
UNE CLARIFICATION LOGIQUE ET MATHÉMATIQUE NÉCESSAIRE DE L'INFÉRENCE PHYLOGÉNÉTIQUE EN SVT

Damien AUBERT¹

Résumé. Du cycle 3 aux classes préparatoires, les élèves sont conduits à construire des arbres phylogénétiques à partir de matrices binaires de présence/absence de caractères en SVT. Pris au sérieux, cet exercice est un authentique problème d'optimisation combinatoire : il s'agit de chercher, parmi tous les arbres possibles, ceux qui minimisent le nombre total de transformations de caractères le long de leurs branches. Or les recettes actuellement enseignées ne résolvent pas ce problème. Elles introduisent des asymétries arbitraires, confondent l'inférence de la topologie et l'enracinement de l'arbre, et ne produisent les résultats attendus que sur des matrices soigneusement choisies.

Cet article explicite d'abord le cadre mathématique minimal de l'inférence phylogénétique par parcimonie. Il montre ensuite en quoi la méthode des groupes emboîtés des cycles 3 et 4 et les méthodes par polarisation à priori du lycée et des classes préparatoires s'écartent de ce cadre et pourquoi elles produisent des arbres généralement incorrects.

Deux méthodes alternatives cohérentes avec le problème posé sont enfin présentées : l'addition pas à pas, qui compare explicitement les positions d'insertion au regard de la fonction objectif, et l'analyse de compatibilité, qui extrait un squelette d'arbre non conflictuel avant d'effectuer des optimisations locales sur les éventuelles polytomies restantes. L'ensemble conduit à plaider pour une plus grande interdisciplinarité : l'enseignement scolaire de la phylogénétique gagnerait en cohérence si les outils mathématiques qu'il mobilise étaient formulés plus explicitement et introduits progressivement à la fois dans les programmes de mathématiques et de SVT.

Mots-clés. Arbre de Steiner, matrice binaire, distance de Hamming, théorie des graphes, interdisciplinarité, phylogénie, parcimonie.

Introduction

Les arbres phylogénétiques sont des représentations de l'histoire évolutive du vivant. À ce titre, ils occupent une place centrale en SVT. Par ailleurs, leur construction mobilise directement des domaines mathématiques variés comme la théorie des graphes, l'optimisation combinatoire et l'algorithmique (Hwang, Richards et Winter, 1992 ; Felsenstein, 2004). L'ambition de cet article est de porter sur les pratiques scolaires actuelles en SVT un regard critique éclairé par les mathématiques, et de montrer que celles-ci pourraient jouer un rôle essentiel dans l'amélioration de l'enseignement de l'inférence phylogénétique, à condition d'assumer pleinement la dimension inter-

disciplinaire de ce problème, c'est-à-dire de mobiliser le cadre mathématique pertinent pour atteindre cet objectif biologique.

L'analyse proposée ici n'a pas pour but de caricaturer ou de moquer une discipline. Elle vise à rappeler une exigence minimale : une procédure ne peut être qualifiée de cohérente que si elle s'applique aux cas généraux et qu'elle respecte les propriétés fondamentales (fonction objectif, symétries et invariances principales) du problème qu'elle prétend résoudre. Le fait qu'elle fournisse parfois la bonne réponse sur des exemples favorables ne suffit pas à la valider. Deux exemples simples permettent de préciser cette idée.

¹ damien.aubert@ac-clermont.fr

UNE CLARIFICATION LOGIQUE ET
MATHÉMATIQUE NÉCESSAIRE DE
L'INFÉRENCE PHYLOGÉNÉTIQUE EN SVT

En géométrie, si on affirme que, pour tout triangle, la somme des carrés des deux plus petits côtés est égale au carré du plus grand, on énonce un résultat vrai dans un cas particulier (triangle rectangle) mais faux en général. Le fait de ne proposer ensuite aux élèves que des triangles rectangles ne rend pas l'énoncé correct, il masque simplement le domaine de validité.

En logique, considérons le raisonnement suivant :

- (1) Socrate est mortel.
- (2) Tous les hommes sont mortels.
- (3) Donc Socrate est un homme.

Les prémisses sont vraies, la conclusion est vraie, mais l'inférence est invalide. Là encore, s'exercer uniquement sur des exemples où la conclusion tombe juste ne transforme pas une inférence invalide en inférence valide. Cela entraîne seulement à confondre la vérité des énoncés et la validité de la déduction.

Comme nous le verrons, les exercices de construction d'arbres phylogénétiques en SVT soulèvent des problèmes du même ordre : un algorithme de résolution peut sembler fonc-

tionner sur des cas soigneusement choisis, tout en ne respectant pas, en général, la structure du problème qu'il est censé résoudre.

1. – Un exemple de situation d'enseignement en SVT

Pour comprendre d'où l'on part, commençons par décrire une situation typique telle qu'elle apparaît dans des activités de SVT au collège (nous verrons plus loin les variantes propres au lycée). On fournit, ou on fait construire par l'observation, un tableau de présence/absence de caractères (des attributs biologiques héréditaires) dans une collection d'espèces. Il est important de noter que même lorsqu'on laisse les élèves compléter le tableau par eux-mêmes, les espèces et les caractères restent toujours fixés par l'enseignant, un point sur lequel insistent les inspecteurs (Lecointre et Camus, 2007 ; Hervé, 2009). Dans l'exemple du tableau 1 ci-dessous, on considère cinq espèces de vertébrés et quatre caractères facilement observables. Ce tableau est équivalent à une matrice binaire où les blancs seraient remplacés par des 0 et les croix par des 1.

	C_1 - Mâchoire	C_2 - Membres à doigts	C_3 - Griffes	C_4 - Plumes
L_1 - Lamproie marine				
L_2 - Truite européenne	X			
L_3 - Petite grenouille verte	X	X		
L_4 - Lézard des murailles	X	X	X	
L_5 - Moineau domestique	X	X	X	X

Tableau 1 : Les caractères d'une collection d'espèces.

À partir de ces données, on demande aux élèves de construire des groupes emboîtés représentant les relations de parenté entre ces espèces. La « méthode des groupes emboîtés » (Chanet et Lusignan, 2010 ; IMT Atlantique,

2023, p. 51) présentée par l'enseignant consiste d'abord à repérer le caractère partagé par le plus grand nombre d'espèces et à dessiner un ensemble regroupant ces espèces. Il s'agit en l'occurrence de la mâchoire, on

forme donc un groupe $\{L_2, L_3, L_4, L_5\}$ et la lamproie L_1 reste à l'extérieur. On regroupe ensuite au sein de l'ensemble ainsi formé les espèces possédant le caractère le plus partagé suivant. On se sert donc à ce stade des membres à doigts pour construire le groupe $\{L_3, L_4, L_5\}$, qui exclut la truite européenne

L_2 . On procède ainsi jusqu'à avoir épuisé tous les caractères.

On obtient alors des « groupes emboîtés », c'est-à-dire un diagramme de Venn, représentant les relations de parenté entre les espèces ainsi classées (cf. figure 1).

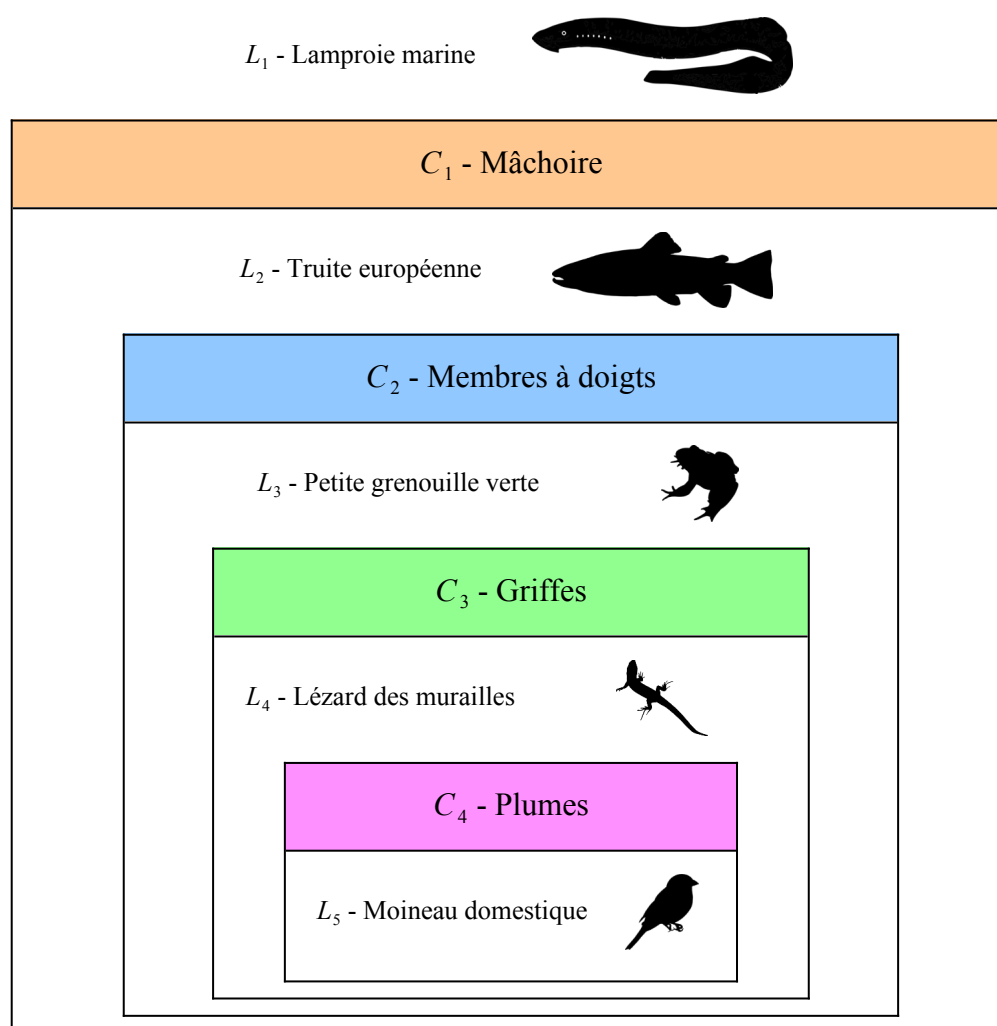


Figure 1 : Groupes emboîtés construits à partir de la matrice de caractères du tableau 1. Silhouettes téléchargées sur <https://www.phylopic.org> (CC0 : *Petromyzon marinus*, *Salmo trutta*, *Passer domesticus* ; CC BY 4.0 Luca Leicht *Pelophylax lessonae* ; CC BY 4.0 Titouan Montessuit : *Podarcis muralis*).

