Analyse à trois éléments (3ia)

Pascal Tassy, Pierre Darlu, Cyrille d'Haese

Analyse à trois éléments (3ia)

Pascal Tassy, Pierre Darlu et Cyrille d'Haese

I – Principe de l'analyse à trois éléments.

L'analyse à trois éléments ("three-taxon analysis" ou "three-item analysis" en anglais, abrégée 3ia dans ce qui suit) est une méthode phylogénétique qui s'applique aux taxons et aux aires biogéographiques. Elle se justifie par l'interprétation suivante des principes de base de la théorie cladistique :

- La 3ia est conçue comme une méthode d'argumentation des relations de parenté entre taxons. Elle permet de contrôler de façon pertinente les idées de parenté émises par un systématicien et s'exempte de celles qui seraient seulement proposées par l'application d'un algorithme.
- Les caractères sont considérés comme des arguments en faveur d'une hypothèse phylogénétique. Ils ne sont pas de simples données ou de simples observations : ce sont des hypothèses d'homologie formellement structurées, c'est-à-dire des hypothèses sur l'identité due à la parenté.
- Le but de l'analyse 3ia est de contrôler si l'hypothèse phylogénétique proposée par le systématicien peut être argumentée à partir des hypothèses d'homologie explicitées sur chacun des caractères.

L'analyse à trois éléments représente donc une forme originale de penser et de pratiquer la phylogénétique et la biogéographie. Les caractères sont considérés comme des arguments phylogénétiques (Hennig 1966 : 90) que le systématicien fournit pour défendre une hypothèse phylogénétique. Les caractères dérivent directement d'hypothèses d'homologie, la notion d'homologie utilisée ici se rapprochant de celle donnée par Richard Owen (1843 : 373). Ce dernier ne définit pas l'homologie mais un homologue. Un homologue constitue un même organe chez différents organismes, sous toute la diversité de ses formes et de ses fonctions. Peut-on dire que le bras chez l'humain est la même chose que la patte antérieure d'un chien ? Il est difficile de répondre par l'affirmative tout autant que par la négative. En revanche, dire que ces structures sont *en même temps* la même tout en étant différentes est plus satisfaisant. L'assertion prend du sens lorsqu'on affirme que le bras humain ressemble *davantage* à une patte avant de chien qu'à une aile de pigeon, c'est-à-dire lorsque l'on introduit un troisième terme.

Les homologues représentent donc des classes d'équivalence de parties d'organismes qui établissent des relations d'identité relative à d'autres classes d'équivalence (Prin 2016 : 432). Ainsi, l'ensemble « bras humain » et l'ensemble « pattes antérieures de chien » peuvent être réunis dans une classe d'équivalence plus inclusive, un nouvel homologue qui exclut l'ensemble « aile de pigeon ». Les trois membres (de l'homme, du chien et du pigeon) sont à leur tour réunis dans une nouvelle classe de degré d'équivalence, celle des membres antérieurs de tétrapodes (figure 1).

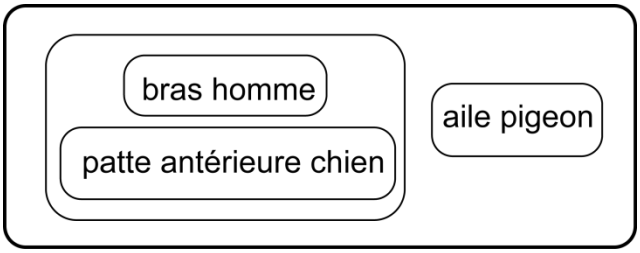


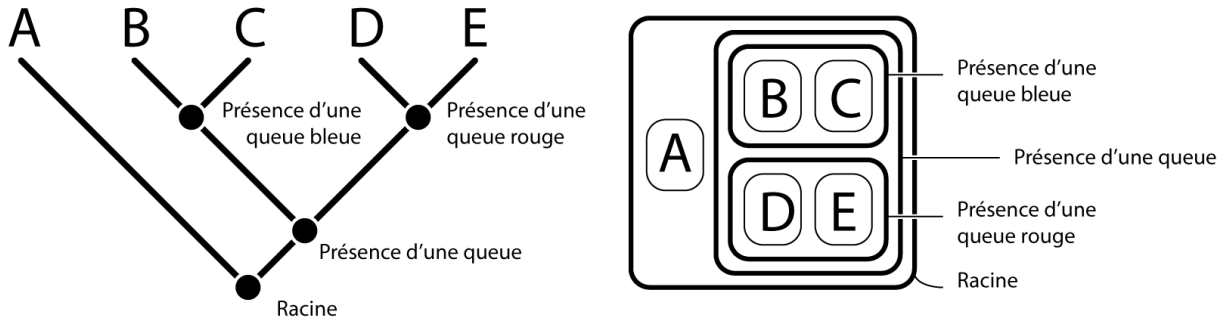
FIGURE 1. *Hypothèse d'homologie sous forme de classes de degré d'équivalence du membre antérieur de tétrapodes.*

Ces classes d'équivalence se présentent sous forme de structures d'ensembles emboités constituant une hiérarchie mathématique isomorphe d'un arbre hiérarchique raciné, c'est-à-dire possédant des nœuds (voir chapitre I.3.3). C'est cette hiérarchie ou arbre qu'on appelle, en 3ia, une *hypothèse d'homologie*.

Le systématicien propose un certain nombre d'hypothèses d'homologie, des classes d'équivalence concernant les taxons terminaux. Pour simplifier, on remplace les homologues (terminaux et internes) par les taxons qui les portent (figure 2). L'arbre hiérarchique résultant est appelé caractère, les nœuds internes correspondent aux états. Dans ce contexte, il est possible mais pas nécessaire, de spécifier un état plésiomorphe. En effet, la racine ne dérive pas nécessairement d'un homologue. Notamment, une "absence de quelque chose" peut difficilement constituer une hypothèse d'identité entre structures, c'est-à-dire un homologue.

