Les méthodes d'analyse des données

Xavier Perrier, Albert Flori, François Bonnot

L'analyse des données de diversité revient, le plus souvent, à l'analyse de tableaux croisant des unités taxonomiques — espèces, populations, cultivars... — et des variables de caractérisation de la diversité. Ces variables sont de nature diverse : caractères agronomiques quantitatifs, descripteurs morphologiques, le plus souvent qualitatifs, marqueurs biochimiques (isoenzymes) ou moléculaires (RFLP, RAPD...), souvent codés en présence ou absence. Les unités taxonomiques sont en général observées individuellement, mais elles peuvent aussi correspondre à des populations, les observations sont alors des moyennes de caractères, des fréquences alléliques... Dans tous les cas, l'analyse de ces données a pour objectif de repérer une éventuelle structuration des unités taxonomiques par l'analyse des ressemblances ou des dissemblances entre ces unités.

Cet objectif relève d'une action de classification, terme vague qui reçoit dans le langage courant deux acceptions différentes. La première concerne l'action de regrouper les objets en ensembles homogènes. Cette ambition a été celle de tous les grands systématiciens, d'Aristote à Linné. La seconde acception relève de l'action d'ordonner logiquement. On s'intéresse alors essentiellement aux relations d'ordre entre les unités taxonomiques, c'est-à-dire aux relations de filiation entre unités au sens de l'évolution, qui relèvent de la phylogénétique.

Ces deux démarches peuvent aboutir à des représentations graphiques sous forme d'arbre. Leur contenu informatif est cependant fondamentalement différent. Dans le premier cas, les nœuds de l'arbre correspondent à des concepts de regroupement. Dans le second, les nœuds sont des unités parentales non observées mais supposées existantes, par exemple des formes ancestrales disparues. Les liens incluent alors la dimension temporelle de l'évolution.

Le problème de la diversité traité ici relève en partie des deux démarches puisque l'on souhaite définir des groupes d'unités comparables, mais aussi décrire les relations de parenté entre ces groupes.

Deux approches peuvent être suivies pour traiter de l'organisation évolutive. La première, dite phénétique, décrit une organisation à partir de mesures objectives de dissemblance entre les unités et reporte à l'interprétation l'introduction d'hypothèses génétiques. Ces dissimilarités sont estimées globalement sur l'ensemble des caractères, c'est la multiplicité des caractères observés qui est censée donner une pertinence à ces mesures. Il est, le plus souvent, extrêmement difficile de détecter une organisation des unités par l'examen direct des dissimilarités. L'objet de l'analyse est alors de chercher un espace de représentation approchée mais facilement lisible de ces dissimilarités : plan factoriel ou structure arborée. Le gain de lisibilité se paie au prix d'une certaine perte d'information, perte que l'on cherchera bien sûr à minimiser.

La seconde approche, dite cladistique, s'intéresse directement aux caractères observés et tente de distinguer les caractères identiques hérités d'un ancêtre commun, qui sont seuls informatifs, des phénomènes de convergence accidentelle. Elle repose sur un modèle de transformation génétique et a été établie initialement sur des caractères morphologiques que l'on savait monogéniques et dont on savait ordonner les différentes modalités en termes d'évolution. On l'applique maintenant le plus souvent à des données de séquençage du génome munies d'un modèle génétique explicite, du type de celui de JUKES et CANTOR (1969) ou dérivé.

Dans le cadre de cet ouvrage, les méthodes cladistiques ou apparentées, comme les méthodes de compatibilité ou de vraisemblance, ne seront pas évoquées, on se référera, par exemple, à DARLU et TASSY (1993). En effet, la nature des marqueurs utilisés ne permet le plus souvent que des approches phénétiques relevant de l'analyse des dissimilarités.

L'objet de ce chapitre est de présenter divers aspects méthodologiques relevant de ces approches. Les données de base seront constituées d'un tableau d'unités taxonomiques décrites par des marqueurs de la diversité. Il pourra aussi s'agir d'un empilement de tableaux lorsqu'un même ensemble d'unités est caractérisé par plusieurs séries de marqueurs de nature différente que l'on ne peut regrouper. Ces tableaux pourront être considérés séparément, mais on cherchera finalement à les analyser simultanément.

Ce chapitre est composé de trois grandes parties.

La première concerne l'étape initiale de toute analyse qui est la définition d'une mesure de la ressemblance ou de la dissemblance entre individus. Le choix de cette mesure est de toute première importance et répond au choix délibéré d'un point de vue sur les données. Lorsque les individus sont des populations et les observations des fréquences alléliques, diverses mesures de dissimilarité, dites génétiques, sont utilisables — distances de Nei, de Gregorius... Leurs propriétés ont été souvent étudiées, on se référera, par exemple, à LEFORT-BUSSON et DE VIENNE (1985). Quand les données sont des individus, décrits par une série de caractères, la nature des variables — quantitatives, qualitatives, binaires — conditionne les types de dissimilarité utilisables. On montrera, sur l'exemple des marqueurs moléculaires de type RFLP ou RAPD, que la connaissance de caractéristiques biologiques des marqueurs permet de justifier le choix d'un indice de dissimilarité.

La deuxième partie traite des méthodes factorielles. Une matrice de dissimilarité entre individus deux à deux étant établie, son analyse a pour objet d'en rechercher une représentation simplifiée mais fidèle. Si la dissimilarité est de nature euclidienne, les méthodes d'analyse factorielle peuvent être utilisées. Les bases des principales méthodes seront rappelées, on s'intéressera en particulier aux techniques d'analyse simultanée de plusieurs tableaux. Si la dissimilarité n'est pas euclidienne, les méthodes de positionnement multidimensionnel *(multidimensional scaling)* permettent de rechercher une décomposition sur un nombre fixé d'axes. Ces méthodes itératives demandent des temps de calcul important lorsque le nombre d'unités à traiter est élevé. Ces méthodes ne seront pas abordées dans le cadre de cet ouvrage, on pourra se référer, par exemple, à ESCOUFIER (1975).

La troisième partie s'intéresse aux représentations arborées. Un processus évolutif, par accumulation de mutations transmissibles, produirait une organisation représentable sous la forme d'un arbre. Cette représentation arborée, par son fondement biologique, est particulièrement pertinente pour l'analyse de la diversité, même si les modalités de l'évolution sont en général plus complexes. Diverses méthodes de création d'arbres seront présentées, ainsi que les techniques qui permettent de construire des arbres synthétiques à partir de plusieurs jeux de marqueurs.

Les méthodes factorielles et les méthodes de représentation arborée constituent deux approches très différentes de la représentation de la diversité. Elles doivent être considérées comme complémentaires plutôt que comme concurrentes. Les méthodes factorielles recherchent avant tout une représentation globale de la diversité, débarrassée autant que possible des effets individuels. A l'inverse, les méthodes arborées tendent, pour les plus utilisées, à représenter fidèlement les relations individuelles. Il s'agit donc bien de deux points de vue différents sur les données.

De nombreux logiciels statistiques proposent les diverses méthodes d'analyse factorielle. Les outils liés à l'analyse des dissimilarités et à leur représentation arborée sont moins fréquents, le logiciel Ntsys est sans doute le plus complet et le plus accessible (ROHLF, 1987). Diverses fonctions non disponibles ou des méthodes en partie remaniées ont été réunies dans un logiciel spécialisé, Darwin *(dissimilarity analysis and representation),* disponible auprès des auteurs (PERRIER, 1998).

Le choix d'un indice de dissimilarité

Il s'agit de définir une mesure de la ressemblance (similarité) entre deux individus ou plutôt, de façon équivalente, de la dissemblance (dissimilarité) puisque l'on peut passer simplement de l'une à l'autre par une transformation linéaire.

Il existe, en fait, un grand nombre de mesures de dissemblance. En fonction de la nature des données, divers choix sont possibles. Certaines de ces dissimilarités ont des propriétés mathématiques intéressantes pour l'analyse ou possèdent des capacités de stabilité face à des erreurs de données, qui constituent des arguments pertinents pour les retenir. Le choix final reste relatif au réel contenu informatif des marqueurs utilisés et au point de vue que l'on souhaite adopter sur cette information.

Enfin, certaines dissimilarités, les ultramétriques et les distances additives, ont des propriétés remarquables de représentabilité sous forme d'arbre. Il n'y a, bien sûr, aucune raison pour que la mesure établie sur les données possède de telles propriétés. L'objet des méthodes de construction d'arbre est de transformer, en la déformant le moins possible, la mesure établie sur les données en une dissimilarité possédant des propriétés de représentabilité.

Définitions et propriétés

L'objet mathématique le plus général traduisant la différence est une dissimilarité, c'est une fonction d de l'ensemble des paires (i,j) d'individus dans l'ensemble des réels positifs ou nuls, symétrique ($d(i,j) = d(j,i)$) et telle que $d(i,i) = 0$ pour tout i. Cette définition est peu contraignante et couvre un grand nombre de mesures possibles. En revanche, elle n'ouvre que peu de propriétés mathématiques et il est nécessaire d'ajouter d'autres contraintes pour acquérir certaines propriétés intéressantes.

Une distance, on dit parfois une métrique, est une dissimilarité particulière obtenue en ajoutant la condition $d(i,j) = 0 \Rightarrow i = j$ et surtout l'inégalité triangulaire entre trois individus, $d(i,j) \leq d(i,k) + d(j,k)$. Cette condition, naturelle, tra-

duit simplement la possibilité de représenter tout triplet de points par un triangle dans le plan. Cette propriété, très intéressante, permet, en particulier, d'éviter le problème de longueurs d'arêtes négatives dans les méthodes de construction d'arbre.

Distance de Manhattan et distance euclidienne

Parmi les distances, une famille importante est constituée des distances dites de Minkowski d'ordre p ($p \geq 1$). Une distance appartient à cette famille s'il existe un entier K et une série de K valeurs x_{ik} affectables à chaque individu i, la distance s'écrivant alors $d(i,j) = (\sum_k | x_{ik} - x_{jk} |^p)^{1/p}$.

Les seuls cas utilisés en pratique correspondent aux valeurs 1 et 2 de p. Quand $p = 1$, l'indice est connu sous le nom de distance de Manhattan ou *city-block*, $d(i,j) = \sum_k | x_{ik} - x_{jk} |$. Pour $p = 2$, on retrouve l'habituelle distance euclidienne, $d(i,j) = [\sum_k (x_{ik} - x_{jk})^2]^{1/2}$. Bien sûr, des distances de Manhattan ou euclidiennes sont obtenues, par construction même, lorsque, à partir d'un tableau individus × variables, on calcule une distance en sommant, sur les variables, les valeurs absolues ou les carrés des écarts. Cependant, certains indices, par exemple ceux qui sont calculés sur des données de présence ou d'absence, peuvent être de l'un de ces deux types sans que cela découle de manière évidente du mode de construction.

Distance de Manhattan et distance euclidienne appartiennent à la même famille et sont liées par certaines relations. On sait montrer, par exemple, que toute distance euclidienne est une distance de Manhattan. On montre aussi que la racine carrée d'une distance de Manhattan est une distance euclidienne.

Transformation puissance

Les transformations puissance d'une dissimilarité permettent, pour certaines puissances, de lui donner les propriétés d'une distance ou même d'une distance euclidienne. En préalable, il faut rappeler que deux dissimilarités d et d' sont équivalentes en ordre si $d(i,j) \leq d(k,l) \Leftrightarrow d'(i,j) \leq d'(k,l)$. Ces deux dissimilarités ordonnent les paires d'individus de la même façon et sont donc d'interprétation comparable. En particulier, si d' peut s'écrire comme une fonction monotone croissante de d alors d' et d sont équivalentes en ordre. C'est le cas des fonctions puissance α ($\alpha \geq 0$), d^α et d sont équivalentes en ordre.

On sait montrer qu'il est toujours possible de trouver une valeur α comprise entre 0 et 1 telle que, si d est une dissimilarité, alors pour tout λ ($0 \leq \lambda \leq \alpha$), d^λ est une distance qui, par la propriété précédente, est équivalente en ordre avec d. Donc, si le contenu informatif des marqueurs amène à choisir une mesure qui n'est qu'une dissimilarité, on sait par une transformation puissance lui conférer les propriétés d'une distance dont on a signalé l'intérêt pour les méthodes de classification.

