Introduction générale aux méthodes comparatives phylogénétiques

Yves Desdevises

Introduction générale aux méthodes comparatives phylogénétiques

Yves Desdevises

Sorbonne Université, CNRS, Biologie Intégrative des Organismes Marins, BIOM, Observatoire Océanologique, F-66650 Banyuls/Mer, France

Résumé

La méthode comparative phylogénétique (PCM pour *Phylogenetic Comparative Method*) est maintenant un outil classique en biologie de l'évolution, où son utilisation est allée croissante depuis le début des années 1980. Cela se matérialise par la diversité d'approches disponibles actuellement. Cet article explique brièvement les concepts et utilisations des PCMs principales, en tant que méthodes d'analyse des relations entre traits phénotypiques, ou entre phénotype et environnement, dans un ensemble d'espèces liées par leurs relations phylogénétiques dont la prise en compte est au cœur des PCMs. Seul le cas des variables quantitatives est abordé ici. Les deux grands types d'approches, avec ou sans prise en compte d'un modèle explicite d'évolution phénotypique, sont présentées, en particulier la méthode des contrastes indépendants, la Régression Phylogénétique des Moindres carrés Généralisés, ou PGLS (*Phylogenetic Generalized Least Squares*) et la PVR (*Phylogenetic Eigenvector Regression*). Le futur des PCMs verra sans doute une incorporation toujours plus grande des méthodes Bayesiennes, et la prise en compte de modèles évolutifs de plus en plus complexes, parfois tenant compte simultanément de plusieurs arbres phylogénétiques.

Abstract

The Phylogenetic Comparative Method (PCM) is now a classical tool in evolutionary biology, where it has been increasingly used since the beginning of the 80's. This is reflected is the wide diversity of approaches available today. This article briefly explains the concepts and uses of the main PCMs, referred here as analytical methods of the links between phenotypic traits, or between phenotypic and environmental traits, in a set of species, taking into account their phylogenetic relationships. The consideration of species phylogeny is at the heart of PCMs. Only quantitative variables are addressed in this article. The two main kinds of approaches, with or without an explicit phenotypic

evolutionary model, are presented, in particular the independent contrasts method, the PGLS (*Phylogenetic Generalized Least Squares*) regression and the PVR (*Phylogenetic Eigenvector Regression*). Future PCMs will certainly see the incorporation of more Bayesian approaches with more complex evolutionary models, sometimes considering simultaneously several phylogenetic trees.

Introduction

Depuis son introduction formelle dans les années 1970 et surtout 1980 en biologie de l'évolution, la méthode comparative phylogénétique (PCM, pour *Phylogenetic Comparative Method*) n'a cessé de gagner du terrain et d'être utilisée par la communauté des chercheurs en biologie de l'évolution bien sûr (voir Fig. 1 dans Cooper et al., 2016), mais pas seulement. Cela ne s'est cependant pas fait d'une manière instantanée, et il n'était pas rare pendant les 2 voire 3 décennies suivantes de devoir expliciter la nécessité de prendre en compte l'inertie phylogénétique lors de la comparaison de variables mesurées sur plusieurs taxons. Cela arrive encore.

Il convient d'abord de définir ce qu'on appelle « méthode comparative » en biologie de l'évolution. Pendant longtemps (Felsenstein, 1985), cela s'appliquait aux situations dans lesquelles on désirait étudier la relation entre des traits phénotypiques en prenant comme objets un ensemble de taxons, le plus souvent des espèces (niveau taxonomique qui sera donc utilisé dans cet article), avec pour objectif de rechercher des coadaptations, des compromis évolutifs (*trade-offs*) ou de tester des hypothèses fonctionnelles. Les relations évolutives entre ces espèces les rendant non indépendantes, celles-ci, via leur phylogénie, doivent être prises en compte lors de l'analyse (d'où l'appellation de méthode comparative « phylogénétique »). Cette non-indépendance est liée au concept d'*inertie* phylogénétique, qui se reflète dans la présence de *signal* phylogénétique (la tendance des espèces proches à se ressembler, voir plus bas) qui peut être explicitement pris en compte par certaines PCMs. Si les variables explicatives sont des descripteurs environnementaux, alors on étudie comment le phénotype est lié à l'environnement, ce qui conduit les PCMs à être également utilisées pour étudier l'adaptation (Martins, 2000). L'intérêt de les utiliser est maintenant bien établi, mais il n'en a toujours pas été de même (Leroi et al., 1994).

Cependant, la notion de PCM a pris récemment un sens plus large, incluant d'autres contextes analytiques dans lesquels il est indispensable de prendre en compte explicitement les relations phylogénétiques des organismes, comme l'étude de la diversification taxonomique (Cooper et al., 2016 ; Cornwell & Nakagawa, 2017). Ainsi, cette définition englobe parfois toute situation analytique dans laquelle la phylogénie des espèces est considérée et les « PCMs comprise a collection of statistical methods for inferring *history* from piecemeal information, primarily combining two types of data: first, an estimate of species relatedness, usually based on their genes, and second,

contemporary trait values of extant organisms »[1] (Cornwell & Nakagawa, 2017). Notez que les relations entre espèces ne nécessitent pas d'être basées sur des « gènes » mais comme il est largement préférable de disposer de distances évolutives fiables (e.g. longueurs de branches) entre organismes, les données moléculaires sont spécialement pertinentes ici.

Dans ce volume, nous nous tiendrons à la définition originelle des PCMs, soit l'étude de l'évolution corrélée entre traits phénotypiques ou entre traits et variables environnementales, en prenant bien sûr en compte les relations évolutives entre les espèces considérées, le plus souvent à travers leur phylogénie. Nous ne considèrerons également que le cas des variables quantitatives, mais il existe également de nombreuses approches pour prendre spécifiquement en compte des variables qualitatives ou binaires (e.g. Maddison, 1990 ; Maddison, 2000 ; Pagel, 1994 ; Paradis & Claude, 2002 ; Hadfield & Nakagawa, 2010), sachant que les méthodes présentées ici peuvent parfois s'en accommoder dans une certaine mesure.

Il y a encore peu de temps, considérer dans des analyses les relations évolutives entre espèces n'était pas aisé à cause du peu d'arbres phylogénétiques disponibles dans la littérature scientifique. S'il n'était pas possible de collecter soi-même des données pour cela, on pouvait aller jusqu'à utiliser la taxonomie pour refléter autant que possible les proximités entre espèces dans ce cadre hiérarchique (Garland et al., 2005). Pendant les deux dernières décennies, le nombre de données disponibles, en particulier moléculaires, et le développement de techniques permettant de combiner efficacement des arbres phylogénétiques issus de la littérature (les « superarbres », voir Bininda-Emonds, 2014) ont rendu bien plus facile de disposer de cette information pour mettre en œuvre les PCMs.

