il constitue donc un spécimen typique. Éventuellement, un autre représentant peut être choisi afin de comparer les résultats.

Le taxon exemplaire est le plus couramment employé. C'est même la méthode utilisée implicitement lorsqu'un objet est utilisé tel quel dans une analyse : il représente un groupe évolutif encore non identifié.

Lorsque le groupe évolutif est bien caractérisé par des états dérivés uniques, il est en principe relativement simple de classer n'importe quel objet appartenant à ce groupe. La taille de l'échantillon à analyser peut ainsi grandir très rapidement. Mais n'oublions pas qu'une analyse cladistique n'est jamais terminée, et que de nouveaux descripteurs, de nouveaux spécimens peuvent venir perturber les travaux antérieurs.

## 7.4 Obtention de sous-arbres

À partir d'un échantillon important, si rien n'est connu, prendre un sous-échantillon au hasard a peu de chance de donner un résultat satisfaisant lors de l'analyse cladistique. Imaginez prendre quelque espèces vivantes au hasard en Amazonie, il est peu probable que les liens de parenté puissent être établis d'emblée, les objets ayant toutes

les chances d'être bien trop éloignés en terme évolutif. Nous avons vu plus haut que l'utilisation de critères a priori peut servir de guide tout en introduisant des biais sélectifs.

L'analyse de l'échantillon complet peut fournir une première idée de classification grossière. Elle peut valablement être complétée par une analyse itérative avec calculs des bootstraps afin de détecter et éliminer un à un les objets perturbateurs, c'est-à-dire des objets dont la présence rend l'arbre moins solide. Cette stratégie peut cependant être très longue.

Une approche bien plus efficace utilise la stratégie "diviser pour mieux conquérir". Elle consiste à aborder le problème grâce à des sous-échantillons, l'objectif étant de trouver un ou plusieurs sous-arbres robustes, qui permettra ensuite d'étendre par différents moyens l'analyse à tout l'échantillon. Un ensemble important de sous-échantillons établis au hasard est analysé. Un calcul rapide des bootstraps pour chacun des sous-arbres fournit une indication suffisante sur la robustesse tout en conservant des temps de calculs raisonnables. Il faut garder à l'esprit que le nombre de combinaisons possibles est très grand, et que la présente approche s'apparente à de la pêche à la ligne, mais la multiplication des tirages aléatoires augmente les chances de succès. Le nombre et la taille des sous-échantillons dépend de la force du signal phylogénétique. En toute logique, plus la taille du sous-arbre est grande, moins nombreux seront les sous-arbres robustes. L'optimum est fonction de l'utilisation ultérieure de ces sous-arbres. S'il s'agit de greffer d'autres galaxies (Sect. 7.5.1), une taille la plus importante possible est souhaitable, et il faut examiner plus de sous-échantillons pour augmenter la probabilité d'en trouver un satisfaisant. S'il s'agit de construire un superarbre (Sect. 7.6), c'est le nombre de sous-arbres solides qui va importer bien que la taille ne devra pas être trop petite. Il est même envisageable dans ce dernier cas de combiner des sous-arbres de tailles différentes. Il existe des méthodes mathématiques permettant d'optimiser le tirage aléatoire des sous-échantillons, mais nous n'aborderons pas cette question dans cet ouvrage.

La recherche aléatoire de sous-arbres s'avère être d'une efficacité redoutable. L'analyse de l'évolution des caractères sur ces sous-arbres montre les grandes lignes de la phylogénie globale à condition qu'ils soient compatibles. Si ce n'est pas le cas, alors sans doute l'échantillon complet est-il trop hétérogène vis à vis de l'évolution, trop complexe par rapport aux descripteurs, ou encore l'hypothèse d'un ancêtre commun n'est-elle pas respectée. Par contre si tous les sous-arbres sont compatibles, alors des groupes évolutifs "naturels" doivent déjà apparaître.

## 7.5 Utilisation d'un arbre contrainte

Lorsqu'un arbre, dont la solidité est pleinement satisfaisante, est disponible, il peut sembler judicieux d'étendre l'analyse afin d'inclure d'autres informations (taxons, caractères). Certes, il est toujours possible d'ajouter quelques éléments nouveaux à une matrice ayant déjà fourni un arbre solide, mais cela n'est véritablement utile que si le nombre total de taxons ne devient pas trop grand par rapport au nombre de caractères. Il est donc tentant de vouloir simplement greffer des taxons sur cet arbre. La résolution d'un arbre peut être également améliorée grâce à des données un peu complémentaires. Nous allons donc voir comment imposer, dans une analyse cladistique,

une contrainte sous la forme d'un arbre, c'est-à-dire comment injecter une phylogénie existante comme contrainte supplémentaire dans l'analyse de l'échantillon complet.

### 7.5.1 Greffe de taxons sur un squelette

La stratégie de la greffe consiste à utiliser la phylogénie existante comme une contrainte structurelle de l'arbre final. Cela limite beaucoup le nombre de solutions possibles en leur imposant d'être compatibles avec la contrainte. La compatibilité stipule que si les nouveaux objets sont enlevés, la phylogénie initiale est retrouvée intégralement. On appelle cette contrainte "arbre squelette" (backbone). Elle implique donc que cet arbre ne représente qu'une partie de l'échantillon complet. Cet arbre doit de préférence être totalement résolu (de chaque nœud sont issues deux branches et deux seulement) sans quoi la contrainte n'aurait que peu de sens aux nœuds polytomiques.

Cette greffe peut être utile pour inclure les objets manquants de l'échantillon complet, en particulier si les données sont moins bien renseignées. Bien sûr, la solidité du résultat (en terme d'indices de bootstraps et de decay) sera moindre que celle du sous-arbre, mais plus importante qu'une analyse simple de l'ensemble de l'échantillon. Il faudra donc rester prudent dans l'interprétation du résultat, plus particulièrement dans les détails structurels de l'arbre. Cependant, les groupes évolutifs devraient apparaître assez clairement, et l'analyse du comportement des caractères permettra de les définir.

La méthode de la greffe peut être utilisée également pour classer quelques objets nouveaux, quitte à reconduire ensuite une nouvelle analyse avec un sous-échantillon adapté aux groupes qu'on souhaite étudier plus en détail. C'est une manière d'agrandir l'échantillon traité, en concentrant l'analyse sur des parties seulement de l'arbre, comme un zoom. Elle est bien adaptée lorsqu'on a peu de taxons à ajouter. Mais en étant plus fine, elle est aussi plus longue que la méthode ci-dessus qui traite d'emblée l'analyse de l'échantillon complet contraint par l'arbre squelette. Néanmoins, ce dernier ne va pas permettre d'augmenter considérablement l'échantillon final. C'est une contrainte supplémentaire mais il ne remplace pas l'information manquante.

### 7.5.2 Optimisation d'un arbre non résolu

Il arrive que l'arbre obtenu à la suite d'une analyse présente de nombreuses polytomies, c'est-à-dire des nœuds à plus de deux branches. On dit qu'il n'est pas totalement résolu. Cela peut être le cas des arbres obtenus par consensus strict. Il est possible cependant de l'utiliser comme contrainte dans l'analyse par parcimonie afin d'améliorer sa résolution. On parle alors d'optimisation, car il peut soit servir de guide dans la recherche heuristique par parcimonie, soit imposer une compatibilité s'il est issu par exemple d'une analyse différente.

Ce type de contrainte implique que les taxons de l'arbre soient les mêmes que ceux de la matrice. Nous verrons plus loin l'intérêt de cette optimisation lors de la construction des superarbres.

## 7.6 Construction de superarbres

La méthode de l'arbre contrainte est limité à un seul arbre et trouve donc des applications restreintes. De plus, la greffe sur squelette suppose que les caractères soient presque identiques, et l'optimisation requiert les mêmes taxons. Il est cependant indispensable de pouvoir combiner plusieurs arbres assez différents en un seul, appelé superarbre.

### 7.6.1 Aboutement de plusieurs arbres

Les biologistes disposent de nombreux arbres représentant les phylogénies de nombreuses espèces d'organismes vivants, provenant d'analyses différentes, avec des jeux de descripteurs se recoupant parfois peu voire pas du tout. Pour obtenir une vision synthétique de la diversité biologique, l'idée de construire des superarbres a fait rapidement son apparition, et la nécessité d'abouter tous ces arbres a vu le jour. En effet, afin de construire l'Arbre de la Vie, il est nécessaire d'établir les liens de parenté entre tous les organismes vivants connus, à partir des toutes les études disponibles. L'analyse cladistique complète est illusoire car le volume d'information est gigantesque et les caractères diffèrent parfois radicalement d'un groupe à l'autre.

Jusqu'il y a peu, l'aboutement consistait en un raccordement manuel des arbres grâce aux taxons communs. Seulement, ce travail est parfois très complexe et reste en partie subjectif. Cet aboutement manuel n'est plus guère possible car le nombre d'arbres à combiner est très grand. Il n'y a que peu de raisons de croire que cette approche manuelle puisse être beaucoup plus utile en astrocladistique.

### 7.6.2 Méthodes numériques

Cette approche en superarbre est un sujet en plein développement en biologie évolutive, grâce à des méthodes numériques plus objectives et beaucoup plus puissantes. Puisqu'il n'est pas possible d'utiliser les même descripteurs pour tous les taxons, il faut donc utiliser la structure de l'arbre, c'est-à-dire sa topologie afin de décrire numériquement et objectivement les relations de parenté, et essayer de combiner ensuite toutes ces informations.

La structure de l'arbre est représentée par des caractères fictifs, prenant les valeurs 0 ou 1 pour chacun des taxons, selon leurs positionnements respectifs entre eux. Plusieurs méthodes existent et prennent les taxons par groupes de deux ou quatre ou encore considèrent des distances. La construction des superarbres est un sujet en pleine expansion et aucune méthode ne semble vraiment l'emporter pour le moment. L'une cependant est souvent adoptée car utilisant l'analyse de parcimonie habituelle.

Dans la méthode MRP (Matrix Representation using Parsimony, disponible dans le logiciel Clann décrit au Chapitre 6), chaque nœud de l'arbre est représenté par un caractère. La valeur 1 est attribuée si le taxon est présent dans les branches filles (en aval de ce nœud), 0 sinon. La valeur " ?" est attribué si le taxon ne se trouve pas sur l'arbre. La matrice est ensuite complétée pour chacun des arbres, il y a donc autant de caractères que de nœuds dans tous les arbres. Cette matrice est ensuite utilisée pour une analyse de parcimonie qui construira une phylogénie sous la forme d'un superarbre. Ce dernier est parcimonieux par rapport à tous les sous-arbres. Il s'apparente en réalité

à un consensus, de sorte que si tous les arbres sont compatibles, il sera totalement résolu. Cette méthode semble bien adaptée pour des petits arbres, ce qui est le cas de la stratégie “diviser pour mieux conquérir” (Sect. 7.5.2).

La reconstruction d’un superarbre nécessite que les sous-arbres soient enracinés, sans quoi le problème n’est pas suffisamment contraint. Dans l’approche “diviser pour mieux conquérir” (Sect. 7.4), dans laquelle on recherche des sous-arbres solides aléatoirement, il est indispensable que tous les sous-arbres soient enracinés avec le même taxon. Le choix de ce groupe de comparaison n’est pas toujours évident a priori, et plusieurs itérations sont probablement nécessaires avant de converger vers un enracinement satisfaisant.

Le superarbre obtenu par la méthode MRP est le consensus strict de tous les superarbres les plus parcimonieux. Cependant, cette parcimonie est définie par rapport à la matrice MRP, en d’autres termes par rapport à l’ensemble des sous-arbres : ce sont les superarbres les plus simples et compatibles avec tous les sous-arbres. Ils sont donc pas nécessairement optimaux (plus parcimonieux) par rapport à la matrice de données qui seule contient le véritable singla phylogénétique. Si le superarbre consensus strict n’est pas bien résolu, il est possible de l’optimiser en l’utilisant comme contrainte dans une analyse avec la matrice de données (Sect. 7.4). Ainsi on obtiendra des superarbres optimisés, c’est-à-dire compatibles avec tous les sous-arbres et parcimonieux vis à vis de la matrice de données.

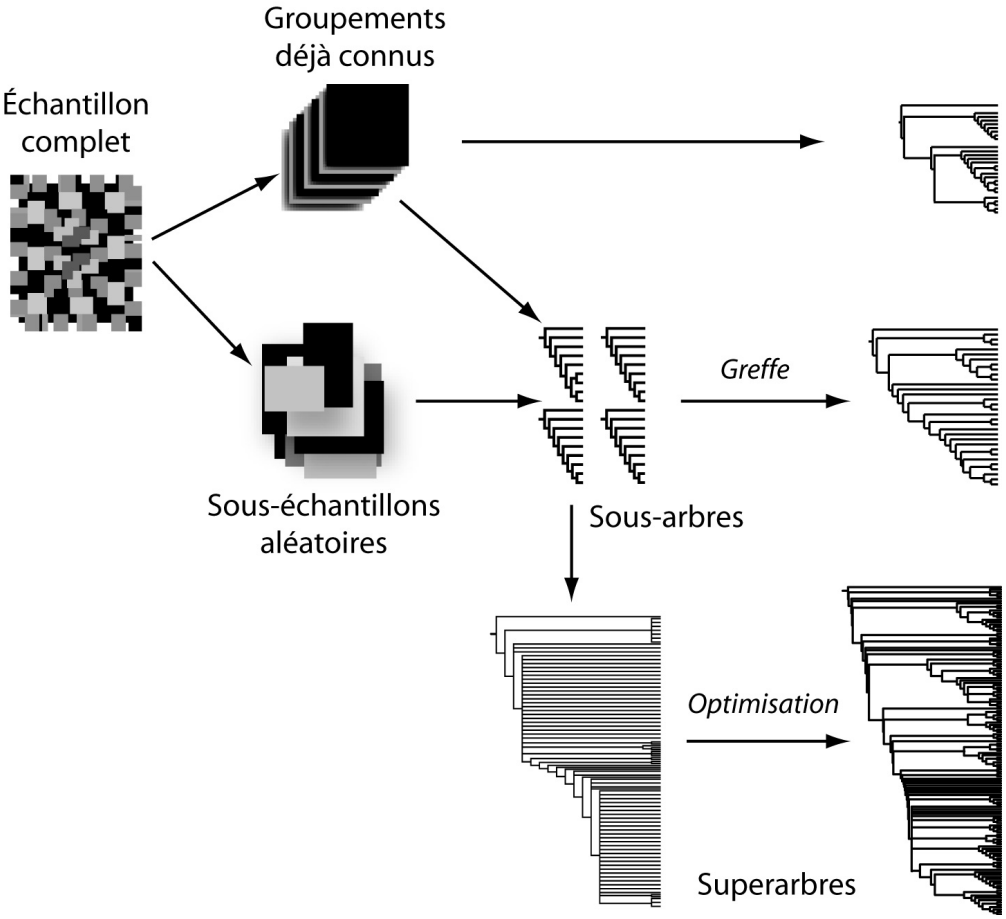
### 7.6.3 Signification et robustesse des superarbres

Autant des outils statistiques ont été inventés pour tester la solidité des arbres, c’est-à-dire pour estimer la force du signal phylogénétique inclus dans les données et représenté par l’arbre, autant le problème n’est pas résolu avec les superarbres. A priori, puisqu’il est compatible avec des arbres robustes, une certaine confiance peut être accordée au résultat. Cependant, l’interprétation du superarbre doit se faire en parallèle avec les arbres afin de détecter d’éventuelles apparitions de clades fictives, même si la méthode MRP ne semble pas trop sujet à de tels artefacts.

N’oublions finalement pas que toute phylogénie n’est qu’une hypothèse soutenue par les données et les contraintes incluses dans l’analyse, et que seule l’interprétation détaillée (Chapitre 8) peut apporter une véritable validation. Le superarbre étant obtenu grâce à une analyse cladistique objective et rigoureuse, les conclusions, les nouvelles questions et la confrontation avec de nouvelles idées et de nouvelles données seront toujours enrichissantes.

## 7.7 Résumé de la stratégie globale

Finalement, l’astrophysicien n’est pas dépourvu face à l’immensité de l’Univers et la multitude des galaxies cataloguées. Il est appelé à un travail méticuleux et progressif, avant de pouvoir proposer une vision d’ensemble de la diversification et de l’évolution des galaxies. Ci-dessous, nous résumons schématiquement la stratégie à adopter pour analyser de grands échantillons de galaxies, et en particulier pour intégrer petit à petit un plus grand nombre d’objets dans un scénario évolutif global.




D'un échantillon de nombreuses galaxies décrites par toutes les observables à notre disposition, soit nous pouvons effectuer un regroupement grâce à nos connaissances antérieures afin d'effectuer une analyse cladistique avec moins de taxons. Soit il nous faut pêcher des sous-échantillons aléatoires qui puissent fournir des sous-arbres robustes. Rapidement, l'aboutement de plusieurs arbres devient nécessaire, et peut se faire par une greffe, ou mieux par la construction d'un superarbre, qui peut être éventuellement optimisé.

# Chapitre 8

# Interprétation d'un arbre

Le cladogramme est une représentation synthétique de l'information phylogénétique présente dans les données, compte tenu des contraintes supplémentaires imposées à l'analyse. Il procède d'une optimisation des comportements évolutifs de tous les caractères. Il met en évidence une organisation hiérarchique de la diversité issue d'une évolution par embranchement. Ce cladogramme montre les liens de parenté, c'est-à-dire les distances respectives, en terme évolutif, entre les différents objets ou taxons. Sa structure est dictée par la cohérence des informations injectées dans l'analyse, cohérence qu'il faut maintenant analyser et comprendre. L'objectif de ce chapitre est d'expliciter les deux lectures possibles d'un arbre, soit en terme de groupes évolutifs, soit en terme d'évolution des caractères le long de la phylogénie, les deux étant finalement indissociables.