De même, on démontre que si d est une distance, on peut toujours trouver une valeur α comprise entre 0 et 1 telle que, pour tout λ ($0 \leq \lambda \leq \alpha$), d^{λ} est une distance euclidienne équivalente en ordre avec d. Une transformation puissance permet donc de passer d'une distance à une distance euclidienne, comme on passe d'une dissimilarité à une distance. Une application intéressante concerne les méthodes d'analyse factorielle. Bien qu'elles ne traitent que des distances euclidiennes, on pourra cependant utiliser ces méthodes à partir d'une dissimilarité quelconque après transformation puissance préalable. Cette transformation paraît préférable à la technique habituelle d'ajout d'une constante, éventuellement très grande, à toutes les valeurs des distances, technique moins respectueuse des données initiales.

Dans certains cas, on connaît les valeurs α de ces deux transformations, par exemple, il a déjà été signalé que la racine carrée ($\alpha = 1/2$) d'une distance de Manhattan est une distance euclidienne. Souvent on ne sait pas établir cette valeur qui doit être estimée numériquement sur les données.

Ultramétrique et distance additive d'arbre

En ajoutant des contraintes supplémentaires à la définition d'une dissimilarité, on peut lui conférer la propriété d'être une distance représentable sous forme d'un arbre. Les méthodes de classification ont toutes pour objectif d'approcher au mieux, au sens d'un certain critère, la dissimilarité observée par une de ces distances représentables.

La plus connue de ces dissimilarités est l'ultramétrique, c'est une distance vérifiant l'inégalité $d(i,j) \leq \max[d(i,k), d(j,k)]$, pour trois individus quelconques i, j et k. Cette propriété exprime que, parmi les trois distances, les deux plus grandes sont égales : si $d(i,j)$ est la plus petite, alors $d(i,j) \leq d(i,k) = d(j,k)$. Tout triplet de points forme donc un triangle isocèle « pointu ». Cette propriété permet une représentation sous la forme habituelle d'un dendrogramme (figure 1a), qui est un arbre possédant un point particulier, la racine, située à égale distance de toutes les feuilles de l'arbre. On montre qu'une ultramétrique est une distance euclidienne.

Dans les années 70, la distance dite additive d'arbre a été proposée. Elle vérifie l'inégalité $d(i,j) + d(k,l) \leq \max[d(i,k) + d(j,l), d(i,l) + d(j,k)]$ pour quatre individus quelconques i, j, k et l. Cette propriété, dite des quatre points, est une extension à quatre points de la condition ultramétrique. Elle exprime que, parmi les trois sommes des distances deux à deux, les deux plus grandes sont égales. Cette propriété permet une représentation sous forme d'un arbre. On vérifie sur un exemple à quatre points (figure 2) que la condition est effectivement respectée puisque $d(i,j) + d(k,l) = a_i + a_j + a_k + a_l$ et $d(i,k) + d(j,l) = d(i,l) + d(j,k) = a_i + a_j + a_k + a_l + 2a_c$. Cet arbre peut être représenté sous forme hiérarchique comme un dendrogramme (figure 1b). Cependant, en l'absence de racine objective, on lui préfère souvent une représentation radiale (figure 1c).

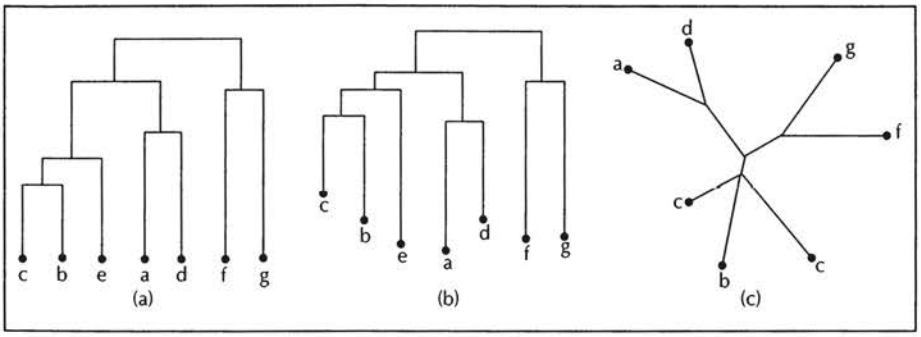

Figure 1. Ultramétrique (a) et distance additive en représentation hiérarchique (b) ou radiale (c).

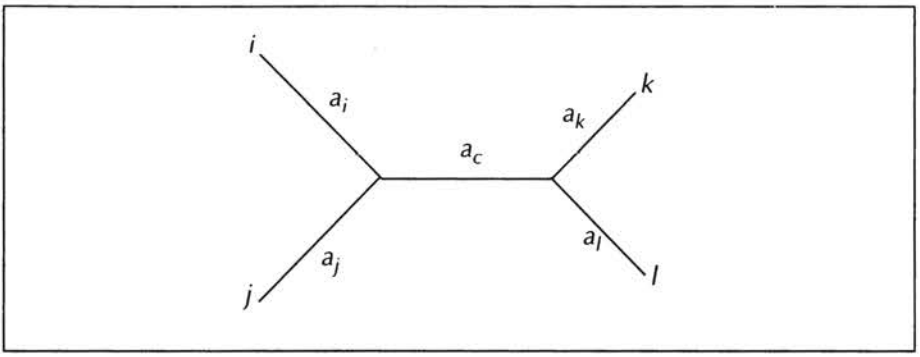

Figure 2. Arbre à quatre feuilles et longueurs des arêtes.

On montre que la condition ultramétrique n'est qu'un cas particulier de la condition des quatre points. L'ultramétrique apparaît alors comme une distance additive d'arbre particulière soumise à une contrainte plus forte. Cette contrainte supplémentaire correspond à l'obligation, sur trois individus, de l'égalité des deux plus grandes distances, contrainte absente d'une distance additive d'arbre, qui autorise des longueurs de branches inégales. Cette moindre contrainte permet donc une représentation plus fidèle des dissimilarités initiales.

Relation d'inclusion entre familles de dissimilarités

Ces différentes familles de dissimilarités ont été établies en ajoutant des contraintes supplémentaires à partir de la définition initiale d'une dissimilarité. Cela induit des relations d'inclusion entre ces différentes familles (figure 3). Ainsi, les ultramétriques sont des distances euclidiennes et des distances additives particulières. Elles sont elles-mêmes des distances de Manhattan, qui, à leur tour, sont des distances, cas particulier de dissimilarités. Quand on parcourt la figure 3 de haut en bas, on rencontre des indices supportant des conditions de plus en plus fortes et donc de moins en moins aptes à décrire

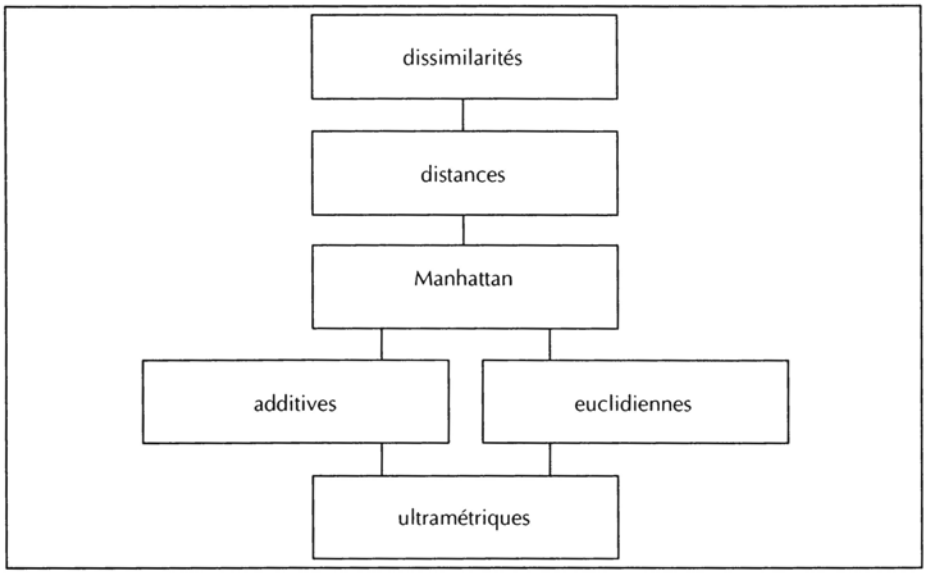

Figure 3. Relations entre familles de dissimilarités, d'après Critchley et Fichet (1994).

fidèlement les relations entre individus. Inversement, du bas vers le haut, les indices perdent leurs propriétés de représentabilité : les ultramétriques et les distances additives sont représentables sous forme d'arbre dans le plan, les distances euclidiennes et les distances de Manhattan sont représentables dans des espaces de dimension élevée. A l'opposé, distances et dissimilarités n'ont aucune propriété intéressante en la matière. Il apparaît alors que la distance de Manhattan est, parmi les distances représentables, celle qui est la moins contrainte. En l'absence d'indication contraire, le choix devrait logiquement se porter sur un tel indice.

Dans la pratique, on mène souvent en parallèle une analyse factorielle sur distance euclidienne et une représentation arborée. On note sur le graphique le niveau équivalent de la distance euclidienne et de la distance additive. La cohérence voudrait donc que la représentation arborée soit fondée sur une distance additive plutôt que sur une ultramétrique comme c'est souvent le cas.

Mesures de dissimilarités sur variables quantitatives

La dissimilarité la plus souvent utilisée sur des données quantitatives est la distance euclidienne $d(i,j) = [\sum_k (x_{ik} - x_{jk})^2]^{1/2}$ (x_{ik} : valeur de la k^e variable pour l'individu i). Elle est très classique en statistique et peut être traitée par les méthodes factorielles. Si les variables sont de nature différente, il est souvent utile de les réduire et, pour rendre la valeur de la distance indépendante du nombre P de variables, on la pondère souvent par $1/P$. D'où l'expression générale $d^2(i,j) = (\sum_k [(x_{ik} - x_{jk}) / \sigma_k]^2) / P$.

L'intérêt mathématique de la distance de Manhattan, $d(i,j) = \sum_k |x_{ik} - x_{jk}|$, a été souligné. On observe que la distance euclidienne, en élevant au carré les différences, donne un poids fort aux écarts importants. Cet *a priori* n'a pas toujours de réelle justification, la distance de Manhattan, qui donne le même poids à toutes les différences, paraît alors préférable. On pondère en général cette distance par le nombre P de variables et, si l'ordre de grandeur des variables est très différent, il est nécessaire de les ramener à une même échelle, par exemple en les rapportant à leur amplitude.

Les mesures de dissimilarités sur variables présence ou absence

Les descripteurs de la diversité génétique sont souvent des variables binaires, qui traduisent l'absence ou la présence d'un caractère. Pour les descripteurs biochimiques ou moléculaires, ces caractères codent en général la présence ou l'absence d'une bande sur un gel d'électrophorèse. Nous noterons conventionnellement 0 l'absence et 1 la présence.

De nombreuses mesures de ressemblance entre individus, ou coefficients d'association (SNEATH et SOKAL, 1973), ont été définies sur les données binaires. Elles ont le plus souvent été proposées par des chercheurs d'une discipline particulière — botanique, zoologie, paléontologie... — et avaient comme principale justification de bien traduire l'idée que se faisaient ces chercheurs de la ressemblance entre leurs objets d'étude. Nous nous limiterons ici à la présentation des indices dont on sait expliciter la logique de construction et qui présentent un comportement raisonnable (BEAULIEU, 1989) ; par exemple, si on ajoute un nouveau marqueur, la dissimilarité entre deux individus doit augmenter (ou diminuer) si les deux individus prennent, pour ce marqueur, des valeurs différentes (identiques).

On note, pour deux individus i et j, a le nombre de marqueurs qui sont présents simultanément chez i et j. De même, d est le nombre d'absences en commun, b le nombre de présences chez i et d'absences chez j et c la situation inverse. Le tableau 1 donne la liste de treize de ces indices. Il est habituel de les exprimer sous la forme de la similarité S, la dissimilarité s'obtenant, pour ces indices, par $D = 1 - S$.