On en fait l'interprétation suivante : si deux espèces appartiennent à un même groupe tandis qu'une troisième n'en fait pas partie, alors ces deux espèces sont plus proches l'une de l'autre que de cette troisième. Cela signifie

qu'elles partagent un dernier ancêtre commun (DAC) plus récent qu'avec cette troisième. Par exemple, le lézard des murailles est plus proche du moineau domestique que de la petite grenouille verte, ou dit autrement le DAC du

UNE CLARIFICATION LOGIQUE ET
MATHÉMATIQUE NÉCESSAIRE DE
L'INFÉRENCE PHYLOGÉNÉTIQUE EN SVT

lézard des murailles et du moineau domestique est plus récent que le DAC du lézard des murailles et de la petite grenouille verte. Ces groupes emboîtés décrivent donc la même hiérarchie de parenté qu'un arbre phylogénétique.

Toutefois, ce dernier a l'avantage d'expliciter l'axe temporel (lorsqu'il est enraciné) et de représenter plus directement l'histoire des divergences des différentes lignées (cf. figure 2).

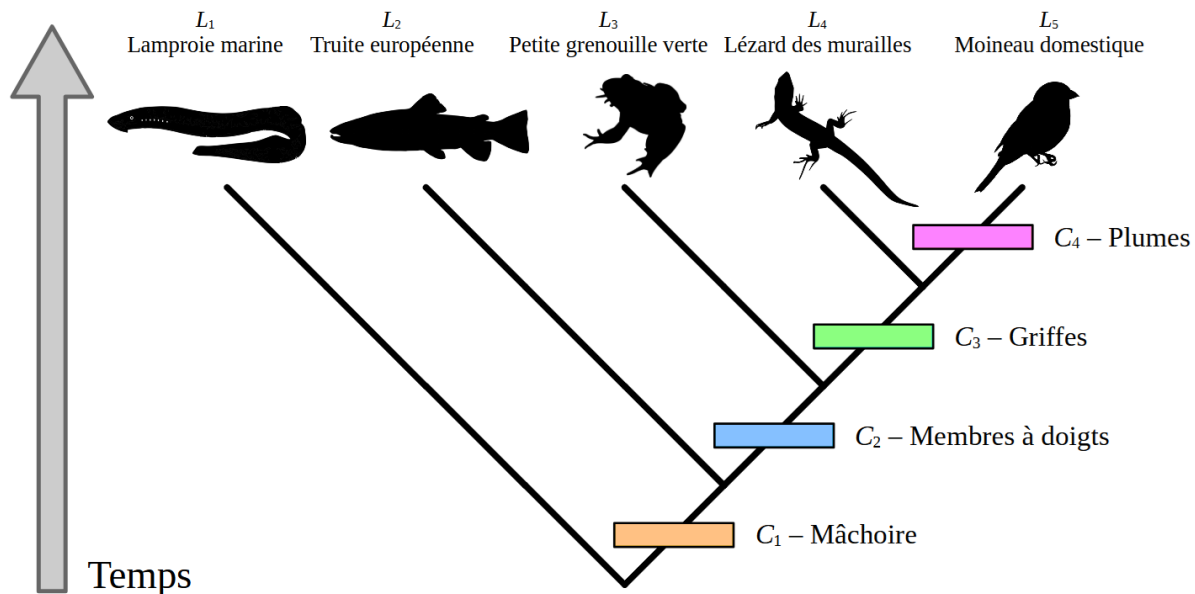


Figure 2 : Arbre phylogénétique équivalent aux groupes emboîtés de la figure 1.

Cette « méthode des groupes emboîtés » a l'avantage d'être assez simple, mais elle ne fonctionne que pour des matrices très simples comme celle du tableau 1. Elle repose en effet sur des hypothèses auxiliaires qui ne sont généralement pas vraies :

- (1) tous les caractères considérés apparaissent au cours de l'évolution (aucune disparition) ;
- (2) ces apparitions ne se produisent qu'une seule fois.

Si certains caractères disparaissent (pattes des serpents), ou si certains apparaissent plusieurs fois (ailes des oiseaux, des chauve-souris et des ptérosaures), alors la méthode ne produit pas les bons regroupements et donne donc un arbre phylogénétique incorrect.

La difficulté principale est que pour savoir si ces hypothèses auxiliaires sont vérifiées pour une matrice particulière, il faut déjà connaître le vrai arbre phylogénétique. Dans la réalité, les caractères qui apparaissent ou disparaissent à plusieurs occasions sont la règle, les caractères qui n'apparaissent qu'une seule fois sont plus rares. Une matrice sans aucun conflit (sans chevauchement entre les regroupements suggérés par différents caractères) est donc complètement irréaliste. Mais le plus grave, c'est que ce type de matrices, volontairement « propres », donne l'impression qu'une recette de regroupement permet de reconstruire l'arbre correct. Or, comme nous allons le voir dans la suite de cet article, la résolution du cas général du problème de l'inférence phylogénétique repose sur une logique très différente.

2. – La définition formelle du problème de l'inférence phylogénétique

Pour évaluer les méthodes scolaires, il faut d'abord préciser formellement le problème qu'elles cherchent à résoudre, depuis les données initiales jusqu'à la solution attendue.

2.1. - Les données du problème et leur codage

Un nombre limité d'espèces forment notre ensemble d'étude. On définit également un ensemble de caractères qui peuvent être présents ou absents chez chacune de ces espèces. Grâce à des observations et des mesures, on construit une matrice binaire $M \in \{0, 1\}^{n \times m}$, composée de n lignes représentant les espèces et m colonnes représentant les caractères (Hwang, Richards et Winter, 1992 ; Kitching et al., 1998 ; Felsenstein, 2004). Les « 0 » représentent habituellement l'état absent et les « 1 » l'état présent (mais ce codage est arbitraire). Nous nous servons de la matrice suivante comme exemple dans toute la suite de l'article :

$$M = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Le travail biologique s'interrompt temporairement ici et l'étude mathématique commence. On peut se représenter l'espace multidimensionnel défini par cette matrice comme un hypercube à m dimensions. Les n lignes sont alors des points de cet espace ; elles constituent un sous-ensemble des sommets de l'hypercube.

2.2. - Le principe de parcimonie et le critère de parcimonie

Le principe de parcimonie ou rasoir d'Ockham est un principe épistémologique qui stipule que pour expliquer des observations, il faut minimiser le nombre d'hypothèses non nécessaires. Suivant ce principe, on préférera toujours l'explication la plus simple qui permet de rendre compte d'un phénomène toutes choses égales par ailleurs. En phylogénétique, on peut donc dire intuitivement que si deux espèces partagent un caractère dans le même état (tous les deux absents ou bien tous les deux présents), il est plus simple de supposer que ce caractère ne s'est pas transformé le long des branches de l'arbre qui relient ces deux espèces. Deux transformations sont possibles, mais c'est une explication moins parcimonieuse. Quatre aussi, mais c'est encore moins parcimonieux. À l'inverse, si deux espèces ont un caractère dans un état opposé (l'un présent, l'autre absent) il est plus simple de supposer que ce caractère s'est transformé exactement une fois le long des branches de l'arbre qui les relient. Trois transformations sont possibles mais c'est moins parcimonieux.

Le principe de parcimonie peut donc se traduire directement en un critère mathématique de parcimonie qui consiste à associer à chaque arbre un score égal au nombre minimal total de transformations nécessaires pour rendre compte des caractères observés, puis à préférer les arbres pour lesquels ce score est minimal. Ce score ayant les propriétés d'une distance, on parle souvent de longueur d'une branche pour désigner le nombre de transformations de caractères se produisant sur celle-ci.

Un arbre phylogénétique est donc un graphe acyclique connexe pondéré, soit $T = (V, E, \ell)$ où V est l'ensemble des sommets de l'arbre (les biologistes parlent couramment

UNE CLARIFICATION LOGIQUE ET
MATHÉMATIQUE NÉCESSAIRE DE
L'INFÉRENCE PHYLOGÉNÉTIQUE EN SVT

de « nœuds » pour désigner les sommets) et E est l'ensemble de ses arêtes (ou « branches »). L'organisation de ces arêtes (indépendamment de leurs longueurs) s'appelle la topologie de l'arbre. Tous les sommets $v \in V$ ont des coordonnées binaires v_j dans l'espace multidimensionnel défini par la matrice des caractères M (ce sont donc tous des sommets de l'hypercube). Les feuilles correspondent aux n lignes de cette matrice, tandis que les sommets internes sont de nouveaux sommets introduits pour structurer l'arbre. On dit que ce sont des sommets de Steiner, en termes biologiques ils jouent le rôle « d'ancêtres communs hypothétiques » (Farris, 1970 ; Kitching et al., 1998 ; Felsenstein, 2004). La fonction ℓ associe à chaque arête $e \in E$ sa longueur $\ell(e)$. La longueur d'une arête est définie simplement comme la somme des valeurs absolues des différences entre les coordonnées des deux sommets qu'elle relie :

$$\ell(e) = \sum_{j=1}^m |u_j - v_j|$$

avec $e = \{u, v\} \in E$; et $u, v \in V$

Autrement dit, c'est une distance de Hamming entre deux vecteurs binaires (Hamming, 1950). Prenons par exemple une arête e qui relie deux sommets $u = (0, 1, 0, 1)$ et $v = (0, 0, 1, 1)$, qu'on peut aussi écrire respectivement (0101) et (0011). La longueur de cette arête sera donc :

$$\begin{aligned} \ell(e) &= |0-0| + |1-0| + |0-1| + |1-1| \\ &= 0 + 1 + 1 + 0 \\ &= 2 \end{aligned}$$

La longueur de l'arbre est alors $L = \sum_{e \in E} \ell(e)$. C'est la *fonction objectif* (ou la fonction de cout, ou encore le critère d'optimalité) qu'on cherche à minimiser quand on parle du critère de parcimonie. Trouver les topologies d'arbres qui minimisent cette fonction objectif particulière est le but principal de la dis-

cipline qu'on appelle la cladistique (Farris, 1970 ; Hwang, Richards et Winter, 1992 ; Kitching et al., 1998 ; Felsenstein, 2004).

Le critère de parcimonie est bien sûr critiquable, d'autres critères plus sophistiqués ont d'ailleurs été mis au point pour inférer les arbres phylogénétiques de manière plus fiable, mais ce n'est pas le sujet de cet article d'explorer ces critères alternatifs. L'objectif des programmes scolaires est d'initier les élèves à l'inférence phylogénétique en se fondant sur ce critère simple et donc pédagogiquement accessible.

2.3. - La longueur d'un arbre ou le problème de « petite parcimonie »

Avant même de pouvoir comparer plusieurs arbres entre eux pour choisir le meilleur selon le critère de parcimonie, il est nécessaire de savoir calculer la longueur de chaque arbre candidat. Ce calcul s'appelle le problème de petite parcimonie. Il consiste à déterminer la longueur minimale d'un arbre de topologie fixée permettant de rendre compte des états observés aux feuilles. Ce problème se résout efficacement (en temps polynomial), grâce à des algorithmes classiques comme ceux de Fitch (1971), Hartigan (1973) ou Sankoff (1975). Le calcul de la longueur de l'arbre ne nécessite pas de déterminer les états des caractères à chaque sommet, on peut donc connaître directement la somme des longueurs de toutes les branches sans connaître la longueur de chaque branche. Cela peut paraître contre-intuitif, mais il s'agit simplement de la conséquence du fait que, pour une topologie donnée, différentes distributions d'états sur les sommets internes sont parfois possibles sans que cela ne change la longueur totale des branches.