C'est la représentation hiérarchique des hypothèses d'homologie qui permet de gérer correctement les caractères, avec n'importe quel nombre d'états.

Un exemple très connu a été proposé par Maddison (1993) : imaginons que, au sein d'un échantillonnage de cinq taxons terminaux, quatre présentent une queue (Q), attribut qui nous permet d'émettre l'hypothèse d'une proche parenté par rapport au cinquième taxon terminal qui en est dépourvu (NQ). Parmi ces quatre terminaux, deux possèdent une queue rouge QR, alors que les deux autres possèdent une queue bleue (QB). Comment représenter toute cette information sous forme d'hypothèse d'homologie qui tiendrait sur une unique colonne d'une matrice de caractères ? Cela est impossible, il faut au moins trois colonnes pour éviter toute ambiguïté (figure 2). En effet, on ne peut pas représenter l'état "présence d'une queue" parce que tous les terminaux présentant l'appendice sont aussi des instances de l'état "rouge" ou de l'état "bleue".



	C1 : Queue Q	C2 : Queue bleue	C3 : Queue rouge
A	Q=0	QB=0	QR=0
B	Q=1	QB=1	QR=0
C	Q=1	QB=1	QR=0
D	Q=1	QB=0	QR=1
E	Q=1	QB=0	QR=1

FIGURE 2. *Représentation du caractère « queue colorée » sous forme d'un arbre et d'une structure emboitée. L'approche cladistique classique ne pouvant pas représenter le caractère initial, chaque état devient un caractère indépendant dans une matrice de caractères sous contrainte (caractères polarisés 0→1). Les relations de dépendance sont perdues (le résultat final peut montrer, par exemple, que le caractère C1 n'est pas une synapomorphie alors que le caractère C2 l'est. Le codage selon la 3ia est donné Tableau 1, cas 2*

En 3ia, la représentation de ce type de caractères est simple (Figure 2). Dans l'exemple de la présence d'une queue et de sa couleur, chaque état ("queue", "rouge" et "bleue") est représenté par une classe ou un ensemble. Les classes indiquant la couleur de la queue sont incluses dans la classe représentant la présence de la queue. L'ensemble est toujours inclus dans une classe qui contient l'ensemble des terminaux, la racine. La racine n'est donc pas nécessairement un état mais, en revanche, elle a une existence : elle peut représenter un état (par exemple si on travaille avec des taxons présentant tous une queue), mais elle peut ne rien représenter. Pour résumer, les différents homologues peuvent présenter des relations définissant une hiérarchie d'ensembles emboîtés, qui peut être représentée par un arbre raciné.

II – La méthode 3ia

a) Définition des 3is

Le traitement des caractères, pour la recherche des arbres optimaux, passe par une décomposition qui peut être automatisée. Chaque caractère va être décomposé en assertions élémentaires à trois éléments, dit « 3is » (pour *three-item statements*). Un 3is représente une unité d'information relationnelle hiérarchique. La hiérarchie la plus simple indique que deux entités sont plus "les mêmes" entre elles qu'aucune ne l'est de la troisième. Un caractère sera donc décomposé en autant de relations minimales qu'il est nécessaire d'en spécifier.

Par exemple, si on émet l'hypothèse que, pour un caractère, le taxon A est plus proche du taxon B que des taxons C et D, on génère le caractère ((A B) C D). Cela revient à exprimer que A et B sont plus proches entre eux qu'aucun ne l'est de C et que A et B sont plus proches qu'aucun ne l'est de D. Le caractère ((A B) C D) sera donc décomposé, de façon équivalente, en deux 3is :

((A B) C D) ⇔ ((A B) C) ⊕ ((A B) D).
où le symbole ⇔ exprime l'équivalence et "⊕" la combinaison de « 3is ».

Le codage transpose ces combinaisons en 0/1 :
((A=1 B=1) C=0 D=0) ⇔ ((A=1 B=1) C=0) ⊕ ((A=1 B=1) D=0).

De même, un caractère tel que ((A B) (C D)) se décompose en 4 « 3is » :
((A B) (C D)) ⇔(A (C D)) ⊕ (B (C D)) ⊕ ((A B) C) ⊕ ((A B) D) .

Le nombre de 3is que génère chaque caractère dépend du nombre total (*t*) de terminaux et du nombre (*n*) de taxons terminaux regroupés par un même état informatif (le codage 1, en l'occurrence), selon la formule (Nelson et Ladiges 1992) :

$$nb(3is_{totaux}) = \frac{1}{2}n(t-n)(n-1)$$

b) la pondération fractionnaire (ou "fractional weight FW)

En appliquant la décomposition précédente, un caractère comme (A(BCD)) produit trois 3is, (A(BC)), (A(BD)) et (A(CD)). Cependant, l'information portée par ces trois 3is est partiellement redondante. En effet, la combinaison de n'importe quel des deux 3is parmi les trois permet de reconstruire le caractère initial :

(A(BCD))⇔(A(BC)) ⊕ (A(BD)) ⇔(A(BC)) ⊕ (A(CD)) ⇔(A(BD)) ⊕ (A(CD))

Ceci est dû à une propriété mathématique des hiérarchies qui impose que l'intersection entre deux entités quelconques parmi les ensembles qui les constituent est ou bien nulle, ou bien l'un de ces ensembles. La combinaison (⊕) de deux 3is partageant des terminaux dans l'ensemble informatif implique donc leur inclusion dans la même classe :

(BCD) ⇔ (BC) ⊕ (BD)
(BCD) ⇔ (BC) ⊕ (CD)
(BCD) ⇔ (BD) ⊕ (CD)

Pour résoudre ce problème de redondance, Nelson et Ladiges (1992) proposent une solution par pondération. Ils considèrent le nombre des 3is résultant de l'analyse d'un caractère comme une mesure de l'information phylogénétique portée par ce caractère. Dans ce cas, le caractère (A(BCD)) contient une information phylogénétique équivalente à celle de deux 3is et non pas de trois 3is. C'est pourquoi ils proposent de pondérer chacun de ces 3 3is par la valeur 2/3. Ils nomment cette mesure "fractional weight" (abrégé *FW* ci-après), attribuant donc une valeur informative réduite mais indépendante à chacun de ces 3is.