Un arbre n'est pas une classification bien qu'il puisse servir à la définition d'une classification, objet du Chapitre 10. La notion d'espèce en biologie est à la fois subjective et variée. Elle a été reconsidérée à plusieurs reprises, et notamment avec l'avènement de la cladistique. Mais peu importe finalement, car une classification dépend étroitement de la méthode employée pour le regroupement des objets, et n'a d'intérêt qu'en fonction de l'usage qu'il en est fait. La notion de clade ou de groupe évolutif est plus objective et étroitement liée à la méthode cladistique utilisée pour cartographier la diversité. Nous allons voir comment aboutir à de telles entités en astrocladistique à travers une interprétation astrophysique de la structure de l'arbre et de l'évolution des caractères qui en découle.

## 8.1 Lecture d'un arbre

Un cladogramme présente un regroupement de taxons au sens de la diversification. Il traduit un scénario évolutif lorsqu'il est enraciné. La lecture d'un arbre consiste à décrypter le sens des branches, des nœuds et des sous-branches, afin d'en dégager une histoire en terme de groupes évolutifs et de lignées qui organisent la diversité. Cette lecture est un peu différente selon qu'il s'agit d'un arbre enraciné ou non.

### 8.1.1 Arbre non enraciné

Des cladogrammes non enracinés très simples ont été montrés au Chapitre 4, et nous présentons un arbre beaucoup plus complexe sur la Fig. 8.1. Les structures et sous-structures traduisent des regroupements de taxons. Ces regroupements n'ont de sens qu'en rapport avec les autres membres de l'échantillon analysé et présents sur l'arbre. C'est donc en terme relatif qu'il faut lire un arbre : certains taxons sont plus proches entre eux qu'avec les autres. La proximité s'entend dans le sens de la diversification, c'est-à-dire d'une histoire évolutive qui est plus ou moins semblable. L'analyse cladistique fait l'hypothèse que tous les taxons de l'échantillon dérivent d'une espèce ancestrale commune. La structure résolue de l'arbre montre que cette hypothèse est correcte mais le fait qu'il soit non enraciné ne permet pas de la situer.

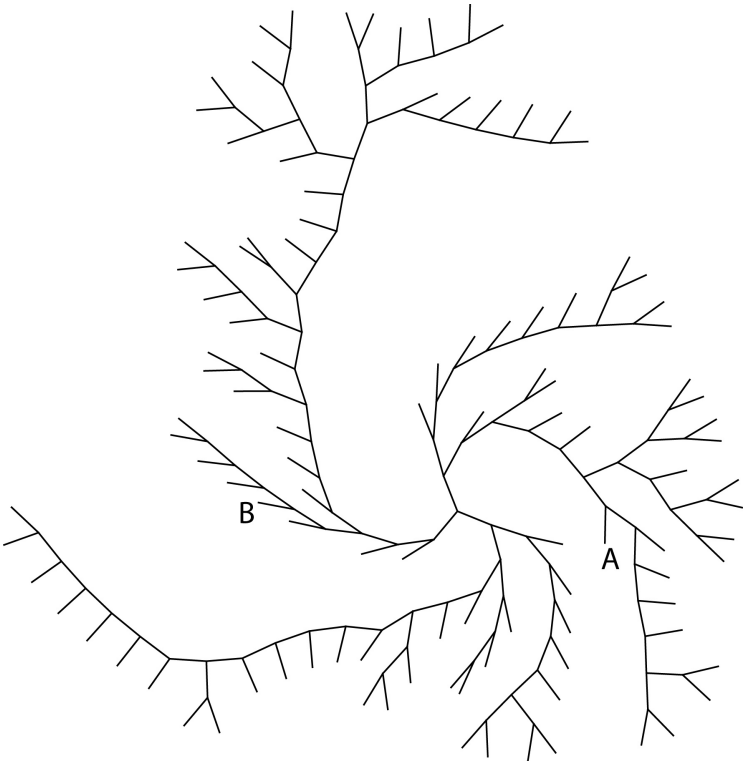



FIGURE 8.1 – Exemple d'arbre non enraciné avec des sous-branches bien visibles. Les taxons A et B identifiés servent de racines aux arbres des Fig. 8.2 et Fig. 8.3 respectivement.

La cladistique n'identifie pas le taxon-ancêtre commun à un groupe, mais le situe en chacun des nœuds de l'arbre. Les branches filles illustrent comment certains taxons se sont différenciés de cet ancêtre commun, parfois en donnant naissance à un ensemble de branches et de sous-branches appelé lignée. Dans un cladogramme non enraciné, des structures ressemblant à des lignées semblent bien visibles. Cependant, elles ne constituent pas a priori des clades ou des groupes évolutifs véritables car la racine de l'évolution peut se situer n'importe où, et pas nécessairement à une extrémité d'une grosse branche. Par exemple, les deux taxons A et B identifiés sur la figure 8.1 nous serviront à enraciner cet arbre de deux manières différentes (Sect. 8.1.2). Le

nœud central qui apparaît parfois (comme sur la figure 8.1) n'a aucun sens particulier. Chaque nœud indique une divergence et pas une convergence de deux branches, de sorte que l'éventuel point de démarrage de l'évolution qu'on pourrait utiliser pour "raconter" l'arbre doit être unique. En effet l'analyse cladistique ne prévoit pas la convergence de deux lignées en une seule. Ceci s'appelle de l'hybridation ou de l'évolution réticulée qui nécessite d'autres outils d'analyse qui ne sont pas abordés dans ce livre (voir Chapitre 3.1.4).

Les conclusions d'ordre phylogénétique dérivées d'un arbre non enraciné sont essentiellement limitées à des notions de distance en terme de diversification à travers l'évolution. Ceci peut s'appliquer à des sous-branches entières, mais il ne faut pas oublier que les feuilles elles-mêmes sont des taxons, c'est-à-dire soit des groupes, soit des individus représentant ces groupes (taxons supraspécifiques, voir Sect. 7.3). Par exemple, les deux taxons A et B identifiés sur la figure 8.1 n'appartiennent certainement pas au même groupe évolutif, mais ils ne semblent pas très éloignés en comparaison de l'ensemble des autres taxons. Ce qui peut se traduire en disant que, pour évoluer de l'un à l'autre, le nombre de bifurcations (nœuds) est plus faible que pour joindre deux extrémités de l'arbre. Nous ne pouvons conclure quant au nombre de changement d'états nécessaires entre A et B, car il nous faut pour cela examiner la projection des caractères sur l'arbre (voir Sect. 8.2), ce qui nous permettra en plus de caractériser certains regroupements.

### 8.1.2 Arbre enraciné

Le choix d'une racine, ou encore d'un groupe de comparaison (outgroup) permet d'orienter le sens de l'évolution. Ce choix n'est jamais simple et repose sur des observations, sur d'autres analyses, sur des modèles, ou même sur une interprétation a posteriori de l'analyse cladistique. En effet cette racine est très fréquemment choisie parmi les membres de l'échantillon étudié lorsque peu d'information externe est disponible. Dans ce cas, le cladogramme est totalement équivalent à l'arbre non enraciné, la racine servant principalement à en faciliter la lecture en terme évolutif. Puisqu'il s'agit d'un simple changement de représentation graphique, il est également très aisé de changer de racine sans réitérer l'analyse très gourmande en temps de calcul.

Nous présentons deux exemples d'enracinement de l'arbre de la figure 8.1 en utilisant les deux taxons A et B (Fig. 8.2 et Fig. 8.3). Cette fois-ci, l'ancêtre commun à tout l'échantillon est situé au nœud en haut à gauche, et le taxon racine (utilisé comme groupe de comparaison) est le plus proche de cet ancêtre, car le moins diversifié, le plus ressemblant en terme d'états évolutifs de caractères. Plus on descend la hiérarchie, plus les taxons sont diversifiés. Cependant, la longueur des branches ne représentent ici pas le nombre de changements d'états, de sorte que certaines branches située en milieu de l'arbre pourraient très bien représenter un nombre important de ces changements, donc une diversification plus grande qu'en bas de l'arbre. Là encore, l'analyse des états des caractères est indispensable pour conclure sur ce point (voir Sect. 8.2).

Le terme de diversification peut avantageusement se comprendre en rapport au nombre de nœuds, donc au nombre de bifurcations, nécessaires pour atteindre une branche donnée. Ces embranchements correspondent en effet à une véritable diversification puisque plusieurs groupes ou lignées apparaissent. Il est important de ne pas confondre la notion de diversification avec celle de différence mesurée par une dis-

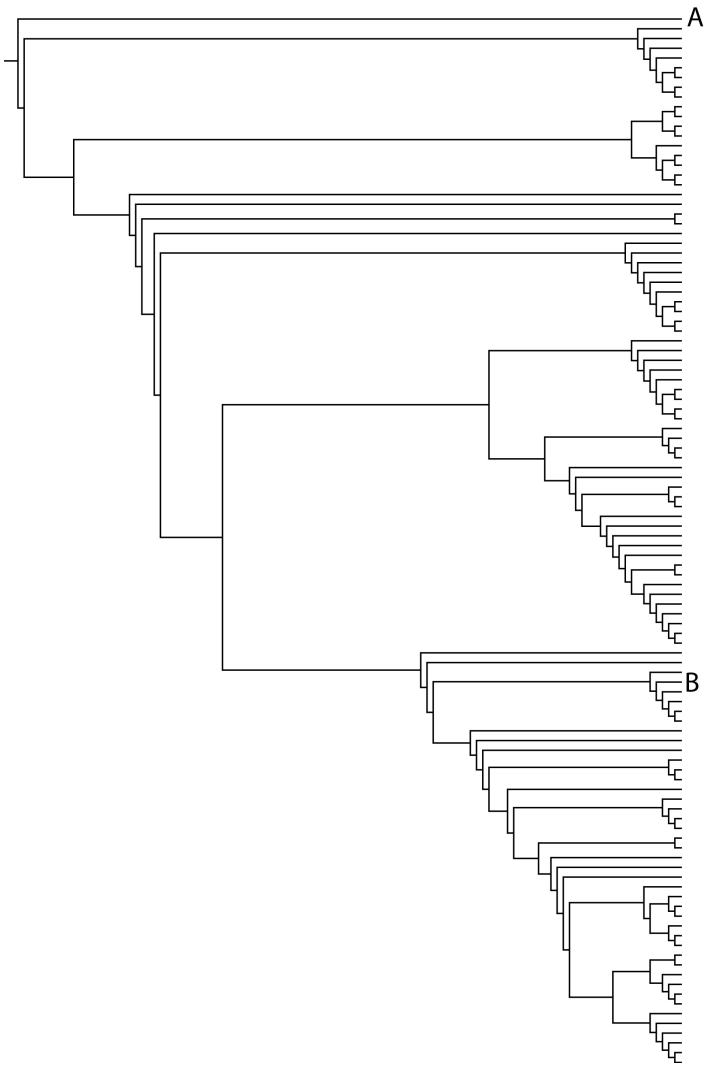



FIGURE 8.2 – Arbre de la Fig. 8.1 enraciné avec A.

tance multivariée, et encore moins avec celle d'évolution qui n'a encore aucun sens à ce stade. Ce point s'éclaircira dans la suite du chapitre, mais le langage doit être manié avec prudence afin d'éviter des interprétations ou surinterprétations trop hâtives.

Les places respectives des deux taxons A et B sur les Fig. 8.2 et Fig. 8.3 sont immédiatement frappantes : ils apparaissent très éloignés l'un de l'autre, et même aux deux extrêmes sur la Fig. 8.3, alors que nous avions conclu de l'arbre non enraciné qu'ils devaient être relativement proches dans la diversification par rapport à l'ensemble de l'échantillon. La prudence s'impose également car la représentation choisie sur ces arbres enracinés (les branches les plus courtes sont placées le plus en bas possible) laissent croire que le sens de l'évolution est du haut vers le bas. En réalité, la diversification irait plutôt de la gauche vers la droite. Il existe d'autres représentations ne rangeant pas verticalement les branches selon leur pseudo-longueur, mais la lecture en est un peu moins aisée.

L'enracinement différent ne change pas les liens de parenté, deux branches jointes

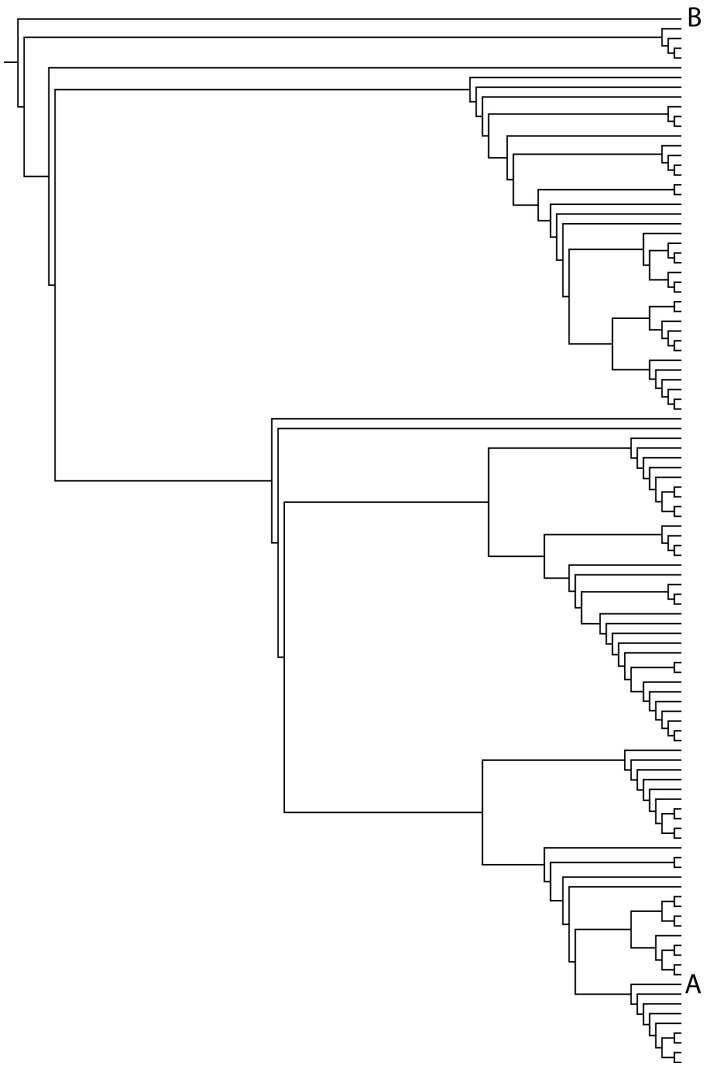



FIGURE 8.3 – Arbre des Fig. 8.1 et Fig. 8.2 enraciné avec B.

à un même nœud restent toujours jointes au même nœud, seul le sens de lecture peut changer. Ainsi la lecture d'un arbre en terme de distance évolutive relative ne dépend pas de l'enracinement. Par contre, l'interprétation globale, en terme de diversification, et la définition des clades et des groupes évolutifs, peuvent changer parfois assez radicalement. Le meilleur choix sera justifié par l'analyse des caractères présentée en Sect. 8.2 et par des comparaisons avec d'autres analyses s'ils elles existent.

Revenons maintenant à la notion d'ancêtre commun à un groupe donné. Cet ancêtre est supposé avoir bénéficié d'une innovation qu'il a ensuite transmise à ces descendants. Par exemple, les galaxies qui ont vu leur gaz balayé par le milieu intergalactique partagent toutes cette innovation qui nécessairement aura des conséquences sur les propriétés des galaxies qu'elles vont engendrer lors des différent processus d'évolution identifiés au Chapitre 5. Les ancêtres en cladistique sont des taxons-ancêtres, fictifs, et situés en chaque nœud. Les embranchements visualisent ainsi bien la transmission, à toutes les branches filles, et les modifications associées, illustrées par le

nombre de ces branches filles.

Sur un cladogramme, quelles pourraient être les galaxies les plus jeunes, les plus vieilles ? Aucune réellement, car ces expressions s'emploient pour des individus, pas pour des taxons qui représentent en réalité un groupe. Même les taxons proches de la racine de l'arbre ne sont pas plus vieux bien que moins diversifiés par rapport à l'ancêtre commun. On pourrait les croire plus jeunes car, formés dans un état ancestral, ils n'auraient pas encore eu le temps d'évoluer. Mais ce raisonnement s'applique à un individu qui lui-même appartient à un groupe évolutif. L'âge d'une galaxie n'a donc pas vraiment de sens (voir Chapitre 5). Il est plus juste de parler de son stade évolutif, c'est-à-dire de l'état de diversification dans lequel on l'observe, et du temps écoulé depuis qu'elle a été formée dans cet état. On peut en effet observer une galaxie formée récemment et appartenant à un groupe évolutif apparu il y a très longtemps, et une autre formée bien avant la première mais appartenant à un groupe apparu plus récemment que le précédent.

Pour illustrer ce point important, il est intéressant d'utiliser un exemple concret emprunté à la biologie, qui a constitué un symbole important dans les premières phases de développement de l'astrocladistique. Le cœlacanthe est un poisson bien connu des pêcheurs africains et des paléontologues. Ces derniers pensaient qu'il avait disparu il y a soixante millions d'années environ, mais il continue de vivre dans les profondeurs des océans. L'espèce moderne ressemble beaucoup à l'espèce ancestrale, dans la limite des données paléontologiques, les fossiles étant incomplets car limités au squelette. De par ses caractéristiques, elle est probablement l'une des espèces intermédiaires lors du passage des animaux marins vers les animaux terrestres. L'espèce humaine est apparue beaucoup plus récemment, pourtant des individus de chaque espèce se côtoient, malgré des espérances de vie de quelques dizaines d'années dans les deux cas. Il existe des individus cœlacanthe plus jeunes que des individus humains. Les deux espèces, vraisemblablement issues d'une espèce ancestrale commune (marine), ont toutes deux accompagné l'évolution des organismes vivants, celle du cœlacanthe s'étant simplement moins diversifiée, moins différenciée, bien qu'étant apparue beaucoup plus tôt dans l'histoire de la vie sur Terre.