La figure 3 illustre un algorithme très simple de résolution du problème de petite parcimonie pour des caractères binaires et une

topologie strictement dichotomique. On place une racine arbitraire sur l'une des branches et on remonte progressivement des feuilles vers cette racine (passe ascendante). Chaque caractère peut être dans l'état « 0 » (absent), « 1 » (présent) ou bien « ? » (inconnu). On calcule alors les états des sommets internes en appliquant les règles présentées dans le tableau 2 pour chaque caractère, de manière indépendante. Il est ensuite possible, si on le souhaite, de lever les ambiguïtés de chaque sommet interne lors d'une seconde passe, descendante cette fois-ci (de la racine vers les feuilles). Celle-ci consiste à choisir arbitrairement les états ambigus de la racine puis à résoudre les états ambigus de chaque sommet interne en lui attribuant l'état de son sommet père, mais cette seconde passe n'est pas nécessaire au calcul de la longueur de l'arbre. L'ensemble de ces deux passes s'appelle l'algorithme ACCTRAN (Faris, 1970 ; Swofford et Maddison, 1987). D'autres algorithmes, comme DELTRAN, peuvent reconstruire des états différents pour certains nœuds internes, et donc des longueurs de branches différentes, mais la longueur totale de l'arbre reste la même (Swofford et Maddison, 1987).

Prenons par exemple les deux feuilles les plus à droite sur la figure 3, dont les coordonnées sont (1001) et (1011). On les combine une à une pour trouver les coordonnées de leur sommet père commun en suivant les règles du tableau 2 : 1 et 1 donnent 1, 0 et 0 donnent 0, 0 et 1 donnent ?, et enfin 1 et 1 donnent 1. Donc leur sommet père a les coordonnées (10?1). Le « ? » signifie que cette coordonnée particulière est encore ambiguë à ce stade de l'algorithme, mais son apparition indique la nécessité d'une transformation dans un sens ou dans l'autre, la longueur de l'arbre est donc incrémentée. Nous pouvons ensuite calculer les coordonnées du sommet qui a pour enfants le sommet (10?1) que nous venons de déterminer et la 3^e feuille en partant de la droite (1 000). En appli-

quant les mêmes règles, on trouve : 1 et 1 donnent 1, 0 et 0 donnent 0, ? et 0 donnent 0, 1 et 0 donnent ?. Donc ce sommet a pour coordonnées (100?). Remarquons que la troisième opération fait disparaître le ? apparu précédemment, mais la quatrième en fait apparaître un nouveau, la longueur de l'arbre doit donc être incrémentée. On poursuit ainsi les opérations jusqu'à la racine pour obtenir la longueur totale de l'arbre.

Entrées		Sorties	
Sommet enfant 1	Sommet enfant 2	Sommet courant	Longueur de l'arbre
1	1	1	+0
0	0	0	+0
0	1	?	+1
0	?	0	+0
1	?	1	+0
?	?	?	+0

Tableau 2 : Règles de calcul de la longueur d'un arbre. Une seule des configurations possibles, ici surlignée, permet d'incrémenter la longueur de l'arbre. Les deux sommets enfants sont interchangeable (règles symétriques).

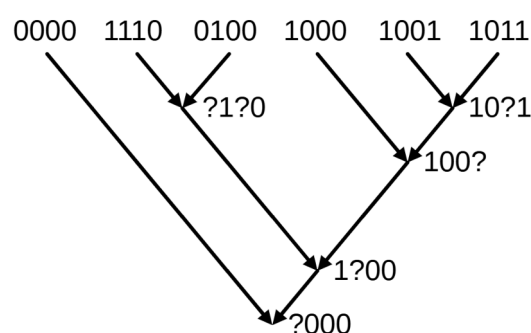


Figure 3 : Exemple de calcul de la longueur d'un arbre dichotomique avec l'algorithme de Fitch. L'arbre a une longueur de 6 pas.

2.4. - Le problème de « grande parcimonie » et ses propriétés

Le problème de grande parcimonie est une variante du problème de l'arbre minimal de Steiner : on cherche, parmi tous les arbres possibles, ceux qui ont une longueur totale minimale (Hwang, Richards et Winter, 1992). Le problème de l'arbre minimal de Steiner est NP-complet, ce qui veut dire qu'on ne connaît pas d'algorithme permettant de garantir qu'on trouve la solution exacte en un temps polynomial pour des entrées arbitraires (Foulds et Graham, 1982). Un temps polynomial signifierait que le nombre d'étapes élémentaires pour la trouver serait majoré par $am^b n^c$, avec m le nombre de caractères, n le nombre d'espèces, et a , b , c trois constantes réelles positives. Mais il n'existe pas de constantes suffisamment grandes qui permettraient de construire une telle borne pour les algorithmes exacts dont nous disposons (en l'occurrence, le problème vient surtout de c).

En pratique, les arbres candidats sont comparés en résolvant de nombreuses fois le problème de petite parcimonie. Comme l'espace des arbres croît exponentiellement avec le nombre d'espèces analysées, il est impossible de vérifier toutes les configurations possibles dès qu'on dépasse seulement une douzaine de feuilles. Toute la difficulté du problème de grande parcimonie réside donc dans la recherche des arbres candidats les plus prometteurs par une exploration astucieuse de l'espace des arbres.

Puisque les lignes de M représentent un sous-ensemble des sommets d'un hypercube de Hamming, alors il est évident que permuter les lignes ou les colonnes, ou inverser toutes les valeurs d'une ou de plusieurs colonnes, sont des opérations de symétrie qui ne changent ni le problème posé ni sa solution (Hwang, Richards et Winter, 1992 ; Felsenstein, 2004).

Pour une matrice M donnée, il peut exister un ou plusieurs arbres solutions de longueur minimale. Une arête peut être de longueur nulle, donc deux sommets de l'arbre peuvent avoir les mêmes coordonnées dans l'espace défini par la matrice (la « distance de Hamming » utilisée est donc dans ce contexte techniquement une pseudo-distance puisqu'elle ne respecte pas l'axiome $d(x, y) = 0 \Rightarrow x = y$). On considère que deux arbres sont la même solution si leurs topologies sont identiques et qu'ils ne diffèrent que par les longueurs de leurs arêtes (classes d'équivalence). Dupliquer une ligne ne peut pas changer la longueur de l'arbre minimal. En effet, les deux sommets de l'arbre correspondant à ces lignes identiques peuvent être reliés à un même sommet de Steiner par deux arêtes de longueurs nulles. On s'attend donc à ce qu'un algorithme de résolution correct ne soit pas sensible à des duplications de ce genre et aboutisse toujours à la même topologie (à la différence près des espèces dupliquées) (Hwang, Richards et Winter, 1992 ; Felsenstein, 2004).

L'ajout ou le retrait d'une colonne qui ne comporte pas au moins deux « 1 » et au moins deux « 0 » ne peut pas changer la topologie de l'arbre solution. On dit donc qu'une telle colonne est cladistiquement non informative. En effet, si une colonne est constante (seulement des « 1 » ou seulement des « 0 »), alors cette dimension ne sera traversée par aucune des arêtes de l'arbre solution. Et si une colonne ne comporte qu'un seul « 1 » ou qu'un seul « 0 », alors cette dimension est nécessairement traversée une seule fois dans l'arbre solution par l'arête pendante qui relie la feuille singularisée par cette valeur au reste de l'arbre (Hwang, Richards et Winter, 1992 ; Felsenstein, 2004).

2.5. - L'orientation de l'arbre

L'arbre non orienté solution du problème défini ci-dessus peut être orienté si on dis-

tingue l'une des espèces de la matrice qui joue le rôle « d'extragroupe ». On fait alors l'hypothèse que cette espèce O s'est séparée des autres espèces étudiées en premier dans l'ordre chronologique. Mathématiquement, cela signifie simplement qu'on doit rajouter un sommet R sur l'arête reliant O au reste de l'arbre. Ce sommet R devient alors la racine de l'arbre. L'orientation de l'arbre n'a aucune influence sur la fonction de cout L et n'entre donc pas en ligne de compte dans l'inférence de la topologie de l'arbre (Farris, 1970 ; Hwang, Richards et Winter, 1992 ; Kitching et al., 1998 ; Felsenstein, 2004). L'information d'extragroupe est en effet extérieure à la matrice de caractères ; elle en est complètement indépendante. L'arbre de la figure 4 est l'unique solution de la matrice M , où c'est la sixième ligne qui sert d'extragroupe. Sa longueur peut être calculée

en utilisant les règles du tableau 2, comme nous l'avons fait pour la figure 3.

Prenons par exemple $L_1=(011110)$ et $L_5=(100111)$, alors $S_4=(???11?)$, soit une augmentation de 4 de la longueur de l'arbre. Ensuite, on a $L_2=(111101)$ et $S_4=(???11?)$, donc $S_3=(1111?1)$, ce qui augmente de 1 la longueur de l'arbre. Comme indiqué dans la sous-section 2.3, une passe descendante peut être effectuée pour lever les ambiguïtés et attribuer des longueurs précises aux différentes branches de l'arbre. Ainsi, les états ambigus de S_4 recevront automatiquement les états non ambigus de S_3 , ce qui permet de reconstituer $S_4=(111111)$. L'application de la formule des longueurs des branches vue dans la sous-section 2.2 permet alors de calculer $\ell(\{S_4, L_1\})=2$ et $\ell(\{S_4, L_5\})=2$.

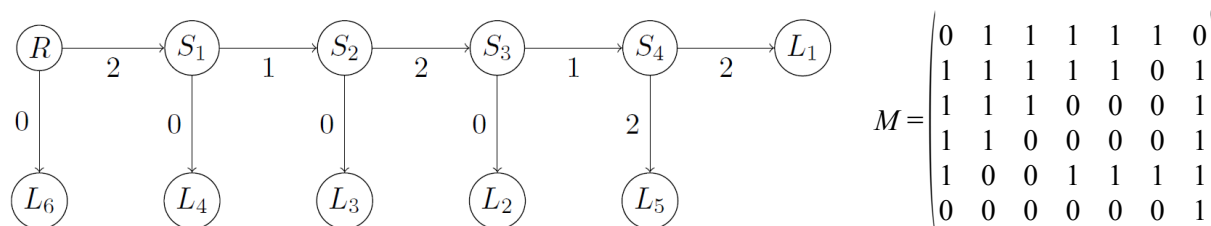


Figure 4 : Unique arbre enraciné solution reliant les espèces de la matrice M .