Le principe général de construction de tous ces indices est le même, la similitude est estimée par le nombre d'accords. Cependant, cette valeur n'a pas de sens absolu et doit être rapportée à une base de comparaison. Les indices divergent dans le mode d'estimation du nombre d'accords et dans le choix de la base de comparaison.

L'estimation des accords dépend du sens que l'on donne à la modalité absence. Si l'on considère que seule la modalité 1 est informative, la modalité 0 traduisant surtout une absence d'information, alors le nombre d'accords est a, le nombre de présences en commun (indices S1 à S9). Si 0 et 1 sont informatifs et

Tableau 1. Principaux indices de similarité sur variables présence ou absence : auteur, exp sion et propriétés ($P = a + b + c + d$).

	Auteurs	Expression	Propriétés*			
			(1)	(2)	(3)	(4)
S1	Russel et Rao	a / P	o	o	o	
S2	Simpson	$a / \min[(a + b),(a + c)]$	n		n	
S3	Braun-Blanquet	$a / \max[(a + b),(a + c)]$				
S4	Dice	$a / [a + (b + c) / 2]$	n		n	S7, S8
S5	Ochiai	$a / [(a + b)(a + c)]^{1/2}$	n		o	
S6	Kulczynsky 1	$(a / 2)([1 / (a + b)] + [1 / (a + c)])$	n		n	
S7	Jaccard	$a / (a + b + c)$	o	o	o	S4, S8
S8	Sokal et Sneath un2	$a / [a + 2(b + c)]$	o	o	o	S4, S7
S9	Kulczynski 2	$a / (b + c)$				
S10	Sokal et Michener	$(a + d) / P$	o	o	o	S11, S
S11	Rogers et Tanimoto	$(a + d) / [a + d + 2(b + c)]$	o	o	o	S11, S
S12	Sokal et Sneath un1	$(a + d) / [a + d + (b + c) / 2]$	n		n	S10, S
S13	Sokal et Sneath un3	$(a + d) / (b + c)$				

* (1) La dissimilarité associée est (o) ou n'est pas (n) une distance, (2) elle est (o) ou n'est pas (n) une distance de Manhattan, (3) sa racine carrée est euclidienne (o) ou ne l'est pas (n), (4) elle est équivalen en ordre avec les indices indiqués (l'absence de mention indique que l'on ne connaît pas aujourd'hui la réponse).

peuvent être considérés comme deux modalités d'une variable qualitative, alors le nombre d'accords est $a + d$, le nombre de présences et d'absences en commun (indices S10 à S13). Le choix entre ces deux attitudes dépend entièrement de la nature des caractères analysés et doit être préalable à toute réflexion sur le choix d'un indice de dissimilarité. Cependant, il n'est pas toujours facile de trancher entre ces deux points de vue. Concernant les marqueurs génétiques de la diversité, on verra que la connaissance biologique des marqueurs considérés permet d'orienter le choix vers un de ces modèles.

Accords estimés par les présences en commun : indices S1 à S9

Le numérateur de ces indices est toujours a. Le dénominateur veut être une estimation du nombre d'accords que présenteraient deux individus identiques. L'indice S1 considère que ce nombre est le nombre de variables P. Ce choix n'est pas judicieux puisque deux individus ne peuvent alors être identiques que s'ils ont la valeur 1 pour toutes les variables. On préférera estimer le dénominateur à partir des nombres de présences chez i, $a + b$, et chez j, $a + c$. Bien sûr, ces deux valeurs ne sont pas égales, et il faut en définir une expression consensuelle. On peut en prendre le minimum (S2), le maximum (S3), ou une série de valeurs entre ces deux extrêmes : la moyenne arithmétique (S4), la moyenne géométrique (S5), la moyenne harmonique (S6). La connaissance des caractères peut parfois guider le choix d'un de ces indices. Cependant, en l'absence de justification claire, on retiendra de préférence le comportement neutre de l'indice de Dice (S4) fondé sur la moyenne arithmétique.

Une autre approche est de considérer que la base de comparaison est le nombre de présences rencontrées chez *i* ou chez *j*, $a + b + c$, d'où l'indice de Jaccard (S7). Cela revient, comme pour S1, à comparer *a* au nombre de variables, mais en traitant les doubles absences comme des données manquantes. Ce point de vue est, dans bien des cas, très raisonnable et explique le succès de cet indice qui est certainement le plus utilisé.

L'indice S8, assez souvent rencontré, et l'indice S9 paraissent difficiles à justifier. On remarque que les indices S4, S7 et S8 peuvent être écrits sous la forme $S = a / [a + \beta(b + c)]$ avec, respectivement, $\beta = 1/2$, $\beta = 1$ et $\beta = 2$. On pourrait construire de nouveaux indices en choisissant d'autres valeurs de β.

Accords estimés par les présences et les absences en commun : indices S10 à S13

Ces indices sont obtenus par extension des indices S1, S8, S4, S9 en remplaçant simplement *a* par $a + d$. A l'évidence, ces indices doivent être symétriques en *a* et *d*, la notation 0 et 1 étant purement arbitraire. L'extension des autres indices conduirait à des indices non symétriques en *a* et *d*, ou parfois non symétriques en *b* et *c*, et donc inacceptables. L'indice S10, extension de S1, mais aussi de S7, l'indice de Jaccard, et S12, extension de S4 fondé sur la moyenne arithmétique, paraissent les mieux appropriés dans beaucoup de cas. Les indices S12, S10 et S11 peuvent être écrits sous la forme $S = (a + d) / [(a + d) + \beta(b + c)]$ avec, respectivement, $\beta = 1/2$, $\beta = 1$ et $\beta = 2$.

Propriétés de ces indices

En complément de la logique de construction, la possession de propriétés particulières peut guider le choix d'un indice. Le tableau 1 rappelle certaines propriétés des dissimilarités associées aux indices proposés. Une première caractéristique importante est le fait d'être une vraie distance. On montre aussi que certains de ces indices sont des distances de Manhattan. Si aucun de ces indices n'est une distance euclidienne, on sait que les distances de Manhattan ont des racines carrées euclidiennes. L'indice S5, sans être une distance de Manhattan, a lui aussi une racine carrée euclidienne. Ces indices, après transformation, peuvent ainsi être traités par des méthodes factorielles.

Enfin, plusieurs de ces indices, qui ont des modes de construction comparables, sont équivalents en ordre. Entre plusieurs indices donnant des ordres comparables, on préférera les indices possédant des propriétés intéressantes, S10 ou S11, plutôt que S12 par exemple.

Les algorithmes de construction d'arbres montrent une forte sensibilité à de petites variations des dissimilarités. Le comportement de ces différents indices face à des perturbations des données constitue donc un critère de choix important. Ce comportement peut être abordé en étudiant les propriétés d'une matrice soumise à une perturbation aléatoire transformant, avec une proba-

bilité *t*, les notations 1 en 0 et inversement. Il est alors possible d'estimer l'espérance *D'* de l'indice considéré pour une perturbation de taux *t* des données.

La figure 4 présente, pour les indices S10, S11 et S12, symétriques en 0 et 1, la valeur de *D'* en fonction de *D*. Le taux d'erreur utilisé est de 0,10, taux que l'on peut rencontrer dans certaines applications, sa valeur ne change d'ailleurs pas le sens des conclusions que l'on peut énoncer. Les erreurs d'observations peuvent entraîner pour l'indice S11 de fortes surestimations des petites valeurs de dissimilarité, alors que les grandes valeurs sont peu modifiées. L'indice S12 a un comportement inverse en laissant les petites valeurs peu modifiées et en sous-estimant les grandes valeurs. L'indice S10 a un comportement plus neutre la distorsion est nulle pour $D = 0,5$ et augmente symétriquement de part et d'autre. Il constitue donc un compromis intéressant. Cependant, les algorithmes de construction d'arbre les plus utilisés procèdent par agrégation ascendante, les plus petites dissimilarités déterminent les premiers regroupements qui conditionneront l'ensemble de l'arbre. Il serait alors préférable de minimiser les perturbations sur les faibles valeurs comme le fait l'indice S12.

L'étude des indices estimant les accords par les seules présences est plus compliquée. Les distorsions observées sont plus fortes que pour les indices symétriques. Elles peuvent être très élevées (jusqu'à 50 % pour les faibles valeurs de *d*, même avec des taux d'erreur de 0,05) lorsque les fréquences de 0 sont élevées (figure 5). Cette situation est fréquente avec les marqueurs moléculaires très polymorphes, pour lesquels le nombre de 1 pour chaque individu est très faible. Cela doit conduire à une certaine prudence face à ce type de matrice, il serait alors nécessaire de vérifier la stabilité des structures en comparant les

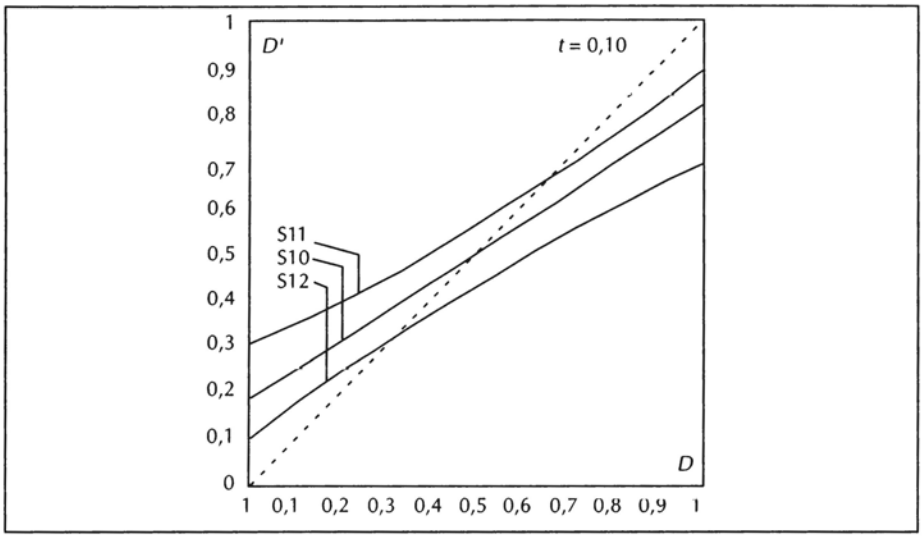

Figure 4. Estimation par simulation des dissimilarités D' *associées à S10, S11 et S12, pour un taux d'erreur de 0,10 en fonction de la dissimilarité initiale* D.

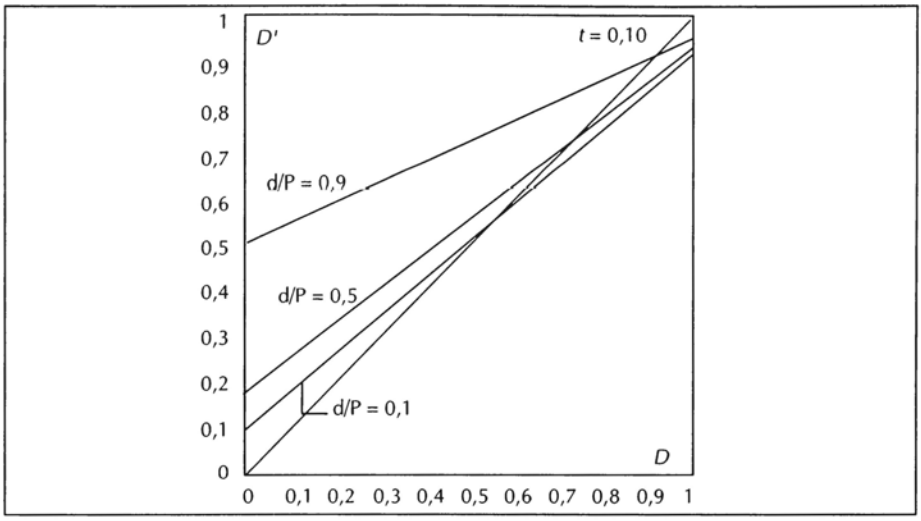

Figure 5. Estimation par simulation des dissimilarités D' *associées à l'indice S7 (Jaccard), pour un taux d'erreur de 0,10 en fonction de la dissimilarité initiale* D.

résultats obtenus sur la matrice complète et débarrassée des marqueurs particulièrement rares. Ils conduisent, en effet, pour la plupart des paires d'individus à augmenter le nombre d'absences en commun. Les indices S4, S5 et S6, correspondant aux trois types de moyennes du dénominateur, ont un comportement très comparable. Ils ont tendance à donner de plus faibles distorsions des petites valeurs de dissimilarité que l'indice de Jaccard, S7, lui-même plus favorable que S8. On les préférera donc pour cette raison.