Cette valeur d'information d'un 3is peut être calculée selon la formule suivante, *n* étant le nombre de terminaux présents sous l'état informatif :

$$FW = \frac{2}{n}$$

La formule qui donne le nombre de 3is indépendants pour un caractère, en entendant par indépendants le nombre de 3is nécessaires pour reconstruire le caractère en absence de toute redondance est donc :

$$nb(3is_{indépendants}) = FW \times nb(3is_{totaux}) = FW \times \frac{1}{2} n(t-n)(n-1) = (t-n)(n-1) = (n-1)(t-n)$$

Exemples :
Le caractère (AB(CDEF)) produit

$$nb(3is_{totaux}) = \frac{1}{2} n(t-n)(n-1) = \frac{1}{2} \times 4(6-4)(4-1) = 12$$

$$nb(3is_{indépendants} = FW \times nb(3is_{totaux}) = \frac{2}{4} \times 12 = 6$$

Il est important de noter que la *FW* s'applique par groupes de 3is en raison de leurs dépendances propres. Dans cet exemple, les 3is résultant de la décomposition du caractère sont :

(AB(CDEF)) se décompose en :
(A(CD)), (A(CE)), (A(CF)), (A(DE)), (A(DF)), (A(EF)), (B(CD)), (B(CE)), (B(CF)), (B(DE)), (B(DF)), (B(EF))

N'importe lequel des trois 3is parmi (A(CD)), (A(CE)), (A(CF)), (A(DE)), (A(DF)), (A(EF)) permet de reconstruire (A(CDEF)), à condition que tous les terminaux se trouvent représentés[1], et n'importe lequel des trois 3is parmi (B(CD)), (B(CE)), (B(CF)), (B(DE)), (B(DF)), (B(EF)) permet de reconstruire (B(CDEF)), avec la même contrainte. Mais nous ne pouvons pas mélanger n'importe lequel des six 3is de ces deux groupes pour reconstruire le

[1] Par exemple, (A(CD)), (A(CF)), (A(DF)) ne permettent pas de classer le terminal E avec les terminaux C, D et F

caractère. Trivialement, les six qui concernent B ne permettent pas de dire quoi que ce soit quant aux relations de A : il en faut trois qui concernent A et trois qui concernent B, en plus d'une représentation complète des terminaux. Ceci aura de l'importance pour la décomposition de caractères à états multiples.

c) décomposition de caractères à états multiples

Une fois la représentation hiérarchique du caractère admise, comment décomposer ce caractère en 3is ? Ce processus nécessite l'utilisation de structures intermédiaires, les composantes (sensu Cao, Zaragüeta Bagils et al. 2007), c'est-à-dire des hiérarchies présentant une seule classe informative, La figure 3 montre l'exemple de deux cas de décomposition de caractères à états multiples et comment le contenu informatif des hypothèses dépend en partie de leur structure.

c.1. Le Cas 1 concerne un caractère dont la structure est dite pectinée : tous les nœuds, sauf le moins inclusif, mènent à un terminal et à un seul nœud interne de rang immédiatement inférieur. La décomposition se fait en trois composantes. Chaque composante dérivée possède un nœud informatif, correspondant à un état du caractère, au sein duquel les terminaux se connectent, formant une polytomie lorsqu'ils sont plus de deux. Chaque composante est décomposée en ses 3is, éventuellement avec une pondération fractionnaire. Cette opération produit une liste de 3is résultant de l'analyse de ces trois composantes. Or, on remarque que, par exemple, le 3is (A(DE)) se trouve trois fois dans la liste. Si l'on regarde le caractère original, il apparait évident que ce 3is ne devrait être présent qu'une seule fois. Deux instances de ce 3is doivent donc être éliminées. Mais lesquelles ? Seuls les 3is (A(DE)) dérivées des composantes 1 et 2 possèdent une pondération fractionnaire ; celui dérivé de la composante 3 n'en possède pas. Ceci veut dire que ce 3is dérivé de la composante 3 ne peut pas être supprimé sans créer une perte d'information, les 3is restants (B(DE)) et (C(DE)) ne permettant pas de reconstruire la composante 3 (ABC(DE)). En revanche, les instances dérivées des composantes 1 et 2 sont potentiellement redondantes. La $FW = 2/3$ des 3is de la composante 2 indiquent que seuls deux parmi les trois 3is dérivés de cette composante ayant A comme terminal sont nécessaires pour reconstruire la composante (AB(CDE)). Le 3is (A(DE)) est donc redondant. En éliminant ce 3is, la pondération fractionnaire affectant le groupe disparaît. Le même raisonnement permet d'éliminer l'instance du même 3is de la composante 1. La méthode 3ia tient donc bien compte de cette dépendance entre états : l'état 3 permet de résoudre une partie de l'incertitude liée à la pondération fractionnaire des états 1 et 2.

Nous remarquons que le 3is (B(DE)) se trouve aussi répété parmi ceux dérivés de la composante 2 et nous procédons à son élimination de la même façon que dans le cas précédent, avec la disparition subséquente de la pondération fractionnaire qui restait dans les 3is dérivés de la composante 2.

Finalement, on remarque que les 3is (A(CD)) et (A(CE)) se trouvent simultanément dans les composantes 1 et 2. Comme toute redondance a été éliminée de la composante 2, ce sont ceux dérivés de la composante 1 qui sont redondants. Nous avions 6 3is dérivés de la composante 1, avec une pondération fractionnaire de ½ et nous avons pu en éliminer 3 : il n'y a plus de *FW* à appliquer parmi les 3is restants. Ce caractère produit un total de 10 3is indépendants sans qu'une *FW* ne doive être appliquée.