Ces PCMs sont de plus en plus nombreuses, et depuis l'article clé de Felsenstein en 1985 (Felsenstein, 1985), beaucoup de méthodes, souvent dérivées les unes des autres, ont été publiées.

On utilise généralement les PCMs afin de tester des hypothèses de *causalité*, c'est-à-dire quand on cherche à savoir si une ou un ensemble de variable(s) explicative(s) (p. ex. environnementales ou phénotypiques) sont à l'origine de la variation d'une ou plusieurs

[1] « Les PCMs se composent d'un ensemble de méthodes statistiques pour inférer l'*histoire* à partir de diverses informations, combinant en priorité deux types de données : premièrement une estimation de la proximité entre espèces, généralement basée sur leurs gènes, et deuxièmement des valeurs actuelles pour des traits chez des espèces existantes. »

variables réponses phénotypiques, mesurée(s) sur un ensemble d'espèces. Ces méthodes sont plus rarement utilisées dans un cadre prédictif pour lequel on recherche plutôt à maximiser la variation expliquée (donc le R^2), ce qui ne rend pas forcément pertinent de « retirer » l'influence de la phylogénie. En outre, la qualité des résultats obtenus à l'aide des PCMs est très liée à celle des phylogénies utilisées (dont les zones d'incertitude ne sont en général pas prises en compte explicitement) et des données récoltées (parfois obtenues dans la littérature scientifique, ce qui peut les rendre de précisions inégales).

Illustrons la méthode comparative avec un exemple (Figure 1) : on connaît à l'heure actuelle plusieurs dizaines de virus à ADN double brin infectant des eucaryotes photosynthétiques (« algues ») unicellulaires, classés dans la famille des Phycodnaviridae. On ne connaît que des virus lytiques dans de tels hôtes, et chaque infection virale conduit à la production de nombreux virions, un nombre appelé *burst size* (BS). On imagine facilement que la sélection tend à maximiser cette BS, elle est directement liée au nombre de copies du virus, donc de ses gènes, qui vont être libérées à chaque cycle de réplication. On peut aussi facilement imaginer que cette BS est liée à la taille de la cellule hôte infectée (plus exactement au ratio taille virus/taille cellule hôte, mais comme on étudie ici des particules virales qui font sensiblement la même taille, on peut donc considérer uniquement la taille de la cellule hôte). En d'autres termes, combien de virus peut-on au maximum faire « entrer » dans la cellule hôte ? Pour tester cette hypothèse, il suffit d'étudier la relation BS = f(taille hôte) chez les Phycodnaviridae pour lesquels ces données sont disponibles. Ce test donne une relation forte et très significativement différente de 0 ($r = 0{,}62$; $P = 0{,}008$). Cependant, les virus proches phylogénétiquement tendent à avoir des BS proches. Si on teste cette relation en contexte phylogénétique via la méthode PGLS (décrite plus bas), on constate que la relation n'est plus significative ($P = 0{,}721$) : la non prise en compte de la phylogénie des virus nous a conduit à voir une relation là où il n'y en avait pas (soit à commettre ce qu'on appelle en statistique une erreur de Type 1, ou de première espèce).

A l'heure actuelle, il existe de nombreuses méthodes d'analyse comparative (souvent dérivées les unes des autres), et beaucoup de logiciels pour les utiliser, souvent sous la forme de *packages* R (R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/). Cela rend encore délicat leur utilisation pour les non spécialistes, qui

doivent choisir la (ou les) méthode appropriée à leurs données, et réussir à l'implémenter et à en interpréter les résultats, en tenant des biais et limites de ces méthodes. Il n'est pas question ici de faire une revue exhaustive des méthodes existantes mais d'en présenter les grands types et leurs principes.

Principaux types de méthodes comparatives

Les méthodes d'analyse comparative peuvent ou non incorporer explicitement un modèle d'évolution phénotypique, formant ainsi deux grandes catégories d'approches. Une controverse existe sur la pertinence d'utiliser des méthodes sans modèle d'évolution phénotypique (Freckleton et al., 2011 ; Diniz-Filho et al., 2012), mais les deux types d'approches peuvent être vues comme complémentaires, ayant des limites et des applications différentes. Les méthodes incorporant un modèle d'évolution sont actuellement de loin les plus utilisées.

Commençons par les méthodes qui utilisent un modèle. Un tel modèle estime comment la variation phénotypique se propage le long de la phylogénie. Cette variation peut dépendre entièrement de la phylogénie (donc la divergence phénotypique entre espèces est directement proportionnelle à leur distance phylogénétique, ce qui revient à dire que plus une espèce est proche de son ancêtre, plus elle en est proche phénotypiquement pour le caractère considéré), ou ne lui être liée qu'en partie (quand une part de la variation du trait peut diverger dans un sens ou dans l'autre indépendamment de la position phylogénétique de l'espèce, éventuellement sous l'action d'un facteur environnemental), voire pas du tout (la variation du trait entre les espèces n'a rien à voir avec leurs divergence évolutive). Il existe différents modèles théoriques pour rendre compte de ces dynamiques.

Modèles

Le modèle le plus simple est celui qui approxime le mouvement Brownien, soit le mouvement aléatoire de particules dans un fluide ; ce modèle Brownien sera appelé ici BM. Dans ce modèle, les valeurs du trait chez deux espèces qui divergent après spéciation évoluent aléatoirement le long des deux nouvelles branches (Figure 2a). Ainsi, la divergence attendue entre les valeurs du trait pour ces deux espèces est en moyenne directement proportionnelle à leur degré de divergence, donc au temps si les longueurs

de branches le représentent bien (Figure 2b). Ce modèle est bien sûr utile pour approximer les situations de dérive génétique, mais également certaines situations de sélection stabilisante (Hansen, 1997). L'utilisation de ce modèle a été critiquée pour l'étude de l'adaptation (Leroi et al., 1994 ; Martins, 2000) où on peut penser que par définition un cadre de dérive génétique n'est pas approprié, mais plutôt une convergence de la valeur du trait vers un optimum. C'est entre autres pour cette raison que l'utilisation d'un modèle plus complexe a été proposée, le modèle de Ornstein-Uhlenbeck (OU, Hansen, 1997). Ce modèle superpose au BM un paramètre qui conduit la valeur des traits dans les différentes espèces à converger vers un optimum (Figure 3), ce qui semble plus réaliste pour tester des hypothèses d'adaptation. En faisant varier les paramètres de ces modèles, on peut en construire de plus complexes, par exemple en permettant plusieurs optimums (« pics ») différents dans deux régions de l'arbre avec un modèle OU (Butler and King, 2004). Le modèle *Accelerating/Decelerating* (ACDC, Blomberg et al., 2003), quant à lui, fait varier la valeur du paramètre de taux d'évolution du BM le long de l'arbre. De cette approche dérive le modèle *Early Burst* (EB, Harmon et al., 2010) dans lequel les valeurs divergent tôt et rapidement, puis se stabilisent. Le but de ces différents modèles est de rendre compte au mieux de l'inertie phylogénétique qui conduit les valeurs de traits phénotypiques à être d'autant plus semblables qu'elles concernent des espèces proches, avec des modalités différentes pour chaque cas particulier. Cela renvoie à la notion de signal phylogénétique.