Cette distinction entre évolution et diversification, entre période d'apparition de l'espèce et niveau de diversification, si bien expliquée par l'histoire du cœlacanthe, a déjà été illustrée simplement en astrophysique par la figure 5.1. Tout l'intérêt des fondements de l'astrocladistique, à travers le découpage, pouvant apparaître comme artificiel, de l'histoire des galaxies en processus de transformation associé à l'identification d'un mécanisme de transmission avec modification, réside dans cette indispensable distinction. L'objectif de l'astrophysique extragalactique n'est pas de comprendre la formation et l'évolution des galaxies, mais leur diversification. Nous ne devons pas expliquer seulement l'évolution d'un type d'objets, nécessairement individuels, mais nous devons appréhender l'évolution d'une population d'objets. Cette apparente subtilité est essentielle à comprendre pour être capable de lire les schémas évolutifs des galaxies que l'astrocladistique présente sous forme d'un arbre.

Un cladogramme enraciné permet donc de visualiser les périodes relatives d'apparition des différents groupes évolutifs et bien sûr de situer les divergences. À ce stade de l'analyse, aucune échelle de temps n'a encore été stipulée. Ces évènements de la diversification ne peuvent donc pas être datés dans l'absolu. Seul un examen minutieux de l'évolution des caractères le long de l'arbre pourrait fournir cette échelle de temps,

à condition cependant de disposer d'une sorte d'horloge évolutive.

## 8.2 Projection des caractères

Pourquoi le cladogramme se présente sous cette forme ? Pourquoi les taxons sont ainsi regroupés ? L'analyse cladistique compare les objets selon leurs états évolutifs partagés. Le principe de parcimonie sélectionne le scénario évolutif le plus simple en minimisant le nombre total de changements d'états. L'évolution de chacun des caractères le long de l'arbre est ainsi la plus régulière possible, selon le critère retenu (voir Sect. 4.3) et compte tenu de l'ensemble des autres caractères. Le résultat présente donc une cohérence maximale que nous allons maintenant exploiter. Toute explication de la structure de l'arbre basée sur la sélection de quelques caractères seulement est nécessairement vouée à l'échec. Il faut donc être prudent afin de ne pas retomber dans le piège de la classification traditionnelle, ce qui ferait perdre la richesse de l'approche cladistique.

Certains caractères sont certainement plus pertinents que d'autres pour décrire l'état évolutif d'une galaxie. Si l'histoire de la diversification des galaxies a un sens et qu'elle peut nous être accessible, il est raisonnable de penser que plusieurs observables vont être pertinentes et vont former un signal phylogénétique fort. La robustesse de l'arbre, estimée par exemple avec des calculs de bootstraps (Chapitre 4), mesure la force de ce signal. L'analyse des comportements des caractères le long de l'arbre nous en indique l'origine.

C'est ainsi que les cladogrammes doivent être examinés caractère par caractère, un peu comme il a été fait à la Sect. 4.2. Le plus commode est d'utiliser des projections à l'aide de codes de couleurs (voir par exemple FRAIX-BURNET ET AL., 2006a). Un exemple en niveau de gris est présenté sur la figure 8.4 pour trois caractères sur un même arbre. Il montre comment une telle visualisation permet une mise en évidence rapide des tendances évolutives de tous les caractères, ainsi que les propriétés de certaines sous-structures de l'arbre. Il faut noter que les couleurs des branches terminales représentent les valeurs des taxons, les couleurs des autres branches étant dérivées en chaque nœud d'après le critère d'optimisation choisi (ici la parcimonie). Deux types d'information peuvent être utilisées : les comportements globaux des états de caractères sur l'arbre, sur les branches et dans les sous-branches, permettant de caractériser les groupes, et les évolutions détaillées de chaque caractère que les modèles devront s'efforcer de reproduire.

Les comportements globaux des états de caractères identifient des groupes présentant des propriétés évolutives uniques héritées d'un ancêtre commun. C'est généralement un ensemble de propriétés qui fera l'unicité d'un groupe évolutif et lui donnera éventuellement son statut dans la hiérarchie évolutive. La rigueur de la taxonomie est ici importante et fait l'objet du Chapitre 10.

L'interprétation de l'arbre doit tenir compte du fait que les variables utilisées sont continues, impliquant des variations au sein d'un même groupe évolutif. Or les plages de variations peuvent se recouper d'un groupe à l'autre. Le codage couleur et les capacités de l'œil humain permettent une détection efficace de valeurs moyennes et leurs déviations, à condition de ne pas se laisser piéger par les contrastes artificiels du codage couleur. Des outils statistiques peuvent certainement être ici employés avec inté-

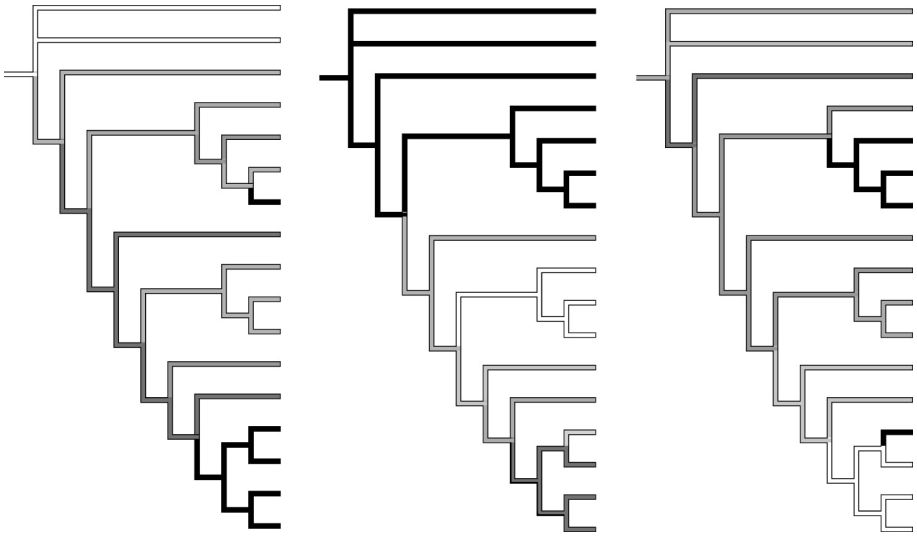


FIGURE 8.4 – Exemple de projection des valeurs de trois caractères sur un arbre, en niveau de gris.

rêt (Sect.8.5).

L'analyse cladistique impose une cohérence entre les évolutions des caractères. Cela signifie qu'il n'est pas possible de changer le comportement d'un caractère sans implications pour les autres. L'examen caractère par caractère permet d'estimer plus en profondeur la pertinence du scénario évolutif schématisé sur l'arbre. Chaque caractère a donc un comportement au cours de la diversification qui peut être de plusieurs types, définissant notamment les notions de synapomorphies, d'homoplasies et d'autapomorphies (voir définitions en Sect. 4.1.2). Le comportement peut déjà être évalué grâce aux indices CI, HI, RI et RCI (Sect. 4.4.3) : un caractère pertinent pour l'évolution et bien compatible avec la phylogénie aura des CI, RI et RCI très élevés. Il se détectera dans sa projection sur l'arbre par un comportement régulier et pouvant caractériser certains groupes évolutifs (visualisés par des sous-structures de l'arbre, voir fig. 8.5 et 8.6). Dans ce dernier cas, il peut même constituer une synapomorphie, base fondamentale de la cladistique, lorsque certains de ses états sont uniques à un groupe.

Les homoplasies apparaissent sur les projections comme des caractères présentant des comportements réguliers, mais rétrogrades ou similaires dans plusieurs branches (évolutions parallèles). Ces caractères sont probablement pertinents pour décrire l'état évolutif des objets, mais, pour une raison à déterminer, ils sont mal adaptés à la phylogénie proposée : ils rendent l'arbre moins robuste. Il est essentiel d'en rechercher les causes d'abord dans l'analyse elle-même (matrice, codage, choix de l'outgroup, contraintes évolutives sur les caractères, incompatibilités éventuelles, ...), puis à travers une interprétation astrophysique, avant de tirer des conclusions définitives quant à l'existence réelle des homoplasies. Les caractères incriminés ne doivent pas être retirés de l'analyse, car ils indiquent un certain comportement évolutif. Par exemple, le caractère "aile" est présent chez les insectes et chez les oiseaux. Il s'agit d'une évolution parallèle qui ne perturbe pas l'analyse de la diversité globale des êtres vivants, parce qu'ils sont décrits avec tous les caractères disponibles. Mais le caractère aile a son importance pour l'étude de la diversification des insectes. De même, la masse des

galaxies peut s'obtenir par différentes successions d'évènements (accrétion, fusion, effondrement monolithique) mais il reste à l'évidence un indicateur de diversification à partir du moment où il n'est pas le seul descripteur.

Enfin, certains caractères peuvent avoir des comportements beaucoup plus aléatoires le long de l'arbre, ou sans intérêt (autapomorphies par exemple). Même si un examen attentif peut confirmer leur peu d'utilité pour décrire l'évolution des galaxies, il n'est pas toujours nécessaire de les éliminer lors d'analyses ultérieures. En effet, comment être certain qu'ils n'auront pas un rôle essentiel avec un échantillon différent et des données nouvelles, en caractérisant peut-être un groupe évolutif encore inconnu ?

La solidité du cladogramme, et donc sa signification réelle, dépend des proportions relatives des différents types de comportements des caractères, la cladistique se basant uniquement sur les synapomorphies. Les autres comportements sont cependant fréquents dans la réalité et peuvent apporter de précieux éclairages sur la physique de la diversification des galaxies. D'une manière générale, il ne faut jamais écarter des caractères sous prétexte qu'ils n'ont pas le comportement idéal pour la méthode utilisée. Si les homoplasies s'avèrent être trop nombreuses pour les galaxies, alors la cladistique ne sera sans doute pas la meilleure approche.

Quoi qu'il en soit, la cladistique est une méthode d'investigation, de sorte qu'il est recommandable de comparer plusieurs résultats obtenus avec des jeux de caractères différents.

Il est bien évident que le choix de l'enracinement a des conséquences importantes pour tous les caractères. C'est donc cet examen de la projection des états des caractères qui permet de décider si le scénario évolutif est satisfaisant du point de vue de la cohérence de l'analyse astrocladistique elle-même, puis de l'interprétation astrophysique qu'on peut en déduire.

## 8.3 Interprétation astrophysique d'un cladogramme

L'ensemble des évolutions des caractères et de leurs propriétés dans les différentes parties de l'arbre nous explique la structure de l'arbre et les raisons des regroupements. Mais cela n'éclaire pas directement sur les différents processus qui ont amené les galaxies a acquérir ces propriétés. Pour cela, il faut utiliser une interprétation astrophysique qui, à partir de modèles plus ou moins élaborés et de connaissances déjà établies par ailleurs, permet d'établir la validité du scénario évolutif proposé et d'en déduire les conséquences. Dans cette section, nous séparons la vérification du résultat, l'examen des corrélations entre caractères, et la déduction des conséquences pour la diversité et l'évolution des galaxies. Il est clair que ces trois parties sont intimement mêlées puisque tout travail de recherche doit si possible fournir des résultats fiables, en conformité avec un certain nombre de connaissances déjà établies, et nouveaux, voire déroutants, quitte à remettre des connaissances en cause.

### 8.3.1 Vérification a posteriori des hypothèses

L'analyse des projections de caractères peut être placée dans le contexte astrophysique afin de vérifier que l'analyse ne fournit pas de résultats à l'évidence aber-

rants, à ne pas confondre avec les résultats nouveaux et éventuellement surprenants. En particulier, le choix de la racine de l'arbre peut être évalué assez rapidement grâce à l'évolution globale de certains caractères particulièrement bien connus. D'autres choix peuvent apparaître clairement, soit parce qu'ils semblent également plausibles, soit qu'ils paraissent plus pertinents. Si la racine fait partie de l'échantillon étudié, il est très facile de changer la disposition de l'arbre. Si la racine est un groupe externe de comparaison, une nouvelle analyse cladistique s'impose.

La comparaison des comportements de plusieurs caractères peut montrer une incompatibilité contredisant les connaissances et modèles actuels. Il peut en être de même avec certains caractères à l'évolution apparemment chaotique. La cause peut être due à des caractères non pertinents pour décrire l'état évolutif des galaxies, à un mauvais choix de la racine, ou encore à la présence de deux chemins évolutifs différents dans l'échantillon, impliquant que les taxons ne dérivent pas d'un seul taxon-ancêtre commun. Ce dernier point est bien illustré dans FRAIX-BURNET ET AL. (2006c). Des analyses complémentaires effectuées en changeant certaines contraintes ou en éliminant certains caractères ou objets, permettent de se faire une idée plus précise de l'origine des contradictions apparentes. Il est bien clair qu'une seule analyse ne suffit généralement pas à obtenir une confiance absolue dans le cladogramme.

La simple façon dont les galaxies sont regroupées sur l'arbre peut également amener certaines questions en rapport avec d'autres analyses cladistiques. Dans tous les cas, une nouvelle analyse peut découler de cette vérification, afin de modifier plus ou moins les contraintes imposées au départ, et d'ainsi jauger la solidité de la phylogénie reconstruite. Cependant, l'astrocladistique étant principalement basée sur des descripteurs, donc des observables, l'arbitraire dû à des paramétrages et des modélisations y occupe une très faible place. Les analyses complémentaires convergeront donc dans la plupart des cas vers une solution satisfaisante sur laquelle une interprétation astrophysique poussée pourra être développée.

### 8.3.2 Cohérences et corrélations au niveau des caractères

Les astronomes sont très habitués aux corrélations entre observables, particulièrement dans le cas des galaxies. Nous avons vu que c'est la seule méthode actuellement utilisable pour intégrer la complexité observationnelle de ces objets. De plus, des corrélations assez fortes ont été établies avec le type morphologique de HUBBLE, ce qui a naturellement conforté les chercheurs à la fois dans cette classification et dans la méthode des corrélations. Cependant, cette approche ne fournit pas une vision globale des comportements évolutifs des caractères.

L'astrocladistique apporte une synthèse de l'évolution de tous les descripteurs disponibles, la projection de leurs états évolutifs (Sect. 8.2) fournissant le comportement pour chacun d'entre eux. Au cours de la diversification des galaxies, chacun des descripteurs évolue, c'est-à-dire se modifie avec le temps. En conséquence, tous les caractères pertinents pour décrire l'évolution des galaxies, c'est-à-dire tous les caractères qui se transforment à travers plusieurs états évolutifs identifiables, sont donc nécessairement corrélés entre eux. Un bon exemple concerne la couleur des étoiles et la masse des galaxies : des étoiles données rougissent avec leur âge en devenant plus froide, et les galaxies ont plutôt tendance à grossir avec l'âge de l'Univers à cause de la gravitation. Les évolutions de ces deux paramètres sont bien corrélées et pourtant sans aucun

lien physique entre eux.

Il semble donc utile de distinguer "corrélation physique" de "corrélation temporelle", comme déjà mentionné à la Sect. 2.2.2. En effet, pour l'analyse cladistique, les caractères doivent être si possible indépendants, excluant a priori un lien physique trop fort, en particulier les redondances entre descripteurs dont nous avons déjà vu qu'elles donneraient un poids artificiellement plus important à un phénomène physico-chimique donnné. Mais comme tous les caractères évoluent, les corrélations temporelles sont inéluctables. Avant de développer des modèles sophistiqués, il est indispensable de décorréler du paramètre temps.

Les transformations des caractères le long d'un cladogramme sont nécessairement rendu cohérentes par l'analyse cladistique et le critère d'optimisation choisi (par exemple la parcimonie). Les galaxies sont des objets complexes en évolution, elles existent, nous les observons. Chacun des caractères servant à les décrire et à caractériser leur état évolutif, évolue a priori librement, puisque nous les supposons indépendants, éventuellement sous certaines contraintes que nous pouvons imposer, comme par exemple le fait de se transformer plutôt régulièrement (critère d'optimisation de WAGNER, Sect. 4.3). Mais la diversité observée des galaxies, incluse dans la matrice de l'analyse astrocladistique, montre que toutes les combinaisons ne se sont pas produites. Le cladogramme est une visualisation synthétique de cet état de fait, une vision cohérente de nos observations. Il nous aide à comprendre pourquoi les caractères n'ont peut-être pas évolué totalement librement. C'est ainsi que l'astrocladistique apporte des indications fortes sur la physique des processus évolutifs des galaxies.

Changer la racine d'un arbre ne modifie pas le comportement évolutif d'un seul caractère, mais de tous. Cette cohérence nécessaire impose des corrélations entre les descripteurs, corrélations "physiques" ou "temporelles". C'est sans doute ici que réside toute la richesse de l'analyse cladistique, en étant à la fois multivariée et évolutive. Citons l'exemple de la masse des galaxies de Virgo (FRAIX-BURNET, 2006, et fig. 8.5) qui semble devoir augmenter au cours de l'évolution. Ce résultat n'est pas une hypothèse, en particulier il n'est pas du tout lié au modèle hiérarchique de formation des galaxies. C'est un résultat de pure cohérence observationnelle entre tous les caractères, et une conséquence de l'enracinement de l'arbre basé sur plusieurs paramètres dont, entre autres, le rougissement des étoiles et l'augmentation de la métallicité du milieu interstellaire. Nous tenons là probablement la preuve du modèle hiérarchique (ne serait-ce que dans le sens où les galaxies grossissent au cours du temps), à moins qu'on renonce à des pans entiers de physique stellaire et de chimie interstellaire.