Notons aussi que, comme expliqué plus haut, la composante 2 produit 4 3is contre 3 seulement pour les autres composantes. Ceci est dû au fait qu'elle concerne trois terminaux parmi les 5 présents et que sa structure est celle dont l'obtention par hasard est le moins probable. En d'autres termes, la composante 2 est la plus riche en information.

Le Cas 2 de la figure 3 présente la décomposition d'un autre caractère à cinq terminaux, parfaitement résolu. A la différence du caractère précédent, celui-ci présente un nœud

symétrique, c'est-à-dire un nœud qui contient plus qu'un autre nœud interne de rang immédiatement inférieur (indépendamment des terminaux qui y sont connectés). Il représente l'hypothèse phylogénétique de la figure 2.

En appliquant la procédure de décomposition utilisée précédemment, on remarque que seuls deux parmi les six 3is de la composante 1 peuvent être éliminés. La pondération fractionnaire doit être recalculée : puisque $nb(3is_{totaux}) = \frac{4}{2}(5-4)(4-1) = 6$ et FW=2/4, alors seuls trois 3is sont nécessaires pour reconstruire la composante. Or, on dispose maintenant de 4 3is. La nouvelle valeur de pondération fractionnaire sera donc de ¾ parce que 4 x ¾ = 3.

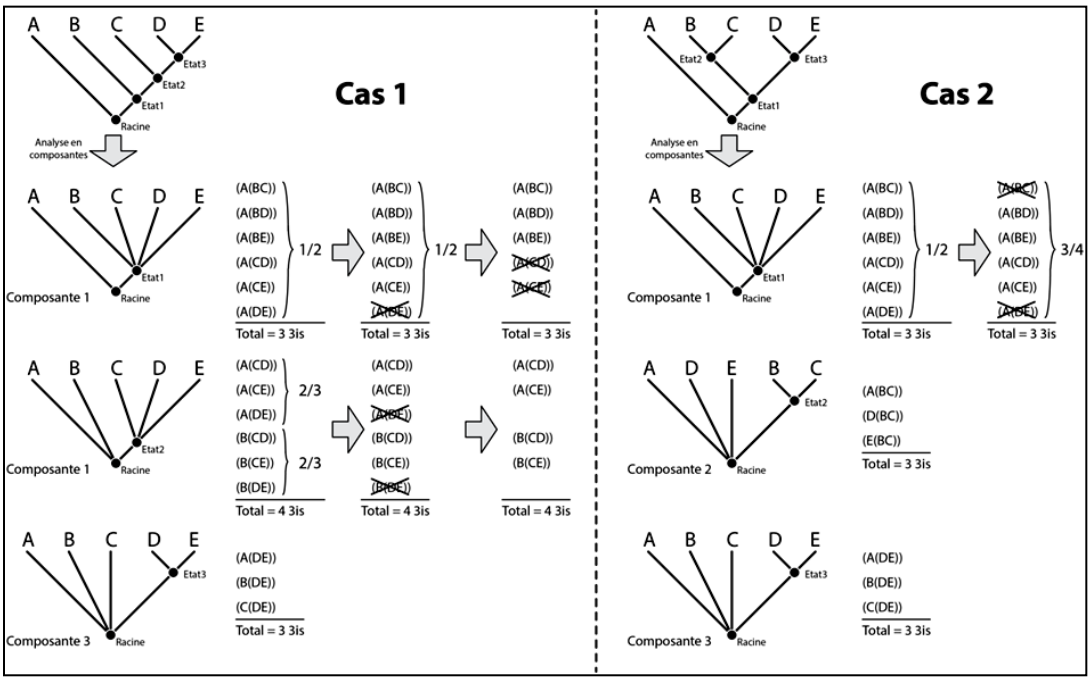


FIGURE 3. *Décomposition en 3is de deux caractères (cas 1 et 2) à cinq terminaux avec une résolution maximale et trois états informatifs.*

Cas 1	Composante 1			Composante 2				composante 3		
FW	1	1	1	1	1	1	1	1	1	1
A	0	0	0	0	0	?	?	0	?	?
B	1	1	1	?	?	0	0	?	0	?
C	1	?	?	1	1	1	1	?	?	0
D	?	1	?	1	?	1	?	1	1	1
E	?	?	1	?	1	?	1	1	1	1

Cas 2	Composante 1				Composante 2			Composante 3		
FW	3/4	3/4	3/4	3/4	1	1	1	1	1	1
A	0	0	0	0	0	?	?	0	?	?
B	1	1	?	?	1	1	1	?	0	?
C	?	?	1	1	1	1	1	?	?	0
D	1	?	1	?	?	0	?	1	1	1
E	?	1	?	1	?	?	0	1	1	1

TABLEAU 1. *Codage et pondération des 3is dans le cas 1 et 2 de la figure 3.*

c.2. Le Cas 2 montre que, lorsque des nœuds symétriques sont présents (et qu'au moins cinq terminaux sont concernés par le caractère), la pondération fractionnaire ne peut pas être complètement éliminée et qu'une certaine incertitude reste présente. Plus étonnant, alors que le caractère pectiné produit 10 3is, le caractère à un nœud symétrique n'en produit que 9, ces deux caractères ayant le même nombre de taxons terminaux et une résolution maximale

(c'est–à-dire aucune polytomie). L'explication de ce phénomène dépasse le cadre de ce texte. Cependant, il a été montré que les notions d'orthologie et de paralogie, définies dans le cadre de la génétique moléculaire (Fitch 1970) et appliquées à la morphologie (Nelson 1994), à la biogéographie historique ou à la cladistique (Nelson and Ladiges 1996), admettent des définitions faisant appel à la forme de la relation hiérarchique entre terminaux (Zaragüeta-Bagils, Lelièvre et al. 2004), qu'il s'agisse de gènes, de traits morphologiques ou d'aires biogéographiques, voire d'âges de taxons terminaux. Ainsi, les arbres hiérarchiques pectinés ne contiennent-ils que des nœuds orthologues, alors que les nœuds symétriques peuvent être assimilés à des instances de paralogie. Or, la notion de paralogie a toujours été associée à une certaine perte d'information. Si, comme nous le défendons ici, le nombre de 3is (en tenant compte de leurs *FW*) est une mesure de l'information hiérarchique, on s'attend donc à en trouver une quantité minorée lorsque des instances de paralogie sont présentes.

d) Données manquantes et données non applicables.