Mesures des dissimilarités sur variables qualitatives

Les caractères sont ici des variables qualitatives présentant un nombre fini de modalités : couleur des pétales, forme du stigmate... Intuitivement, deux individus sont proches s'ils possèdent la même modalité pour un grand nombre de variables.

Les indices utilisables correspondent à la généralisation des indices présentés pour les variables binaires symétriques en 0 et 1, qui sont en fait des variables qualitatives à deux modalités.

On note m le nombre de variables en accord, u le nombre de variables en désaccord, P étant toujours le nombre de variables. Les indices de similarité correspondants s'écrivent :

– $m / (m + u) = m / P$ (Sokal et Michener) ;
– $m / (m + 2u) = m / (P + u)$ (Rogers et Tanimoto) ;
– $2m / (2m + u) = 2m / (P + m)$ (Sokal et Sneath).

La dissimilarité s'obtient directement par $d = 1 - s$.

On utilise souvent la distance de χ^2 sur un tableau disjonctif complet. Pour une variable, on crée pour chacune de ses modalités une variable binaire prenant la valeur 1 si l'individu présente cette modalité, 0 sinon, et on calcule le χ^2 classique sur cette nouvelle matrice de données. Cette distance ne donne pas un même poids à toutes les modalités, en particulier les modalités rares prennent un poids très fort. Même si cet effet peut être parfois recherché, il paraît souvent indésirable.

Le choix d'un indice de similarité sur marqueurs biochimiques et moléculaires

La ressemblance génotypique entre deux individus peut être mesurée par une similarité génétique définie à partir des formes alléliques des gènes observés. Certains types de marqueurs (isoenzymes, microsatellites...) donnent, en général, une information génétique directe et permettent un codage en allèles par identification de la composition allélique de chaque locus. La similarité génétique, telle qu'on la définit, est alors directement estimable. D'autres marqueurs ne permettent pas d'accéder à la totalité de l'information génétique et seul un phénotype est observé. La similarité génétique ne peut être calculée directement, mais les caractéristiques de ces marqueurs permettent de définir l'indice de similarité le plus pertinent. Il est intéressant d'évaluer de plus l'ordre de grandeur de l'erreur liée à cette perte d'information.

Les marqueurs multiallèliques

On fixe *a priori* une définition de la similarité génétique T_{ij} entre deux individus i et j comme la moyenne sur les L locus du rapport du nombre n_{ls} d'allèles du locus l présents simultanément chez les deux individus et du nombre n_{lc} d'allèles comparés : $T_{ij} = (\sum_l n_{ls} / n_{lc}) / L$. Si π est la ploïdie, n_{lc} vaut π pour tous les locus. En posant $n_s = \sum_l n_{ls}$ et $n_c = \pi L$, alors $T_{ij} = n_s / n_c = n_s / (\pi L)$.

Pour une espèce diploïde, un locus peut être codé par autant de variables qu'il a d'allèles, chaque variable prenant les valeurs 2, 1 ou 0. Suivant ce codage, pour deux individus i et j, la combinaison des génotypes de chaque locus appartient à un des sept groupes, notés A à G (tableau 2). Le nombre de locus de chaque groupe est noté l_A à l_G. On peut alors écrire T_{ij} à partir de $n_s = 2l_A + l_B + 2l_E + l_F$ et $n_c = 2(l_A + l_B + l_C + l_D + l_E + l_F + l_G)$.

On définit les paramètres $n_{r,s}$ ($r \geq s$) comme le nombre d'allèles présents r fois chez un individu et s fois chez l'autre. Chacun des groupes A à G contribue à ces paramètres. Par exemple, pour un locus à a_l allèles, une paire du groupe D participe une fois à $n_{2,0}$, deux fois à $n_{1,0}$ et $a_l - 3$ fois à $n_{0,0}$.

Tableau 2. Les sept combinaisons possibles des génotypes de deux individus diploïdes.

	A	*B*	*C*	*D*	*E*	*F*	*G*
Individu *i*	20..	200..	200..	2000..	110..	110..	1100..
Individu *j*	20..	110..	020..	0110..	110..	101..	0011..

On peut exprimer les effectifs I_A à I_G à partir de ces paramètres $n_{r,s}$, d'où l'écriture de T_{ij} comme rapport de $n_s = 2n_{2,2} + n_{2,1} + n_{1,1}$ et $n_c = (2n_{2,2} + n_{2,1} + n_{1,1}) + (n_{2,1} + 2n_{2,0} + n_{1,0}) / 2$.

On reconnaît alors dans l'expression de T_{ij} une généralisation de S4, l'indice de Dice. Le nombre d'accords est effectivement de 1 pour les génotypes contribuant à $n_{2,1}$ et $n_{1,1}$ et de 2 pour ceux contribuant à $n_{2,2}$. De même, les désaccords sont au nombre de 1 pour $n_{2,1}$ et $n_{1,0}$ et de 2 pour $n_{2,0}$. On note que pour une espèce haploïde, certains micro-organismes par exemple, on retrouverait directement l'indice de Dice.

L'indice de Dice appartient à la famille des indices qui ne prennent pas en compte l'information apportée par les doubles absences. Ce point de vue est logique puisque le nombre de doubles absences n'apporte aucune information sur la proximité des individus et ne dépend que du nombre d'allèles des locus.

Les marqueurs codominants codés en bande

Pour des marqueurs de type RFLP ou les isoenzymes, il est parfois impossible de reconnaître les allèles appartenant à un même locus et seul un codage en bandes reste disponible. Pour une espèce diploïde, la présence d'une bande, codée 1, peut alors correspondre à un locus homozygote, qui devrait être codé 2, ou à un des allèles d'un locus hétérozygote, normalement codé 1. Les différents groupes de génotypes définis précédemment ne sont plus identifiables et le phénotype observé diffère du génotype réel (tableau 3).

Comme dans le cas précédent, l'information des doubles absences n'est pas pertinente et par analogie avec l'indice de Dice, on peut retenir, comme mesure de ressemblance, un indice de la famille $S_{ij} = n_{1,1} / (n_{1,1} + \beta n_{1,0})$. On

Tableau 3. Les sept combinaisons possibles des génotypes et les phénotypes observés par marqueurs codominants.

		A	*B*	*C*	*D*	*E*	*F*	*G*
Génotype	*i*	20..	200..	200..	2000.	110..	110..	1100..
	j	20..	110..	020..	0110.	110..	101..	0011..
Phénotype	*i*	10..	100..	100..	1000.	110..	110..	1100..
	j	10..	110..	010..	0110.	110..	101..	0011..

peut alors rechercher la valeur de β telle que, sous certaines hypothèses, S_{ij} approche au mieux T_{ij}.

Pour une espèce diploïde autogame qui serait homozygote pour tous les locus, $S_{ij} = T_{ij}$ pour $\beta = 1/2$, S_{ij} est alors directement l'indice de Dice.

On connaît, en général, pour une espèce donnée, une valeur de t, le taux d'hétérozygotie moyen. En admettant que cette valeur puisse s'appliquer à chaque locus, on peut définir les valeurs de β, fonction de t, qui annulent $S_{ij} - T_{ij}$. L'expression de cette différence est complexe et les relations entre β et t ont été recherchées numériquement pour différentes distributions du nombre d'allèles par locus, le maximum étant fixé à 7 (figure 6). Les distributions I, II et IV correspondent à des distributions extrêmes, la distribution III reproduit une distribution observée sur l'hévéa (M. Seguin, comm. pers.).

On remarque le cas des locus à faible nombre d'allèles (I ou II) où la similarité génétique n'est approchée, pour des hétérozygoties faibles, que pour des valeurs de β proches de 1.

Si l'on prend comme base de référence la distribution III, on pourrait proposer de fixer la valeur de β à 0,5 pour $t \leq 0{,}1$, à 0,6 pour $0{,}1 < t \leq 0{,}3$, à 0,7 pour $t > 0{,}3$.

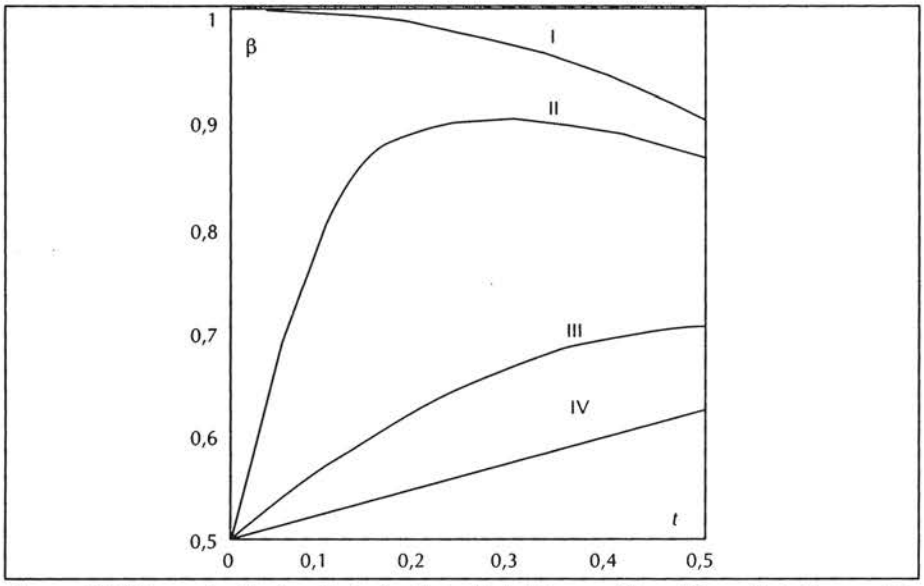

Figure 6. Valeurs de l'indice β de la dissimilarité D = a / [a + β(c + d)] telles que D_{ij} est égal à la dissimilarité génétique T_{ij} calculée à partir de l'information génétique complète, en fonction du taux d'hétérozygotie t et pour diverses distributions du nombre d'allèles :

I : tous les locus ont 2 allèles ;

II : locus à 2 ou 3 allèles dans les proportions 0,9 et 0,1 ;

III : locus de 2 à 7 allèles dans les proportions 0,27, 0,32, 0,19, 0,11, 0,07 et 0,03 ;

IV : tous les locus ont 7 allèles.

L'ordre de grandeur de la différence a été évalué par simulation, pour différents taux d'hétérozygotie et une série de valeurs de β entre 1/2 et 1, à partir d'une population de 40 individus décrits par 200 locus. L'écart $S_{ij} - T_{ij}$ est composé d'un biais systématique, sans conséquence sur les méthodes d'analyse, et d'une erreur vraie qui seule nous intéresse ici. On exprime cette erreur par le demi-écart entre la plus faible et la plus forte des valeurs observées dans les simulations, elle est exprimée en pourcentage de la similarité génétique moyenne. Pour une distribution de type III, cette erreur maximale est de l'ordre de 3 % pour *t* voisin de 0,2, mais peut atteindre 5 % pour des espèces très hétérozygotes.

On montre, de plus, que cette erreur dépend surtout de *t* et peu de β. En pratique, on pourra le plus souvent retenir β = 0,5, correspondant à l'indice de Dice.

Les marqueurs dominants codés en bandes

Les marqueurs dominants, RAPD ou AFLP, peuvent être considérés comme des locus à deux allèles particuliers puisqu'un seul allèle est repérable, l'autre étant un allèle nul non matérialisé sur le gel. Ces marqueurs sont dits dominants car il est impossible de connaître le nombre de copies d'un allèle pour un locus particulier et donc de distinguer les homozygotes des hétérozygotes.