Signal phylogénétique

La notion de signal phylogénétique reflète l'idée que des espèces se ressemblent d'autant plus (i.e. possèdent des valeurs proches pour un caractère phénotypique) qu'elles sont phylogénétiquement proches (Figure 4). Une absence de signal pour un trait signifie que ses valeurs dans les différentes espèces varient librement dès après un évènement de spéciation, sans plus de lien avec la valeur du trait chez l'ancêtre immédiat (on parle de labilité) ; de tels traits sont de bons candidats à l'adaptation. En revanche, des traits possédant un signal fort ont des valeurs qui peuvent être au moins en partie prédites directement à partir de la phylogénie des espèces qui les portent.

Il existe différents types de mesures de signal phylogénétique (Münkemüller et al., 2012), dont certains incorporent un modèle d'évolution phénotypique. L'estimateur le plus connu est sans doute λ (*lambda*) de Pagel (Pagel, 1999). Ce coefficient varie entre 0

(pas de signal phylogénétique) et 1 (où l'évolution phénotypique dépend directement de la divergence évolutive, soit un modèle BM). Le K de Blomberg (Blomberg et al., 2003) permet quant à lui des valeurs supérieures à 1, où les espèces se ressemblent encore davantage que cela est prédit par un BM. Il existe d'autres coefficients qui prennent en compte différents types de dynamiques évolutives (taux d'évolution variable, divergence rapide ou tardive, …) aboutissant à un lien plus ou moins fort des valeurs des traits en considérant l'arbre phylogénétique. Il est possible d'estimer les valeurs de ces paramètres en maximum de vraisemblance, en trouvant la valeur qui s'ajuste le mieux aux données et même en testant si elle est significativement différente de 0. Attention : la présence d'un signal significativement différent de 0 dans un caractère ne signifie pas qu'il y a un signal phylogénétique dans la corrélation de ce caractère avec un autre (même si les deux contiennent individuellement un signal). Pour un caractère, la mesure du signal peut par exemple être utile pour l'estimation d'états de caractères ancestraux ou étudier les tempos évolutifs.

Méthodes comparatives

Il existe différentes PCMs qui incorporent des modèles d'évolution phénotypiques. Toutes se basent sur une matrice de variance-covariance qui matérialise la structure de l'arbre, donc la dépendance phylogénétique des espèces entre elles (Figure 5). Dans une telle matrice, les covariances représentent l'histoire phylogénétique partagée des espèces depuis la racine de l'arbre, les variances (diagonale) matérialisent le chemin évolutif que chaque espèce a suivi depuis la racine (Figure 5).

La première est la méthode des contrastes indépendants, proposée en 1985 par Joe Felsenstein (Felsenstein, 1985), dans un article clé qui a proposé la première véritable PCM et a ainsi véritablement donné le coup d'envoi à cette discipline de la biologie de l'évolution. La méthode des contrastes indépendants (FIC pour *Felsenstein Independent Contrasts*) se base sur l'hypothèse que l'évolution phénotypique suit un modèle BM. Dans ce cas, la divergence entre les valeurs des traits chez deux espèces est directement proportionnelle au temps évolutif depuis lequel elles ont divergé, approximé par la somme des longueurs de branches qui les sépare. Cette approche est basée sur l'idée que si les valeurs des traits pour chaque espèce ne sont pas indépendantes, au niveau de chaque nœud les divergences mesurées pour un trait sont indépendantes des divergences mesurées pour le ou les autre(s) nœuds (Figure 6). Ce sont ces divergences

qui constituent les contrastes, et pour mieux comprendre la méthode, on peut se dire que chaque nœud, suivi d'une spéciation, représente une « expérience » de la nature, répétée indépendamment aux autres nœuds. Il s'agit de voir par exemple si à chaque fois qu'un trait A génère un contraste positif entre deux espèces, un trait B génère pour ces mêmes espèces un contraste également positif. Cette répétition d'« expériences naturelles », observée indépendamment à chaque nœud, permet de soutenir l'hypothèse d'une association positive entre ces traits. L'utilisation d'un BM permet également d'estimer la valeur des traits aux nœuds non terminaux en remontant vers la racine de l'arbre, et ainsi de calculer les mêmes contrastes à chaque nœud interne. Pour n espèces et un arbre parfaitement résolu, on obtient n-1 contrastes (sans perte de puissance statistique car les régressions doivent se faire par l'origine, sans terme constant (Legendre & Desdevises, 2009)). Notez que la méthode FIC incorpore un facteur de correction des contrastes calculés pour les valeurs estimées aux nœuds afin de prendre en compte l'incertitude générée. Avant d'être utilisés pour des analyses statistiques, comme des régressions linéaires, qui requièrent l'homoscédasticité, les contrastes doivent être standardisés, c'est-à-dire divisés par leur écart-type. Comme cela a été dit plus haut, le présupposé du BM implique que la variance de chaque contraste est égale à la somme des longueurs de branches entre les espèces « contrastées », ainsi la standardisation implique simplement de les diviser par la racine carrée de cette somme. Les contrastes (standardisés) peuvent ensuite être utilisés pour toutes sortes d'analyses, qui sont en général des régressions linéaires.

D'autres méthodes permettent d'utiliser des modèles plus complexes que le BM, comme le OU. L'approche la plus connue et sans doute la plus utilisée est la Régression Phylogénétique par les Moindres carrés Généralisés, ou PGLS (*Phylogenetic Generalized Least Squares*, (Grafen, 1989 ; Pagel, 1994 ; Martins, 1994). Ici la matrice de variance-covariance, contenant l'information phylogénétique, est utilisée pour corriger le terme d'erreur de la régression, qui ne contient donc pas des résidus indépendants et normalement distribués comme dans une régression des moindres carrés classique, mais avec une structure phylogénétique connue que la méthode peut prendre en compte. En appliquant différentes transformations à la matrice de variance-covariance, la PGLS permet d'accommoder différents modèles, comme les BM, OU, ACDC, dont le choix peut s'effectuer en sélectionnant la régression la mieux ajustée aux données, par exemple en utilisant le Critère d'Information d'Akaike (AIC). On peut également incorporer

explicitement un paramètre de signal phylogénétique, comme λ, qui prend en compte le signal existant dans la corrélation entre les caractères étudiés en servant de coefficient multiplicateur des covariances dans la matrice de variance-covariance. L'intérêt de la GLS est sa versatilité, elle englobe la méthode FIC si on paramètre un BM mais contrairement à celle-ci ne nécessite pas de considérer des valeurs uniques (souvent des moyennes) pour chaque espèce. Elle permet ainsi de considérer la variabilité intraspécifique des traits, avec la facilité d'usage de la régression classique.