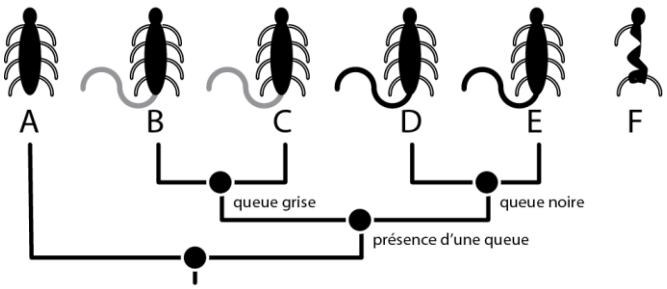
D'après Kluge (1993), la 3ia (dans ses développements initiaux, utilisant une représentation en matrice comme celle du tableau 1) produirait une distorsion dans les caractères pour deux raisons. La première concerne la multiplication de colonnes correspondant chacune à un 3is, considérée erronément par cet auteur comme une multiplication de caractères alors que les 3is sont logiquement dépendants, ce qui violerait l'un des principes de la représentation des caractères en analyse phylogénétique pour laquelle les colonnes de la matrice doivent indiquer des hypothèses indépendantes. Nous avons vu que les 3is ne sont pas des caractères, mais des hypothèses élémentaires d'homologie. Or, les méthodes utilisant des matrices violent ce même principe d'indépendance des caractères lorsqu'il existe des relations entre eux (voir par exemple la Figure 2). La deuxième critique se focalise sur l'ajout de points d'interrogation dans la matrice transformée (comme dans le tableau 1), alors qu'aucune donnée manquante n'est présente dans la matrice originale. Platnick (1993) a essayé de répondre à cette dernière critique en indiquant que les "?" de la matrice transformée ne représentent pas des données manquantes mais des données non applicables. Le problème soulevé est donc celui d'un codage différentiel pour les données manquantes et pour les données non applicables.

Revenons à l'exemple des animaux avec ou sans queue et la couleur de celle-ci pour comprendre de quoi on parle. En effet, on utilise le symbole "?" pour représenter et traiter les données manquantes et non applicables (ou d'autre symboles comme "-", "N", "NA", "*", etc). L'ensemble des algorithmes courants interprètent et traitent les données manquantes et les états non applicables de façon identique. Or une "donnée manquante" est une information en soi, qui concerne un ou plusieurs états d'un caractère d'un taxon terminal. La structure n'est observable sur aucun des spécimens disponibles pour des raisons diverses : mauvaise préservation de spécimens fossiles ou impossibilité d'accès à l'information qui peut être d'ordre technique (i.e. séquençage des spécimens conservés dans le formol), scientifique, patrimoniale (i.e. dissection de spécimen type), financier, etc.

Au contraire, les symboles indiquant la non-applicabilité d'un état signifient la présence de dépendances (par exemple la couleur de la queue pour des terminaux n'ayant pas de queue). Dans ce cas, l'état n'est pas observable sur un terminal car la présence de ce trait (c'est-à-dire ce sur quoi on émet une hypothèse au sens de Vignes-Lebbe (, c'est-à-dire ce sur quoi on émet une hypothèse, ici la couleur.2000) dépend d'un autre trait plus inclusif,, ici la présence de la queue qui dépend bien entendu de la présence de cet organe, cet état ne s'appliquant pas aux terminaux ne présentant pas de queue.

Supposons un spécimen fossile incomplet, F, (figure 4) que l'on peut attribuer, malgré son état de préservation, au taxon supraspécifique regroupant les terminaux (A à E) traités dans

l'exemple de la figure 4. Est-il possible que ce spécimen F ait eu une queue ? Certainement, cela est possible comme il est possible que cette queue soit grise ou noire. Comme on ne peut nier aucune de ces éventualités, ce manque d'information équivaut à la possibilité que n'importe lequel des états définis soit présent. En d'autres termes et pour respecter la nature des hypothèses phylogénétiques cladistiques, on ne peut rejeter aucune hypothèse face à l'impossibilité d'observer une structure.



	queue	queue noire	queue grise
EG	0	0	0
A	0	?	?
B	1	0	1
C	1	0	1
D	1	1	0
E	1	1	0
F	?	?	?

FIGURE 4.

FIGURE 4. *Représentation des données manquantes et non-applicable. Exemple du caractère hiérarchique (présence ou absence de queue. En 3ia à gauche : (A((BC)(DE)) et en parcimonie à droite, par une matrice. (voir texte)*

Supposons maintenant une série de spécimens appartenant à un taxon ne présentant pas de queue (spécimens appartenant au taxon A de la figure 4). De toute évidence, puisqu'ils n'ont pas de queue, il est définitivement impossible qu'elle soit grise ou noire. En d'autres termes, en cas d'absence d'un caractère duquel dépendent d'autres états, nous avons la certitude de l'absence de chacun des états dépendants.

La méthode de parcimonie, représentant les caractères par des matrices (figure 4, à droite) , traite les données manquantes et les états inapplicables de façon identique, alors que leur signification est différente. Les traiter de manière identique ne peut donc être qu'erroné. En analyse à 3ia, les données manquantes impliquent l'exclusion du ou des taxons pour lesquels la partie concernée par l'hypothèse d'homologie est inconnue. Ainsi, dans la figure 4, le caractère ne met en relation que cinq des six terminaux, rien ne pouvant être dit à propos de F dont la partie postérieure est manquante, d'où le codage (A((BC)(DE) qui signifie bien que rien n'est dit et par conséquent que rien n'est nié à propos de F: lorsqu'on ne dit rien on ne rejette rien. . L'utilisation d'un système de représentation hiérarchique à la place de matrices permet de faire des hypothèses partielles. En excluant les terminaux pour lesquels la partie n'est pas observable, la 3ia n'exclue aucune relation de parenté possible pour ces terminaux.