Parmi les types *A* à *G* définis précédemment pour une espèce diploïde, seuls *A, B, C* et *E* existent puisque les locus n'ont que deux allèles. Les deux allèles ne jouant pas un rôle identique, les génotypes *A* et *B* doivent être subdivisés suivant les phénotypes, en distinguant l'allèle repérable de l'allèle nul. L'impossibilité du codage en allèles conduirait, comme pour les marqueurs RFLP, à un phénotype noté théorique dans le tableau 4. Comme l'allèle nul, noté par un point, n'est pas repérable, un seul des états alléliques est donc lisible, d'où le phénotype réellement lu.

Le nombre *L* de locus est le nombre *P* de bandes différentes observées dans la population. Il faut remarquer immédiatement que $n_{0,0'}$ le nombre de doubles absences, a ici un sens très clair. Il correspond au nombre de locus homozy-

Tableau 4. Génotypes, phénotypes théoriques et phénotypes observés par marqueurs dominants.

		A1	*A2*	*B1*	*B2*	*C*	*D*
Génotype	*i*	20	02	20	02	20	11
	j	20	02	11	11	02	11
Phénotype théorique	*i*	10	0.	10	0.	10	1.
	j	10	0.	1.	1.	0.	1.
Phénotype lu	*i*	1	0	1	0	1	1
	j	1	0	1	1	0	1

gotes pour l'allèle nul chez les deux individus et apporte autant d'information que le nombre de doubles présences. Il sera logique de retenir un indice symétrique en 0 et 1. Le dénominateur de la similarité génétique T_{ij} est égal, dans ce cas particulier, à P. Cette valeur conduit à retenir l'indice de Sokal et Michener, il s'exprime ici par $S_{ij} = (n_{1,1} + n_{0,0}) / P$.

La différence $S_{ij} - T_{ij}$, qui mesure l'erreur d'estimation de la similarité génétique, s'exprime simplement par $(I_{B1} - I_{B2}) / P$. Cette différence est nulle si les fréquences des génotypes (2,0) et (0,2) sont égales, c'est-à-dire si les fréquences des deux allèles sont identiques pour tous les locus, elle restera faible si elles tendent à l'être, en moyenne, sur les locus. D'autre part, si l'espèce est très homozygote, les proportions de génotypes B sont faibles et $I_{B1} - I_{B2}$ restera faible.

Pour des espèces plus hétérozygotes, cette différence peut être forte puisque les génotypes B peuvent représenter jusqu'à 50 % des effectifs. Comme précédemment, des simulations numériques ont permis d'en estimer l'ordre de grandeur pour différents niveaux d'hétérozygotie et pour différentes proportions des deux allèles d'un locus. L'erreur vraie est, comme précédemment, la demi-amplitude des extrêmes observés rapportée à la similarité génétique moyenne. Elle croît avec l'hétérozygotie et atteint 13 % pour $t = 0,5$ et aucun déficit allélique, elle dépasse 18 % si les allèles non identifiables sont majoritaires et tombe à 10 % dans le cas contraire. Dans tous les cas, l'imprécision introduite par la perte d'information génétique est nettement plus importante que pour les marqueurs codominants.

On montre, sur ces exemples, que le contenu informatif des marqueurs permet de justifier le choix d'un indice de similarité, justification qui est le plus souvent absente des publications dans ce domaine.

Les méthodes d'analyse factorielle

Les méthodes d'analyse factorielle ont pour objectif de repérer les structures les plus fortes dans les populations que l'on étudie et d'éliminer les particularités ponctuelles qui gênent la perception générale des phénomènes. Elles peuvent ainsi être d'une grande utilité lorsque l'on aborde l'étude de la diversité d'une espèce ou d'une population.

L'objet de ces méthodes est de produire une représentation géométrique des mesures de dissemblance entre unités. Cette représentation permet de manipuler de manière concrète la notion de diversité. On aboutit ainsi à une décomposition hiérarchisée de la diversité, qui permet de distinguer les tendances fondamentales des variations particulières.

L'utilisation de ces techniques nécessite de connaître les principes sur lesquels elles reposent et les conventions qu'elles utilisent. Après une présentation générale de l'analyse factorielle mettant l'accent sur ses aspects conventionnels, on verra comment elle s'intègre spécialement dans trois applications bien connues : l'analyse factorielle sur tableau de distances (AFTD), l'analyse en composantes principales (ACP) et l'analyse des correspondances multiples (ACM).

On abordera ensuite le problème du traitement des tableaux multiples, qui se rencontre en particulier en étudiant une population décrite par plusieurs types de marqueurs. La méthode de l'analyse factorielle multiple (AFM) permet, lorsque l'on dispose sur les individus d'informations provenant de différents points de vue, de mettre en évidence les caractères principaux, qui se manifestent de manière consensuelle, et ceux qui au contraire sont spécifiques d'un type de mesures.

L'analyse factorielle : présentation générale

Nuage de points et dispersion

Si la distance choisie pour mesurer les dissemblances entre les unités étudiées est euclidienne, l'ensemble des unités peut être représenté sous la forme d'un nuage de points dans un espace suffisamment vaste. Dans cet espace on peut distinguer un point B, appelé point moyen, ou barycentre, qui est défini comme étant le plus près possible de tous les points du nuage. Il correspond à un « type moyen » neutre dont chaque unité réelle diffère par ses caractéristiques propres.

Dans ces conditions, la diversité totale de la population est représentée par la dispersion du nuage autour de B. L'analyse factorielle convient de quantifier cette dispersion par l'inertie du nuage par rapport à B, c'est-à-dire par la somme des carrés des distances de chaque point à B.

Axes principaux d'inertie et coordonnées des individus

En général, l'espace dans lequel se trouve le nuage est un espace de dimension élevée. Pratiquement, cela signifie qu'en partant de B, on peut se diriger dans une multitude de directions. Autrement dit, il existe de très nombreuses façons pour un individu de se distinguer du type moyen et toutes les unités observées en diffèrent chacune à sa manière. Le but de l'analyse factorielle est de ramener approximativement toute particularité individuelle à la composition d'un petit nombre de directions très usitées et indépendantes les unes des autres.

Pour juger de la « fréquence d'usage » d'une direction *U*, on construit un critère en sommant pour tous les individus une valeur d'autant plus importante que l'angle que fait *U* avec la direction spécifique de l'individu est aigu et que

l'individu est à une distance élevée du point moyen B. Ce critère s'appelle l'inertie du nuage le long de *U*. La direction qui maximise ce critère s'appelle l'axe principal d'inertie du nuage. L'analyse factorielle produit successivement les autres axes en choisissant pour chacun la direction qui maximise le critère de l'inertie parmi les directions orthogonales aux axes déjà identifiés. L'orthogonalité des axes assure l'indépendance des variations résumées par ces axes.

Il résulte de ce processus un système d'axes indépendants expliquant successivement le maximum de variation. On sait calculer les coordonnées de chaque unité sur ces axes et ainsi les placer dans ce système orthogonal. L'ensemble des coordonnées de tous les individus sur un axe s'appelle un facteur.

Les coordonnées factorielles permettent de reconstituer la dispersion totale de la population. Cependant, et c'est l'intérêt essentiel des analyses factorielles, la sélection des premiers facteurs seulement permettra en général de reconstruire une grande part de la dispersion totale. Plusieurs critères plus ou moins empiriques ont été proposés pour déterminer le nombre d'axes à conserver (SAPORTA, 1990). Il convient cependant de rester circonspect en réalisant cette sélection et le choix des axes doit être raisonné dans chaque cas.

Sans autre information, les coordonnées des individus sur les axes n'ont guère de signification. Pour qu'une analyse factorielle soit interprétable, il faut disposer d'une source d'information « externe » que l'on pourra relier aux axes factoriels. Cette information externe peut ne pas être explicite et n'apparaître que lorsque l'on constate que tel axe rassemble ou, au contraire, oppose certains individus reconnaissables. Le plus souvent pourtant, un axe sera interprété en constatant qu'il est lié à des variables mesurées sur les individus, qu'elles aient ou non participé au calcul de la matrice de distance.

Ce n'est que lorsqu'un axe est interprété et que la part de dispersion qui lui est attachée est compréhensible que l'on considérera qu'il s'agit d'un des facteurs qui rentrent en ligne de compte dans l'élaboration de la diversité.

Représentations graphiques

Pour avoir une représentation visuelle de la diversité, mais aussi pour faciliter l'interprétation des axes, les méthodes d'analyse factorielle proposent des représentations graphiques des individus et le cas échéant des variables. Ces graphiques constituent des apports déterminants de ces méthodes. Ils permettent de voir, au sens propre du terme, des oppositions, des regroupements et des tendances difficiles à percevoir avec des statistiques chiffrées. Chaque type de graphique a cependant ses règles de lecture propres qu'il convient de bien maîtriser pour éviter toute interprétation hasardeuse.

Les méthodes classiques d'analyse factorielle

En pratique, l'analyse factorielle telle qu'elle vient d'être présentée ne constitue qu'une étape dans un processus qui part d'un tableau de données, opère

d'éventuels codages, calcule s'il le faut les distances, trouve les axes factoriels et produit des aides à l'interprétation adaptées à la nature des données. C'est l'ensemble de ce processus qui forme une « méthode » d'analyse factorielle. Selon le type de données dont on dispose, les méthodes les plus couramment employées sont : l'AFTD pour un tableau de distances, l'ACP pour un tableau de variables quantitatives et l'ACM pour un tableau de variables qualitatives.

L'ANALYSE FACTORIELLE SUR TABLEAU DE DISTANCES

L'analyse factorielle sur tableau de distances est utilisée pour traiter une matrice de distances calculées avec un indice original, choisi de manière adaptée à ses propres données. Il est cependant nécessaire que la distance obtenue soit une distance euclidienne ou rendue euclidienne par transformation.

Dans ce cas, deux individus i et j peuvent être représentés par deux points e_i et e_j d'un espace de dimension K inconnue tels que le carré de la distance d_{ij} peut s'écrire : $d_{ij}^2 = (e_i - e_j)'(e_i - e_j)$.

D'un point de vue calculatoire, on montre que les axes successifs de l'analyse factorielle et l'inertie du nuage sur chaque axe correspondent aux vecteurs propres et aux valeurs propres de la matrice W des produits scalaires entre éléments. Si l'origine des coordonnées est placée au point moyen du nuage, les produits scalaires $w_{ij} = < e_i ; e_j >$ sont entièrement déterminés par les d_{ij} par la formule de Torgerson : $w_{ij} = -(d_{ij}^2 - d_{i.}^2 - d_{.j}^2 + d^2) / 2$.

La diagonalisation de cette matrice W permet d'extraire les vecteurs propres et les valeurs propres associées.

Le nombre de vecteurs propres associés à une valeur propre non nulle est la dimension K de l'espace contenant le nuage de points.

L'AFTD constitue la mise en œuvre la plus simple de l'analyse factorielle. Dans ces conditions, la seule aide à l'interprétation disponible est le graphique des individus sur un choix quelconque de deux axes factoriels.

L'ANALYSE EN COMPOSANTES PRINCIPALES

L'analyse en composantes principales traite de variables quantitatives et impose une distance euclidienne entre individus. Les distances euclidiennes qui peuvent être calculées ne diffèrent que par les poids que l'on accorde aux variables. Le système de pondération le plus usuel consiste à ramener toutes les variables à une même échelle en les pondérant par l'inverse de leur écart type. On dit alors que l'ACP est normée, ou centrée-réduite.

De même que pour l'AFTD, les coordonnées des individus sur les composantes principales peuvent se calculer à partir de la matrice des produits scalaires. Celle-ci s'obtient immédiatement à partir du tableau des données centrées.

L'interprétation des axes fournis par une ACP est fortement enrichie puisque les corrélations entre les facteurs et les variables du tableau peuvent être calculées. L'examen de ces corrélations pour chaque axe permet de mettre en évidence la ou les variables les plus liées à cet axe. On met ainsi en évidence des faisceaux de variables qui varient conjointement et dont la variation conjointe provoque une forte discrimination des individus.