### 8.3.3 Diversification et évolution des galaxies

Une fois affirmées l'orientation du sens de la diversification et la structure, l'interprétation de l'arbre peut s'attacher aux différents groupes et leurs caractéristiques qu'il faut ensuite comprendre par des modèles. En particulier, le lien entre la diversification et l'évolution doit être impérativement précisé, car le cladogramme dont nous avons parlé jusqu'à présent dans ce chapitre révèle la diversification, pas l'évolution proprement dite. Pour ce faire, il nous faut essayer de déterminer une sorte d'horloge évolutive basée sur un nombre très restreint de caractères. Concrètement cela implique d'estimer les échelles relatives de temps des changements d'états, apportant ainsi un éclairage sur le taux de diversification : est-ce que le degré de diversification est plus

ou moins proportionnel au temps ou est-ce que cela dépend grandement des caractères impliqués ? Le terme d'horloge semble également indiquer une base temporelle qui permet de chiffrer dans l'absolu le rythme de l'évolution. Cet idéal nécessite un caractère dont le taux de changement de ces états évolutifs serait assez régulier, fiable et bien connu. Il permettrait une datation des divergences apparues dans la diversification des galaxies. Sans aller aussi loin, il est probablement déjà suffisant d'utiliser un caractère un peu global dont l'évolution d'ensemble nous paraît bien connue. Par exemple, la couleur B-V d'une galaxie a tendance à augmenter, mises à part des épisodes de sursauts de formation d'étoiles dans certains groupes de galaxies. La convergence de plusieurs de ces critères permet en principe de clarifier le lien entre diversification et évolution sur le cladogramme. Mais nous voyons encore une fois la difficulté d'utilisation de ce concept d'évolution que nous avons déjà mentionné à plusieurs reprises. Par ailleurs, il ne faut pas confondre cette horloge évolutive avec le redshift d'une galaxie, donc avec l'âge de l'Univers. Nous parlons d'évolution des galaxies, pas de l'évolution de l'Univers : des groupes de galaxies peuvent avoir évolué de la même manière, mais à des époques différentes de l'Univers.

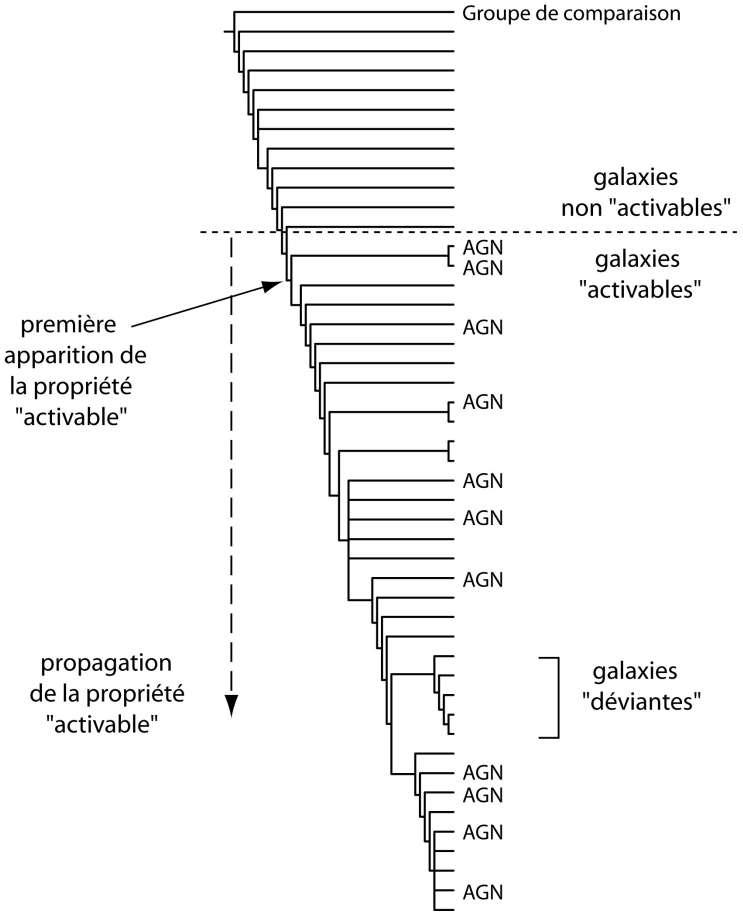



FIGURE 8.5 – Cladogramme schématique des galaxies de l'amas de Virgo (d'après FRAIX-BURNET, 2006).

L'interprétation astrophysique d'un arbre en terme de diversification s'effectue en

gardant à l'esprit que chaque groupe a acquis ses propriétés par transmission lors de processus évolutifs successifs. La séquence exacte de ces évènements est sans doute difficile à établir et restera peut-être inaccessible. Mais les objets d'un même groupe ont certainement eu des histoires similaires puisqu'ils ont hérités de propriétés communes. C'est ce qui définit les liens de parenté entre les galaxies. Le cladogramme de la figure 8.5 est dit déséquilibré car la diversification s'effectue d'une manière assez régulière avec une seule sous-structure (groupe des galaxies dites "déviantes") à partir d'une espèce ancestrale commune située en haut à gauche. Le groupe de comparaison est donc le plus ressemblant à cet ancêtre. Chacun des nœuds qui s'en suit correspond à un ancêtre commun à toutes les branches filles, c'est-à-dire toutes les branches situées "en aval". Chaque nœud et toutes les branches filles correspondantes (donc les taxons aux terminaisons) forment un groupe monophylétique, ou encore clade, qui partagent une même propriété héritée de leur ancêtre commun.

Afin d'illustrer la notion de transmission lors de la diversification, nous avons choisi sur la figure 8.5 un caractère : la présence d'un noyau actif (AGN pour "Active Galactic Nucleus"). Le cladogramme montre que les AGNs n'apparaissent qu'à un certain stade de la diversification, donc à un certain stade évolutif. Mais il apparaît également qu'ils ne sont pas toujours présents ensuite, étant dispersés d'une manière apparemment aléatoire sur la deuxième partie de l'arbre, étant simplement absents du groupe de galaxies "déviantes". Il existe donc une propriété qui se transmet, qui n'est pas la présence d'un noyau actif, mais la possibilité que celui-ci se manifeste. Nous parlons donc de galaxies "activables". Nous avons bien identifié un groupe évolutif, sans être certain de la position de l'ancêtre : le nœud correspondant est situé au plus tard dans l'évolution au niveau du premier AGN, mais cette propriété "activable" a pu apparaître plus tôt. Des analyses ultérieures, avec d'autres objets et d'autres descripteurs apporteront certainement des éléments de réponse. En particulier, il faudra mieux caractériser ce groupe afin de déterminer si ce caractère AGN ne cache pas des descripteurs plus fondamentaux (par exemple la masse du trou noir central). Le nom de ce groupe est donc laissé entre guillemets car il n'est pas satisfaisant puisque lié à une propriété particulière.

De même, le groupe "déviant" est ainsi dénommé provisoirement, d'après les comportements évolutifs de ces caractères. L'étude complète (FRAIX-BURNET, 2006) montre en effet que ces galaxies sont petites, alors qu'elles sont bien diversifiées puisqu'elles se situent loin de l'espèce ancestrale. L'ensemble des caractères indiquent donc qu'elles ont évolué comme les autres, mais elles ont une masse faible pour ce stade évolutif, et n'ont pas d'AGN parmi elles. Est-ce que l'échantillon est trop restreint ? Est-ce qu'il s'agit d'un cas de renversement dans l'évolution du caractère (elles auraient perdu de la masse et le caractère "activable") ? N'ont-elles simplement pas pu grossir à cause de leur environnement particulier ? Est-ce que leurs AGNs ne peuvent se manifester parce qu'il leur manque une propriété que nous ignorons ? Autant de questions posées non pas sous l'angle de particularités individuelles, mais de particularités collectives de groupes évolutifs entiers, dans un schéma global de diversification.

## 8.4 Clades, espèces, groupes évolutifs

Pour faciliter la description de l'histoire évolutive des galaxies, en utiilisant les propriétés statistiques de chaque groupes évolutifs, il est indispensable de disposer d'une classification adaptée, tout au moins de regrouper les galaxies à partir d'arguments évolutifs. Le cladogramme est un outil idéal pour cela. Dans le Chapitre 10, nous aborderons la question de la taxonomie dans le contexte des galaxies et de l'astrocladistique. Mais comment peut-on identifier concrètement des groupes à partir d'un cladogramme ?

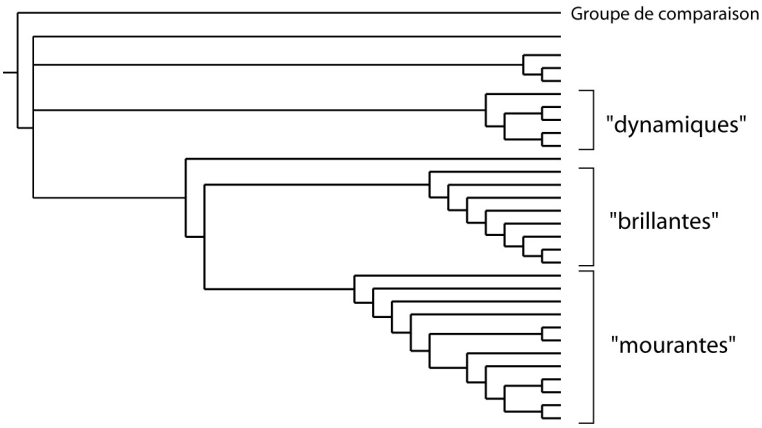



FIGURE 8.6 – Cladogramme schématique des galaxies naines du Groupe Local (d'après Fraix-Burnet et al. 2006c).

Nous avons déjà vu que des sous-branches peuvent constituer des groupes évolutifs s'ils peuvent être caractérisés par des états de caractères particuliers. Ces sous-branches peuvent être une grosse partie de l'arbre comme sur la figure 8.5. Ils peuvent également en être une toute petite partie comme le groupe des galaxies "déviantes" de la même figure.

Un autre arbre est montré sur la figure 8.6. À partir du nœud ancestral, une branche mène au groupe de comparaison, tandis que l'autre mène à une polytomie, c'est-à-dire à un nœud non résolu d'où partent quatre branches. Les liens relatifs de diversification ne sont pas donc pas déterminés. Une branche mène à un taxon isolé, une autre à un petit groupe (non identifié sur la figure), une autre à un groupe marqué "dynamiques", enfin la quatrième branche caractérise une lignée complexe, avec entre autre deux groupes bien séparés, "brillantes" et "mourantes". Les dénominations provisoires des groupes correspondent à un ensemble de propriétés bien caractéristiques et ne répondent pas entièrement à des critères rigoureux de taxonomie.

Les trois groupes identifiés sont des groupes évolutifs, et si on y inclut l'ancêtre de chacun, ils constituent très probablement des clades, ce que des études ultérieures devront cependant confirmer. Rien n'interdit de définir d'autres groupes, pour des raisons de commodités, mais ces groupes ne seront pas monophylétiques. Ils ne correspondront pas à un groupement naturel issu de la diversification des galaxies. La notion d'espèce est d'ailleurs suffisamment floue (voir Sect 3.3.3 et Chapitre 10) pour permettre d'identifier simplement un groupe ayant une certaine spécificité. Néanmoins, elle ne repose pas sur des critères évolutifs, de sorte que la notion d'espèce n'a pas de

rapport avec la cladistique.

La définition de clades et de groupes évolutifs parmi les galaxies constitue la base du travail de classification qu'il faudra un jour effectuer. Ceci ne sera possible que lorsqu'une vision assez large de la diversité connue des galaxies aura été établie dans un contexte évolutif, c'est-à-dire lorsqu'une ébauche d'un superarbre des galaxies pourra être construit.

## 8.5 Propriétés statistiques des clades

La définition des groupes permet de mettre en évidence des caractéristiques statistiques propres à chacun d'entre eux. En effet, la particularité des variables continues est d'être soumises à une dispersion naturelle parfois appelée variance cosmique dans le cas astrophysique. Comparer deux groupes revient à comparer les distributions des variables au sein de chacun. Plutôt que de définir une frontière rigide et le plus souvent arbitraire, il est plus réaliste de s'attendre à un chevauchement des distributions. Nous allons donc illustrer dans cette section comment une classification plus pertinente, même si moins facile à visualiser, permet une interprétation physique plus pertinente grâce à deux exemples.

### 8.5.1 Exemple 1 : le plan fondamental des galaxies

Dans cette étude (FRAIX-BURNET ET AL., 2010), quatre paramètres seulement ont été utilisé pour analyser un échantillon de 699 galaxies. Trois d'entre eux (brillance de surface, dispersion centrale de vitesse et rayon effectif) constituent un espace dans lequel les galaxies de types précoces (elliptiques et lenticulaires) se rangent sur un plan assez plat par endroit. C'est la fameuse relation dite du plan fondamental connue depuis plus d'une vingtaine d'années. Le quatrième paramètre est l'indice Mg2 qui caractérise approximativement la métallicité.

Cet exemple d'astrocladistique peut paraître étonnant étant donné le très faible nombre de caractères choisi. Cependant, il faut noter que les 4 paramètres choisis sont nécessairement indépendants, c'est-à-dire non liés causalement. C'est du reste la raison pour laquelle ce plan fondamental a résisté à une explication simple jusqu'à présent. Nous avons en effet là affaire à un cas exemplaire de corrélation temporelle, historique, paramétrique ou fortuite, de variables indépendantes dépendant toutes du temps ou d'une certaine mesure équivalente de diversification. Il s'agit donc d'un cas parfait pour l'astrocladistique, et le résultat de FRAIX-BURNET ET AL. (2010) le démontre assez nettement. Bien entendu, il faudrait allonger la liste des paramètres, mais comme nous l'avons vu précédemment, cela n'est pas aisé et ne peut se faire d'emblée, à cause de l'évolution cosmique ou évolution stellaire redondante dans les observables et essentiellement absente de ce jeu de caractères.

Malgré le rapport défavorable du nombre de caractères par rapport au nombre de galaxies, l'analyse astrocladistique a convergé rapidement vers un arbre structuré assez robuste (fig. 8.7). Sept groupes ont été identifié, mais ce nombre est quelque peu arbitraire. En effet, en regardant l'arbre en détail, on peut définir d'autres groupes plus petits. Cela montre encore une fois la différence entre une analyse phylogénétique et

une classification, la cladistique ne traitant que de la première. Le regroupement pourrait donc être modifié, et les anlayses suivantes refaites.

Le plan fondamental, presque vu de face (plan dispersion de vitesse – brillance de surface), et montrer sur la figure 8.7 également, avec l'arbre projeté sur ce plan. Cette projection illustre deux aspects nouveaux apportés par l'astrocladistique. Tout d'abord, elle fournit les relations évolutives entre les groupes, relations visualisables sur n'importe quelle projection, ce qui est une indication cruciale pour l'interprétation physique de l'origine du plan fondamental et de ces structures. Ensuite, elle permet de garder à l'esprit que n'importe quel diagramme binaire, n'importe quel corrélation, n'est qu'une projection d'un espace de plus grande dimensionalité. Ainsi, même le plan fondamental n'est qu'une projection d'une structure dans un espace à plus de trois dimensions. Dans l'exemple présent, les groupements et le schéma évolutif (l'arbre) sont obtenus dans un espace à cinq dimensions, comprenant les quatre paramètres utilisés pour l'analyse plus l'état de diversification.

Il en résulte une dispersion due à cette projection, en plus de la dispersion naturelle. Il est donc essentiel de traiter les groupes de manière statistiques comme illustré sur la figure 8.9 où des représentations du type "boxplot" montre la dispersion de chacun des paramètres pour chacun des groupes, et permet de comparer facilement les groupes entre eux. Aussi bien les dispersions, que les valeurs moyennes et médianes que les recouvrements, sont utiles pour les interprétations physiques de l'origine de ces groupes, de leurs ressemblances et différences réelles, et de leurs histoires.

Les simulations numériques apportent beaucoup d'aide pour l'interprétation. En effet, celles-ci prennent en compte la physique des processus évolutifs des galaxies, alors que les travaux plus observationnels découpent systématiquement le plan fondamental sur des critères morphologiques. Ainsi, de nombreuses corrélations bien connues prennent des aspects différents avec les groupes obtenus par cladistique (fig. 8.8), présentent des ressemblances parfois étonnantes avec des résultats de simulations numériques (pour plus de détail voir FRAIX-BURNET ET AL., 2010). Nous pouvons en particulier constater sur la figure 8.8, diagramme Mg2 – dispersion de vitesse (log-sigma), que la corrélation bien visible et bien connue pour l'ensemble de l'échantillon disparaît complètement au sein de chaque groupe. Par contre, les groupes se distribuent le long de cette corrélation globale en correspondance avec le niveau de diversification. Faut-il une preuve plus nette que cette corrélation n'est pas physique, mais fortuite (ou évolutive ou historique) due à la diversification ?

Comme l'affirme l'astrocladistique, les galaxies sont des objets réellement complexes. Une preuve en est que les simulations numériques ne peuvent pas échantillonner toutes les possibilités de formation et de transformation des galaxies, sachant que les objets que nous observons sont le résultat d'une multitude de processus successifs. Les simulations finissent donc par produire une population de galaxies simulées, à comparer donc avec la population de galaxies réelles. Grâce à ces travaux de plus en plus nombreux, de grandes lignes se dégagent. Nous trouvons par exemple que trois de nos groupes présentent un plan fondamental bien défini et sont donc le résultat d'au moins une fusion dissipative. Pour l'un d'entre eux, ces évènements doivent être assez anciens puisque la métallicité est faible. Deux d'entre eux ont subi en plus des accrétions ou fusions non-dissipatives pour expliquer leur grande masse. Le troisième groupe, regroupant des galaxies moins massives donc, occupent dans l'espace des phases le même emplacement que les bulbes de galaxies spirales. Ceci n'est pas

totalement nouveau, mais démontre encore une fois que la morphologie n'est pas un critère très pertinent vis à vis de l'évolution. De plus, ceci démontre également le danger d'établir une classification a priori, basée sur un critère totalement arbitraire. L'astrocladistique', bien au contraire, se base uniquement sur la physique, le plus complètement possible puisque multivariée.

Cette analyse astrocladistique montre également que notre classification ne dépend pas de l'amas ni du redshift dans l'échantillon. C'est-à-dire que ce n'est pas le plan fondamental qui est universel, mais plutôt la classification trouvée, autrement dit la diversification est la même partout dans les amas à notre disposition. Bien sûr, ce résultat ne concerne que l"échantillon étudié et sa plage de redshifts. En conséquence, le plan fondamental est en réalité une collection de régions dans l'espace des phases correspondant, régions définissant plus ou moins bien un plan. En particulier, les groupes les moins diversifiés sont très dispersés, et leurs galaxies n'ont donc probablement pas été formées par fusions, mais plutôt par effondrement monolithique et accrétions. On peut également affirmer que les galaxies de l'un de ces groupes semblent avoir été balayées de leur gaz.