Une autre catégorie de méthodes ne fait pas appel explicitement à un modèle d'évolution phénotypique, mais a pour objectif d'extraire des données la part de la variation phylogénétique (« inertie ») et la part non héritable dans laquelle on va chercher un support à des hypothèses d'adaptation. Ces méthodes prennent en compte la non-indépendance des données à travers l'utilisation d'une matrice d'association phylogénétique, utilisant une mesure de dissimilarité évolutive comme la distance patristique (Figure 7) calculée à partir d'un arbre phylogénétique, donc une mesure matérialisant la distance évolutive qui sépare les espèces. Cette fois ce n'est pas le chemin évolutif commun entre les espèces qui est considéré comme dans une matrice de variance-covariance, mais bien la quantité d'évolution qui les sépare. On peut simplement prendre la somme des longueurs de branches entre les espèces, mais aussi les distances génétiques (si on dispose par exemple d'un alignement de séquences) ou d'autres types de mesures.

La première de ces méthodes est basée sur une application des techniques utilisées pour prendre en compte la structure spatiale des données en écologie, à travers des indices d'autocorrélation spatiale, comme le I de Moran (voir Legendre, 1993). Ici, il s'agit d'autocorrélation phylogénétique (Cheverud et al., 1985 ; Gittleman & Kot, 1990). Cette approche dite autorégressive n'est actuellement plus utilisée, mais des méthodes apparues par la suite (Diniz-Filho et al., 1998), considérées comme plus performantes, ont proposé d'extraire l'information sur la non indépendance des données contenues dans ces matrices de distance à l'aide d'une autre technique : l'analyse en coordonnées principales (PCoA pour *Principal Coordinate Analysis*, voir Legendre & Legendre, 2012). Une PCoA correspond à une Analyse en Composantes Principales (ACP, voir Legendre & Legendre, 2012) effectuée sur une matrice de distance. Pour visualiser de façon simple comment fonctionne une PCoA, alors qu'une ACP va permettre de générer un nouveau

repère (les axes principaux, formés par des combinaisons des différentes variables) dans lequel seront placés les objets (« points ») originaux, la PCoA, qui ne travaille pas sur les variables mesurées sur les objets mais seulement sur les distances qui les relient, va permettre d'obtenir des coordonnées pour ces objets (ici les espèces) qui à l'origine n'en possèdent pas. En effet, au niveau phylogénétique chaque espèce n'est définie que par la distance qui la sépare des autres. Ils obtiennent ici des coordonnées (Figure 8) qui vont ensuite pouvoir être utilisées comme des variables (les coordonnées principales, PC, représentant la phylogénie si elles sont issues d'une matrice de distance patristique) dans n'importe quelle analyse statistique. Cette méthode a été nommée PVR pour *Phylogenetic Eigenvector Regression*. Comme on obtient jusqu'à n-1 PCs pour n espèces, le nombre de variables devient rapidement trop élevé (en particulier si on y ajoute d'autres variables, e.g. environnementales) pour des analyses statistiques qui deviennent alors surparamétrées. Il est donc nécessaire de sélectionner les PCs à utiliser pour représenter la phylogénie ; on perd ainsi une partie de l'information, mais cela n'empêche pas ces techniques de fonctionner correctement (Diniz-Filho et al., 2012). Plusieurs approches ont été proposées pour opérer cette sélection (Diniz-Filho et al., 1998 ; Desdevises et al., 2003), une des plus simple étant d'utiliser le modèle du bâton brisé (*broken stick model*). On peut maintenant étudier comment le (ou les) trait phénotypique réponse est lié à la phylogénie (PCs) et à d'autres traits phénotypiques et/ou des variables environnementales. On utilise pour cela les outils classiques de la régression multiple, linéaire ou non, qui permettent de quantifier explicitement la part due à la phylogénie de la part « non héritable », par exemple pour étudier l'adaptation. Il est possible d'aller un peu plus loin et d'affiner les interprétations en quantifiant la part de la variation expliquée à la fois par la phylogénie et l'environnement (Figure 9), qui correspond à ce qu'on appelle le conservatisme phylogénétique de niche (Westoby et al., 1995 ; Münkemüller et al., 2015). Cela passe par une série de régressions dont on combine ensuite les R^2 (Desdevises et al., 2003) pour en extraire la variation expliquée par la phylogénie et celle expliquée par une autre composante comme l'environnement, à travers un partitionnement de la variation phénotypique. On peut même partitionner ainsi plus de deux composantes (Cubo et al., 2008), et décomposer la variation d'un trait en composantes historique, fonctionnelle et structurale.

Il existe d'autres méthodes, plus complexes, aboutissant à décomposer la variation phénotypique en différentes fractions, comme le modèle phylogénétique mixte (PMM)

de Michael Lynch (Lynch, 1991 ; Housworth et al., 2004) qui partitionne la variation en composantes héritable et non historique en utilisant des calculs de génétique quantitative. Ces méthodes ont récemment été appliquées à un contexte Bayesien (Hadfield & Nakagawa, 2010) permettant de les rendre plus versatiles.

Comme cela a été mentionné plus haut, l'utilisation de méthode sans modèle explicite d'évolution phénotypique est débattue. Cependant, une méthode est apparue récemment, basée sur la transformation d'une matrice de distances évolutives en coordonnées principales, tout en y incorporant un modèle d'évolution. Il s'agit de la méthode des *Phylogenetic Eigenvector Maps* (PEM, Guénard et al., 2013), qui jette un pont entre les deux types de méthodes comparatives.

Le futur

Cet article ne fait que survoler l'éventail des méthodes comparatives actuellement appliquées (voir Garamszegi, 2014 ; Cooper et al., 2016 ; Cornwell & Nakagawa, 2017), et de nouvelles sont proposées continuellement, pour différents types de données (quantitatives, qualitatives, binaires). La quantité toujours plus importante de données disponibles, notamment de données moléculaires qui permettent d'élaborer des phylogénies toujours plus grandes et plus précises, permet une exploration de plus en plus fine des patrons et processus d'évolution du vivant.