L'ANALYSE DES CORRESPONDANCES MULTIPLES

L'analyse des correspondances multiples (ACM) permet d'étudier des individus décrits par des variables qualitatives à modalités. Cette méthode calcule les distances entre individus par la distance du χ^2 sur le tableau disjonctif complet (voir ci-dessus). C'est-à-dire que deux individus sont d'autant plus distants l'un de l'autre qu'ils présentent de nombreux désaccords et que leurs désaccords leur font adopter à l'un ou à l'autre des modalités rares.

Cette méthode est aussi connue sous les noms d'analyse factorielle des correspondances multiples (AFCM), d'analyse factorielle des correspondances sur tableau disjonctif complet (AFC sur TDC), voire simplement d'analyse factorielle des correspondances (AFC).

L'interprétation des axes fournis par l'analyse factorielle de cette matrice de distances se déduit des relations qu'ils possèdent avec trois types d'objets : les individus, les variables et les modalités.

❑ Les individus fournissent les mêmes interprétations que dans les autres méthodes d'analyse factorielle.

❑ Comme en ACP, on peut évaluer le lien entre les variables du tableau et chaque axe. Pour une variable qualitative, ce lien s'apprécie par le rapport η^2, qui est égal à la part de l'inertie totale de l'axe reconstituée par la somme des carrés des écarts intermodalités. Un facteur peut alors être perçu comme une synthèse numérique d'un certain nombre de variables qui varient conjointement.

❑ Une variable qualitative est un objet plus complexe qu'une variable quantitative et la diversité qu'elle induit se manifeste sur d'autant plus d'axes que la variable a de modalités. Une interprétation plus fine des axes et des liaisons entre variables doit donc être recherchée au niveau des modalités. Cependant, le choix de la distance du χ^2 fait que la forte contribution d'une modalité à un axe peut n'être due qu'à la rareté de cette modalité et donc à l'excentricité du petit nombre d'individus qui la possède. Lorsqu'il n'en est pas ainsi, un axe révèle un ensemble de modalités en association mutuelle, c'est-à-dire présentes ou absentes simultanément chez un grand nombre d'individus.

En ce qui concerne les représentations graphiques, la distance du χ^2 fait qu'en disposant sur un même graphique les individus et les modalités, un individu est proche des modalités qu'il possède et une modalité est proche des individus qui la possèdent.

Le traitement des tableaux multiples et l'analyse factorielle multiple

L'ANALYSE CONJOINTE DE PLUSIEURS TABLEAUX

Face à des données de nature différente, on peut souhaiter combiner l'information apportée par les différents tableaux de façon à mettre en évidence des structures plus fines dans la population.

Il peut arriver en effet qu'une organisation des individus qui apparaît lors de l'analyse d'un certain tableau et qui est clairement due à des variations de certains caractères importants de ce tableau ne se manifeste pas lors de l'analyse d'un autre type de données. Une approche pourrait être de procéder à l'analyse conjointe d'un supertableau juxtaposant les différentes données. On ferait ainsi apparaître toutes les oppositions possibles, et on repérerait celles qui sont communes à tous les types de données et celles qui, au contraire, sont spécifiques à un tableau donné.

Deux obstacles majeurs s'opposent cependant à l'analyse immédiate d'un tel supertableau :
– le type des variables — qualitatif ou quantitatif — n'est pas nécessairement le même pour tous les tableaux, il conviendrait *a priori* d'analyser les tableaux qualitatifs par l'ACM et les tableaux quantitatifs par l'ACP ;
– tous les tableaux ne possèdent pas nécessairement la même inertie et n'expriment pas forcément la même organisation de la diversité. Les tableaux les plus fortement structurés risquent donc d'influencer les résultats de façon prépondérante. L'apport des tableaux les moins structurés passerait alors inaperçu.

Un résultat intéressant permet de contourner le premier obstacle : on démontre en effet que l'on peut obtenir les mêmes facteurs que ceux qui sont produits par l'ACM en réalisant une ACP, pondérée de manière adéquate, du tableau formé par les indicatrices des variables qualitatives, c'est-à-dire du tableau codé sous forme disjonctive complète. Dans le cas des données moléculaires, cela signifie que chaque marqueur doit être représenté par deux colonnes : l'une indiquant sa présence et l'autre son absence.

Le second obstacle se contourne en adoptant une pondération des tableaux qui réduit l'influence des tableaux ayant une structure forte. Pour cela il faut quantifier *a priori* la part de l'inertie totale d'un tableau qui correspond à sa structure interprétable. Cette quantification est assez arbitraire et c'est l'un des points sur lesquels se distinguent les différentes méthodes d'analyse des tableaux multiples. La méthode Statis (LAVIT, 1988) choisit comme mesure la somme des carrés des valeurs propres de l'analyse séparée du tableau. L'AFM (ESCOFIER et PAGES, 1993) préfère la plus grande de ces valeurs propres. Sans entrer dans le détail, on peut dire que la première peut réduire l'importance d'un tableau dont, par hasard, les dernières valeurs propres seraient élevées.

La seconde au contraire peut donner plus de poids qu'il n'en faudrait à un tableau dont les premières valeurs propres ne sont pas très différentes les unes des autres.

L'ANALYSE FACTORIELLE MULTIPLE

L'analyse factorielle multiple est donc une ACP particulière, dans laquelle toutes les variables d'un tableau sont pondérées par l'inverse de la variance de l'axe principal d'inertie de l'analyse séparée du tableau.

En tant qu'ACP, elle permet de repérer les individus qui se ressemblent pour l'ensemble des variables et donc de faire une typologie des individus. Elle permet aussi d'attribuer le passage d'une classe de cette typologie à une autre, à la variation conjointe de quelques variables. Ce sont ces directions communes de variation que l'AFM produit comme axes d'inertie, sous le nom d'axes de l'analyse globale.

Cependant, l'organisation des données sous forme de juxtaposition de tableaux enrichit l'interprétation de l'AFM. D'une part, l'AFM fournit pour chaque axe de l'analyse globale des indices permettant de déterminer s'il n'est dû qu'à des variables d'un seul tableau ou s'il est commun à plusieurs types de données. D'autre part, l'AFM calcule les corrélations entre les axes de l'analyse globale et les axes que produiraient les analyses séparées de chaque tableau. Lorsque le critère précédent indique qu'un certain axe de l'analyse globale est commun à plusieurs tableaux, cela permet, pour aller plus loin dans l'analyse, de déterminer à quel caractère précisément au sein de chaque type de mesures, l'axe commun est lié.

Enfin, ces corrélations permettent de connaître, pour un certain axe d'une analyse séparée dont on saisit parfaitement le sens, le numéro de l'axe de l'analyse globale auquel il contribue essentiellement. Si cet axe global est commun à plusieurs tableaux, cela signifie que l'axe partiel se manifeste de manière consensuelle. Au contraire, si l'axe global est spécifique, cela signifie que l'axe partiel correspond à un caractère qui est propre au type d'observations du tableau séparé.

L'AFM possède par ailleurs un certain nombre de propriétés intéressantes pour ce qui est de la représentation des individus. En particulier, on montre qu'elle offre un compromis intéressant entre d'une part, la qualité de représentation des nuages correspondants à chaque tableau séparé et d'autre part, la proximité pour un même individu de ses représentations partielles.

La représentation arborée des dissimilarités

Le principe de toute méthode de représentation arborée est d'approcher au mieux la dissimilarité δ, choisie pour sa pertinence à décrire les relations entre

individus, par une distance d représentable sous forme d'un arbre, c'est-à-dire une ultramétrique ou une distance additive d'arbre (BARTHELEMY et GUENOCHE, 1988).

La recherche de la solution exacte nécessite l'énumération de toutes les configurations d'arbre possibles sur les n individus. Pour chaque structure d'arbre possible, on estime au sens des moindres carrés les longueurs des arêtes de l'arbre. On retient finalement l'arbre qui permet de minimiser la somme des carrés des écarts entre dissimilarité initiale et distance reconstituée dans l'arbre. On montre, par récurrence, que le nombre d'arbres binaires différents construits sur n individus est $\prod_{i=3,n} (2i - 5)$. Pour $n = 10$, on dénombre plus de $2\ 10^6$ arbres différents et plus de $2\ 10^{20}$ pour $n = 20$. L'énumération des tous les arbres devient donc irréalisable, même avec les moyens informatiques les plus puissants, dès que n dépasse quelques dizaines.

Dans ces situations, la seule possibilité est de construire, à partir d'heuristiques raisonnables, des solutions qui seront les meilleurs possibles mais que l'on ne pourra jamais assurer comme optimales. Diverses heuristiques ont été proposées, elles permettent de construire des algorithmes de plus ou moins grande complexité, la complexité d'un algorithme mesurant la croissance du temps de calcul lorsque n augmente. Dans le cadre de cet ouvrage, ne seront évoquées que les méthodes de regroupement qui sont de complexité suffisamment faible pour traiter quelques centaines d'individus.

Les individus étudiés sont souvent décrits par plusieurs types de variables qui ne peuvent être réunis dans le calcul d'une dissimilarité unique. Chaque jeu de variables est alors traité séparément et l'on s'interroge sur la présence de structures identiques dans les différents arbres. Deux méthodes de construction d'arbres synthétiques, les arbres consensus et les sous-arbres communs, seront présentées.

Les méthodes de regroupement

Ce sont des méthodes itératives qui procèdent par agglomérations ascendantes successives, en construisant l'arbre pas-à-pas. Initialement, la matrice traitée a autant d'éléments que d'individus de la population étudiée et on fixe à l'arbre une structure en étoile. A chaque itération, deux éléments, individus ou groupes déjà formés, définis comme voisins, sont réunis pour former un groupe. Ils sont choisis pour que l'arbre qui découle de leur regroupement optimise un critère fixé. Ce groupe devient un nouvel élément fictif qui remplace les deux éléments réunis, la matrice est mise à jour et diminue alors d'une unité. Le processus est réitéré jusqu'à ce que tous les individus soient réunis dans un même groupe.

Les diverses méthodes de ce type se caractérisent par des choix différents aux trois points clés de chaque itération : la sélection des éléments à réunir, qui

dépend de la définition du voisinage retenue, la mise à jour de la matrice de dissimilarité par calcul d'une dissimilarité entre le groupe formé et les autres éléments, la construction des longueurs des deux arêtes issues des deux éléments réunis.

La définition du voisinage

La définition la plus naturelle est de choisir comme voisins les deux individus, ou groupes, présentant la plus petite dissimilarité. Les éléments i et j sont définis comme voisins si $\delta(i,j)$ est la plus petite des valeurs de dissimilarité.

Ce critère permet de retrouver l'arbre solution dans un cas théorique où la distance initiale serait déjà une ultramétrique. En revanche, il ne permet pas forcément de retrouver la bonne structure si la dissimilarité initiale est une distance additive d'arbre. Sur l'exemple de la figure 2 sur quatre points seulement, on peut imaginer que la distance $d(i,k) = a_i + a_c + a_k$ soit la plus petite des distances bien que i et k soient dans deux paires opposées par l'arête centrale. Ce critère regroupera obligatoirement i et k et ne permettra pas de retrouver l'arbre vrai. Pour cela d'autres critères, qui prennent en compte l'ensemble des distances pour juger d'un voisinage, doivent être utilisés.

Saitou et Nei (1987), dans un contexte de diversité génétique, proposent un critère de voisinage fondé sur le principe de parcimonie, qui est à la base de l'approche phylogénétique. L'objectif est de créer un arbre qui soit de longueur totale minimale. On considère qu'une arête représente un nombre d'événements mutationnels et en vertu d'un principe de base qui veut que l'évolution procède toujours par modification génétique la plus simple possible, le nombre d'événements, et donc la longueur totale, doit être minimal. Ces considérations conduisent à une définition de voisinage relatif, défini en minimisant un critère $Q(i,j)$, fonction de $\delta(i,j)$ et de la moyenne des dissimilarités de i et j aux $n-2$ autres éléments k :

$$Q(i,j) = \delta(i,j) - (\Sigma_k [\delta(i,k) + \delta(j,k)]) / (n-2).$$

On montre que ce critère possède des propriétés d'optimalité en termes de moindres carrés et que son domaine d'application dépasse la cadre de la reconstruction phylogénétique. Il peut s'interpréter, très généralement, comme une pondération de la dissimilarité entre deux individus par leurs dissimilarités aux autres individus. Deux individus proches, qui diffèrent fortement des autres individus, sont plus ressemblants que deux individus proches, mais proches également des autres individus. Cette attitude est, dans bien des cas, justifiée et explique le succès de cette méthode.