L'interprétation de ce genre d'analyses est donc extrêmement riche, grâce notamment à l'objectivité de la classification, obtenue sans aucun modèle ou découpage a priori, et grâce à sa pertinence en terme de physique puisque nous n'utilisons que des paramètres adéquats. À l'évidence, l'astrocladistique apporte une nouvelle manière de comprendre ce fameux plan fondamental, qui reste un peu énigmatique depuis plus de 20 ans, en y introduisant la notion d'évolution, mais également en collant davantage à la physique.

### 8.5.2 Exemple 2 : les amas globulaires de notre Galaxie

Les amas globulaires sont finalement des galaxies simples, sans gaz ni poussière, dont les propriétés sont presque essentiellement définies par l'environnement dans lequel ils se sont formés. Il était connu depuis longtemps que les amas globulaires de notre Galaxie pouvait se répartir en quelques groupes, entre 2 et 4, mais que surtout leur diversité nécessitait un second paramètre en plus de la seule métallicité Fe/H. Selon la philosophie de l'astrocladistique, c'est évidemment cet environnement de formation qui constitue ce fameux "second paramètre".

Sur un échantillon de 54 amas globulaires de notre Galaxie (FRAIX-BURNET ET AL., 2009), trois paramètres caractérisant l'état physico-chimique des amas ont été choisis. dans la même philosophie que pour le plan fondamental précédemment. Leur âge a été ajouté mais il permet seulement de ranger les amas au sein de chaque groupe évolutif. Dans l'analyse elle-même, l'âge s'est vue attribué un poids d'un demi, car il est bien clair qu'il représente l'âge des étoiles et ne caractérise donc pas les conditions physico-chimique dans laquelle les amas observés se sont formés. L'âge ne peut pas discriminer entre deux populations puisqu'il évolue de la même manière pour tous. Par contre, au sein d'un même groupe évolutif, la prise en compte de ce paramètre dans l'analyse, avec un poids moindre, permet de ranger les objets selon une séquence chronologique. La preuve de ceci est que l'analyse effectuée sans l'âge retrouve exactement les mêmes groupes, mais ces groupes sont alors non résolus (polytomies).

Le résultat de l'analyse cladistique est montré sur la figure 8.10. L'échantillon se découpe aisément en 3 groupes dont les propriétés, analysées de manière statistique

comme pour le plan fondamental précédemment, indiquent sans ambiquité des conditions de formation différentes. Sur les diagrammes binaires obtenus avec les quatre paramètres de l'analyse (Fig. 8.10), les groupes se distinguent nettement et indiquent à l'évidence des propriétés différentes. En particulier, la métallicité en fonction de l'âge montre que les trois groupes se sont formés avec des métallicités moyennes différentes, à des périodes différentes, et sur des durées différentes. Il est également frappant dans cet exemple de constater que les groupes se distinguent selon des paramètres totalement différents de ceux utilisés pour l'analyse. La table 8.1 montre les valeurs moyennes (ainsi que la dispersion entre parenthèses) au sein de chaque groupe pour quelques paramètres essentiellement d'ordre géométrique.

| | Groupe 1 | Groupe 2 | Groupe 3 |
|---|---|---|---|
| Nombre d'amas | 25 | 11 | 18 |
| Distance au centre galactique (kpc) | 9.4 (7.4) | 12.9 (8.0) | 4.2 (2.9) |
| Hauteur au-dessus du disque galactique (kpc) | 4.8 (4.5) | 8.6 (7.6) | 1.9 (2.0) |
| Métallicité Fe/H | -1.40 (0.35) | -1.92 (0.16) | -0.92 (0.35) |
| Magnitude V | -8.5 (0.7) | -7.6 (0.6) | -7.1 (0.9) |
| Vitesse de rotation dans le plan de notre Galaxie (km/s) | -7. | +46. | +119. |
| Dispersion des vitesses radiales (km/s) | 120 (107) | 151 (107) | 69 (74) |
| Âge | 9.98 (0.96) | 11.17 (0.70) | 10.18 (0.48) |

TABLE 8.1 – Valeurs moyennes (et dispersion) de quelques propriétés pour chaque groupe d'amas globulaires.

Tout ces points démontrent encore une fois qu'une classification pertinente (multivariée et évolutive) amène à une interprétation physique pertinente. Cette analyse démontre également que la diversité des objets, ici les amas globulaires, est bien due principalement à leur histoire, celle de leur assemblage dans le cas présent. Il est alors possible de décrire assez précisément les scénarios de formation de ces amas dans notre Galaxie, apportant naturellement des contraintes sur la manière dont notre Galaxie s'est assemblée. En effet, en utilisant plusieurs caractéristiques de chacun des groupes, nous pouvons déduire que les amas du groupe 2 sont situés principalement dans le halo externe et se sont formés en premier lors d'une phase non dissipative de l'effondrement de notre Galaxie. Les amas du groupe 1, plutôt du halo interne, se sont formés plus tard dans un environnement un peu plus turbulent. Enfin, les amas du groupe 3 se sont formés essentiellement dans le disque épais de notre Galaxie dans une période intermédiaire et plus courte.

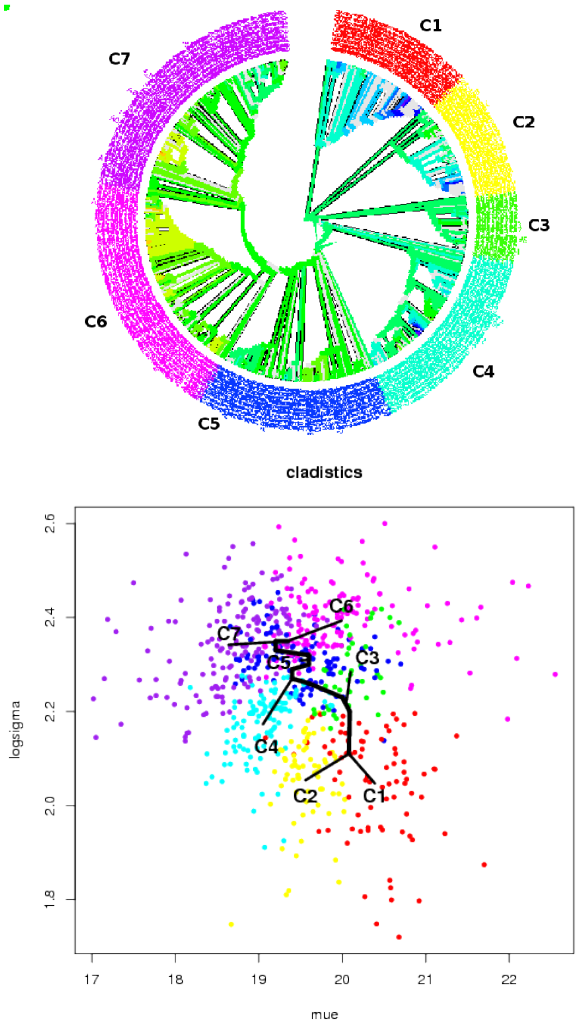



FIGURE 8.7 – En haut : arbre comprenant 699 galaxies obtenu dans l'analyse cladistique du plan fondamental à partir duquel les groupes sont définis. Chaque groupe est représenté par une couleur. En bas : projection de l'arbre (traits noirs) et des groupes dans le plan.

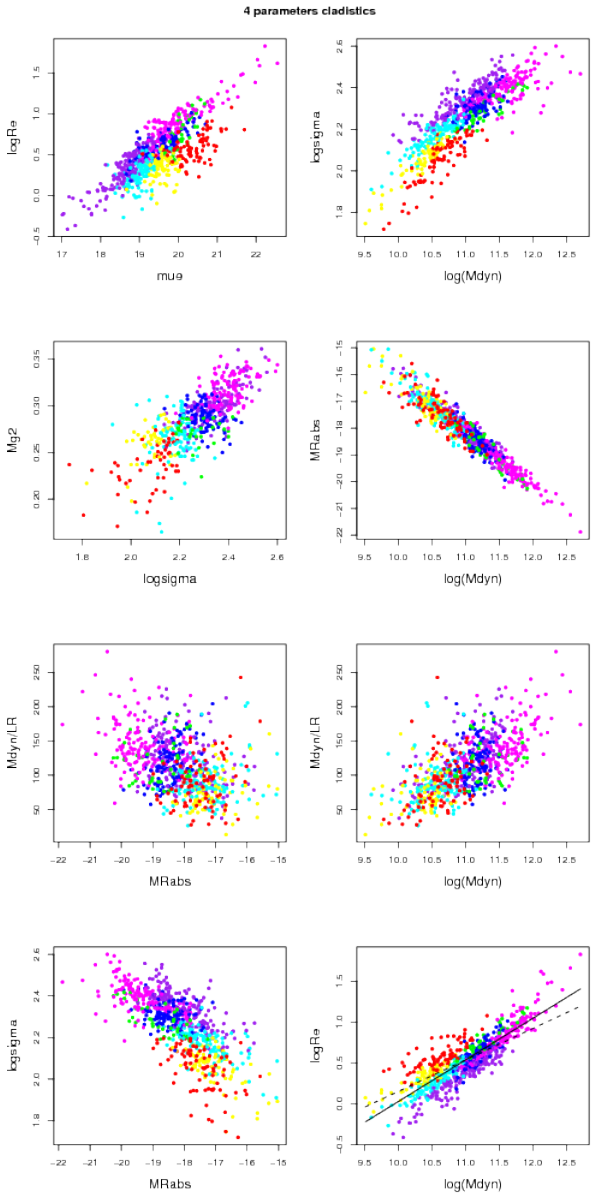



FIGURE 8.8 – Relations entre différents paramètres.

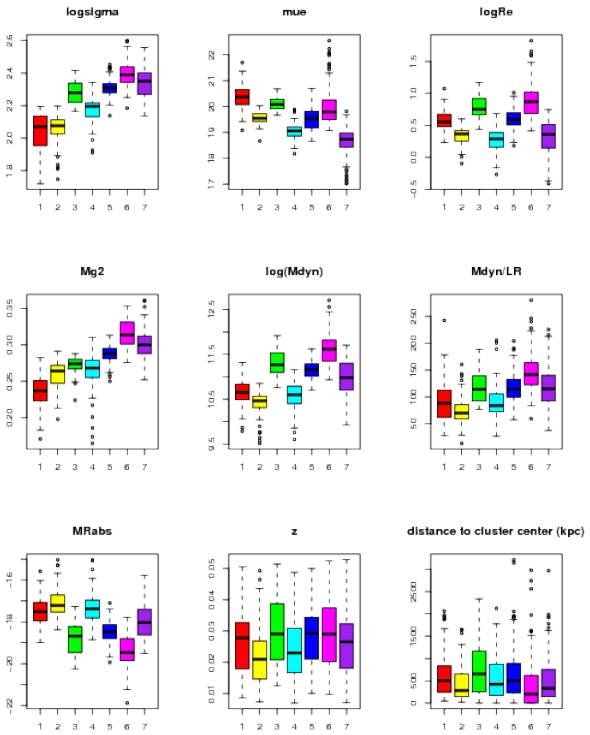



FIGURE 8.9 – Dispersions des paramètres au sein de chaque groupe, représentées sous la forme de "boxplot".

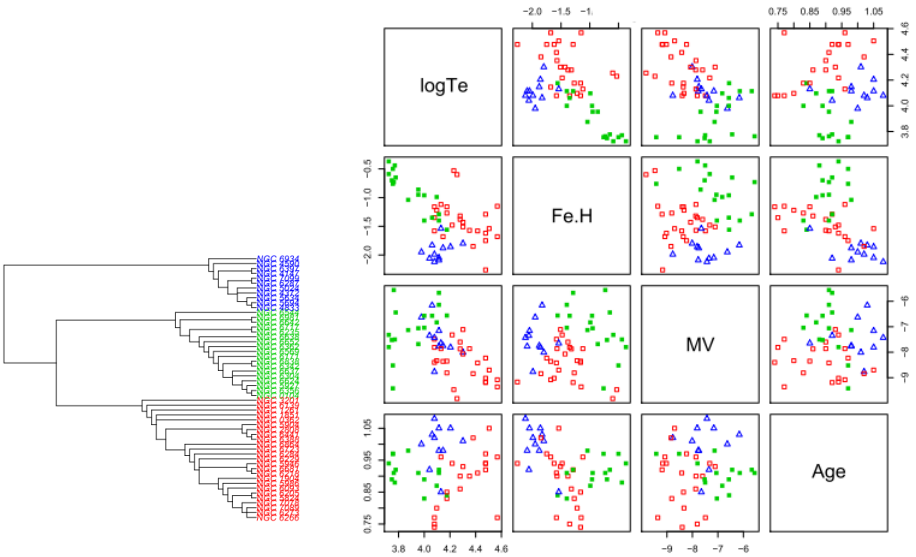



FIGURE 8.10 – Classification cladistique des amas globulaires de notre Galaxie. Chaque groupe est représenté par une couleur. À gauche : cladogramme à partir duquel les groupes sont définis. À droite : diagrammes reliant les 4 paramètres utilisés pour l'analyse cladistique.

# Chapitre 9

# Méthodes phylogénétiques : vue d'ensemble

Nous allons voir dans ce chapitre que la cladistique fait partie des méthodes phylogénétiques qui couvrent l'ensemble des approches permettant de reconstruire un arbre évolutif. En réalité, on distingue les méthodes basées sur les caractères (comme la parcimonie maximum évoquée tout au long de ce livre) des méthodes basées sur une distance. De plus, ces méthodes peuvent être non-paramétriques (parcimonie maximum, évolution minimum) ou paramétriques (dites probabilistes comme le maximum de vraisemblance ou les méthodes Bayésiennes), ces dernières imposant un modèle d'évolution. Nous donnons ici un aperçu général de leurs différences sans entrer dans les détails mathématiques. Les développements théoriques, algorithmiques et applicatifs sont assez foisonnants dans ce domaine, l'exploration des limites de chaque méthode étant en perpétuelle expérimentation. Un aperçu de ces différentes méthodes peut être trouvé dans DARLU AND TASSY (1993), CROFT (2008), ou encore SEMPLE AND STEEL (2003) pour les aspects mathématiques.

En fin de chapitre, nous examinerons les analyses par regroupement ("cluster analyses" ou analyses de distances multivariées) qui en dehors de toutes considérations évolutives permettent de regrouper les objets à partir d'une matrice de distances. Ces analyses multivariées commencent elles aussi à être utiliser en astrophysique. Elles sont un complément intéressant aux analyses cladistiques.

Ce chapitre, très bref, vise à montrer que l'astrocladistique n'est qu'une porte ouverte sur tout un monde extrêmement vaste de méthodes phylogénétiques. Par l'adoption du paradigme de population en évolution, l'astrocladistique rapproche l'astrophysique d'un autre champ de recherche très dynamique.

## 9.1 Méthodes de caractères et méthodes de distances

La cladistique telle que nous l'avons pratiquée dans ce livre se fait appeler généralement méthode de parcimonie maximum. Elle se base sur la matrice de caractères qui sont des données issues plus ou moins directement des observables et caractérisant l'état évolutif de chaque taxon. L'arbre final choisi, le plus parcimonieux, comporte un nombre de changements d'état de caractères minimum. La quantité optimisée est donc le nombre d'évènements total apparaissant sur le schéma évolutif représenté par

l'arbre. Ce nombre n'a a priori pas les propriétés d'une distance au sens mathématique, et ne fournit a priori aucune indication temporelle.

Les méthodes de caractères comprennent le maximum de parcimonie (cladistique, non paramétrique) et le maximum de vraisemblance ("maximum likelihood", paramétrique). Cette dernière méthode définit un modèle d'évolution a priori sur lequel la phylogénie va être ajustée au maximum.

À partir d'une matrice de caractères, il est aisé de construire une matrice de distances, les caractères représentant des coordonnées dans un espace multidimensionnel. La métrique de cet espace est caractérisée par le choix de la distance qui n'est pas unique. Toute la difficulté réside donc dans ce choix qui dépend nécessairement du type d'objets et de données mais reste quelque peu arbitraire.

À partir de la matrice de distances, il est généralement impossible de retrouver la matrice de caractères, ce qui prouve qu'une certaine quantité d'information est perdue. En effet, la distance s'intéresse à la différence globale entre les objets sans se préoccuper du chemin nécessaire pour passer d'un objet à l'autre. La mesure des distances tend à court-circuiter les évènements évolutifs, lissant les changement trop fréquents ou apparaissant sur une faible fraction des caractères. Elles supposent donc nécessairement que les homoplasies sont négligeables sans avoir de réel moyen de vérifier cette hypothèse sous-jacente. En somme les analyses de distances s'intéressent aux distances dans un espace à la métrique prédéfinie, alors que les analyses de caractères s'intéressent au cout évolutif (en quelque sorte l'effort à fournir) du chemin à parcourir dans un relief inconnu. Ce chemin passe par les nœuds de l'arbre qui représentent des points intermédiaires dans l'évolution de chacun des caractères. Les analyses de distances ignorent ces points intermédiaires et n'indiquent donc pas l'état des caractères aux nœuds de l'arbre phylogénétique ni les évolutions trop complexes des caractères.

Ceci dit, lorsque les caractères, à partir desquels les distances sont calculées, sont bien pertinents vis à vis de la diversification, c'est-à-dire essentiellement dans le cas où ils comportent peu d'homoplasies, les résultats des deux types d'analyses ne diffèrent pas toujours dramatiquement. Cela signifie que l'information perdue n'est pas toujours importante et peut sans doute s'expliquer par un comportement adéquat des caractères. L'avantage des méthodes de distances est que les algorithmes sont généralement bien plus rapide que pour la parcimonie maximum, ce qui explique leur usage très répandu.

## 9.2 Méthodes paramétriques et non-paramétriques

Le méthodes paramétriques présupposent un certain modèle d'évolution permettant de contraindre la reconstruction de la phylogénie. Parmi les méthodes paramétriques, on trouve les méthodes d'ajustement comme le maximum de vraisemblance qui est basée sur les caractères, alors que les méthodes de distances paramétriques sont très nombreuses. On peut mentionner à titre indicatif les méthodes dites Bayesiennes développées assez récemment qui peuvent utiliser les caractères.