Les développements présents et à venir des PCMs sont notamment centrés sur l'incorporation de modèles d'évolution phénotypique plus réalistes donc plus complexes, et hétérogènes, c'est-à-dire qui peuvent varier dans l'arbre phylogénétique pour représenter des modalités d'évolution différentes.

Dans la plupart des PCMs, l'arbre phylogénétique est considéré comme une donnée, et les résultats de l'analyse dépendent ainsi de sa précision, sa solidité, et bien sûr de la justesse avec laquelle il représente les relations évolutives réelles entre les espèces étudiées. Or les phylogénies sont souvent élaborées avec un degré d'incertitude qu'il est important pouvoir prendre en compte dans les PCMs (de telles approches émergent, dans un cadre Bayesien). Egalement, les PCMs, en particulier celles qui incorporent un modèle d'évolution phénotypique, utilisent des arbres phylogénétiques, c'est-à-dire une hypothèse évolutive que l'évolution suit un chemin fait d'une succession de divisions binaires sans hybridation. Or, pour de nombreuses entités biologiques, en particulier les

microorganismes tels que les bactéries et les virus, il est fort probable qu'un arbre phylogénétique ne reflète pas la complexité de leurs relations évolutives (McInerney et al., 2008 ; Koonin & Dolja, 2014) et que leurs relations évolutives soient bien représentées par des réseaux phylogénétiques. Il est donc nécessaire que des PCMs soient développées (Bastide et al., 2018) pour pouvoir utiliser ce type d'information phylogénétique (qui ne pose cependant pas de problème pour les méthodes basées sur une matrice de distances évolutives).

Les PCMs décrites jusqu'ici se concentrent sur un groupe d'organismes afin d'en disséquer les mécanismes évolutifs. Cependant, il est fréquent que plusieurs groupes de espèces soient en interaction étroite, comme chez les associations symbiotiques (e.g. hôte-parasite) où les processus évolutifs et co-adaptatifs deviennent si entremêlés, à travers des patrons cophylogénétiques (e.g. Bellec et al., 2014), qu'il est critique de considérer simultanément leurs arbres phylogénétiques pour des analyses comparatives duales. De telles méthodes comparatives « cophylogénétiques » sont nécessaires et commencent tout juste à être développées (Adams & Nason, 2018).

La profusion de méthodes existantes peut laisser les biologistes non spécialistes en difficulté quant au choix de la meilleure approche comparative à utiliser pour un problème particulier. Si l'utilisation des méthodes elles-mêmes ne pose en général pas de problème tant l'utilisation de logiciels d'analyse est devenue courante, Cooper et al. (2016) mentionnent avec pertinence qu'il manque des outils accessibles pour permettre aux non experts de choisir la méthode la plus appropriée au problème à traiter, de s'assurer que les conditions de son utilisation sont bien remplies, et d'interpréter les résultats en tenant compte des limites de la méthode utilisée. Cela permettrait une utilisation toujours plus pertinente des PCMs et une connaissance toujours plus fine des dynamiques évolutives au sein de l'arbre de la vie.

Références

Adams D.C. & Nason J.D., 2018. A phylogenetic comparative method for evaluating trait coevolution across two phylogenies for sets of interacting species. *Evolution* 72 : 234-243.

Bastide P., Solís-Lemus C., Kriebel R., Sparks K.W. & Ané, C., 2018. Phylogenetic comparative methods on phylogenetic networks with reticulations. *Systematic Biology*. 67 (5): 800-820.

Bellec L., Clerissi C., Edern R., Foulon E., Simon N., Grimsley N. & Desdevises, Y., 2014. Cophylogenetic interactions between marine viruses and eukaryotic picophytoplankton. *BMC Evolutionary Biology* 14 : 59.

Bininda-Emonds O., 2014. An introduction to supertree construction (and partitioned phylogenetic analyses) with a view toward the distinction between gene trees and species trees, *in* L.Z. Garamszegi (ed.), *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology*, Springer, 49-76.

Blomberg S. & Garland T. Jr, 2002. Tempo and mode in evolution: phylogenetic inertia, adaptation and comparative methods. *Journal of Evolutionary Biology* 15 : 899-910.

Blomberg S., Garland T. Jr & Ives A., 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* 57 : 717-745.

Butler M.A. & King A.A., 2004. Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution. *The American Naturalist* 164 : 683-695.

Cheverud J.M., Dow M.M. & Leutenegger W., 1985. The Quantitative Assessment of Phylogenetic Constraints in Comparative Analyses: Sexual Dimorphism in Body Weight Among Primates. *Evolution* 39 : 1335-1351.

Cooper N., Thomas G.H. & FitzJohn R.G., 2016. Shedding light on the "dark side" of phylogenetic comparative methods. M*ethods in Ecology and Evolution* 7 : 693-699.

Cornwell W. & Nakagawa S., 2017. Phylogenetic comparative methods. *Current Biology* 27 : R333–R336.

Cubo J., Legendre P., de Ricqles A., Montes L., de Margerie E., Castanet J. & Desdevises Y., 2008. Phylogenetic, functional, and structural components of variation in bone growth rate of amniotes. *Evolution & Development* 10 : 217-227.

Desdevises Y., Legendre P., Azouzi L. & Morand S., 2003. Quantifying phylogenetically structured environmental variation. *Evolution* 57 : 2647-2652.

Diniz-Filho J.A.F., Bini L.M., Rangel T.F., Morales-Castilla I., Olalla-Tárraga M.Á., Rodríguez M.Á. & Hawkins B.A., 2012. On the selection of phylogenetic eigenvectors for ecological analyses. *Ecography* 35 : 239-249.

Diniz-Filho J.A.F., Ramos de Sant'Ana C. & Bini L., 1998. An eigenvector method for estimating phylogenetic inertia. *Evolution* 52 : 1247-1262.

Felsenstein J., 1985. Phylogenies and the comparative method. *The American Naturalist* 125 : 1-15.

Freckleton R.P., Cooper N. & Jetz W., 2011. Comparative Methods as a Statistical Fix: The

Dangers of Ignoring an Evolutionary Model. *The American Naturalist* 178 : E10-E17.

Garamszegi L.Z. (ed.), 2014. *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology*. Springer.

Garland T., Bennett A.F. & Rezende E.L., 2005. Phylogenetic approaches in comparative physiology. *Journal of Experimental Biology* 208 : 3015–3035.

Gittleman J.L. & Kot M., 1990. Adaptation - Statistics and a Null Model for Estimating Phylogenetic Effects. *Systematic Zoology* 39 : 227-241.

Grafen A., 1989. The Phylogenetic Regression. *Philosophical Transaction of the Royal Society of London, B, Biological Sciences* 326 : 119-157.

Guénard G., Legendre P. & Peres-Neto P., 2013. Phylogenetic eigenvector maps: a framework to model and predict species traits. *Methods in Ecology and Evolution* 4 : 1120-1131.