Sattath et Tversky (1977) adoptent une démarche très différente. Ils partent de la caractérisation d'une distance additive d'arbre par la condition des quatre points. Pour quatre individus quelconques i, j, k et l, si $d(i,j) + d(k,l)$ est la plus petite des trois sommes des distances deux à deux alors les deux plus grandes sont égales : $d(i,k) + d(j,l) = d(i,l) + d(j,k)$. La dissimilarité initiale δ n'est pas

une distance additive mais il est toujours possible de former les sommes des dissimilarités deux à deux. Parmi ces trois sommes, l'une est la plus petite, $\delta(i,j) + \delta(k,l)$ par exemple, on considère alors que les paires de points (i,j) et (k,l) sont de bons candidats pour être des voisins. On leur attribue un score de 1 alors que les autres paires de points, (i,k), (j,l), (i,l) et (j,k), se voient attribuer un score nul. On balaie tous les quadruplets de points que l'on peut former sur les n points et on retient finalement comme voisins les individus de la paire qui a le score total le plus grand. Cette définition de voisinage topologique est de nature ordinale puisque seule importe l'ordre des trois sommes et non pas directement les valeurs des dissimilarités qui les composent. Elle paraît particulièrement adaptée lorsque les valeurs des dissimilarités sont plus ou moins entachées d'erreur.

La mise à jour des dissimilarités

Deux groupes, composés éventuellement d'un seul individu, étant réunis pour former un nouveau groupe, il est nécessaire de définir une dissimilarité entre le nouvel élément fictif créé et les autres éléments présents à cette itération. Soient i et j les groupes réunis pour former un nouvel élément s, c_i et c_j les effectifs de ces groupes, et k un autre élément, la définition la plus naturelle de $\delta(s,k)$ est la moyenne arithmétique des dissimilarités entre k et les individus composant i et j : $\delta(s,k) = [c_i\, \delta(i,k) + c_j\, \delta(j,k)] / (c_i + c_j)$. Elle correspond à un critère non pondéré dans la mesure où tous les individus jouent le même rôle.

On utilise souvent aussi une moyenne arithmétique que l'on dit pondérée, au sens où un poids différent doit être attribué aux individus des groupes pour que la moyenne ne dépende pas des effectifs des groupes et s'écrive : $\delta(s,k) = [\delta(i,k) + \delta(j,k)] / 2$. Le choix entre ces deux critères dépend de la nature de la population étudiée. Si l'ensemble relève d'un réel processus d'échantillonnage dans une structure de diversité donnée, alors le nombre d'individus dans chaque élément a un sens et doit être pris en compte dans le calcul de la dissimilarité. Si l'ensemble n'est qu'une réunion circonstancielle d'unités sans caractère de représentativité, l'effectif de chaque groupe n'est que le fait du hasard et n'a pas à intervenir dans le calcul de la diversité.

Dans le cas d'ajustement à une distance additive, on utilise plutôt la formule $\delta(s,k) = [\delta(i,k) + \delta(j,k) - \delta(i,j)] / 2$, ou son équivalent pondéré. Cette formule permet, lors de la reconstruction des arêtes, de retrouver les bonnes longueurs d'arêtes lorsque l'arbre initial est déjà une distance additive. Cette modification est sans influence sur les choix des pas ultérieurs.

Il a parfois été proposé d'utiliser, pour l'ajustement à une ultramétrique, les critères dits du lien simple et du lien complet qui correspondent respectivement à $\delta(s,k) = \min[\delta(i,k), \delta(j,k)]$ et $\delta(s,k) = \max[(i,k), (j,k)]$. Sauf cas très particulier, il est difficile de justifier ces critères dans des problèmes de diversité. Si le lien simple a été beaucoup étudié, c'est essentiellement parce qu'il entraîne des propriétés mathématiques remarquables de l'ultramétrique produite.

La reconstruction des arêtes

Les éléments i et j sont réunis pour former le nœud s de l'arbre, il reste à fixer les longueurs des arêtes $l(i,s)$ et $l(j,s)$. On peut simplement répartir également la dissimilarité $\delta(i,j)$ entre les deux arêtes : $l(i,s) = l(j,s) = \delta(i,j) / 2$. Ce mode de calcul correspond à un ajustement de la dissimilarité initiale par une ultramétrique.

Si l'on n'impose pas la condition ultramétrique, on accepte que les deux arêtes puissent avoir des longueurs différentes pour, qu'en moyenne, les dissimilarités des $n - 2$ éléments k à i et à j soient représentées au mieux. La différence des longueurs d'arêtes, notée e, est la somme sur tous les éléments k de $[\delta(i,k) - \delta(j,k)] / (n - 2)$. D'où les formules de calcul des longueurs des arêtes : $l(i,s) = [\delta(i,j) + e] / 2$ et $l(j,s) = [\delta(i,j) - e] / 2$. Ce mode de calcul correspond à un ajustement de la dissimilarité initiale par une distance additive d'arbre.

Ces formules de reconstruction des arêtes ne garantissent pas que ces estimations soient les meilleures et, par ailleurs, ces longueurs d'arêtes n'interviennent pas dans les étapes suivantes. Il est donc souvent préférable de réestimer *a posteriori* les longueurs d'arêtes lorsque la topologie de l'arbre a été établie. En effet, on sait estimer globalement, au sens des moindres carrés, la longueur de chaque arête d'un arbre de topologie donnée. Pour cela, on exprime la distance dans l'arbre entre deux individus i et j comme la somme des longueurs des arêtes appartenant au chemin de i à j et on souhaite que cette distance soit la plus proche, au sens des moindres carrés, de la dissimilarité initiale. On résout alors un système de $n(n - 1) / 2$ équations à autant d'inconnues que d'arêtes. Cette réestimation fait perdre la propriété ultramétrique.

Quelques algorithmes connus

Les algorithmes les plus connus correspondent à certaines combinaisons compatibles parmi les différentes combinaisons possibles des modes de définition du voisinage, des diverses formules de mise à jour des dissimilarités et des modes d'estimation des longueurs des arêtes.

L'ensemble des méthodes souvent qualifiées de classification ascendante hiérarchique correspond à la définition du voisinage selon la dissimilarité minimale, à un ajustement à une ultramétrique et propose les diverses formules de mise à jour. Parmi celles-ci, les formules de type moyenne ou moyenne pondérée sont les plus utilisées et les méthodes correspondantes sont fréquemment qualifiées par les abréviations anglo-saxonnes UPGMA et WPGMA pour *unweighted (ou weighted) pair group method using average.*

La méthode NJtree (pour *neighbor-joining*) proposée par Saitou et Nei (1987) est souvent utilisée en diversité génétique. Elle utilise le critère de voisinage relatif, le critère de mise à jour des dissimilarités est celui de la moyenne non pondérée et les dissimilarités sont ajustées à une distance additive d'arbre. La complexité calculatoire est de l'ordre de n^3 et permet de traiter des tableaux de

quelques centaines d'individus. De nombreuses études par simulation ont montré l'efficacité de cette méthode pour retrouver l'arbre vrai. GASCUEL (1997) propose une version UNJ *(unweighted neighbor-joining)* qui utilise un critère de moyenne pondérée. Le choix entre ces deux méthodes dépend, comme cela a déjà été signalé, de la nature de l'échantillon étudié.

SATTATH et IVERSKY (1977) proposent l'algorithme Addtree (ou méthode des scores) qui utilise la moyenne comme critère d'agrégation et fait un ajustement à une distance additive d'arbre. La notion de voisinage est celle du voisinage topologique dont on a déjà souligné l'intérêt. La complexité en n^5 est nettement plus élevée que celle de NJtree. Cette méthode, dont on a également démontré l'efficacité, pourra être préférée pour son point de vue topologique.

Arbres consensus et sous-arbres communs

Le même ensemble d'individus est fréquemment décrit par plusieurs types de marqueurs de la diversité. A chaque type de marqueur correspond une matrice de dissimilarité et un ajustement à un arbre. Chaque type de marqueur apporte une information différente, les arbres produits sont donc différents. Cependant, on peut espérer en dégager une structure forte, un arbre consensus, qui serait commune aux différents arbres.

Le problème de la comparaison d'arbres se rencontre également dans les approches de rééchantillonnage (*bootstrap* ou *jackknife*). L'objet de ces approches est de tester la stabilité des représentations arborées obtenues. Le principe est de comparer les arbres établis à partir de nombreux échantillons tirés aléatoirement dans l'ensemble des marqueurs observés (FELSENSTEIN, 1985). Si les marqueurs sont nombreux et de contenu informatif équivalent, des marqueurs moléculaires par exemple, chacun de ces échantillons est censé donner la même image de la diversité. Si tous les arbres obtenus sont très proches, on conclura à la stabilité de la structure mise en évidence. Si, au contraire, différents échantillons aléatoires des marqueurs donnent des arbres très différents, alors la structure obtenue devra être regardée avec la plus grande prudence. Une mesure de dispersion autour d'un arbre consensus synthétique quantifie les écarts entre ces arbres.

Considérant que la principale information sur la diversité relève beaucoup plus de la structure de l'arbre que de la longueur des arêtes, les méthodes présentées ici construisent des arbres synthétiques en terme de structure, sans que la longueur des arêtes soit prise en compte.

LES ARBRES CONSENSUS

Diverses méthodes de consensus ont été proposées sur des arbres racinés. Une généralisation directe aux arbres non racinés ne pose pas de problème majeur.

La méthode du consensus strict ne retient que les arêtes qui sont présentes dans tous les arbres comparés. Cette méthode pouvant être considérée comme trop sévère, il a été proposé la règle du consensus majoritaire, qui retient les arêtes présentes dans plus de 50 % des arbres. Si l'on compare deux arbres seulement, ces deux consensus sont équivalents.

Une mesure de la différence entre deux arbres peut être définie comme la somme des différences de chacun à leur arbre consensus. L'arbre consensus est toujours plus simple en terme d'organisation que les arbres initiaux. On quantifie cette organisation par un indice de complexité v, la dissimilarité entre deux arbres H et H', de consensus H_c, est alors :

$d(H,H') = v(H) + v(H') - 2v(H_c)$.

Diverses définitions de complexité peuvent être considérées. La plus habituelle est de mesurer la complexité d'un arbre par son nombre d'arêtes. Pour les arbres binaires qui n'ont que des nœuds de degré 3, la complexité de H et H' vaut $2n - 3$, la complexité est la somme du nombre d'arêtes externes, n, et du nombre d'arêtes internes, I_c, d'où $d(H,H') = 2n - 6 - 2I_c$.

La distance, dite distance d'arêtes, varie donc suivant le nombre d'arêtes internes conservées dans l'arbre consensus. Si I_c vaut 0, le consensus est une étoile, les arbres comparés ne présentent aucune structure commune, à l'opposé si I_c vaut $n - 3$, les arbres initiaux sont identiques. On normalise souvent cette distance par $2n - 6$ pour lui garder une variation entre 0 et 1.

Une interprétation de la valeur d'une distance ne peut se faire que par référence à la valeur qui serait obtenue sur des arbres tirés aléatoirement, donc sans autre structure commune que celle liée au hasard. On ne connaît la distribution de cet indice, pour des arbres aléatoires, qu'asymptotiquement, c'est-à-dire lorsque le nombre d'individus tend vers l'infini. On peut cependant estimer ces distributions, pour des effectifs finis, par simulation, et donc construire un test statistique approché. Le tableau 5 donne les valeurs seuils pour différentes valeurs de n et différents seuils de probabilité. Par exemple, pour un arbre sur 40 individus, il y a 5 % de chances seulement que le hasard donne des arbres de distance inférieure à 0,973. Cette distance apparaît très peu discriminante et on refusera le plus souvent l'hypothèse aléatoire même si les arbres n'ont en commun qu'un très petit nombre d'arêtes internes.