Les problématiques "?" associés aux états inapplicables (comme chez A) disparaissent aussi par la méthode 3ia, parce que les "données non-applicables" concernent des nœuds internes du caractère hiérarchique : un état n'est appliqué qu'aux taxons terminaux pour lesquels il est informatif. Dans la représentation hiérarchique, ils définissent un sous-ensemble de l'ensemble représentant l'état parent : dans l'exemple de la figure 4 l'état "queue noire" est un sous-ensemble de l'état "queue" et ne concerne donc que les terminaux présentant une queue, le terminal A étant exclu de ces états. Ainsi, les critiques de Kluge évoquées plus haut (1993) ne sont pas pertinentes pour la 3ia, alors qu'elles restent pertinentes pour les méthodes utilisant des matrices (Zaragüeta-Bagils and Bourdon 2007).

e) Recherche des arbres optimaux

Une fois que sont décrites les hypothèses d'homologie suite à la décomposition en 3is, comment passer à l'étape suivante de reconstruction d'un cladogramme ? Deux méthodes sont possibles : la recherche de l'arbre qui maximise la compatibilité des 3is (en maximisant la

somme des FW des 3is compatibles), ou la recherche de l'arbre le plus parcimonieux, c'est-à-dire celui qui minimise la longueur de l'arbre, en terme de nombre de changements au niveau des 3is.

Hennig (1966) préconise de maximiser la congruence entre caractères pour le choix ou le calcul de l'arbre optimal. La notion de congruence maximale a été historiquement interprétée de deux façons : comme la méthode de parcimonie (Farris 1970, chapitre V) ou comme la maximisation de la compatibilité entre caractères (Estabrook, Johnson et al. 1976, chapitre X). Or, dans le cas particulier de la 3ia (lorsque ces critères s'appliquent aux hypothèses d'homologie primaire, c'est-à-dire aux 3is), compatibilité et parcimonie sont deux méthodes qui ne peuvent donner que des résultats identiques puisque les 3is ne représentent des hiérarchies élémentaires du type (X=0(Y=1,Z=1)) qui ne requièrent qu'un ou deux pas pour être compatibles ou incompatibles, respectivement, avec n'importe quel arbre. (Felsenstein, 1981 ; Wilkinson 1994).

En effet, si le nombre total de caractères est *c* et le nombre de caractères compatibles avec le cladogramme formant une clique est *u*, alors la longueur de l'arbre est de $L = u+2(c-u)=2c-u$ qui est minimal quand *u* est maximal. Il en résulte que l'arbre qui maximise la compatibilité de l'ensemble des 3is est aussi l'arbre le plus parcimonieux. Il s'agit d'un résultat directement lié aux représentations des hypothèses en 3ia, un résultat particulièrement élégant montrant que les deux interprétations méthodologiques de la notion de congruence de Hennig sont identiques dans le cadre de la 3ia.

Un exemple de la procédure des 3is.

A titre d'exemple, prenons les données du tableau 3 où la morphologie de 7 espèces imaginaires sont décrites par trois caractères : les ailes, les pattes et les cornes. Ces caractères sont codés en deux états pour les ailes et les cornes et en 3 états pour les pattes, exemple tiré du manuel d'utilisation du logiciel Nelson05 (Cao et al., 2005), ancêtre de Lisbeth (Zaragüeta i Bagils, 2012)..L'organisation hiérarchique des états de ces caractères sont proposées sous formes de parenthèse emboitées comme indiqué plus haut (par exemple (1,(2)) pour les ailes). En retenant les organisations de chacun de ces caractères, une description des taxons est alors proposées en transposant, dans le système de parenthèses, les états des caractères par le ou les taxons qui les possèdent. L'étape de recherche du ou des arbres optimaux se fait à partir de ces descriptions.

Taxons	**[1] Aile**	**[2] Patte**	**[3] Corne**
A = Licorne	1 absence	1 non	1 absente
B = Yeti	2 présence	2 deux	2 présente
C = L'Eole	**(1,(2))**	3 quatre	**(1,(2))**
DD = Serpent à plume		**(1,(2,(3)))**	
E = Griffon			
FF = Pegase			
G = Ange			
Descriptions			
[**1**] (A B (C E FF G) DD)			
[**2**] (C DD (B E G (A FF)))			
[**3**] ((A) B C DD E FF G)			

Tableau 3. Exemple de codage des données selon 3ia (voir texte).

Tous les arbres dichotomiques possibles que l'on peut construire à partir de l'ensemble des taxons sont examinés. Pour chacun de ces arbres il est noté le nombre de 3is repérés comme présents et compatibles, ou comme absents et incompatibles, tout en tenant compte des pondérations fractionnaires s'il y a lieu (pas dans l'exemple du tableau 3).
Les arbres qui présentent le nombre maximum de 3is compatibles et/ou le minimum de 3is incompatibles sont sélectionnés. Comme l'examen de tous les arbres peut être une étape longue si le nombre de taxons est élevé, des méthodes de recherche par « branch and bound » peuvent améliorer la faisabilité de cette étape.

f) Mesures de soutien en 3ia.

Dans le contexte 3ia, la question se pose, comme avec toutes les autres méthodes, de la qualité de l'ajustement des caractères sur l'arbre optimal, c'est-à-dire à quel point ces caractères permettent de défendre une hypothèse phylogénétique (e.g. Farris 1969). Deux mesures principales ont été proposées en analyse cladistique : l'indice de cohérence (*IC*), et l'indice de rétention (*IR*) (voir Chapitre V). La transposition en 3ia peut être la suivante :

f.1. L'indice de cohérence.

Rappelons que l'indice de cohérence est, en parcimonie classique, le rapport I.C.=*R/L* entre le nombre minimum théorique de pas (changements 0-1) pour un ensemble *c* de caractère (*R = c)* et le nombre *L* de transformations 0-1 effectivement observées dans l'arbre.