Les nombres sur les arêtes sont leurs longueurs. La longueur totale de cet arbre est de 10.

La solution et son unicité ont été vérifiées avec le logiciel PAUP* (Swofford, 2003).

3. – La logique défectueuse des « groupes emboîtés » au collège

Le cadre mathématique minimal étant maintenant posé, nous pouvons évaluer la méthode des groupes emboîtés utilisée au collège : d'abord en la formalisant, puis en comparant son résultat à l'optimum et, enfin, en critiquant la validité du raisonnement au regard du problème qu'il est censé résoudre.

3.1. - La formalisation de la procédure des groupes emboîtés

Pour résoudre le problème de grande parimonie défini dans la section précédente, les professeurs de SVT enseignent au collège la procédure dite des « groupes emboîtés » (Leconte (dir.) et al., 2008 ; Hervé, 2009 ; DGESCO, 2020). Cette procédure est aussi enseignée par les professeurs des écoles en CM1/CM2 (DGESCO, 2023a). L'idée est de regrouper successivement les espèces qui partagent le plus de caractères présents. Le résultat

UNE CLARIFICATION LOGIQUE ET
MATHÉMATIQUE NÉCESSAIRE DE
L'INFÉRENCE PHYLOGÉNÉTIQUE EN SVT

tat est d'abord dessiné sous la forme d'une classification hiérarchique (formellement, un diagramme de Venn) qui est ensuite transformée en un arbre phylogénétique. On peut traduire formellement cet algorithme en termes de théorie des graphes pour construire directement l'arbre de la manière suivante :

1. on commence par créer un sommet de Steiner racine auquel on relie toutes les feuilles (correspondant aux lignes de la matrice M , c'est-à-dire les espèces) ;
2. pour le sommet de Steiner actuellement considéré, on cherche dans M le caractère qui contient le plus de « 1 », et au moins un « 0 », parmi les feuilles qui sont connectées à ce dernier (sous-colonne) ;
3. on laisse les feuilles qui possèdent un « 0 » dans cette sous-colonne connectées au sommet de Steiner courant, et on regroupe les autres en les connectant à un nouveau sommet de Steiner enfant du précédent ;
4. si au moins deux feuilles sont encore directement connectées au sommet de Steiner actuellement considéré, alors on continue à énumérer les colonnes pour essayer de les regrouper ;
5. s'il n'y a pas au moins deux feuilles encore directement connectées au sommet de Steiner courant ou si on a fini d'énumérer les colonnes, alors on passe aux sommets de Steiner enfants qui ont été générés et on recommence à l'étape 2 ;
6. l'algorithme s'arrête lorsque plus aucun nouveau sommet de Steiner ne peut être créé.

Avec notre matrice d'exemple M , on commence par créer le sommet racine R auquel toutes les feuilles représentant les espèces sont reliées. On constate que, dans la matrice M , la colonne 7 est celle qui comporte le plus de « 1 » ; on en compte 5. On regroupe donc L_2 à L_6 en les reliant à un nouveau sommet de Steiner S_1 , et on laisse L_1 connectée à R . Pour les 5 espèces connectées à S_1 , on constate que la colonne 1 est celle qui comporte le plus de « 1 » ; on en compte 4 (seulement 3 dans la colonne 2 puisque la ligne 1 est exclue). On regroupe donc L_2 à L_5 en les connectant à S_2 et on laisse L_6 connectée à S_1 . Ensuite, pour les 4 espèces reliées à S_2 , c'est la colonne 2 qui a le plus de « 1 », ce qui exclut L_5 du regroupement sur S_3 . Pour les 3 espèces restantes, c'est la colonne 3 qui a le plus de « 1 », on regroupe donc L_2 et L_3 . L'algorithme est terminé, l'arbre généré est celui de la figure 5. Les longueurs de branches sur cet arbre sont issues de la procédure expliquée dans la sous-section 2.3 et déjà utilisée dans la sous-section 2.5. Ainsi, en remontant les branches à partir de $L_2=(1111101)$ et $L_3=(1110001)$, on trouve $S_4=(111??01)$. De même, à partir de S_4 et $L_4=(1100001)$, on trouve $S_3=(11?0001)$. Une fois cette passe ascendante complétée, la passe descendante reconstruit donc $S_4=(1110001)$, d'où l'on déduira $\ell(\{S_4, L_3\})=0$ et $\ell(\{S_4, L_2\})=2$. Les mêmes raisonnements poursuivis sur l'ensemble de l'arbre permettent ainsi de calculer sa longueur totale et les longueurs des branches illustrées dans la figure 5.

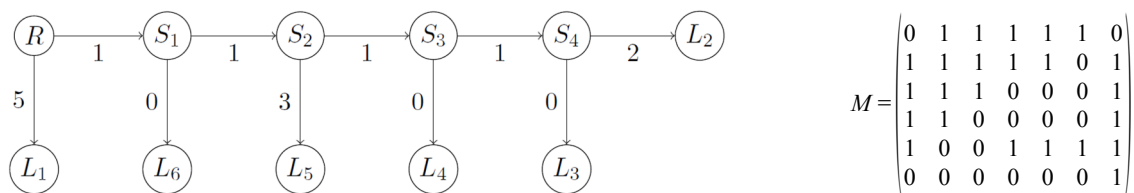


Figure 5 : Arbre enraciné produit par l'algorithme des groupes emboîtés. R est la racine imposée par l'algorithme malgré le fait que l'extragroupe est censé être L_6 . La longueur totale de cet arbre est 14.

3.2. - Une critique de la procédure des groupes emboîtés

Comme on peut le voir facilement, les arbres des figures 4 et 5 ne correspondent pas. Non seulement l'emplacement de la racine contredit l'hypothèse de l'espèce L_6 comme extragroupe, mais même si on ne tient pas compte de la racine, les deux arbres n'ont pas la même topologie. La méthode des groupes emboîtés n'aboutit donc pas à la bonne solution avec notre matrice d'exemple M . L'arbre qu'elle génère est même très loin de l'optimalité en termes de parcimonie avec une longueur totale de 14 au lieu de 10.

Cela est dû au fait que cet algorithme introduit une asymétrie illégitime en regroupant les « 1 », mais pas les « 0 ». Or, les « 1 » et les « 0 » jouent un rôle parfaitement symétrique dans la fonction de coût qu'on cherche à minimiser. On constate également que le premier regroupement est effectué grâce à la colonne 7 alors qu'elle est censée être non informative. Si une colonne ne peut pas changer la solution optimale, alors il est incohérent qu'elle soit mobilisée pour justifier un regroupement.

De surcroît, le fait même de regrouper en comptant les « 1 » dans les différentes colonnes a pour conséquence que le résultat de l'algorithme est sensible aux duplications de lignes, puisque celles-ci ajoutent des « 1 » dans certaines colonnes, mais pas dans d'autres (concrètement pour dupliquer une ligne il suffit d'ajouter une espèce très similaire à une espèce déjà présente dans la matrice). On peut ainsi obtenir à peu près n'importe quel arbre en multipliant habilement certaines lignes.

Par exemple, la matrice M remplit les conditions permettant une telle manipulation du résultat au moment du troisième regroupement : L_5 était exclu du regroupement de L_2 ,

L_3 et L_4 sur la base des trois « 1 » de la colonne 2 ; ajoutons deux copies de L_5 , et maintenant les colonnes 4 et 5 comportent chacune quatre « 1 » indiquant toutes les deux le regroupement de L_2 et L_5 .

Ces trois énormes défauts montrent que le problème n'est pas un manque de précision, mais un manque de justesse. Par précision, on entend la fréquence à laquelle le bon résultat est obtenu quand on répète la procédure avec différentes matrices. Une méthode peut être tolérablement imprécise si, malgré une certaine variabilité, elle cible bien la quantité à optimiser. Par justesse, on entend l'absence de biais vis-à-vis du vrai objectif : en moyenne, la méthode doit pointer vers l'optimum recherché.

Or, ici, la méthode des groupes emboîtés ne vise pas la bonne fonction objectif, augmenter la quantité de données fait converger vers des topologies fausses. Autrement dit, ce n'est pas une euristique approximative de la parcimonie mais une procédure biaisée qui optimise autre chose, ce qui est rédhibitoire.

Pour aboutir au bon résultat malgré un algorithme totalement inadapté, les professeurs de SVT construisent la matrice (comme celle du tableau 1) en choisissant sciemment les espèces et les caractères de telle manière que :

- (1) la ligne de l'extragroupe ne comporte que des « 0 » ;
- (2) l'arbre optimal attendu ne comporte qu'une seule traversée de chaque dimension (c'est-à-dire, en termes biologiques, une seule transformation pour chaque caractère dans l'ensemble de l'arbre).

On part donc de la réponse qu'on veut obtenir pour sélectionner les données permettant de *sauver* la méthode. Il est ainsi légitime de questionner la scientificité d'une telle pratique et la pertinence de son enseignement.

4. – La logique défectueuse de la « polarisation à priori » au lycée

Examinons à présent les méthodes utilisées au lycée et en CPGE, qui ajoutent une étape de polarisation préalable des caractères avant la construction de l'arbre pour remédier aux erreurs d'enracinement comme celle que nous avons repérée dans la section précédente.

4.1. - La nouvelle étape de polarisation à priori

Pour essayer de remédier aux défauts évidents de la méthode des groupes emboîtés que nous avons mis en lumière dans la section précédente, concernant notamment l'extragroupe, les deux méthodes utilisées au lycée (Lecointre (dir.) et al., 2008 ; Hervé, 2009, 2010 ; DGESCO, 2023b ; Collectif d'enseignants, 2024, p. 285) et en CPGE (MENJS, 2021) imposent une étape préalable de recodage de la matrice appelée la « polarisation à priori ». Cela consiste à considérer que l'état d'un caractère, absent ou présent, est forcément « ancestral » (c'est-à-dire le plus ancien) chez l'extragroupe, tandis que l'autre état est « dérivé » (le plus récent). Mathématiquement, cela veut dire qu'on inverse les colonnes qui contiennent un « 1 » sur la ligne de l'extragroupe de manière qu'elle ne comporte plus que des « 0 ». Cette polarisation est dite « à priori » car les caractères se trouvent ainsi orientés avant la construction de la topologie de l'arbre. En prenant L_6 comme extragroupe on recode la matrice M en inversant seulement la 7^e colonne, on obtient alors la matrice polarisée M' :

$$M' = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

4.2. - La méthode 1 : colonne par colonne

La première méthode est en réalité identique à la méthode des groupes emboîtés utilisée au collège, elle ne diffère que par l'ajout de l'étape de polarisation à priori. On crée une racine à laquelle on relie toutes les espèces, puis on regroupe successivement les espèces reliées à chaque sommet interne en cherchant le caractère (colonne) qui comporte le plus de « 1 » pour ce sous-groupe d'espèces.