Parmi les méthodes non-paramétriques, moins nombreuses, dont le maximum de parcimonie ou cladistique fait partie, on trouve les méthodes agglomératives de reconstruction d'arbres dont les plus connues sont UPGMA (unweighted pair-group method of arithmetic averages) et Neighbor-Joining qui sont des méthodes de distances.

Les méthodes paramétriques convergent généralement plus rapidement, et per-

mettent finalement de tester notre compréhension des processus de diversification. Mais les méthodes non-paramétriques sont plus indépendantes de ces hypothèses a priori.

Néanmoins, la cladistique, qui reste une méthode de choix en ce qui concerne la phylogénie, n'est pas dénuée de présupposé quant au modèle d'évolution puisque l'analyse retient les arbres les plus parcimonieux. Cela est un choix de nature probabiliste, par rapport à un modèle d'évolution implicite, incluant non seulement les paramètres des taxons, mais également ceux, induits, des nœuds de l'arbre. Il a été démontré que dans certains cas précis, l'analyse peut ainsi renforcer une branche erronée avec l'augmentation du nombre de caractère, rendant ainsi la méthode non statistiquement consistante.

## 9.3 Méthodes probabilistes

Les méthodes dites probabilistes présupposent un modèle d'évolution des caractères, et cherchent l'arbre dont les branches correspondent le mieux à ce modèle. Ce sont donc des méthodes généralement paramétriques, qui peuvent être très efficaces, mais nécessitent une bonne connnaissance des processus de diversification. Il existe de nombreux modèles d'évolution chez les biologistes, les plus classiques reposant sur des fréquences aléatoires de mutation de gènes.

La transposition à l'astrophysique concernerait a priori seulement la méthode, pas les modèles. En effet, on peut imaginer que ce genre d'approches serait bien adaptée car nous avons une bonne connaissance des phénomènes physico-chimiques à l'œuvre dans les processus de transformation des galaxies. Nous devrions même pouvoir formaliser les processus de transmission avec modification d'un point de vue statistique. Les méthodes probabilistes n'ont encore pas été utilisées en astrophysique, la cladistique ayant constitué la méthode la plus logique et la plus objective pour entrer dans le monde de la phylogénétique extragalactique.

## 9.4 La cladistique en caractères continus

Contrairement à ce qu'on pourrait imaginer, les biologistes ont aussi des observables quantitatives à disposition. Ce sont les données morphométriques comme par exemple des longueurs de membres, des volumes d'organes, ou des variables décrivant des formes géométriques plus ou moins complexes. Autant ce genre de données sont facilement utilisables dans des analyses de distances, voire même dans des analyses de caractères probabilistes et paramétriques, autant elles soulèvent des difficultés pour la cladistique. Ce qui finalement a engendré beaucoup de questions et de réticences à les utiliser pour des études phylogénétiques.

Ces difficultés sont de deux ordres. Le premier est conceptuel, le deuxième algorithmique. Ce dernier n'a sans doute pas pu se développer beaucoup du fait du premier.

Si une variable est réellement continue, elle ne semble pas utile pour la phylogénie parce que dans l'idéal de la cladistique, les clades sont identifiables par des états de caractères bien spécifiques, qui peuvent se transmettre ou se modifier lors de la transmission. Ces états de caractère sont appelés à identifier et définir les groupes évolutifs. La question qui se pose dans le cas des variables continues est l'identification de ces

états ou de leur équivalent. La solution la plus simple est la discrétisation qui a été largement utilisée, notamment en astrocladistique jusqu'à présent. Plusieurs problèmes se posent pour la discrétisation.

Premièrement, quelle est la signification de la limite des bins ? On retrouve ce problème dans bien d'autres situations, comme par exemple la classification morphologique des galaxies de DE VAUCOULEURS, les noyaux actifs de galaxies avec la dichotomie FRI/FRII, ou même le découpage des couleurs (bleu, vert, rouge, ...) alors que chaque couleur n'est que la délimitation arbitraire d'un continuum spectral. Il n'y a pas de réponse absolue à cette question, et il est indispensable de garder à l'esprit que la modification de ces limites a nécessairement une incidence sur les résultats.

Deuxièmement, deux objets appartenant au même bin ne sont pas nécessairement identiques. Troisièmement, deux objets appartenant à deux bins différents ne sont pas nécessairement différents, et peuvent même être plus semblables que deux objets d'un même bin.

Quatrièmement, l'intérêt de l'analyse cladistique, ou de toute analyse phylogénétique, est de ranger les objets non pas dans des boites, mais sur un arbre, c'est-à-dire sur un schéma continu de relations évolutives. Il semble donc un peu contradictoire de ranger les caractères dans des boites, d'autant plus qu'on risque de perdre une information précieuse en agissant ainsi.

Ces problèmes jettent un doute sur la discrétisation de variables continues en cladistique, même si en pratique les résultats obtenus sont rarement aberrants. Quelle est donc la nature des caractères quantitatifs en cladistique ?

Pour les variables quantitatives en général, on s'attend qu'au sein d'un même groupe il y ait une distribution de valeurs pour un paramètre donné. C'est donc cette distribution qui caractérise ce groupe, et cette distribution qui permet d'établir les liens évolutifs entre les groupes. C'est exactement ce qui ressort de l'analyse du plan fondamental des galaxies telle que présentée sur la figure 8.9. Cependant, ce résultat à été obtenus avec des individus comme taxons, et l'étape suivante est de caractériser chacun des groupes afin de pouvoir effectuer une analyse cladistique avec les sept groupes identifiés. Cela revient évidemment à choisir un taxon supraspécifique (Sect. 7.3), en tenant compte du fait que les distributions de deux groupes pour un paramètre donné sont susceptibles de se chevaucher.

D'un point de vue statistique, le fait qu'un groupe soit caractérisé par des distributions de valeurs, même se recoupant, ne pose pas de problème. Par contre, la question reste de savoir ce qui est transmis lors d'une réplication, ou plutôt à quel point la modification va faire sortir le nouvel objet du groupe du progéniteur ? Il semble certain que des caractéristiques quantitatives sont bien transmises et caractérisent une lignée, que ce soit dans le cas des organismes vivants ou dans le cas des galaxies. Par exemple, une galaxie massive "engendrera" nécessairement une autre galaxie du même ordre de grandeur de masse. Le caractère "masse" est donc bien transmis, avec modification (la masse ayant toute probabilité d'augmentée au cours du processus de transformation), tout au moins dans une valeur moyenne pour le groupe, car il n'est pas impossible que la distribution soit elle modifiée.

Conceptuellement, il n'apparaît pas de contre-indication majeure à considérer les variables continues comme des caractères pour une analyse cladistique, bien que cela soit encore l'objet de certaines controverses.

Du point de vue algorithmique, on pourrait être tenté d'augmenter le nombre de

bins de manière conséquente afin de se rapprocher de la philosophie des calculs en éléments finis. Nous avons vu que le choix de 30 bins semble bien adapté pour les galaxies compte tenu des incertitudes de mesure standard (Sect. 6.5) et des limites du logiciel utilisé. Il est cependant possible d'aller au-delà, mais cette approche sous-entend l'hypothèse que chaque valeur appartient à un bin unique représentant un état évolutif censé caractériser le groupe. Or nous savons que le groupe est représenté par une distribution de valeurs, donc il doit être représenté par une distribution de bins. Ce n'est que très récemment qu'un algorithme spécifique a été établi dans cette optique. Il s'agit du logiciel TNT présenté en Sect. 6.8 (GOLOBOFF ET AL., 2006). Aucune application n'a encore été effectuée en astrocladistique, mais cela ne saurait tarder.

## 9.5 Analyses par regroupement

Les analyses de distances multivariées (analyses de regroupement ou "cluster analysis") ne sont pas des méthodes phylogénétiques à proprement parler car elles n'intègrent pas la notion d'évolution. Elles sont très répandues notamment dans les problématiques étudiant des "populations" au sens large du terme (médecine, sociologie, économie, biologie). La classification obtenue est objective et multicritère. Une abondante recherche en statistique est dévolue à cette activité. Des applications astrophysiques commencent à voir le jour, elles sont très complémentaires de l'astrocladistique. Dans ces méthodes il y a deux approches.

La première approche est hiérarchique et construit un arbre appelé dendogramme. La première chose à faire est de choisir une mesure de distance. La plus naturelle pour un physicien est la distance euclidienne. Mais elle n'est pas nécessairement justifiée pour n'importe quel type de regroupement. Il existe beaucoup d'autres distances, dont une bien connue est celle de Manhattan, qui sont souvent spécifiques du type de données (comme les variables discrètes ou génétiques).

La distance correspond à la métrique de l'espace des phases. Il y a ensuite plusieurs manières de définir la distance entre groupes, c'est-à-dire la manière de les relier ("linkage") : par exemple on peut prendre la moyenne de toutes les distances entre objets pris deux à deux, ou celle des objets les plus éloignés ou les plus proches. De ces deux choix, distance et linkage, résulteront des dendogrammes différents.

Enfin, une fois que le dendogramme est construit, il reste à définir les groupes. Pour cela, on peut utiliser des mesures de similitudes. En somme, cela revient à utiliser directement les longueurs de branches qui reflètent les distances entre objets et groupes potentiels, et à couper l'arbre à un niveau donné : si on coupe à un niveau bas (dimension des groupes petites), on obtient autant de groupes que d'objets, si on coupe à un niveau élevé (grande dimension) on n'obtient qu'un seul groupe. Contrairement au cas des cladogrammes, les branches des dendogrammes ont des longueurs calibrées, de sorte que le nombre et la composition des groupes sont imposées par le seul choix du niveau où l'on coupe, la structure de l'arbre dépendant du choix de la distance et du liinkage.

La deuxième approche est du type agglomérative-divisive (K-means) et cherche le nombre optimum de groupes selon des critères d'optimisation. Il existe là aussi plusieurs méthodes qui peuvent être plus ou moins bien adaptée au type de données. Il faut ensuite les compléter par des méthodes qui regroupent les objets dans le nombre de

groupes ainsi trouvé. On peut mentionner les analyses discriminantes qui permettent de grouper des objets en un nombre prédéfini de classes.

Ces analyses par regroupement ne sont encore pas très utilisées en astrophysique, mais commencent à prouver leur utilité (par exemple CHATTOPADHYAY AND CHATTOPADHYAY, 2007; CHATTOPADHYAY ET AL., 2007, 2009a,b; SÁNCHEZ ALMEIDA ET AL., 2010) même en complément de l'astrocladistique (FRAIX-BURNET ET AL., 2010).

Un dernier mot sur l'analyse en composantes principales (ACP ou PCA), méthode un peu utilisée en astrophysique. Il ne s'agit pas d'une méthode de classification ou de regroupement, mais d'une méthode de réduction de la dimensionalité. Elle est basée sur la matrice de corrélation, et cherche un espace orthonormé de vecteurs propres obtenus à partir de combinaisons linéaires des variables. Son intérêt est de limiter le nombre de paramètres permettant de décrire la diversité de l'échantillon, en éliiminant les variables corrélées, plus particulièrement en éliminant ainsi l'effet taille. On espère dégager ainsi les axes les plus discriminants. À partir de ces nouvelles coordonnées (les composantes principales), il est possible d'effectuer des analyses de groupement. Il est donc quelque peu incohérent de faire des ajustements de modèles dans l'espace ACP ou de comparer à d'autres objets à moins que ces modèles ou ces objets n'aient eux-mêmes été inclus dans l'échantillon pour l'analyse ACP. Il est en effet inutile de chercher un sens physique à ces vecteurs propres, combinaisons linéaires de variables réelles qui n'ont de justification que purement statistique. La signification physique doit toujours être effectuée avec les variables physiques dans le même esprit que les exemples donnés en Sect. 8.5.

# Chapitre 10

# Vers une nouvelle taxonomie des galaxies

L'astrocladistique propose de regrouper les galaxies non pas sur des critères observationnels partiels, donc d'apparence, mais à partir de phylogénies représentées sur des arbres. Les structures des cladogrammes permettent de définir de manière claire et objective les concepts de groupes évolutifs et de clades. Une classification basée sur une histoire évolutive de la diversité des galaxies est ainsi envisageable. Les biologistes utilisent encore la nomenclature de LINNÉ qui, bien que conceptuellement hiérarchique, n'en est pas pour autant justifiée par des considérations évolutives. Elle connaît du reste quelques conflits avec les résultats issues des analyses cladistiques. Le projet Phylocode (CANTINO AND DE QUEIROZ, 2010) vise à inventer une nouvelle classification des espèces vivantes fondée sur les phylogénies issues des cladogrammes et sur des règles taxonomiques rigoureuses. Cette initiative suscite bien entendu de nombreux débats vis à vis de la nomenclature linnéenne (NIXON ET AL., 2003; BENTON, 2007).

Il est bien évident qu'une telle tentative en astrophysique ne pourra aboutir que lorsque des phylogénies extrêmement robustes, fiables et suffisamment larges auront été établies, ce qui n'est pas le cas à ce jour. Néanmoins, la nécessité de nommer des groupes évolutifs afin de regrouper les millions de galaxies connues sur des bases cladistiques va devenir rapidement indispensable. Alors autant commencer sur des bases solides. Ce chapitre se veut être une introduction à la taxonomie et s'inspire largement des réflexions et des règles élaborées dans le cadre du Phylocode. Il ne s'agit pas encore d'ébaucher les éléments d'une classification, mais plutôt d'ouvrir des pistes de réflexion.

## 10.1 Notions de taxonomie

La taxonomie est l'art de nommer les choses, les organismes vivants plus particulièrement. LINNÉ, ADANSON et d'autres de leurs contemporains ont identifié les problèmes des classifications jusqu'au Moyen Âge. Il est apparut nécessaire de distinguer trois éléments : un nom, une définition et une description. Par exemple, un oiseau n'est pas défini comme un animal avec des ailes, cela incluerait les chauves-souris et bon nombre d'insectes. Les plumes non plus ne suffisent pas car nous ne savons pas ce que les découvertes paléontologiques nous réservent, sans parler même des nombreuses

espèces actuellement vivantes et encore inconnues. La définition complète d'un oiseau est un "vertébré tétrapode à sang chaud, au corps recouvert de plumes, dont les membres antérieurs sont des ailes, les membres postérieurs des pattes, dont la tête est munie d'un bec corné dépourvu de dents, et qui sont en général adaptés au vol". De même, les mammifères sont définis comme des "vertébrés à sang chaud et température constante, à respiration pulmonaire, dont les femelles allaitent leurs petits à la mamelle". La description quant à elle peut être beaucoup plus détaillée puisqu'elle est est censée rassembler toutes nos connaissances.

En tout état de cause, le nom doit être détaché le plus possible des caractères et ne peut à lui seul suggérer une définition, encore moins une description. C'est pour cela que la nomenclature de LINNÉ a eu tant de succès à travers les siècles en résolvant les problèmes des classifications du Moyen-Âge. De plus, cette nomenclature reflète bien la hiérarchisation de la diversité du monde vivant, elle a donc résisté à la théorie de DARWIN. Le plus remarquable est qu'elle soit toujours présente et fort utile malgré l'avènement de la biologie moléculaire avec ses descripteurs complètement inimaginables au XVIIIième siècle.

Cependant, la cladistique a changé un peu l'organisation de cette hiérarchie et a ancré l'idée que les cladogrammes, donc les classifications, évoluent avec les connaissances. En conséquence, les genres et les espèces ont été un peu redistribués, et leurs noms avec, de sorte que les rangements linnéens se retrouvent légèrement mélangés sur un cladogramme. Les grandes questions concernant la notion d'espèce, évoquées plus haut, se répercutent naturellement sur la nomenclature. Le système linnéen atteindrait-il ses limites ? Un cladogramme présente les liens de parenté entre les différents organismes vivants, et toute classification, devant nécessairement prendre en compte l'évolution, puisque cette dernière est responsable de la diversité, doit être compatible avec celui-ci. Un des objectifs de la biologie contemporaine est de réfléchir à la meilleure façon de nommer les lignées révélées par le cladogramme d'une manière pérenne, tout en autorisant une évolution des différents embranchements au gré des découvertes futures. Formidable défi dont le succès ne pourra être évalué que dans quelques siècles.

## 10.2 Problèmes de la taxonomie actuelle des galaxies

La qualité principale d'une classification est de pouvoir désigner simplement une multitude d'objets en englobant de multiples descripteurs. Résumons ici les problèmes liés aux classifications des galaxies.

Actuellement, les galaxies sont invariablement présentées comme appartenant à trois espèces, celles de la classification de HUBBLE. Puis viennent ensuite la descirption plus réaliste des galaxies avec foule de détalis observationnels et physico-chimiques. Pourtant, la classification de HUBBLE est basée sur la morphologie et présente plusieurs défauts de principe. C'est une classification d'abord visuelle, dans le sens où elle se fait encore essentiellement à l'œil car les logiciels de reconnaissance de forme n'arrivent pas à égaler la puissance de cet organe de l'être humain. L'ordinateur est dépourvu de la part de subjectivité nécessaire à cette classification puisque même les rares spécialistes ne sont pas toujours d'accord entre eux pour ranger les galaxies. Mais elle est visuelle aussi dans le sens où elle n'est définie, et donc valable, que dans le domaine visible uniquement. Cela pose une question philosophique intéressante :

pourquoi la diversité des galaxies, à travers plusieurs milliards d'années d'évolution, ne serait-elle caractérisée non seulement par un seul paramètre, la morphologie, mais qui plus est dans un domaine de longueur d'onde très restreint qui correspond au maximum de sensibilité de l'œil de l'être humain ? Ne serait-ce pas en contradiction avec le principe cosmologique ou anthropique qui suppose que l'homme n'occupe pas de place particulière dans l'Univers ?

Enfin, la morphologie est un paramètre est purement qualitatif, le seul parmi la totalité des observables à notre disposition. Ceci n'est peut-être pas un défaut important, mais ce sont tout de même les constituants fondamentaux (étoiles, gaz, poussière) qui dictent l'apparence et l'histoire d'une galaxie (voir Chapitre 3). Souvent, on semble justifier les propriétés fondamentales à partir de la morphologie. Par exemple, on explique la faible rotation ou la grande dispersion des vitesses ou encore la faible quantité d'hydrogène neutre par le fait qu'une galaxie est elliptique. Pourtant c'est l'inverse puisque c'est la cinématique qui dicte la forme. On voit ici le rôle de la nomenclature, et qu'inconsciemment on aimerait que le terme "elliptique" désigne bien autre chose qu'une simple forme.