La distance d'arêtes porte le même jugement sur toutes les arêtes quelle que soit leur position dans l'arbre. La complexité d'un arbre ne dépend que de son nombre d'arêtes indépendamment de son organisation interne. Une idée intuitive serait de donner un poids différent aux arêtes suivant leur aptitude à « structurer » l'arbre. Par exemple, une arête externe qui concerne un seul élément est peu structurante, à l'inverse une arête qui partage l'ensemble en deux sous-groupes d'effectif $n / 2$ est la plus structurante. On associe donc à une arête un poids qui est le produit des effectifs de chacun des sous-groupes définis par cette arête. La complexité est donc la somme de ces poids sur toutes les arêtes. On exprime alors, comme précédemment, une distance, dite

de bipartitions, en fonction de la complexité des arbres et de leur consensus strict. Le maximum atteint ne s'exprime pas simplement mais peut être calculé à partir de *n*, la distance est alors normalisée par ce maximum. Comme pour la distance en arêtes, la distribution sous un modèle aléatoire n'est pas connue mais on peut l'approcher par simulation. Le tableau 6 donne les valeurs seuils pour différentes probabilités d'un test statistique approché. Il s'avère nettement plus discriminant.

Tableau 5. Distance d'arêtes entre arbres. Valeurs seuils pour différents niveaux de probabilité. Test empirique établi à partir de la simulation de 20 000 paires d'arbres aléatoires à *n* feuilles.

	Probabilité (%)			
n	1	5	10	20
20	0,882	0,941	0,941	1
40	0,973	0,973	0,973	1
60	0,982	0,982	0,982	1
80	0,987	0,987	0,987	1
100	0,990	0,990	0,990	1

Tableau 6. Distance de bipartition entre arbres. Valeurs seuils pour différents niveaux de probabilité. Test empirique établi à partir de la simulation de 20 000 paires d'arbres aléatoires à *n* feuilles.

	Probabilité (%)			
n	1	5	10	20
20	0,727	0,753	0,766	0,782
40	0,580	0,600	0,611	0,626
60	0,492	0,511	0,522	0,537
80	0,435	0,453	0,464	0,477
100	0,394	0,412	0,421	0,434

Les sous-arbres communs

Il n'est pas rare, dans certaines applications, d'obtenir des arbres consensus d'aspect très étoilé, donc sans structure marquée, alors que l'examen direct des arbres initiaux révèle des analogies de structure évidentes.

L'arbre consensus fait l'hypothèse que tous les individus sont correctement représentés, le déplacement d'un individu d'un groupe à un autre est une information forte qui dénie une réelle séparation de ces groupes. Pour de nombreuses applications cette hypothèse n'est pas entièrement acceptable. Par exemple, on a montré précédemment que les dissimilarités estimées à partir de marqueurs moléculaires pouvaient être entachées d'une certaine erreur. Si la population étudiée regroupe des unités proches génétiquement, à l'échelle infraspécifique par exemple, l'imprécision de mesure des dissimilarités peut suffire à expliquer le caractère erratique de certaines unités.

Si l'on considère que quelques individus seulement peuvent masquer une structure commune, il devient intéressant de chercher à identifier ces individus « fluctuants » plutôt que de leur laisser jouer un rôle équivalent à celui des autres dans la détermination de la structure commune. On est alors amené à reformuler le problème de la recherche des structures communes à plusieurs arbres. Le problème devient celui de la détermination du plus petit sous-ensemble qu'il suffit d'élaguer dans chacun des arbres pour obtenir des arbres identiques ou, inversement, du plus grand sous-ensemble d'individus ayant la même structure dans tous les arbres. Ces individus forment le sous-arbre commun. La figure 7 permet de comparer, sur un exemple simple, ce sous-arbre commun et l'arbre consensus.

Cette approche est beaucoup moins répandue que l'approche consensus et a fait l'objet de peu de travaux. La simplicité d'énoncé du problème dissimule un problème algorithmique relativement complexe. Récemment, KUBICKA *et al.* (1995) ont publié les bases d'un algorithme qui donne une solution exacte, tout en restant de complexité suffisamment faible pour être utilisable en pratique. L'algorithme repose sur l'énumération de toutes les solutions possibles mais en limitant sur un critère d'arrêt la profondeur d'exploration des branches d'où le gain important en complexité.

L'ordre o du sous-arbre commun, c'est-à-dire le nombre d'individus conservés, peut être considéré comme une mesure de la ressemblance entre arbres. L'ordre maximum est n, il est obtenu pour deux arbres identiques. A l'inverse la valeur minimale de l'ordre est 3 puisqu'il n'y a qu'une topologie possible de trois points distincts, elle est donc commune aux deux arbres. D'où une définition de dissimilarité entre arbres $D = n - o$, que l'on peut normaliser par $n - 3$ pour garder des variations entre 0 et 1. L'utilisation pratique de ce critère nécessite de connaître la distribution de o sous l'hypothèse d'indépendance des arbres. Cette distribution peut être approchée par simulation et a été

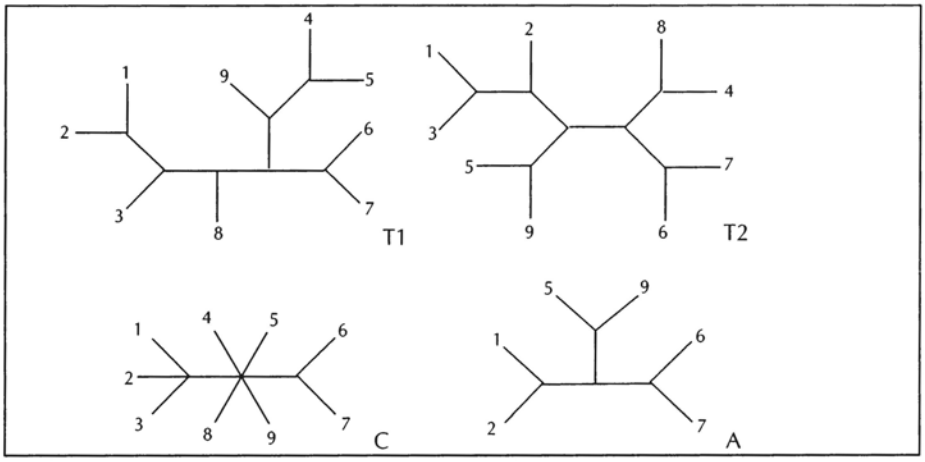

Figure 7. Arbre concensus (C) et sous-arbre commun (A) des arbres T1 et T2.

calculée pour des arbres binaires de 20 à 100 feuilles (tableau 7). Elle peut être utilisée pour fixer une règle d'interprétation. En fixant un seuil à 5 % par exemple, on considère que le hasard donnera moins de 5 fois sur 100 des arbres d'ordre supérieur ou égal à 10, 13, 16, 18 et 19 lorsque *n* varie de 20 à 100 par pas de 20.

Il faut retenir qu'arbre consensus et sous-arbre commun n'ont pas le même point de vue sur les données. Ils se justifient l'un et l'autre en fonction de l'interprétation que l'on fait des déplacements d'individus dans les arbres initiaux. S'il s'agit du déplacement accidentel de quelques individus au sein d'une structure commune forte alors le sous-arbre commun paraît le plus adapté. Si au contraire les déplacements traduisent réellement des différences de structure fondamentales alors l'arbre consensus est logiquement plus adapté. En pratique, il est rarement possible d'opter franchement pour une de ces deux hypothèses et les deux approches pourront être mises en œuvre de manière complémentaire.

Tableau 7. Distribution du nombre de feuilles des sous-arbres communs de 1 000 paires d'arbres binaires aléatoires à 20, 40, 60, 80 et 100 feuilles.

	6	7	8	9	10	11	12	13	14	15	16	17	18	19	Moyenne
20	1	40	50	9											7,67
40				5	34	43	17	0	1						10,76
60						2	30	48	17	3					12,89
80								3	25	39	24	7	2		15,13
100									2	19	36	28	11	4	16,39

Conclusion

Pour conclure ce chapitre, il faut souligner le dynamisme de ce secteur de recherche. On croyait, avec l'ouvrage de Sneath et Sokal (1973), que tout était dit, ou presque, en taxonomie numérique. Les récents développements théoriques, l'augmentation considérable des moyens informatiques, l'apparition de nouveaux types de marqueur ont complètement remis en question cette discipline, qui est actuellement un secteur de recherche très actif à l'interface entre les mathématiques et la biologie.

On aura compris à la lecture de ce chapitre la nécessaire intervention conjointe de ces deux domaines. La question naturelle qui surgit d'une telle présentation — certaines méthodes sont-elles meilleures que d'autres ? — n'a pas, dans ce domaine, de sens réel. Il n'existe pas de solution mathématique universelle, valable en toutes circonstances. Parmi la diversité des choix possibles, seule la connaissance de l'espèce végétale étudiée — niveau de diversité, mode de reproduction, hétérozygotie... —, de la nature de l'échantillon analysé, des caractéristiques des marqueurs utilisés permet de retenir les approches les plus pertinentes dans chaque contexte particulier.

Références bibliographiques

BARTHELEMY J.P., GUENOCHE A., 1988. Les arbres et les représentations des proximités. Paris, France, Masson, 239 p.

BEAULIEU F.B., 1989. A classification of presence/absence based dissimilarity coefficients. Journal of Classification, 6 : 233-246.

DARLU P., TASSY P., 1993. Reconstruction phylogénétique. Paris, France, Masson, 245 p.

CRITCHLEY F., FICHET B., 1994. The partial order by inclusion of the principal classes of dissimilarity on a finite set and some of their basic properties. *In :* Classification and dissimilarity analysis, B. van Custem éd., New York, Etats-Unis, Springer, Lecture Notes in Statistics, p. 5-65.

ESCOFIER B., PAGES J., 1993. Analyses factorielles simples et multiples : objectifs, méthodes et interprétation. Paris, France, Dunod.

ESCOUFIER Y., 1975. Le positionnement multidimensionnel. Revue de statistique appliquée, 23(4) : 5-14.

FELSENSTEIN J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution, 39(4) : 783-791.

GASCUEL O., 1997. Concerning the NJ algorithm and its unweighted version, UNJ. *In :* Mathematical hierarchies and biology, DIMACS workshop. American Mathematical Society, Series in Discrete Mathematics and Theoretical Computer Science 37, p. 149-170.

JUKES T.H., CANTOR C.R., 1969. Evolution of protein molecules. *In :* Mammalian protein metabolism, H.N. Munro éd., New York, Etats-Unis, Academic Press, p. 21-132.

KUBICKA E., KUBICKI G., MCMORRIS F.R., 1995. An algorithm to find agreement subtrees. Journal of Classification, 12 : 91-99.

LAVIT C., 1988. Analyse conjointe de tableaux quantitatifs. Paris, France, Masson.

LEFORT-BUSSON M., DE VIENNE D., 1985. Les distances génétiques, estimations et applications. Paris, France, Inra, 181 p.

PERRIER X., 1998. Analyse de la diversité génétique : mesures de dissimilarité et représentation arborée. Thèse de doctorat, université Montpellier II, Montpellier, France, 192 p.

ROHLF F.J., 1987. Ntsys-pc numerical taxonomy and multivariate analysis system. New York, Etats-Unis, Applied Biostatistics Inc., Setauket.

SAITOU N., NEI M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution, 4(4) : 406-425.

SAPORTA G., 1990. Probabilités, analyse de données et statistiques. Paris, France, Technip, 493 p.

SATTATH S., TVERSKY A., 1977. Additive similarity trees. Psychometrika, 42(3) : 319-345.

SNEATH P.H.A., SOKAL R.R., 1973. Numerical taxonomy. San Francisco, Etats-Unis, Freeman, 573 p.