Appliquons cette méthode sur la matrice M' . Pour commencer, on constate que les colonnes 1 et 2 sont à égalité avec quatre « 1 ». Aussi surprenant que cela puisse paraître, aucune des méthodes enseignées ne prévoit comment gérer une égalité discordante. Nous allons donc suivre les deux chemins possibles à chaque fois que la situation se présente.

Si on suit la colonne 1, alors on regroupe les quatre feuilles L_2 à L_5 qui partagent le même état supposément dérivé 1, alors que L_1 et L_6 , qui sont dans l'état 0, restent attachés directement à R sans possibilité de regroupement. Ensuite, c'est la colonne 2 qui comporte le plus de « 1 » (soit trois « 1 » dans le sous-groupe L_2 à L_5), on regroupe donc les trois feuilles L_2 , L_3 et L_4 . Enfin, c'est la colonne 3 qui comporte le plus de « 1 », on regroupe donc L_2 et L_3 (soit deux « 1 » dans le sous-groupe L_2 à L_4). On obtient ainsi un arbre de longueur 14.

Si on suit d'abord la colonne 2, alors on regroupe les quatre feuilles L_1 à L_4 alors que L_5 et L_6 restent attachés directement à R sans possibilité de regroupement. Il y a ensuite encore une égalité avec trois « 1 » dans les colonnes 1 et 3. Donc on regroupe soit de L_2 à L_4 avec la colonne 1, soit de L_1 à L_3 avec la colonne 3. Dans le premier cas, on regroupe finalement L_2 et L_3 avec les deux « 1 » de la colonne 3. Dans le second cas, on regroupe fina-

lement L_1 et L_2 avec les deux « 1 » des colonnes 4 et 5, qui sont identiques et permettent donc de rejeter le regroupement alternatif de

L_2 et L_3 avec les deux « 1 » de la colonne 1. Les trois arbres différents obtenus par cette méthode sont présentés dans la figure 6.

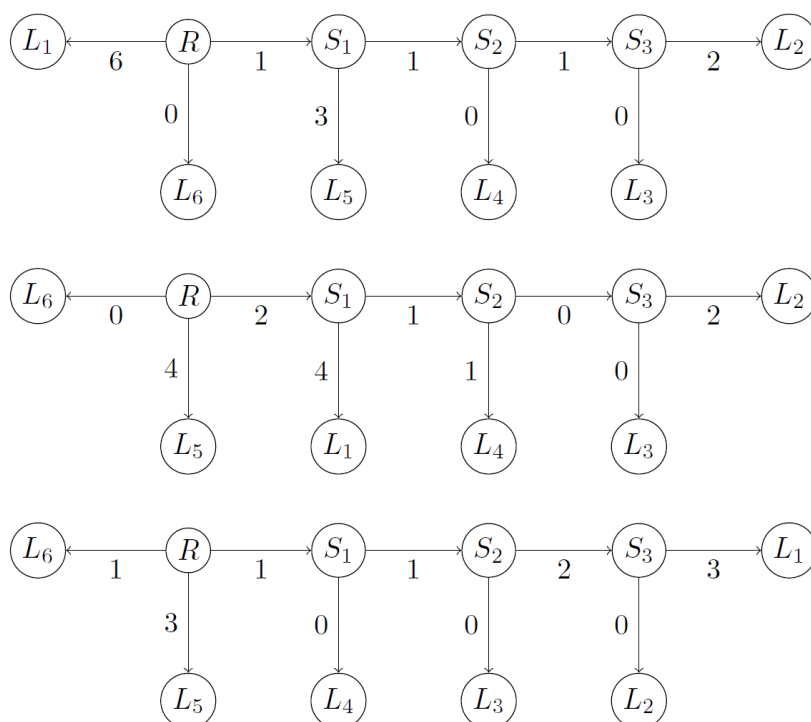


Figure 6 : Arbres enracinés produits par la variante 1 de la méthode par polarisation à priori. La longueur des deux premiers arbres est 14, le 3^e a pour longueur 11.

4.3. - La méthode 2 : ligne par ligne

La deuxième méthode consiste à ajouter les espèces une par une dans l'ordre croissant du nombre de « 1 » de chaque ligne. On place alors l'espèce choisie aussi près que possible de la racine, en la regroupant avec l'espèce déjà placée ou le groupe d'espèces déjà placé avec lequel elle partage le plus de « 1 ». Appliquons maintenant cette méthode sur notre exemple M' . Les effectifs de « 1 » de chaque ligne par ordre croissant sont :

$$|L_6|=0 ; |L_4|=2 ; |L_3|=3 ; |L_5|=4 ; |L_2|=5 ; |L_1|=6.$$

$$M' = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

On part donc de L_6 (extragroupe) accroché à R , puis on ajoute successivement L_4 , L_3 , L_5 , L_2 , L_1 . Notez bien qu'il n'y a pas d'ambiguïté sur l'ordre à suivre dans cet exemple mais, s'il y en avait eu, on aurait dû essayer toutes les possibilités. On commence par ajouter L_4 à R (seule position possible). Puis L_3 partage deux « 1 » avec L_4 (colonnes 1 et 2), on regroupe donc L_3 et L_4 . Ensuite, L_5 partage un « 1 » (colonne 1) avec L_3 et L_4 , on l'insère

UNE CLARIFICATION LOGIQUE ET
MATHÉMATIQUE NÉCESSAIRE DE
L'INFÉRENCE PHYLOGÉNÉTIQUE EN SVT

donc à la base de ce groupe. Après, L_2 partage trois « 1 » avec L_3 (colonnes 1, 2 et 3) et avec L_5 (colonnes 1, 4 et 5), il y a donc deux positions d'insertion possibles. Enfin on ajoute L_1

qui partage quatre « 1 » avec L_2 (colonnes 2, 3, 4 et 5). Les deux arbres différents obtenus par cette méthode sont présentés dans la figure 7.

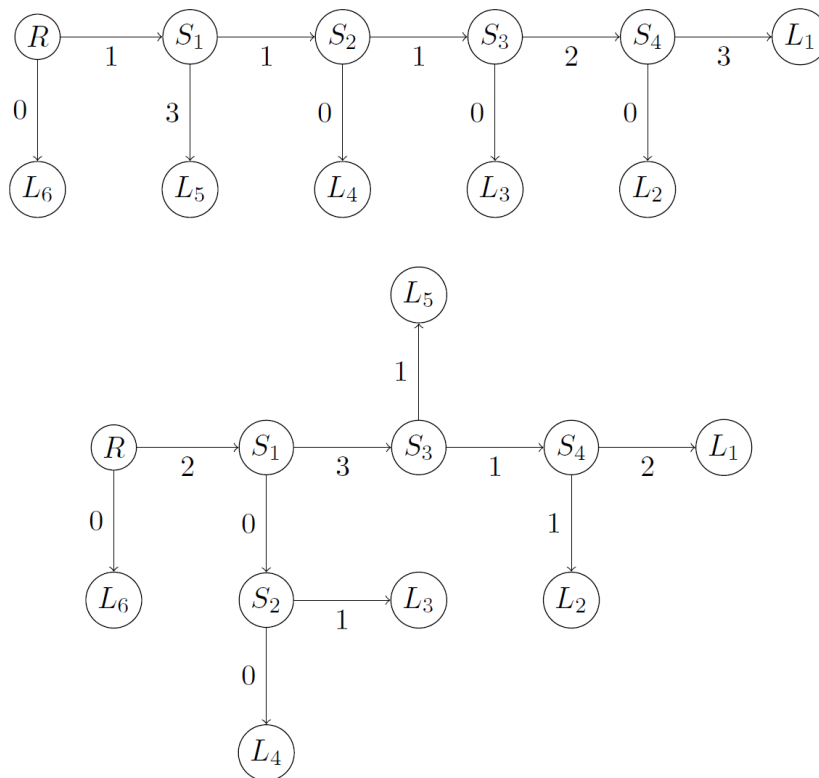


Figure 7 : Arbres enracinés produits par la variante 2 de la méthode par polarisation à priori. Le premier arbre est équivalent au 3^e arbre de la figure 6 (à la configuration de la racine près) et sa longueur totale est 11 comme ce dernier. La longueur totale du deuxième arbre est 11 également.

4.4. - Une critique des procédures par « polarisation à priori »

Aucune des deux variantes de la méthode par polarisation à priori n'aboutit à l'arbre solution de longueur minimale. Polariser les caractères et regrouper en fonction des « 1 » partagés revient à restreindre les réversions ($1 \rightarrow 0$), on limite donc de manière injustifiée l'espace des arbres atteignables. Or le problème consiste à rechercher, dans l'hypercube de Hamming $\{0, 1\}^m$, un arbre de Steiner reliant les n lignes de la matrice (qui forment n sommets de cet hypercube parmi les 2^m qui le composent) avec une longueur totale minimale

(cf. section 2) ; rien n'exclut en général plusieurs traversées d'une même dimension sur l'arbre optimal, et donc des retours vers l'état 0. Imposer une orientation privilégiée des transformations force des détours inutiles qui augmentent le coût. Ces deux procédures mêlent en fait enracinement et inférence topologique. L'extragroupe devrait uniquement être considéré pour placer la racine après optimisation de la topologie non enracinée (Meacham, 1984 ; Nixon et Carpenter, 1993, 2013 ; Kitching et al., 1998). Ici, on utilise l'extragroupe pour orienter chaque colonne afin de placer les espèces. Les deux variantes de cette

méthode se trompent donc d'objectif, la règle « on regroupe là où il y a le plus de « 1 » » ne vise pas à minimiser la longueur totale de l'arbre, même approximativement.

Fondamentalement, ces deux procédures violent l'invariance par recodage des colonnes. Pour un arbre donné, le nombre de changements produits par une colonne C est égal à celui de sa colonne complément $\bar{C}$: échanger 0 et 1 ne change pas le cout de parcimonie. La polarisation à priori casse cette symétrie en déclarant « ancestral » l'état de l'extragroupe. Cela a pour conséquence d'empêcher les regroupements justifiés par des réversions ($1 \rightarrow 0$), d'exclure de leurs groupes les espèces qui en ont subi ou, à l'inverse, de regrouper artificiellement des espèces ayant subi des convergences ($0 \rightarrow 1$). En fait, c'est le principe même d'attribuer de manière absolue une polarité à un caractère qui n'a aucun sens : un caractère peut bien se transformer dans le sens $0 \rightarrow 1$ sur une arête, puis dans le sens $1 \rightarrow 0$ sur une autre arête, l'état 0 sera alors ancestral dans le premier cas et dérivé dans le second cas, et inversement pour l'état 1, alors qu'il s'agit toujours du même caractère. La polarité est donc une propriété d'une transformation d'un caractère et non une propriété d'un caractère. Ajoutons que la variante 1 de cette méthode par polarisation à priori est en fait équivalente à la méthode des groupes emboîtés au collège, elle possède donc tous ses défauts comme la sensibilité aux duplications de lignes.