Quelle que soit la méthode de classification adoptée pour les galaxies jusqu'à présent, la même philosophie est appliquée. Il n'y a pas de distinction entre nomenclature et description, le nom incluant nécessairement la propriété. En effet, lorsqu'on essaie de définir une galaxie spirale, on butte sur une évidence : une galaxie spirale est une galaxie qui présente une structure spirale. Point. De même pour les galaxies ultralumineuses, les galaxies actives ou les galaxies naines. Dire qu'une galaxie est spirale ne suffit pas à décrire l'ensemble des caractéristiques de cette galaxie, loin s'en faut. Cette confusion entre nom et description perdurera tant qu'on ne prendra pas tous les paramètres en compte pour décrire les classes de galaxies, qu'on ne leur attribuera pas des noms génériques détachés des propriétés et que ces classes ainsi nommées ne seront pas proprement définies. Pour espérer raconter l'histoire des galaxies sans avoir à parler en détail de toutes les observables et leurs processus évolutifs associés, il nous faudra donc inventer une nouvelle taxonomie.

## 10.3 Principes de base

À la lumière de l'histoire de la classification en biologie, les problèmes de la taxonomie actuelle des galaxies sont plus facilement identifiables. Ils proviennent à la fois d'une classification inadaptée mais aussi d'une nomenclature sclérosante. Le premier point peut certainement être résolu par des analyses du type cladistique grâce à des définitions précises et objectives de classes qui peuvent englober différentes entités identifiables sur une phylogénie : clade, groupe évolutif, lignée, ou toute autre notion qui permet de simplifier la description de la diversité des galaxies au cours de l'évolution de l'Univers.

La nomenclature est nécessairement liée à la classification, elle doit être capable d'en désigner les différentes entités. Une classification n'est que le reflet de la méthodologie de regroupement utilisée. L'usage possible d'une classification, et donc son utilité, est donc intimement lié à la méthodologie d'analyse de la diversité. Le diagramme de HUBBLE est certainement très utile pour décrire les différentes morphologies des galaxies, mais elle ne sert à rien pour parler de leur diversité : utiliser la

nomenclature de la classification de HUBBLE pour décrire la formation, l'évolution, et la diversification des galaxies, est nécessairement limitatif et inadapté. Ceci est encore vrai pour toute classification du type catalogue, c'est-à-dire basé sur des critères observationnels en nombre très limité.

Comparons avec la biologie afin d'illustrer cette confusion entre nomenclature et descripteurs. Tout d'abord, voici ce que pourrait être un parallèle de la classification de HUBBLE en biologie :

| | |
|---|---|
| galaxies spirales | animaux ailés |
| galaxies spirales barrées | animaux ailés ovipares |
| galaxies elliptiques | animaux sans membres |

Il est évident qu'aucune description raisonnable de la diversité des animaux n'est possible de cette manière, de même que les galaxies ne sont pas limitées à leur simple forme. Voici maintenant les matrices de descripteurs de trois classes d'objets pris dans chacune des deux disciplines :

| | ailes | oeufs |
|---|---|---|
| oiseaux | oui | oui |
| reptiles | non | oui |
| singes | non | non |

| | barre | bras |
|---|---|---|
| spirales barrées | oui | oui |
| spirales | non | oui |
| elliptiques | non | non |

Les deux matrices sont rigoureusement identiques, mais il y a une grande différence dans la nomenclature. Une galaxie spirale barrée est une galaxie qui a des bras et une barre. Dans cette nomenclature, qui colle à la morphologie, il n'y a que deux descripteurs possibles, toutes les autres (nombreuses) caractéristiques des galaxies en sont automatiquement exclues. Elles sont totalement inutiles puisque, par définition, elles ne caractérisent pas les "classes" morphologiques : ajouter la présence ou l'absence d'un trou noir central imposerait de modifier le nom de la classe, par exemple en "elliptique avec trou noir" et "elliptique sans trou noir". Mais ce faisant, on change la classification, donc la nomemclature en la transformant en une suite de caractéristique. Au contraire, un oiseau est défini par "oeufs, plumes, ailes, vertébré tétrapode à sang chaud, bec sans dent", la matrice pourrait donc être étendue afin de les distinguer davantage parmi la diversité des animaux, notamment des animaux à ailes comme les chauves-souris. Mais le nom "oiseau" et sa définition associée résume à lui tout seul l'ensemble des descripteurs propres à cette classe. Les matrices de descriptions sont donc attachées et sous-entendues.

Chaque entité d'une classification doit comporter trois éléments : un nom, une définition et une description. Le nom permet simplement de désigner la classe et ne doit de préférence pas suggérer une caractéristique dont nous venons d'en voir le danger. À ce nom correspond une définition qui peut être assez générale, comme celle donnée ci-dessus pour l'oiseau, et éventuellement faire référence à un spécimen. Enfin la description énumère toutes les caractéristiques qui permettent d'identifier les individus appartenant à cette classe et qui détaillent l'état de nos connaissances.

Il y a donc deux parties bien distinctes dans l'établissement d'une classification : le regroupement des objets qui va permettre de préciser les définitions et les descriptions des classes, et la nomenclature qui va les nommer selon des règles taxonomiques dépendant en grande partie de la méthode utilisée pour le regroupement.

## 10.4 Une classification adaptée

### 10.4.1 Une histoire évolutive n'est pas une classification

Avant toute chose, une classification doit avoir une utilité, donc un objectif. Que veut-on classer, et pourquoi ? Nous voulons classer les galaxies dans toute leur diversité afin de synthétiser nos connaissances sur ces objets. Il faut donc prendre en compte tous les descripteurs sans exceptions. Nous savons de plus que ces objets évoluent puisque nous observons que leur environnement (l'Univers) évolue et que les galaxies les plus lointaines, donc observées à une époque plus reculée de l'histoire de l'Univers, sont différentes de celles que nous observons autour de nous. La classification des galaxies doit donc impérativement tenir compte de leur histoire évolutive.

L'astrocladistique répond à ces deux conditions à la fois : elle ne choisit pas les descripteurs et inclut l'évolution dans le processus même de l'analyse. Mais une phylogénie, une histoire évolutive, ne constitue pas une classification. Un cladogramme ne fait que regrouper les différents taxons en se basant sur leurs liens de parenté. Il nous offre une visualisation du processus de diversification. La définition des classes est une démarche indépendante qui permet la désignation commode de ces regroupements.

Par exemple, le Phylocode définit l'espèce comme un segment d'une lignée qu'on peut identifier comme différent d'un autre segment selon un critère au choix. Cette définition est précise mais laisse une assez grande latitude et un certain arbitraire dans la désignation des taxons associés. La notion d'espèce issue d'un cladogramme n'est donc pas beaucoup moins floue que les autres notions d'espèce en biologie. Elle a cependant l'avantage de désigner un ensemble de taxons sur l'arbre, sans préjuger de la propriété monophylétique du groupe. La notion de groupe évolutif est un peu plus précise en rassemblant des objets proches dans la diversification sans pour autant être un segment précis d'une lignée.

Ces deux concepts servent surtout à décrire simplement l'arbre. Mais aucune classification évolutive ne saurait se fonder sur des définitions aussi floues et arbitraires. En effet, les taxons d'une espèce donnée n'ont pas tous rigoureusement la même histoire et ceux qui la partagent n'appartiennent pas nécessairement tous à cette espèce. Cependant, les noms servant à la désigner devront respecter les principes de bases de la taxonomie.

En revanche le clade, défini comme un groupe monophylétique, donc incluant un ancêtre et tous ses descendants, représente certainement une entité pertinente pour la diversification des objets, et semble donc parfaitement adapté comme base objective pour une classification. Concrètement un clade est représenté sur un cladogramme par un nœud et toutes les branches filles. Les clades peuvent donc être imbriqués, un clade pouvant contenir d'autres clades. Dans les cas d'hybridation, les clades peuvent même se chevaucher. Il n'y a pas de relation hiérarchique entre les clades et les espèces, une espèce peut appartenir à plusieurs clades.

### 10.4.2 Une classification évolue avec les connaissances

L'analyse cladistique s'appuie sur l'ensemble des descripteurs disponibles à une époque donnée. La phylogénie qui en découle est donc toujours provisoire, toujours susceptible d'évoluer en même temps que les connaissances et les progrès techniques.

Nous avons vu comment une analyse cladistique qui aurait été effectuée à l'époque de HUBBLE peut produire le diagramme en diapason (Sect. 6.1). Mais la diversité des galaxies a pris une autre dimension aujourd'hui, et aussi bien ce diagramme que la classification associée sont à reprendre. En dehors de tout problème de nomenclature, on comprend bien que la définition et la description des classes doit s'adapter. Si on se contente de décrire un oiseau comme un animal ailé, que fait-on le jour où on découvre les chauve-souris ? Si on ajoute les plumes, que faire d'un fossile de dinosaure à plumes et à dents pointues ? HUBBLE ne pouvait pas se douter que la barre n'est pas l'élément le plus discriminant dans la diversité des galaxies spirales.

Si les galaxies avaient été désignées autrement, alors il semble probable que le diagramme de HUBBLE aurait évolué en quelque chose ressemblant à un arbre aux multiples ramifications. Il est frappant de constater que le système de LINNÉ, bien plus vieux que celui de HUBBLE, a survécu à toutes les découvertes et même à l'avènement de la biologie moléculaire. Il s'agit ici de nomenclature, mais comme LINNÉ l'avait bâtie à partir d'un système hiérarchique de classification, la hiérarchie n'ayant jamais été remise en cause, la nomenclature a naturellement accompagné les changements de classification sans la bloquer. La cladistique a légèrement bouleversé l'arbre de la vie, de sorte que quelques contradictions de nomenclature apparaissent. De même que la classification doit évoluer, la nomenclature doit changer de temps en temps, en particulier lorsque la méthode utilisée pour la classification change. D'où les réflexions sur le Phylocode en biologie et les quelques pistes esquissées dans ce chapitre.

L'astrocladistique se présente donc non seulement comme un outil puissant pour synthétiser nos connaissances en cartographiant la diversité des galaxies sous le regard de l'évolution, mais aussi comme une méthodologie naturellement adaptée aux progrès de nos observations et de notre compréhension de ces objets. Cependant, réussir en parallèle à rendre la classification aussi souple qu'un cladogramme n'est certainement pas chose aisée. Car une classification doit être également aussi pérenne que possible afin de conserver la mémoire de nos connaissances acquises petit à petit. Autant la phylogénie peut changer fréquemment, autant la désignation d'un taxon doit être plus stable. C'est bien évidemment la grande force du système linnéen, qui pourtant a peut-être aujourd'hui trouvé ses limites.

## 10.5 Quelques règles possibles de taxonomie des galaxies

La règle adoptée par LINNÉ, stipulant que le nom d'une classe doit être totalement détaché des caractéristiques des objets, a fait largement ses preuves. LINNÉ recommandait d'utiliser des noms communs, des noms de lieu ou de personnes.

Les astrophysiciens connaissent par exemple les galaxies de SEYFERT ou de Markarian. Malheureusement, les définitions associées sont liées à des propriétés observationnelles très restreintes et ne correspondent donc pas à des groupes évolutifs, c'est-à-dire à des entités ayant une place spécifique dans la diversification des galaxies.

Il est également acceptable d'utiliser le nom commun d'un objet typique de la classe. Pour les galaxies cela correspondrait le plus souvent à un nom peu poétique, comme les noyaux actifs dits BLLacs, ou, bien pire, un numéro de catalogue. Là encore la catégorie associée est actuellement définie par des propriétés restreintes.

Enfin, l'invention d'un nom ou l'emprunt d'un mot du vocabulaire général ne sont

pas interdits. Le Phylocode précise tout de même que les noms doivent comprendre uniquement des lettres latines, respecter la grammaire latine, et être prononçables en ayant toujours une voyelle par syllabe. On y voit l'influence de la tradition biologiste avec l'usage du latin, langue officielle des scientifiques au XVIIIème siècle, alliée au soucis d'un langage si possible universel. Si l'anglais est certainement de nos jours la langue de communication en astrophysique, le choix de cette langue ne serait certainement pas très judicieux, le latin ayant incontestablement un caractère plus neutre.

La nomenclature de LINNÉ et les développements ultérieurs ajoutent une terminaison précise aux noms, d'abord pour latiniser même les noms propres, ensuite pour ranger grossièrement un organisme rien qu'à la vue de cette terminaison. C'est sans doute là que réside la puissance de la nomenclature linnéenne, car elle a un côté infiniment pratique : le nom, composé d'un genre puis d'une espèce, chacun étant terminé par une déclinaison spécifique, permet d'un seul coup d'œil de connaître le grand type d'organisme dont il s'agit. Seulement, cette nomenclature riche ne fonctionne que dans le cadre de la méthode de classification utilisée, en l'occurrence une organisation hiérarchique basée sur la phénétique ou analyse de distance multivariée.

En bouleversant quelque peu cette organisation hiérarchique, la cladistique a introduit un décalage entre la nomenclature et la structure de l'arbre. Puisque la cladistique elle-même s'attend à ce que les connaissances futures puissent chambouler encore cette organisation, le Phylocode a décidé de renoncer à ce que le nom d'une espèce ou d'une clade traduise ses relations de parenté. C'est certainement une perte de simplicité, mais il est ennuyeux de voir plusieurs espèces d'un même genre éparpillées sur plusieurs branches parmi des espèces d'autres genres. C'est exactement ce qui se passe par exemple pour les galaxies elliptiques ou spirales, ou encore les galaxies de SEYFERT, qui ne sont plus regroupées sur les arbres issues d'analyses astrocladistiques.

Le Phylocode est extrêmement précis pour la nomenclature des clades qui sont au bout du compte les seules entités concernés par la classification. En effet, dans l'esprit d'ADANSON ((ADANSON, 1763), elles correspondent à un regroupement naturel lié à l'histoire évolutive des taxons. Par exemple, il est ainsi stipulé que le nom d'un clade doit comporter un seul mot, commençant par une majuscule, accompagné par une définition phylogénétique, en anglais ou en latin. Cette définition est résolument basée sur un cladogramme, elle implique donc une hypothèse phylogénétique. Pour que l'existence de ce clade soit reconnu, et que son nom prenne une importance dans la classification, il est donc évident que de nombreuses analyses indépendantes doivent avoir été effectuées.

Des définitions possibles des clades sont basées sur :

— les nœuds de l'arbre : le clade comprend le nœud le moins inclusif contenant à la fois les taxons A et B et tous les descendants de ce nœud ;
— les branches : le clade comprend le nœud le plus inclusif ne contenant pas le taxon Z, et contenant tous les descendants de ce nœud ;
— des apomorphies : le clade comprend tous les taxons ayant hérité du taxon A l'état de caractère dérivé ;
— des combinaisons des définitions précédentes.

Cette liste n'est absolument pas exhaustive. Le Phylocode reste encore un projet, mais il peut être précieux pour les astrophysiciens de s'en inspirer pour élaborer une classification basée sur l'astrocladistique ou toute autre approche similaire afin de renouveler la classification désormais inadaptée des galaxies selon leur seul type

morphologique.

Nous n'irons pas plus loin dans cet ouvrage. Le chemin est encore long, mais il semble aujourd'hui nécessaire de s'y engager, sur des bases solides.

# Chapitre 11

# Conclusion

J'espère que le lecteur aura suivi le fil logique qui mène à la nécessaire refonte de notre manière de concevoir la diversité des galaxies. Les rudiments de cladistique ébauchés dans ce livre sont en principe suffisants pour comprendre comment il est possible d'aborder nos connaissances actuelles sur l'Univers avec un regard neuf et rigoureux. Et les explications concernant l'application concrète aux galaxies ont pour objectif de motiver le lecteur aux efforts indispensables pour s'imprégner de l'approche.

Car plusieurs obstacles sont à franchir de la part de l'astrophysicien curieux. Le côté technique, le jargon et la philosophie de la cladistique ne sont encore pas culturellement intuitifs pour l'être humain. L'interprétation d'un cladogramme n'est pas immédiate, elle nécessite d'apprendre à lire un arbre. Obtenir un cladogramme de galaxies est déjà une étape importante. Mais le comprendre et en exploiter toute sa richesse demande encore beaucoup d'efforts. La grande nouveauté vient également de la nécessité de devoir raisonner statistiquement dans un espace à plusieurs dimensions, chose à laquelle les physiciens ne sont pas formés. Il y a donc un travail conséquent à fournir que le présent ouvrage espère faciliter.

Pour en arriver là, il ne faut pas que l'astrophysicien soit bloqué par l'idée fausse que cette approche ne peut s'appliquer qu'aux organismes vivants et qu'elle n'aura jamais aucune utilité en astrophysique. C'est malheureusement une réaction souvent rencontrée qui ne révèle pas d'un esprit scientifique très honnête. La cladistique est une méthodologie statistique et mathématique, au même titre que l'analyse de distance multivariée, la décomposition en composantes principales, ou l'analyse de Fourier, bien connues des astrophysiciens et des biologistes, entre autres.

Il y a enfin un obstacle culturel spécifique à l'étude des galaxies, qui consiste au renoncement à l'héritage de HUBBLE et à quatre-vingts ans de travaux de la part de la quasi-totalité de la communauté autour du diagramme en diapason et ses types morphologiques. Démarche difficile, d'autant plus qu'elle est étroitement associée à la nécessité de renoncer à une certaine simplicité, simplicité de la classification de HUBBLE, mais aussi simplicité de la vision des galaxies qu'elle implique.

Nos observations nous poussent inéluctablement vers une telle révolution culturelle, et nombre de mes collègues le savent très bien, au point que certains ont déjà cherché des pistes possibles. L'astrocladistique est la première approche radicalement nouvelle et la plus avancée à ce jour. Sa véritable percée réside dans le fait que non seulement elle appréhende la diversité des galaxies, mais elle étudie directement la

diversification des galaxies, englobant ainsi tous les problèmes de synthèse des observations, de formation et d'évolution des galaxies, ainsi que de classification.