Dans le cadre du 3ia, l'indice de cohérence serait le rapport entre deux valeurs : 1) le nombre minimal théorique de changements 0-1 nécessaire pour expliquer tous les 3is s'ils étaient compatibles avec l'arbre, ce qui par définition n'exige qu'un seul changement par 3is, soit $u=c$, c étant le nombre de 3is et u le nombre de 3is compatibles, et 2) le nombre effectif de changements 0-1 lorsque les 3is ne sont pas compatibles, soit, comme décrit plus haut, $L= u+2(c-u)=2c-u$, puisqu'un 3is non compatible change deux fois dans l'arbre, mais pas plus. Dans ces conditions, en tenant compte des *FW*, *IC* est égal à :

$$IC = \frac{c}{2c - u}$$

Rappelons que *IC* fait référence ici à un rapport de quantités de changements, par analogie avec la parcimonie classique. Une autre mesure de cohérence pourrait s'envisager dans le cadre du 3ia, simplement en faisant le rapport entre le nombre de 3is effectivement compatibles avec l'arbre et le nombre total de 3is.

$$IC = \frac{u}{c}$$

Nous allons montrer que ce rapport est celui fourni par l'indice de rétention.

f.2. L'indice de rétention.

L'indice de rétention est, selon Farris (chapitre V), égal à :

$IR = \frac{G - L}{G - R}$ où G est le nombre de pas qui seraient nécessaires si tous les états ne changeaient que sur les branches terminales (donc selon un arbre en « étoile », sans aucune structure) ; L est la longueur de l'arbre optimal et R le nombre minimal de transformations 0/1.
La transposition de *IR* dans le contexte de la 3ia nécessite de proposer un équivalent à G qui traduise l'absence de structure et demande à se placer dans la situation la moins parcimonieuse pour mesurer le coût que représente cette situation. Pour cela, considérons que

tous les 3is, au nombre de *c*, sont tels que leurs états sont tous du même type ((1,1),0) et que les synapomorphies qu'ils définissent sont toutes erronées (tout comme l'arbre est supposé sans structure, dans le cas de la parcimonie classique). Dans ce cas, le nombre de changements maximum auquel on s'attend pour chacun des 3is est de 2, l'origine étant dans l'état 0. Au total, *G* dans l'optique du 3ia serait donc équivalent à 2*c*.

Sous ces conditions, *L* étant égal à [*u*+2(*c*-*u*)] et *R à 2c* on a :

$$IR = \frac{2c - [u + 2(c-u)]}{2c - c} = \frac{u}{c}$$

c'est à dire :

$$IR = \frac{nb(3is_{compatibles})}{nb(3is)}.$$

IR représente donc la proportion de 3is compatibles avec le cladogramme (Williams, in Kitching, Forey et al. 1998), alors que l'indice de cohérence représente la proportion de 3is compatibles avec le minimum de changements 0-1. Notons cependant que cette transposition de l'indice de rétention est considérée comme n'ayant aucune signification en 3ia (Archie 1989, Farris 1989, Farris 1989, Archie 1996).

IV –Exemples comparatifs

Pour finir, voici deux exemples simples qui permettent d'illustrer les différences entre les approches 3ia et parcimonie.

a) Différences dans la transcription des observations

Imaginons des organismes appartenant à six taxons. Ceux appartenant aux deux taxons A et B sont rouges ; ceux appartenant à C et D sont verts et les membres de E et F sont bleus. Pour représenter ces observations en 3ia permettant de regrouper des taxons ensemble selon la couleur, il suffit d'un seul caractère avec trois états (l'état AB, l'état CD et l'état EF) :

((A B) (C D) (E F))

L'arbre optimal est celui dont la structure est celle de l'unique caractère :

((A B) (C D) (E F))

Les « synapomorphies » de ces taxons, *dans le contexte 3ia*, doivent être comprises comme des hypothèses sur le partage de la possession d'une couleur particulière, sans se prononcer sur la couleur d'origine.

En parcimonie, les couleurs seraient codées sur une seule colonne avec, par exemple, le codage "r" pour A et B, "v" pour C et D et "b" pour E et F. Selon la couleur du ou des groupes externes, un certain nombre d'arbres différents pourront être obtenue et leur consensus strict irrésolu :

(A B C D E F).

Ce résultat montre que le codage en parcimonie ne représente ni les observations ni les hypothèses d'homologie, mais des hypothèses supposant que ce que l'on observe correspond à une monophylie ou à une paraphylie (de Pinna 1994).

En revanche, la 3ia permet de montrer les relations de parenté entre les taxons sur la base d'hypothèses concernant l'identité des couleurs. Elle n'en génère pas d'autres que celles fournies par le systématicien, contrairement à la parcimonie, qui génère des hypothèses de monophylie ou paraphylie des états de caractère.

b) Différences en ce qui concerne la précision et l'interprétation

Le premier exemple de 3ia a été produit pour montrer ce qui, à l'époque, était considéré comme une augmentation de la précision de la méthode de parcimonie (Nelson and Platnick 1991).

L'exemple concerne trois caractères et quatre taxons A, B, C, D, plus un groupe externe X. Le taxon A ne présente aucun état qui le regroupe avec B, C ou D. Son codage est identique à celui de X, le groupe externe.

Chaque caractère groupe deux terminaux différents parmi B, C et D, mais jamais les trois ensemble (Tableau 2 ?).

	C1	C2	C3
X	0	0	0
A	0	0	0
B	1	1	0
C	1	0	1
D	0	1	1

TABLEAU 2. *Matrice de 3 caractères et 5 taxons dont un taxon externe (X)*..

La question est la suivante : cette matrice, qui peut être répétée à l'identique autant de fois que l'on voudra, contient-elle une information quelconque ? Si oui, laquelle ?