Ces deux procédures ne sont garanties de tomber sur le bon résultat que lorsque toutes les colonnes de la matrice polarisée forment une unique famille laminaire, c'est-à-dire lorsque toutes les colonnes sont soit disjointes soit incluses les unes dans les autres vis-à-vis des « 1 ». Mais alors le problème est déjà résolu car il n'y a rien à optimiser. Dès qu'on quitte ce cas trivial, c'est-à-dire dès qu'on tra-

vaille sur des données réelles ou sur des matrices non truquées, les méthodes scolaires actuellement pratiquées ne convergent plus vers l'arbre optimal à cause des nombreux biais que nous avons détaillés ci-dessus.

Toutes ces méthodes encouragent donc une dérive pédagogique : on sélectionne les matrices *pour que ça marche*. On choisit un extragroupe tout-zéro, on écarte les lignes ou les colonnes qui violent la laminarité, bref, on construit des exemples depuis la réponse souhaitée. Ce n'est pas former à l'inférence ; c'est apprendre à contourner un problème d'optimisation en le remplaçant par une présentation truquée. L'excuse de la simplification pédagogique n'est pas recevable car ces algorithmes n'ont pas un problème de précision, mais un problème de justesse : ils ne visent pas le bon objectif puisqu'ils contredisent les invariances essentielles du problème de Steiner sur l'hypercube de Hamming.

La voie correcte consiste à :

- (1) optimiser d'abord la topologie non enracinée sur M en utilisant la véritable fonction objectif (parcimonie) ou une approximation ;
- (2) placer ensuite la racine en se servant de l'information d'extragroupe.

Dans les deux sections suivantes, nous décrirons deux méthodes appropriées qui permettent de trouver l'arbre de longueur minimale et qui devraient donc être enseignées à la place des trois méthodes inappropriées décrites précédemment.

5. – La méthode standard de l'addition pas à pas

Après cette analyse critique, nous présentons une première méthode simple et didactisable qui utilise explicitement la fonction ob-

jectif pour résoudre le problème de l'inférence phylogénétique : l'addition pas à pas.

5.1. - Le principe de l'addition pas à pas

On a déjà souligné que le problème de l'arbre minimal de Steiner était NP-complet (Foulds et Graham, 1982). Il existe cependant des heuristiques qui peuvent trouver cet arbre en un temps raisonnable. L'algorithme de l'addition pas à pas (ou *stepwise addition*) est la méthode la plus classique permettant d'y parvenir (Farris, 1970 ; Kitching et al., 1998 ; Swofford, 2003). C'est une méthode très généraliste, qu'on utilise aussi pour résoudre d'autres problèmes d'optimisation combinatoire, comme le fameux problème du voyageur de commerce (ou *travelling salesman problem*) (Jünger, Reinelt et Rinaldi, 1995).

On commence par une paire de feuilles reliées par une arête. Puis on ajoute à l'arbre partiel les autres feuilles une par une sur l'une des arêtes. On choisit la position de la feuille en cours d'insertion de manière à ce que la longueur totale de l'arbre partiel soit minimisée. Il s'agit donc d'un algorithme glouton : plutôt que de minimiser directement la longueur de l'arbre entier, on minimise la longueur de l'arbre partiel à chaque étape en choisissant la position d'insertion optimale (sans chercher à réorganiser les feuilles déjà placées). Si plusieurs positions sont également optimales il est conseillé de suivre tous les chemins possibles.

5.2. - Le choix de l'ordre d'insertion

La méthode de l'addition pas à pas étant sensible à l'ordre d'insertion des feuilles, il est impératif de choisir cet ordre de manière judi-

cieuse afin d'augmenter les chances de tomber sur l'arbre optimal. L'ordre de « l'insertion du plus proche » (ou *closest/nearest insertion*) est un ordre à la fois simple à déterminer et très efficace. Le principe est d'ajouter dans l'arbre la ligne de la matrice qui est la plus proche (en distance de Hamming) de l'une des lignes déjà placées. La justification de ce choix est intuitive : plus deux lignes sont proches et plus il y a de chances que la distance de Hamming entre ces deux lignes soit égale à la somme des distances des arêtes reliant les deux feuilles correspondantes sur l'arbre optimal. À l'inverse, plus la distance de Hamming entre deux lignes de la matrice est grande, plus elle a de chances d'être une sous-estimation de la distance entre les deux feuilles correspondantes sur l'arbre optimal. On traite donc ces paires le plus tard possible pour éviter d'être coincé précocement dans un minimum local, nous empêchant d'atteindre le minimum global. Si plusieurs ordres d'insertion sont possibles, il est préférable de tous les essayer. Précisons bien que la distance entre deux lignes ne détermine que l'ordre d'insertion, et non l'emplacement de l'insertion. Toutes les positions sont essayées et seule la meilleure est retenue.

5.3. - Un exemple d'application de l'addition pas à pas

Appliquons cet algorithme à notre matrice d'exemple M . On construit d'abord une matrice D de distances de Hamming entre les lignes de M . Pour construire cette matrice D , il suffit d'appliquer pour toutes les paires de lignes la formule de la distance de Hamming utilisée pour les longueurs de branches (voir sous-section 2.2) comme si toutes les espèces étaient reliées directement à toutes les autres. Prenons par exemple $L_1=(0111110)$ et $L_2=(1111101)$, alors :

$$\begin{aligned} D_{1,2} &= |0-1|+|1-1|+|1-1|+|1-1|+|1-1|+|1-0|+|0-1| \\ &= 1+0+0+0+0+1+1 \\ &= 3 \end{aligned}$$

Ainsi, dans la matrice D , le nombre situé à l'intersection de la première ligne et de la deuxième colonne est 3. La formule de la distance de Hamming est symétrique, donc la matrice D l'est également.

$$D = \begin{pmatrix} 0 & 3 & 5 & 6 & 4 & 6 \\ 3 & 0 & 2 & 3 & 3 & 5 \\ 5 & 2 & 0 & 1 & 5 & 3 \\ 6 & 3 & 1 & 0 & 4 & 2 \\ 4 & 3 & 5 & 4 & 0 & 4 \\ 6 & 5 & 3 & 2 & 4 & 0 \end{pmatrix}$$

On voit que la distance minimale entre L_3 et L_4 est 1. Puis on trouve une distance de 2 entre L_2 et L_3 ou entre L_4 et L_6 . On ajoute donc dans l'arbre soit L_2 soit L_6 , puis on insère l'autre. Ensuite, on insère soit L_1 soit L_5 qui

sont tous les deux à une distance 3 de L_2 , et enfin on insère l'autre. Les quatre ordres d'insertion (ou addition) possibles sont donc :

1. $L_3/L_4, L_2, L_6, L_1, L_5$.
2. $L_3/L_4, L_2, L_6, L_5, L_1$.
3. $L_3/L_4, L_6, L_2, L_1, L_5$.
4. $L_3/L_4, L_6, L_2, L_5, L_1$.

Si on suit l'ordre 1 ou 2, l'arbre composé uniquement de L_3, L_4 et L_2 a pour longueur 3. Si on place L_6 sur l'une des arêtes pendantes menant à L_2 ou à L_3 , la longueur de l'arbre est 6, mais si on la place sur l'arête menant à L_4 , la longueur de l'arbre est 5, c'est donc la position optimale (cf. figure 8). Si on avait choisi de suivre l'ordre 3 ou 4 en plaçant d'abord L_6 puis L_2 , alors la position optimale de L_2 aurait été sur l'arête menant à L_3 .

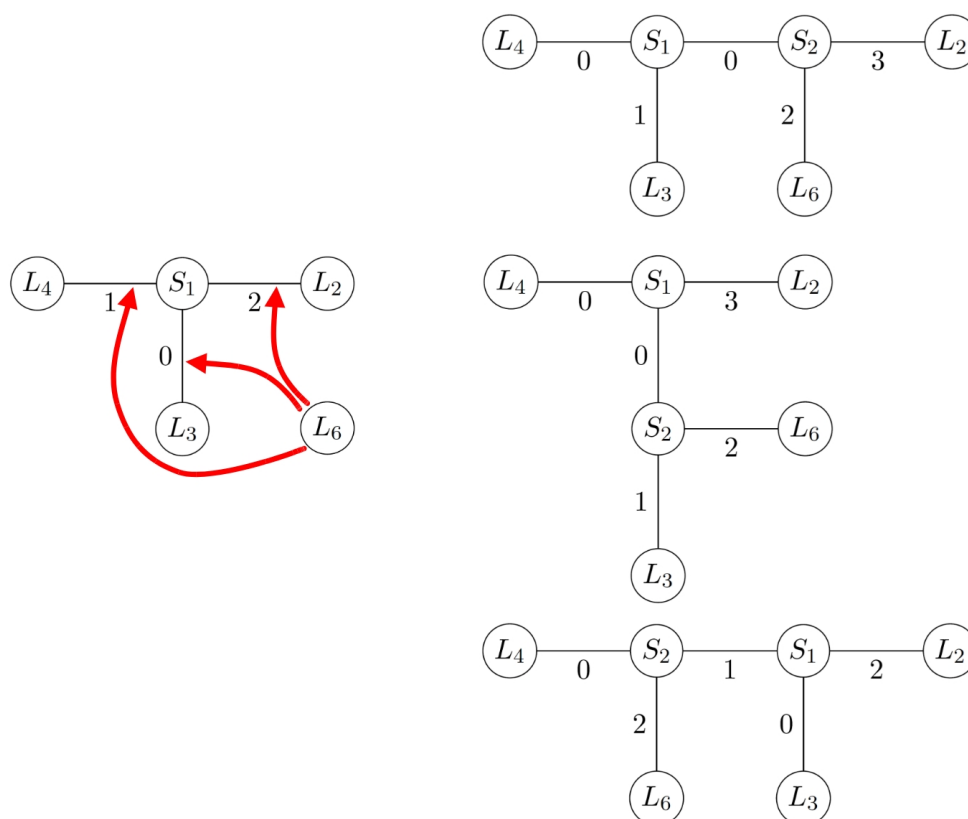


Figure 8 : Exemple d'insertion de la feuille L_6 sur un arbre partiel non enraciné, et les trois arbres possibles qui résultent de ces choix d'insertion. Les longueurs des branches ont été calculées avec

l'algorithme ACCTTRAN en choisissant S_1 comme racine arbitraire (cf. sous-section 2.3).