Les premiers résultats montrent que l'évolution des galaxies semble se faire essentiellement par embranchement, avec relativement peu d'homoplasies. Même si l'exploration n'est pas terminée, cela suffit grandement à justifier les investissements déjà mis en œuvre dans cette voie radicalement nouvelle et à encourager l'élargissement du champ d'exploration autour des méthodes phylogénétiques en général. Cela aide également à poser de nouvelles questions, sous un jour inédit, et à clarifier des notions, qui pourraient parfois sembler triviales, comme il a été fait dans ce livre. C'est ainsi que notre compréhension du monde peut progresser.

La cladistique est donc une approche qui aujourd'hui semble adaptée à l'étude de la diversification des galaxies. Néanmoins, il n'est pas dit qu'elle restera définitivement la meilleure. Son intérêt majeur réside d'abord dans le renouvellement de notre conception du monde des galaxies, ouvrant ainsi un nouveau et vaste champ de recherche. Il apporte tous les outils développés par ailleurs pour étudier des populations, c'est-à-dire tout ensemble complexe d'individus complexes interagissant et évoluant. Dans cet ouvrage, nous nous sommes limités à ce qui fonctionne aujourd'hui autour de l'astrocladistique, à ce qui a déjà été utilisé avec succès. De nombreuses pistes restent à explorer.

Le chemin vers une meilleure compréhension de la diversification des galaxies et une manière adéquate de la raconter, à travers une classification totalement renouvelée, est en encore très long. Il nécessitera beaucoup de réflexions, de débats, d'échecs et, espérons-le, de réussites. L'astrocladistique, présentée dans cet ouvrage, est déjà un premier succès qui ouvre désormais la voie. L'objectif est immense, puisqu'il s'agit ni plus ni moins de raconter l'histoire de l'Univers et la vie des galaxies, véritable révolution de notre vision du Monde qui ne fait qu'accompagner l'explosion de nos connaissances extragalactiques.

# Bibliographie


ADANSON, M., 1763. Famille Des Plantes. chez Vincent, impr.-libraire de Mgr le Comte de Provence (Paris).

BAUGH, C.M., CROTON, D.J., GAZTAÑAGA, E.E.A., 2004. The 2df galaxy redshift survey : Hierarchical galaxy clustering. *Monthly Notices of the Royal Astronomical Society* 351, L44–L48. arXiv:astro-ph/0401405.

BENTON, M.J., 2007. The phyoclode : Beating a dead horse ? *Acta Palaeontologica Polonica* 52, 651–655.

BROWER, A. V.Z., D.R., VOGLER, A., 1996. Gene trees, species trees, and systematics : A cladistic perspective. *Annu. Rev. Ecol. Syst* 27, 423–450.

CABANAC, R.A., DE LAPPARENT, A., HICKSON, P., 2002. *Astronomy & Astrophysics* 389, 1090–1116.

CANTINO, P., DE QUEIROZ, K., 2010. Phylocode, international code of phylogenetic nomenclature.

CHATTOPADHYAY, A., CHATTOPADHYAY, T., DAVOUST, E., MONDAL, S., SHARINA, M., 2009a. Study of ngc 5128 globular clusters under multivariate statistical paradigm. *Astrophysical Journal* 705, 1533. arXiv:0909.4161.

CHATTOPADHYAY, T., BABU, J., CHATTOPADHYAY, A., MONDAL, S., 2009b. Horizontal branch morphology of globular clusters : A multivariate statistical analysis. *Astrophysical Journal* 700, 1768.

CHATTOPADHYAY, T., CHATTOPADHYAY, A., 2006. Objective classification of spiral galaxies having extended rotation curves beyond the optical radius. *The Astronomical Journal* 131, 2452–2468.

CHATTOPADHYAY, T., CHATTOPADHYAY, A., 2007. Globular clusters of local group – statistical analysis. *Astronomy & Astrophysics* 472, 131–140.

CHATTOPADHYAY, T., MISRA, R., NASKAR, M., CHATTOPADHYAY, A., 2007. Statistical evidences of three classes of gamma ray bursts. *Astrophysical Journal* 667, 1017. arXiv:0705.4020.

CHATTOPADHYAY, T., MONDAL, S., CHATTOPADHYAY, A., 2008. Globular clusters in the milky way and dwarf galaxies - a distribution-free statistical comparison. *Astrophysical Journal* 683, 172.

CONNOLLY, A.J., SZALAY, A.S., BERSHADY, M.A., KINNEY, A.L., CALZETTI, D., 1995. Spectral classification of galaxies : an orthogonal approach. *Astronomical Journal* 110, 1071. astro-ph/9411044.

CREEVEY, C.J., MCINERNEY, J.O., 2005. Bioinformatics 21 (3), 390.

CROFT, W., 2008. Evolutionary linguistics. *Annual Review of Anthropology* 37, 219–234.

DARLU, P., TASSY, P., 1993. La reconstruction phylogénétique : concepts et méthodes /.

DARWIN, C., 1859. The Origin of Species. John Murray, London.

DEGNAN, J.H., ROSENBERG, N.A., 2006. Discordance of species trees with their most likely gene trees. *PLoS Genet* 2, e68.

DISNEY, M.J., ROMANO, J.D., GARCIA-APPADOO, D.A., WEST, A.A., DALCANTON, J.J., CORTESE, L., 2008. Galaxies appear simpler than expected. *Nature* 455, 1082–1084. http://arxiv.org/abs/0811.1554.

DUTIL, Y., 2001. *Astrophysics and Space Science* 277 (Suppl.), 165–168.

ELLIS, S.C., DRIVER, S.P., ALLEN, P.D., LISKE, J.B BLAND-HAWTHORN, J., DE PROPRIS, R., 2005. The millennium galaxy catalogue : on the natural subdivision of galaxies. *Monthly Notices of the Royal Astronomical Society* 363, 1257–1271. arXiv:astro-ph/0508365.

FRAIX-BURNET, D., 2006. Determining the evolutionary history of galaxies by astrocladistics : some results on close galaxies, in : D. Barret, F. Casoli, S.C.F.C.T.C., Pagani, L. (Eds.), Journées de la SF2A, Paris (France), Société Française d'Astronomie et d'Astrophysique (SF2A). http://hal.archives-ouvertes.fr/ccsd-00104352.

FRAIX-BURNET, D., 2009. Evolutionary Biology Concept, Modeling, and Application. Springer Berlin Heidelberg. chapter Galaxies and Cladistics. Biomedical and Life Sciences, pp. 363–378. arXiv:0909.4164.

FRAIX-BURNET, D., CHOLER, P., DOUZERY, E., 2006a. Towards a Phylogenetic Analysis of Galaxy Evolution : a Case Study with the Dwarf Galaxies of the Local Group. *Astronomy and Astrophysics* 455, 845–851. astro-ph/0605221.

FRAIX-BURNET, D., CHOLER, P., DOUZERY, E., VERHAMME, A., 2006b. Astrocladistics : a phylogenetic analysis of galaxy evolution I. Character evolutions and galaxy histories. *Journal of Classification* 23, 31–56. astro-ph/0602581.

FRAIX-BURNET, D., DAVOUST, E., CHARBONNEL, C., 2009. The environment of formation as a second parameter for globular cluster classification. *MNRAS* 398, 1706–1714. arXiv:0906.3458.

FRAIX-BURNET, D., DOUZERY, E., CHOLER, P., VERHAMME, A., 2006c. Astrocladistics : a phylogenetic analysis of galaxy evolution II. Formation and diversification of galaxies. *Journal of Classification* 23, 57–78. astro-ph/0602580.

FRAIX-BURNET, D., DUGUÉ, M., CHATTOPADHYAY, A., CHATTOPADHYAY, T., DAVOUST, E., 2010. Structures in the fundamental plane of early-type galaxies. *Monthly Notices of the Royal Astronomical Society* accepted for publication. http://fr.arxiv.org/abs/1005.5645.

GOLOBOFF, P.A., MATTONI, C.I., QUINTEROS, A.S., 2006. Continuous characters analyzed as such. *Cladistics* 22, 589–601.

GUTH, A., 2001. Eternal inflation, in : Miller, J.B. (Ed.), Cosmic Questions xii. The New York Academy of Sciences, 14-16 April 1999 in Washington, D.C.,, p. 66. arXiv:astro-ph/0101507.

HENNIG, W., 1965. Phylogenetic systematics. *Annual Review of Entomology* 10, 97–116.

HOLMES, S., 2003. Bootstrapping phylogenetic trees : Theory and methods. *Statistical Science* 18, 241–255.

HUBBLE, E.P., 1922. A general study of diffusive galactic nebulae. *Astrophysical Journal* 56, 162–199.

HUBBLE, E.P., 1936. The Realm of Nebulae. New Haven :Yale Univ. Press.

KASTELLE, T., 2005. A classification method for evolutionary economics, in : 4th European Meeting on Applied Evolutionary Economics, 19-21 May 2005.

KROUPA, P., 2002. The initial mass function and its variation, in : Grebel, E., Brandner, W. (Eds.), Modes of Star Formation and the Origin of Field Star Populations. volume 285 of *ASP Conference Series*.

LIPSCOMB, D., 1998. Basics of Cladistics Analysis.

LU, H., ZHOU, H., WANG, J., WANG, T., DONG, X., ZHUANG, Z., LI, C., 2006. Ensemble learning independent component analysis of normal galaxy spectra. *The Astronomical Journal* 131, 790–805. astro-ph/0510246.

MADDISON, W.P., MADDISON, D.R., 2004. Mesquite : a modular system for evolutionary analysis.

MAKARENKOV, V., KEVORKOV, D., LEGENDRE, P., 2006. Phylogenetic network construction approaches, in : Arora, D.K., Berka, R., Singh, G.B. (Eds.), Bioinformatics. Elsevier. volume 6 of *Applied Mycology and Biotechnology*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.86.5285&rep=rep1&type=pdf.

NAKHLEH, L., RUTHS, D., INNAN, H., 2006. Meta-analysis and Combining Information in Genetics and Genomics. Chapman & Hall / CRC. chapter Gene trees, species trees, and species networks. 1 edition. pp. 1–27. http://www.phylo.org/pdf_docs/40_(42)Nakhleh_NakhlehRuthsInnan.pdf.

NICHOLS, R., 2001. Gene trees and species trees are not the same. *Trends in Ecology & Evolution* 16, 358 – 364.

NIXON, K.C., CARPENTER, J.M., STEVENSON, D.W., 2003. The phylocode is fatally flawed, and the "linnaean" system can easily be fixed. *The Botanical Review* 69, 111–120.

POLLARD, D.A., IYER, V.N., MOSES, A.M., EISEN, M.B., 2006. Widespread discordance of gene trees with species tree in drosophila : Evidence for incomplete lineage sorting. *PLoS Genet* 2, e173.

ROBERTS, M.S., HAYNES, M.P., 1994. Physical parameters along the hubble sequence. *Annual Review of Astronomy and Astrophysics* 32, 115–152.

ROBINSON, P.M.W., ROBERT, J.O., 1996. *Research in Humanities Computing* 4, 115.

SÁNCHEZ ALMEIDA, J., AGUERRI, J.A.L., MUÑOZ-TUÑÓN, C., DE VICENTE, A., 2010. Automatic unsupervised classification of all sloan digital sky survey data release 7 galaxy spectra. *ApJ* 714, 487–504. http://arxiv.org/abs/1003.3186.

SEMPLE, C., STEEL, M.A., 2003. Phylogenetics. Oxford University Press.

SPERGEL, D.N., BEAN, R., DORE', O.E.A., 2006. Wilkinson microwave anisotropy probe (wmap) three year results : Implications for cosmology. *Astrophysical Journal* . arXiv:astro-ph/0603449.

SWOFFORD, D.L., 2003. Paup* : Phylogenetic analysis using parsimony (*and other methods).

TREU, T., STIAVELLI, M., BERTIN, G., CASERTANO, S., MØLLER, P., 2001. *Monthly Notices of the Royal Astronomical Society* 326, 237–254.

TRIAY, R., 2005. A solution to the cosmological constant problem. *International Journal of Modern Physics D* 14, 1667. arXiv:gr-qc/0510088,http://hal.ccsd.cnrs.fr/ccsd-00012304.

VAN DEN BERGH, S., 1998. Galaxy Morphology and Classification. Cambridge University Press.

VIGNAIS, P., 2001. La Biologie Des Origine à nos jours. EDP Sciences.

WATANABE, M., KODAIRA, K., OKAMURA, S., 1985. Digital surface photometry of galaxies toward a quantitative classification. iv - principal component analysis of surface-photometric parameters. *Astrophysical Journal* 292, 72–78.

WELLS, R.S., 1987. in : Hoenigswald, H.M. & Wiener, L.P. (Ed.), Biological Metaphor and Cladistic Classification : An Interdisciplinary Perspective, p. 39.

WHITMORE, B.C., 1984. An objective classification system for spiral galaxies. i the two dominant dimensions. *Astrophysical Journal* 278, 61–80.

WILEY, E., SIEGEL-CAUSEY, D., BROOKS, D., FUNK, V., 1991. The Compleat Cladist : A Primer of Phylogenetic Procedures. The University of Kansas, Museum of Natural History, Special Publication No. 19.

WOESE, C.R., 2000. Interpreting the universal phylogenetic tree. *Proceedings of the National Academy of Science* 97, 8392–8396.

# Glossaire

**analogie**
: propriété de structures qui ont la même fonction ou qui sont similaires mais ont des origines différentes.

**ancêtre**
: espèce hypothétique à l'origine des innovations (états dérivés) caractérisant la lignée.

**arbre déséquilibré ou linéaire**
: arbre régulier ne présentant pas ou très peu de sous-ensembles de branches, c'est-à-dire ne définissant qu'une seule lignée.

**arbre enraciné**
: arbre dont le sens de la diversification a été polarisé grâce à une racine définie par un extragroupe (ou "outgroup").

**arbre résolu**
: un arbre résolu est un arbre dans lequel de chaque nœud partent deux branches et deux seulement. Il ne comprend donc que des bifurcations et aucune polytomie. Un arbre (totalement) non résolu est dit arbre en étoile ("star-like" ou "star tree").

**autapomorphie**
: caractéristique unique à un seul taxon. Les autapomorphies ne sont pas utiles en phylogénie.

**caractère**
: descripteur ou propriété pouvant caractériser différents états évolutifs d'un taxon.

**clade**
: mot venant du grec *klados* qui signifie "rameau". Un clade comprend l'ancêtre commun et tous les groupes ou lignées issus de l'ancêtre commun. Un clade est un groupe monophylétique.

**cladisme ou systématique phylogénétique**
: classification systématique des organismes vivants fondée sur les relations phylogénétiques. Une analyse cladistique étudie le sens des transformations évolutives de caractères.

**état évolutif**

attribut discret ou continu, qualitatif ou quantitatif, décrivant une étape particulière dans la tranformation évolutive d'un caractère donné.

**extragroupe ou outgroup ou groupe de comparaison**
taxon en principe extérieur à l'échantillon étudié ("ingroup") et partageant avec lui un même ancêtre commun. Permet de définir une racine à l'arbre donc de préciser le sens de la diversification.

**homologie**
similitude par ancêtre commun. Propriété de structures qui ont la même origine mais que peuvent avoir des fonctions différentes.

**homoplasie**
caractéristique apparaissant indépendamment dans plusieurs lignées d'ancêtres différents. Regroupe les convergences, les évolutions parallèles ainsi que les régressions.

**intraspécifique**
propre à une espèce.

**lignée**
ensemble de taxons issus d'un même ancêtre. Souvent représentée par une seule branche sur les arbres, une lignée n'est pas un groupe monophylétique car elle n'inclut pas l'ancêtre ni tous les descendants appartenant aux autres lignées issues du même ancêtre.

**monophylétique**
caractérise un regroupement de taxons ayant tous le même ancêtre commun et incluant cet ancêtre.

**paraphylétique**
caractérise un regroupement d'une partie seulement des taxons ayant le même ancêtre commun.

**parcimonie**
principe d'optimisation pour le choix de l'arbre parmi tous les arrangements possibles, rejoignant un principe général en Sciences connu sous le nom de rasoir d'OCKHAM. La méthode du maximum de parcimonie sélectionne l'arbre qui minimise le nombre total de changements des états des caractères (nombre de pas). Cet arbre le plus parcimonieux représente la phylogénie la plus simple, ou encore le schéma évolutif le plus simple,

**pas**
nombre total de changements des états des caractères sur un arbre. C'est ce nombre qui est minimisé dans les méthodes de maximum de parcimonie.

**phénétique**
méthode de classification basée sur l'analyse de distance multivariée, donc sur la similitude globale.

**phylogenèse**

histoire évolutive (-genèse) des espèces (-phylo). Par "genèse" on entend formation au sens large, pas uniquement la formation initiale (primitive) car la formation des espèces actuelles implique l'évolution. La phylogenèse retrace l'histoire de la diversification.

**polyphylétique**

caractérise un regroupement de taxons plus ou moins similaires mais d'ancêtres différents.

**réticulation**

schéma de diversification provoqué par les hybridation ou les tranferts horizontaux de gènes. La phylogénie ne se représente plus sous la forme d'un arbre mais d'un réseau dit réticulogramme ou "split network".

**supraspécifique**

se dit d'un taxon représentant un clade ou un groupe évolutif.

**synapomorphie**

caractéristique nouvelle (innovation) et distinctive partagée par un groupe d'organismes et qui en définit le clade ou le groupe évolutif.

**systématique**

science se donnant pour objectif l'étude et la description de la diversité des êtres vivants, la recherche de la nature et des causes de leurs différences et de leurs ressemblances, la mise en évidence de relations de parenté existant entre eux et l'élaboration de classifications traduisant ces relations de parenté. La cladistique est aussi appelée systématique phylogénétique.

**taxon**

objet qu'on cherche à insérer dans une phylogénie, constituant une feuille (bout de branche) du cladogramme. Un taxon peut être une espèce ou un individu ou encore une sous-espèce, un groupe, etc.

# Index