Du point de vue de la 3ia, il est évident qu'il existe une information : il n'y a aucune raison de grouper le taxon A avec les autres taxons terminaux B, C ou D. Les relations entre B, C et D ne peuvent pas être résolues car on a autant d'arguments pour grouper ensemble B et C, C et D ou B et D que de regrouper ensemble les trois taxons B, C et D, sachant que leurs relations sont hiérarchiques. L'analyse à trois éléments trouve comme solution (X A (B C D)). La parcimonie trouve plusieurs arbres équiparcimonieux dont le consensus strict est totalement irrésolu. Etonnamment, Nelson et Platnick ne s'interrogent pas sur les synapomorphies qui soutiennent ces relations. Si l'on s'en tient à trois caractères, pour simplifier : (X A D (B C)), (X A C (B D)), (X A B (C D)), lesquels soutiennent le clade (B C D) ? De toute évidence, les trois. En conclusion, aucune homoplasie n'existe dans ce cas, c'est-à-dire que nous ne pouvons pas rejeter l'identité des homologues permettant de grouper (B C), (B D) et (C D). Nous devons cependant accepter qu'il y a eu une perte de chacun de ces états dérivés pour un des terminaux.

Bibliographie à ajuster

Archie, J. W. (1989). "Homoplasy excess ratios: New indices for measuring levels of homoplasy in phylogenetic systematics and a critique of the consistency index." Systematic Zoology **38**: 253-269.

Archie, J. W. (1996). Measures of homoplasy. Homoplasy: the resurrence of similarity in evolution. M. J. Sanderson and L. Hufford. San Diego, CA, Academic Press**:** 153-188.

Cao, N., et al. (2007). NELSON05. Paris, Published by the authors.

Cao, N., et al. (2007). "A hierarchical representation of the hypothesis of homology." Geodiversitas **29**(1): 5-15.

de Pinna, M. C. C. (1994). Ontogeny, rooting and polarity. Models in phylogeny reconstruction. R. W. Scotland, D. J. Siebert and D. W. Williams. Oxford, The Systematics Association / Clarendon Press**:** 157-172.

Farris, J. S. (1969). "A successive approximations approach to character weighting." Systematic Zoology **18**: 374-385.

Farris, J. S. (1989). "The retention index and the rescaled consistency index." Cladistics **5**: 417-419.

Fitch, W. M. (1970). "Distinguishing homologous from analogous proteins." Systematic Zoology **19**: 99-113.

Hennig, W. (1966). Phylogenetic systematics. Chicago, Illinois, University of Illinois Press.

Kluge, A. G. (1993). "Three-taxon transformation in phylogenetic inference: ambiguity and distortion as regards explanatory power." Cladistics **9**: 246-259.

Maddison, D. R. (1993). "Missing data versus missing characters in phylogenetic analysis." Systematic Biology **42**: 576-581.

Nelson, G. and P. Y. Ladiges (1992). "Information content and fractional weight of three-item statements." Systematic Biology **41**(4): 490-494.

Nelson, G. J. (1994). Homology and Systematics. The hierarchical basis of comparative biology. B. K. Hall. San Diego, Academic Press**:** 101-149.

Nelson, G. J. and P. Y. Ladiges (1996). "Paralogy in Cladistic Biogeography and Analysis of Paralogy-Free Subtrees." American Museum Novitates **3167**: 1-58.

Owen, R. (1843). Lectures on the comparative anatomy and physiology of the invertebrate animals, delivered at the Royal College of Surgeons in 1843. London, Longman, Brown, Green, and Longmans.

Platnick, N. I. (1993). "Character optimization and weighting: differences between the standard and three-taxon approaches to phylogenetic inference." Cladistics **9**: 267-272.

Prin, S. (2016). The relational view of phylogenetic hypotheses and what it tells us on the phylogeny/classifi cation relation problem. The Future of Phylogenetic Systematics: The Legacy of Willi Hennig. D. W. Williams, M. Schmitt and Q. Wheeler. London, The Systematics Association Special volume series, vol 86. Cambridge University Press. **86:** 431-468.

Zaragüeta-Bagils, R. and E. Bourdon (2007). "Three-item analysis: Hierarchical representation and treatment of missing and inapplicable data." Comptes Rendus Palevol.

Zaragüeta Bagils, R., et al. (2012). "LisBeth: New cladistics for phylogenetics and biogeography." Comptes Rendus Palevol **11**: 563-566.

Archie, J. W. (1996). Measures of homoplasy. Homoplasy: the resurrence of similarity in evolution. M. J. Sanderson and L. Hufford. San Diego, CA, Academic Press**:** 153-188.

Cao N., Ducasse L & Zaragueta Bagils R. 2005 *NELSON05*. Publié par les auteurs, Paris.

de Pinna, M. C. C. (1994). Ontogeny, rooting and polarity. In R. W. Scotland, D. J. Siebert and D. W. William (eds). Models in phylogeny reconstruction. The Systematics Association / Clarendon Press, Oxford, pp. 157-172.

Prin, S. (2016). The relational view of phylogenetic hypotheses and what it tells us on the phylogeny/classifi cation relation problem. The Future of Phylogenetic Systematics: The Legacy of Willi Hennig. D. W. Williams, M. Schmitt and Q. Wheeler. London, The Systematics Association Special volume series, vol 86. Cambridge University Press. **86:** 431-468.

Zaragüeta-Bagils, R. and E. Bourdon (2007). "Three-item analysis: Hierarchical representation and treatment of missing and inapplicable data." Comptes Rendus Palevol 6 : 527-534.

Zaragüeta Bagils E, Ung V., Grand A. Vignes-Lebbe R., cao N. & Ducasse J. 2012 Lisbeth: a new cladistics for phylogenetics and biogeography. *Comptes Rendus Palevol* 11 : 563-566.

Archie, 1989

Archie, 1996

Cao, et al., 2007

Farris, 1969

Farris, 1989

Fitch, 1970

Hennig, 1966

Kluge, 1993

Maddison 1993

Nelson, 1994

Nelson et Planick, 1991

Nelson et Ladiges, 1992

Nelson et Ladiges, 1996

Owen, 1843

Platnick, 1993