Hadfield J.D. & Nakagawa S., 2010. General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of Evolutionary Biology* 23 : 494-508.

Hansen T.F., 1997. Stabilizing selection and the comparative analysis of adaptation. *Evolution* 51 : 1341-1351.

Harmon L.J., Losos J.B., Jonathan Davies T., Gillespie R.G., Gittleman J.L., Bryan Jennings W., Kozak K.H., McPeek M.A., Moreno-Roark F., Near T.J., Purvis A., Ricklefs R.E., Schluter D., Schulte Ii J.A., Seehausen O., Sidlauskas B.L., Torres-Carvajal O., Weir J.T., & Mooers A.Ø., 2010. Early bursts of body size and shape evolution are rare in comparative data. *Evolution* 64 : 2385-2396.

Housworth E.A., Martins E.P., & Lynch M., 2004. The phylogenetic mixed model. *The American Naturalist* 163 : 84-96.

Koonin E.V. & Dolja V.V., 2014. Virus world as an evolutionary network of viruses and capsidless selfish elements. *Microbiology and Molecular Biology Reviews* 78 : 278-303.

Legendre P., 1993. Spatial Autocorrelation: Trouble or New Paradigm? *Ecology* 74 : 1659-1673.

Legendre P. & Desdevises Y., 2009. Independent contrasts and regression through the origin. *Journal of Theoretical Biology* 259 : 727-743.

Legendre P. & Legendre L., 2012. *Numerical Ecology*. Elsevier.

Leroi A.M., Rose M.R. & Lauder G.V., 1994. What does the Comparative Method Reveal

About Adaptation? *The American Naturalist* 143 : 381-402.

Lynch M., 1991. Methods for the Analysis of Comparative Data in Evolutionary Biology. *Evolution* 45 : 1065-1080.

Maddison W., 2000. Testing Character Correlation using Pairwise Comparisons on a Phylogeny. *Journal of Theoretical Biology* 202 : 195-204.

Maddison W.P., 1990. A method for testing the correlated evolution of two binary characters: are gains or losses concentrated on certain branches of a phylogenetic tree? *Evolution* 44 : 539-557.

Martins E., 2000. Adaptation and the comparative method. *Trends in Ecology & Evolution* 15 : 296-299.

Martins E.P., 1994. Estimating the Rate of Phenotypic Evolution from Comparative Data. *The American Naturalist* 144 : 193-209.

McInerney J.O., Cotton J.A. & Pisani D., 2008. The prokaryotic tree of life: past, present... and future? *Trends in Ecology & Evolution* 23 : 276-281.

Münkemüller T., Boucher F.C., Thuiller W. & Lavergne S., 2015. Phylogenetic niche conservatism - common pitfalls and ways forward. *Functional Ecology* 29 : 627-639.

Münkemüller, T., Lavergne, S., Bzeznik, B., Dray, S., Jombart, T., Schiffers, K. & Thuiller W., 2012. How to measure and test phylogenetic signal. *Methods in Ecology and Evolution* 3 : 743-756.

Pagel M., 1999. Inferring the historical patterns of biological evolution. *Nature* 401 : 877-884.

Pagel M., 1994. Detecting Correlated Evolution on Phylogenies - a General-Method for the Comparative-Analysis of Discrete Characters. *Proceedings of the Royal Society of London B: Biological Sciences* 255 : 37-45.

Paradis E. & Claude J., 2002. Analysis of comparative data using generalized estimating equations. *Journal of Theoretical Biology* 218 : 175-185.

Westoby M., Leishman M. & Lord J., 1995. On misinterpreting the “phylogenetic correction.” *Journal of Ecology* 83 : 531-534.

Légendes des figures

Figure 1 : Exemple d'utilisation des PCMs. a. virus (dont 2 sont identifiés par « v ») émergeant de la microalgue verte unicellulaire *Ostreococcus tauri*. b. *Burst size* (nombre de particules virales émise lors d'une lyse) en fonction de la taille de la cellule hôte chez 17 espèces de Phycodnavirus : cette relation est très significative à l'aide d'une régression linéaire classique (ligne continue), mais ne l'est pas en contrôlant la non-indépendance des données à travers la phylogénie (c.) de ces virus (ligne pointillée)

Figure 2 : a. Modèle Brownien : simulation de l'évolution d'un caractère quantitatif chez deux espèces A et B après spéciation. Les caractères évoluent aléatoirement mais leur divergence tend à augmenter avec le temps. La moyenne du trait reste la même, sa variance augmente proportionnellement à la divergence. b. Divergence phénotypique entre A et B en fonction du temps

Figure 3 : Modèle de Ornstein-Uhlenbeck. Le phénotype des espèces A et B converge vers un optimum, et la divergence phénotypique après spéciation tend à se stabiliser.

Figure 4 : Signal phylogénétique : a. signal fort, b. pas de signal

Figure 5 : Matrice de variance-covariance dérivée la phylogénie de 5 espèces A-E (les chiffres représentent les longueurs de branches). Les valeurs hors diagonale sont les covariances, obtenues à partir de la quantité d'évolution commune aux espèces impliquées, c'est-à-dire le chemin évolutif effectué par leur ancêtre commun (identifié par les flèches pour 2 exemples, entre A et C, et entre B et D). La diagonale contient les variances, c'est-à-dire la distance qui sépare chaque espèce de la racine de l'arbre. Cette matrice représente la ressemblance attendue entre les valeurs des traits pour ces espèces

Figure 6 : Méthode des contrastes indépendants. La taille corporelle est-elle liée au diamètre de l'œil dans un clade de téléostéens ? Les valeurs aux espèces (e.g. entre A et C) sont proches car les espèces ne sont pas indépendantes, mais les différences au niveau de chaque nœud le sont : par exemple à partir du nœud *F* dont dérivent A et C, on calcule une différence de +4 pour la taille entre ces espèces, liée à une différence de +3 pour l'œil ; similairement, une différence de +2 entre les tailles de B et D s'accompagne d'une différence de +1 pour le diamètre de l'œil. Cela suggère une relation car ces deux « expériences naturelles » sont indépendantes. Le calcul des contrastes (différences entre espèces) peut également s'effectuer en remontant vers la racine (contrastes en

pointillés). On peut en effet estimer les valeurs des traits aux ancêtres (l'incertitude liée à cette estimation est prise en compte dans le calcul) en utilisant le modèle d'évolution Brownien sur lequel se base la méthode. Avant d'être utilisés pour des analyses, les contrastes doivent être standardisés par leur écart-type (√(somme des longueurs de branches), voir Figure 3).