Ces deux arbres partiels à 4 feuilles sont donc identiques, les ordres 3 et 4 donneront donc le même résultat final que les ordres 1 et 2, respectivement. Continuons avec l'ordre 1 et insérons L_1 . En testant les 5 positions possibles, on trouve que l'unique position optimale est sur l'arête pendante menant à L_2 ; la longueur de l'arbre est maintenant 8. On teste enfin les 7 positions possibles pour l'insertion de L_5 et on trouve que l'unique position optimale se trouve sur l'arête pendante menant à L_1 . La longueur de l'arbre final est 10. On l'enracine en ajoutant un sommet racine sur l'arête pendante menant à L_6 . Cet arbre est le même que l'arbre solution de la figure 4. Avec l'ordre 2, on insère d'abord L_5 , et on a deux positions optimales possibles, soit sur l'arête pendante menant à L_2 , soit sur l'arête pendante menant à L_6 pour une longueur totale de 8 dans les deux cas. Dans le premier cas, l'insertion de L_1 nous fait bien retomber sur l'arbre solution de longueur 10. Dans le second cas, l'insertion de L_1 est optimale sur l'arête pendante menant à L_2 , on obtient alors un arbre final sous-optimal de longueur 11 identique au premier arbre de la figure 7. On rejette donc cet arbre et on ne conserve que le précédent.

5.4. - Les avantages et les limites de l'addition pas à pas

L'addition pas à pas respecte la fonction objectif du problème et toutes ses propriétés fondamentales. L'exploration de tous les ordres d'insertion retenus et de toutes les positions optimales à chaque étape d'insertion nous a permis de trouver l'arbre de longueur minimale correspondant à la matrice M sans avoir besoin d'essayer les 105 arbres dichotomiques possibles pour 6 feuilles. Il faut cependant avouer que ces manipulations demandent une certaine quantité de tâtonnements. On peut, bien sûr, ne conserver dès le départ qu'un

seul ordre d'insertion : les chances de trouver l'arbre le plus court restent « excellentes ». L'exemple précédent montre d'ailleurs que les quatre ordres essayés mènent bien à l'arbre optimal. C'est aux positions d'insertion également parcimonieuses qu'il faut prêter davantage attention. Le fait de devoir calculer la longueur de chaque arbre partiel, et donc résoudre pour chacun le problème de « petite parcimonie » (cf. sous-section 2.3), peut sembler très fastidieux, mais des techniques géométriques permettent de dessiner rapidement tous ces arbres à la main en se servant d'un quadrillage orthogonal (typiquement la grille d'une feuille à grands carreaux) tout en simplifiant les calculs de longueurs (Aubert, 2025). Ceci dit, l'utilisation d'un logiciel pédagogique adapté est aussi une option envisageable pour automatiser ces tâches.

6. – L'heuristique de l'analyse de compatibilité

Une seconde approche pour résoudre le problème de l'inférence phylogénétique consiste à exploiter d'abord les caractères compatibles afin de réduire fortement l'espace des solutions à explorer.

6.1. - La description de l'analyse de compatibilité

Chaque colonne C_j de M représente un caractère binaire (soit un vecteur colonne ne comportant que des « 1 » et des « 0 »). Deux colonnes C_p et C_q sont dites compatibles s'il existe un arbre où chacune n'exige qu'un seul changement (les « 1 » et les « 0 » forment deux sous-arbres connexes pour chaque colonne). Pour des colonnes binaires, ce critère est équivalent à une famille non croisée : C_p et C_q sont compatibles (forment une famille non

croisée) s'il n'apparaît pas simultanément les quatre couples $\{00,01,10,11\}$ quand on lit les deux colonnes ligne par ligne (Le Quesne, 1969 ; Meacham, 1981 ; Meacham et Estabrook, 1985). On construit alors le graphe de compatibilité G dont les sommets sont les colonnes et où chaque arête relie une paire compatible. Chercher un sous-ensemble le plus grand possible de colonnes mutuellement compatibles revient à chercher une clique maximum dans G . S'il y a plusieurs cliques maximums, on retient leur intersection. Puisque toute clique définit une phylogénie parfaite pour les colonnes qui la composent, la clique maximum (ou l'intersection des cliques maximums) détermine de manière unique un squelette d'arbre vraisemblable.

La construction de ce squelette aboutit toujours à la même topologie quel que soit l'ordre dans lequel on utilise les colonnes pour effectuer les regroupements successifs. Les polytomies restantes (sommets de degré strictement supérieur à 3) peuvent être résolues en essayant toutes les configurations locales possibles et en retenant celles qui produisent l'arbre de longueur minimale. L'idée générale de cette méthode est donc de partir d'un constat simple : dans un arbre, chaque fois qu'on traverse une dimension (c'est-à-dire qu'une colonne change de 0 à 1 ou de 1 à 0), on ajoute du cout ; et on n'a besoin de retraverser la même dimension plusieurs fois que lorsque des colonnes sont en conflit (incompatibles, croisées). On commence donc par extraire un squelette d'arbre qui est le produit d'un maximum de colonnes non conflictuelles (non croisées) : elles regroupent les lignes en

paquets qu'on peut relier avec une seule traversée de chaque dimension retenue. Ensuite, on utilise les colonnes restantes en veillant à les retraverser le moins possible. En minimisant le nombre de colonnes qui doivent être retraversées, on minimise indirectement le nombre total de traversées multiples et donc, au bout du compte, la longueur totale de l'arbre (Meacham, 1981 ; Meacham et Estabrook, 1985 ; Estabrook, 2008).

6.2. - Un exemple d'analyse de compatibilité

Pour la matrice M , on exclut de l'analyse de compatibilité la colonne C_7 qui est non informative (et donc nécessairement compatible avec toutes les colonnes). Commençons la construction du graphe de compatibilité en prenant par exemple les colonnes C_1 et C_2 . En lisant ces deux colonnes ligne par ligne, on observe successivement les couples 01, 11, 11, 11, 10, 00. Comme les quatre sortes de couples sont présents, on en déduit que les colonnes C_1 et C_2 ne sont pas compatibles, on ne relie donc pas les sommets C_1 et C_2 par une arête dans le graphe de compatibilité. Si on s'intéresse maintenant à la paire de colonnes C_2 et C_3 , on lit dans l'ordre les couples 11, 11, 11, 10, 00, 00. La combinaison 01 est absente de cette liste, donc les colonnes C_2 et C_3 sont compatibles, on relie alors les sommets C_2 et C_3 par une arête dans le graphe de compatibilité. En poursuivant l'analyse jusqu'au bout, on obtient le graphe de compatibilité G présenté dans la figure 9. Son unique clique maximum est $\{C_4, C_5, C_6\}$.

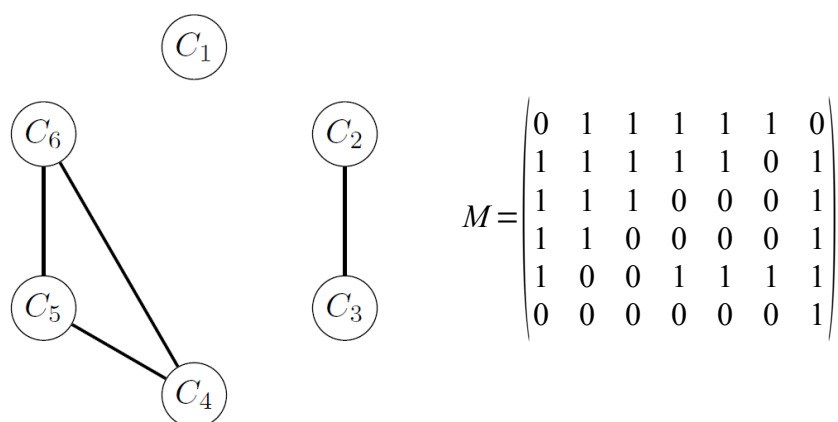


Figure 9 : Graphe de compatibilité G (à gauche) des colonnes de M (à droite).

Comme le montre la figure 10, on commence par créer un sommet de Steiner S_1 auquel on relie directement toutes les feuilles. Les colonnes C_4 et C_5 sont identiques $(110010)^T$, elles permettent de regrouper les lignes L_1 , L_2 et L_5 d'une part, et L_3 , L_4 et L_6 d'autre part. Notez bien qu'on regroupe à la fois en fonction des « 1 » partagés et des « 0 »

partagés, ce qui revient à scinder en fonction des différences ; la nouvelle arête $\{S_1, S_2\}$ sert alors de frontière à cette bipartition (la logique appliquée ici est donc bien distincte de celle qu'on a vue avec les groupes emboîtés et la polarisation à priori). À ce stade, on a donc deux sommets de Steiner reliés entre eux, l'un connecté à L_1 , L_2 et L_5 , et l'autre connecté à L_3 , L_4 et L_6 .

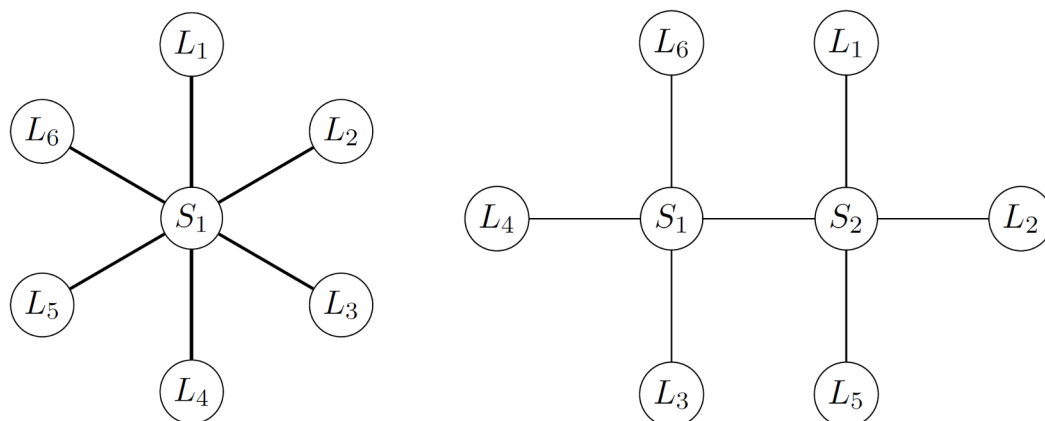


Figure 10 : Les deux premières étapes de construction de l'arbre à partir de la matrice M et de son graphe de compatibilité G . À chaque étape de cette construction, on cherche mentalement à scinder un sommet interne en deux, en redistribuant les arêtes qui y étaient attachées aux deux sommets issus de cette scission, de manière à partitionner les états du caractère considéré à cette étape.