Figure 7 : Matrice d'association phylogénétique dérivée la phylogénie de 5 espèces A-E (les chiffres représentent les longueurs de branches). Les valeurs sont les distances patristiques entre espèces, soit la somme des longueurs de branches qui les séparent (identifié par les flèches pour 2 exemples, entre A et C, et entre B et D)

Figure 8 : Extraction de l'information phylogénétique par une analyse en coordonnées principales (PCoA) : de la phylogénie (a.) est extraite la matrice de distance patristique (b.), les distances entre espèces sont les seules informations ici (c.), dont la PCoA (d.) permet de récupérer des « coordonnées phylogénétiques » (*PCs*, dont seules quelques-unes sont représentées sur les 2 premiers axes principaux, mais on peut utiliser davantage d'axes). Les *PCs* sont ensuite utilisées comme variables représentant la phylogénie dans des analyses statistiques

Figure 9 : Partitionnement de la variation phénotypique en composantes phylogénétique (a), environnementale (c), et la fraction commune (b) entre les deux, d représentant la variation non expliquée. La fraction commune peut être interprétée comme le *conservatisme phylogénétique de niche*, soit les adaptations développées par des espèces-sœurs sous l'effet des mêmes contraintes de l'environnement.

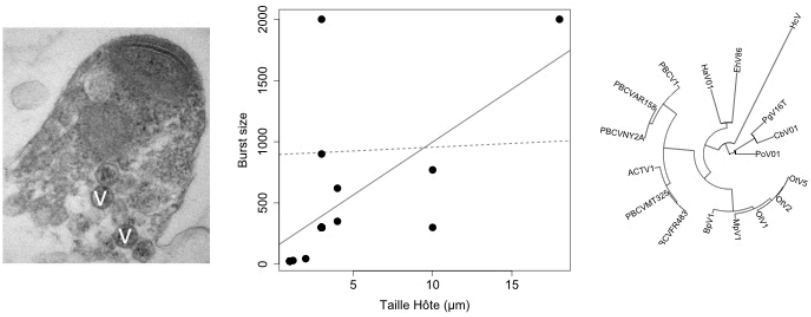
V
V
Burst size
0
500
1000
1500
2000
5
10
15
Taille Hôte (µm)
HcV
EhV86
HaV01
PBCV1
PBCVAR158
PBCVNY2A
PgV16T
CtV01
PoV01
ACTV1
PBCVMT325
CVFR483
BpV1
HaV11
OV1
OV2
OtV5
a.
b.
c.

Figure 1

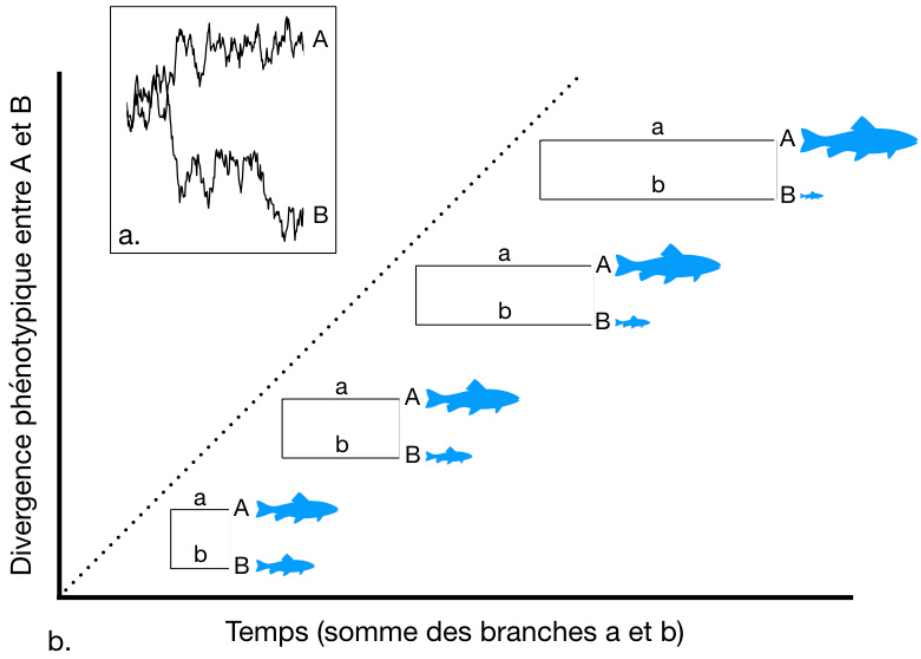
A
B
a.
Divergence phénotypique entre A et B
a
b
A
B
Temps (somme des branches a et b)
b.

Figure 2

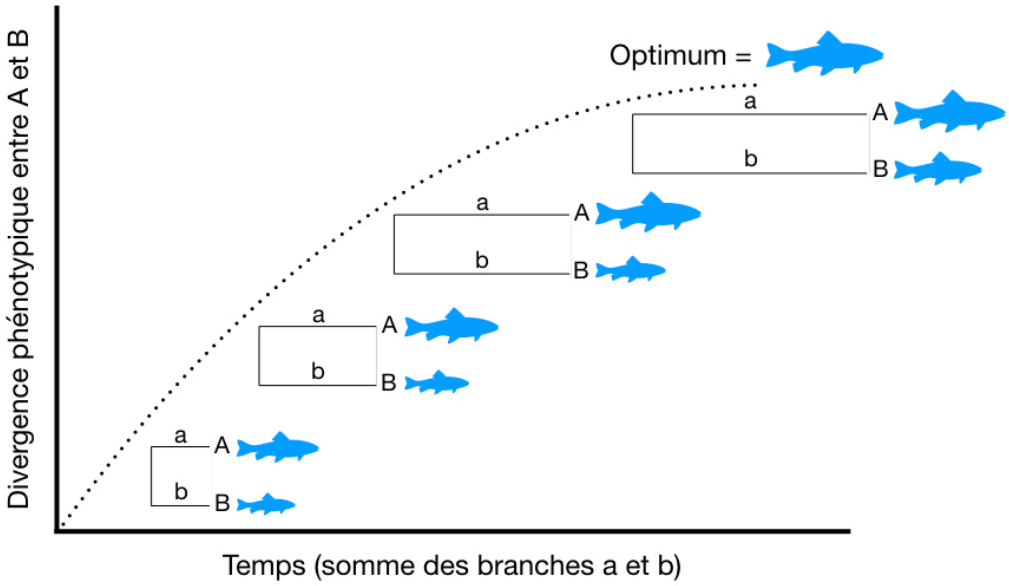
Optimum =
Divergence phénotypique entre A et B
Temps (somme des branches a et b)
a
b
A
B

Figure 3

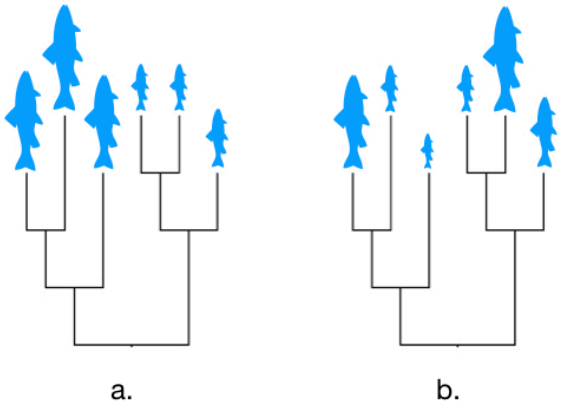

Figure 4

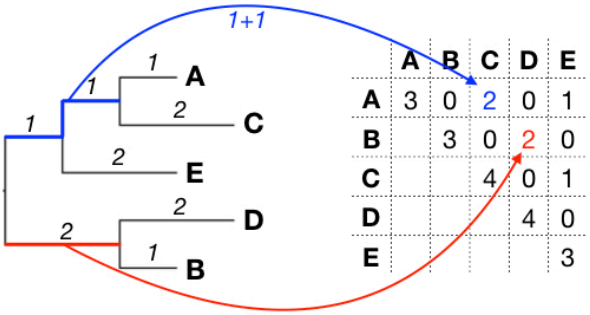

Figure 5

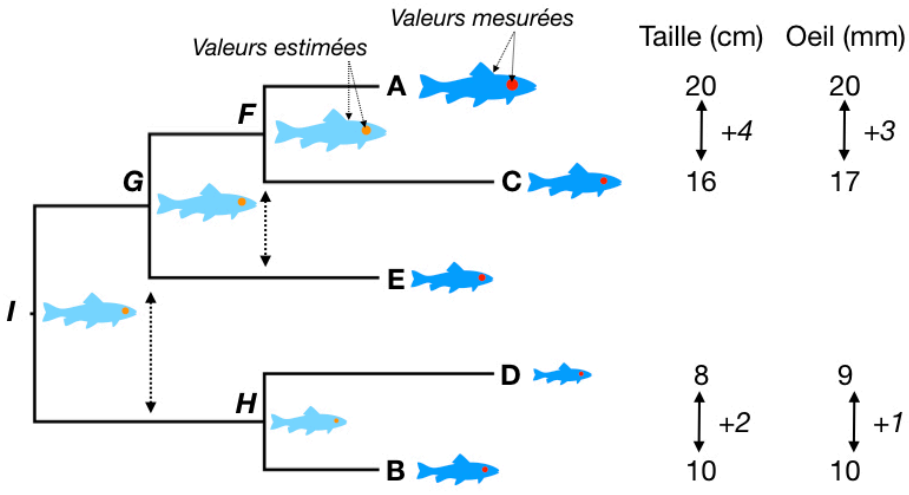
Valeurs mesurées
Valeurs estimées
Taille (cm)
Oeil (mm)
A
F
G
C
E
I
D
H
B
20
+4
16
20
+3
17
8
+2
10
9
+1
10

Figure 6

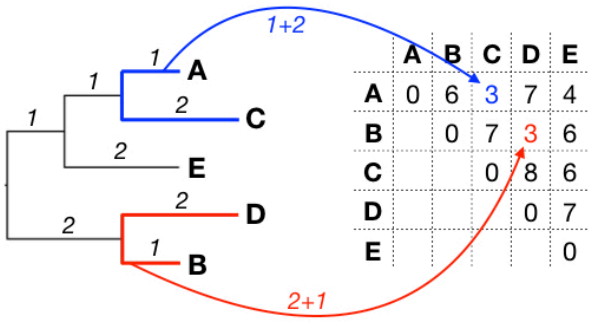

Figure 7

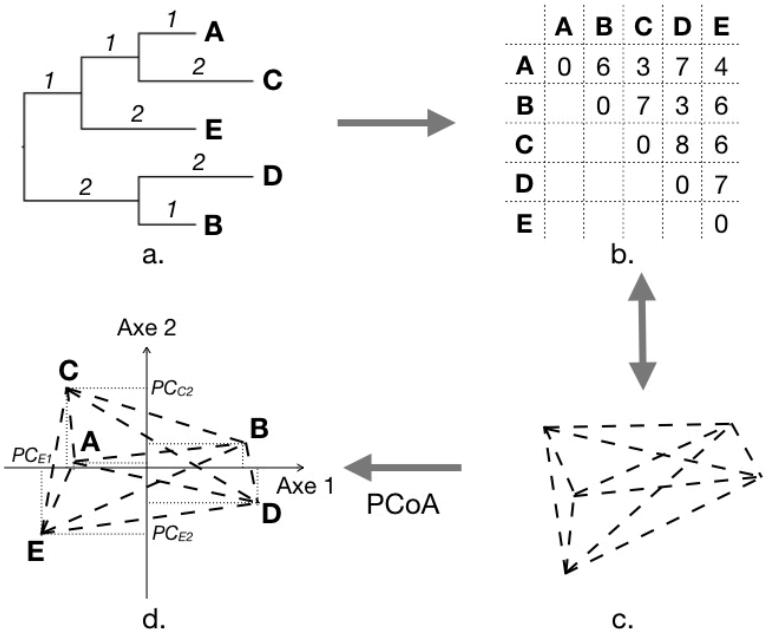
1
A
1
2
C
1
2
E
2
D
2
1
B
a.
A B C D E
A 0 6 3 7 4
B 0 7 3 6
C 0 8 6
D 0 7
E 0
b.
c.
Axe 2
C
PCC2
B
A
PCE1
Axe 1
D
PCE2
E
PCoA
d.

Figure 8

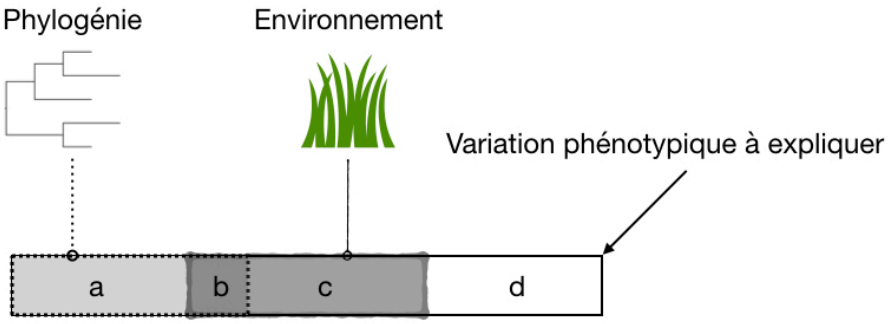
Phylogénie
Environnement
Variation phénotypique à expliquer
a
b
c
d

Figure 9