On utilise ensuite $C_6 = (100010)^T$ pour scinder L_1 et L_5 d'une part, et L_2 , L_3 , L_4 et L_6 d'autre part. On obtient finalement le squelette

d'arbre présenté dans la figure 11. Il n'y a que trois configurations dichotomiques possibles pour les trois feuilles L_3 , L_4 et L_6 autour de S_1 , il suffit donc de les tester toutes les trois et de

retenir l'arbre de longueur minimale. Enfin, on enracine en plaçant un sommet racine R sur l'arête pendante menant à l'extragroupe. On trouve ainsi sans problème l'arbre solution de longueur 10 de la figure 4. Dans des cas plus simples, mais pas nécessairement triviaux, l'analyse de compatibilité peut aboutir directement à l'arbre solution sans qu'aucune polytomie n'apparaisse (Aubert, 2023a, 2024). Enfin, plutôt que d'essayer les différentes configurations possibles des éventuelles polytomies restantes pour les tester directement avec le critère de parcimonie, il reste envisageable de

poursuivre l'application de la logique de compatibilité en sélectionnant par exemple le caractère restant qui est compatible avec le plus grand nombre d'autres caractères. Ici, on aurait le choix entre C_2 et C_3 à égalité, mais C_2 ne permet pas de scinder la polytomie puisque seuls L_6 et L_5 sont dans l'état 0 pour ce caractère. On utilise donc C_3 pour scinder L_6 et L_4 d'une part, et L_3 et S_2 d'autre part (S_2 étant nécessairement dans l'état 1 puisque L_2 et L_1 le sont). On aboutit alors encore une fois à l'arbre solution de la figure 4.

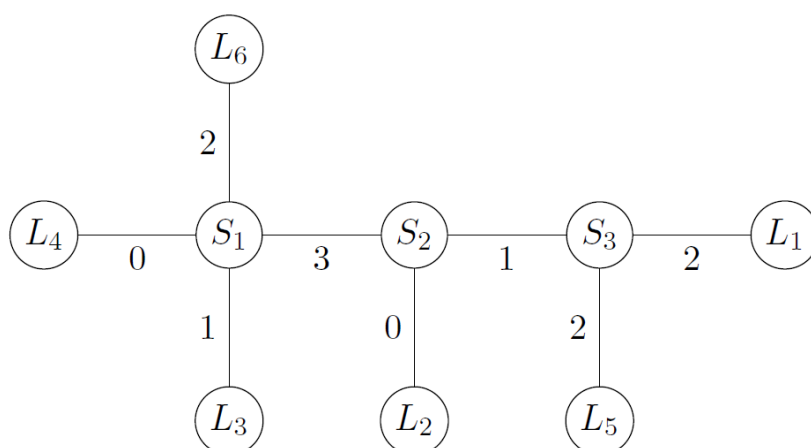


Figure 11 : Squelette d'arbre produit par l'analyse de compatibilité. L'arbre n'est pas encore enraciné à ce stade. Il a une longueur totale de 11, ce qui est normal puisque la polytomie sur S_1 est sous-optimale et doit être réarrangée pour parvenir à l'arbre de longueur minimale.

6.3. - Les avantages et les limites de l'analyse de compatibilité

Cette approche est immunisée contre les errements de la méthode des groupes emboîtés et des méthodes par polarisation a priori puisqu'elle respecte les propriétés fondamentales du problème posé. Elle est, de plus, assez simple à enseigner et à mettre en place en classe. La construction du squelette principal de l'arbre ne nécessite pas d'explorer différents ordres d'insertion ou différentes configurations des feuilles. L'absence de tâtonnements trop fastidieux permet ainsi de réaliser rapidement cette méthode, même à la main (Mea-

cham, 1981 ; Aubert, 2023a, 2024). Il est bien sûr encore plus pratique de laisser à un logiciel pédagogique adapté comme PhyloGraphe (Aubert, 2023b) le soin d'évaluer la longueur d'une topologie, surtout si ces évaluations doivent être répétées pour plusieurs arbres candidats.

La recherche d'une clique maximum dans un graphe donné est un problème NP-complet en théorie, mais pour des graphes d'une taille raisonnable d'une douzaine de sommets au maximum (comme ceux qu'on peut rencontrer dans les problèmes phylogénétiques les plus difficiles de niveau terminale ou en CPGE), la

solution est souvent visuellement évidente. Comme toute euristique, l'analyse de compatibilité ne garantit pas de trouver le minimum global. Il est possible de construire des matrices adversariales avec lesquelles cette méthode se laisse coincer dans un minimum local. En pratique cependant, l'analyse de compatibilité permet presque toujours de trouver l'arbre le plus parcimonieux, même avec des matrices relativement complexes. Il s'agit donc d'une méthode robuste qu'on peut appliquer sur des cas réalistes de manière efficace.

7. – L'introduction progressive des notions mathématiques

Après cet examen critique des méthodes scolaires d'inférence phylogénétique en SVT et une discussion de deux méthodes candidates pouvant les remplacer, nous allons envisager comment les notions mathématiques que ce problème nécessite de mobiliser pourraient être introduites progressivement au cours de la scolarité.

7.1. - L'enjeu didactique de la cohérence interdisciplinaire

Les analyses précédentes ne conduisent pas à exiger des élèves qu'ils maîtrisent d'emblée toute la technicité de l'inférence phylogénétique. Elles montrent plutôt qu'un enseignement cohérent suppose d'installer progressivement quelques idées simples : expliciter une fonction objectif, comparer plusieurs solutions candidates, manipuler et construire des graphes en tenant compte de certaines contraintes, appliquer les étapes d'un algorithme, etc. L'intérêt d'une telle progression est que les notions mathématiques en jeu peuvent être travaillées bien avant d'être réinvesties dans le contexte phylogénétique. Distance, matrice, réseau, arbre, sommet, arête, optimisation, orientation : toutes ces notions

apparaissent dans des problèmes classiques, indépendants de la biologie, mais suffisamment proches pour éclairer ensuite le problème phylogénétique. L'enjeu n'est donc pas d'introduire prématurément toute la formalisation de la cladistique, mais de construire progressivement les outils intellectuels qui permettront d'en comprendre la logique.

7.2. - Les distances dans les rues et à vol d'oiseau

Une première notion utile est celle de distance. Dans la vie courante, les élèves rencontrent très tôt deux façons différentes d'évaluer l'éloignement entre deux points : la distance « à vol d'oiseau », qui correspond à la longueur du segment joignant directement ces points, et la distance parcourue dans un réseau de rues orthogonales, où on doit se déplacer seulement de manière horizontale ou verticale. Dans le premier cas, on mobilise la géométrie euclidienne, notamment le théorème de Pythagore à partir de la classe de 4^e dans les programmes actuels (DGESCO, 2026) ; dans le second, on somme des déplacements élémentaires dans deux directions indépendantes. Cette distinction fournit une excellente entrée vers la distance de Manhattan (dont la distance de Hamming n'est qu'un cas particulier pour les données binaires). Le passage par un quadrillage urbain permet ainsi de préparer une idée essentielle : selon le problème posé, la distance pertinente n'est pas toujours la plus familière, et elle dépend des déplacements autorisés. C'est précisément ce qui se passe ensuite lorsqu'on compte des transformations de caractères. En particulier, la réponse à la question « qui est plus proche de qui ? » dépend de la manière dont les distances sont mesurées. Dans l'exemple de la figure 12, A est plus proche de B que de C en distance euclidienne, mais plus éloigné de B que de C en distance de Manhattan. Une telle activité peut être introduite très tôt, dès le cycle 3, sans faire réfè-

rence au théorème de Pythagore : il suffit de comparer des trajets sur quadrillage et des segments tracés dans le plan mesurés avec une règle graduée.

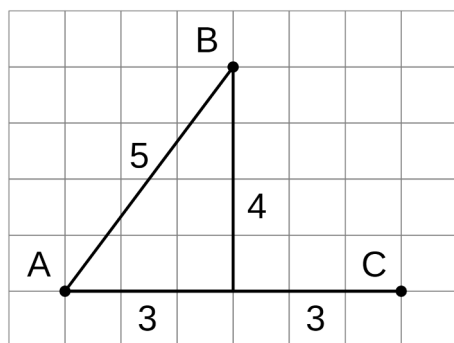


Figure 12 : Un exemple de comparaison géométrique entre les distances euclidienne et de Manhattan. Ici on a $d_M(A, B) = 7 > d_M(A, C) = 6$, mais $d_E(A, B) = 5 < d_E(A, C) = 6$.

Le même problème se pose en phylogénétique, la réponse dépend de la définition présu- mée de la notion de distance impliquée par la question « qui est plus proche de qui ? ». La distance peut ainsi être mesurée par le nombre de différences de caractères (distance de Ham- ming), le nombre de transformations de carac- tères le long du chemin sur un arbre (cette dis- tance est dite patristique, c'est la somme des distances de Hamming entre toutes les paires de sommets adjacents qui composent le che- min), ou bien l'ancienneté relative du dernier ancêtre commun sur l'arbre enraciné (cette dis- tance est dite cladistique). Par exemple, si on reprend notre matrice M et l'arbre solution de la figure 4, on trouve que L_5 est cladistique- ment plus proche de L_1 que de L_2 , mais est plus éloigné de L_1 que de L_2 en termes de dis- tances patristiques et de Hamming (4 contre 3). De même, L_5 a une distance de Hamming plus petite avec L_6 qu'avec L_3 (4 contre 5), mais la distance patristique entre L_5 et L_6 est plus grande qu'entre L_5 et L_3 (8 contre 5). C'est d'ailleurs pour cette raison simple qu'un algorithme naïf du type « regrouper successi-

vement les espèces qui se ressemblent le plus » ne fonctionne pas pour faire de l'inférence phylogénétique.

7.3. - Des optimisations de réseaux routiers

Une deuxième famille de problèmes per- met d'introduire la notion de fonction objectif. On peut, par exemple, considérer plusieurs villes placées sur une carte et comparer diffé- rents projets de réseaux routiers censés les re- lier. Le but est alors de déterminer lequel cou- terait le moins cher à construire, donc celui qui a la plus petite longueur totale. Dans une pre- mière étape, il suffit de comparer quelques ré- seaux déjà proposés. Selon le niveau, les lon- gueurs entre les points peuvent être calculées de manière explicite grâce à leurs coordonnées ou bien simplement mesurées avec une règle graduée.

Les contraintes précises qu'on peut impo- ser correspondent à des problèmes d'optimisa- tion très classiques. Par exemple, si on impose l'absence de cycles et l'interdiction d'ajouter de nouveaux sommets, alors il s'agit du pro- blème de l'arbre couvrant minimal (ou *mini- mum spanning tree*). L'avantage de cette va- riante est qu'il existe des algorithmes simples et rapides, comme celui de Kruskal (1956), permettant de trouver à coup sûr l'arbre opti- mal. Si on autorise l'ajout de sommets inter- médiaires, alors le problème devient celui de l'arbre de Steiner minimal dans le plan eucli- dien, ce qui est très proche du problème de l'inférence phylogénétique (comme expliqué dans la section 2). Il est alors trop difficile de trouver la solution optimale, mais il est tou- jours possible de comparer différentes proposi- tions et de déterminer la meilleure comme le montre la figure 13. Tous les exemples que nous avons abordés témoignent que les notions mathématiques mobilisées par le problème de l'inférence phylogénétique ne sont ni hors de portée des élèves ni réservées à un enseigne- ment spécialisé.

UNE CLARIFICATION LOGIQUE ET
MATHÉMATIQUE NÉCESSAIRE DE
L'INFÉRENCE PHYLOGÉNÉTIQUE EN SVT

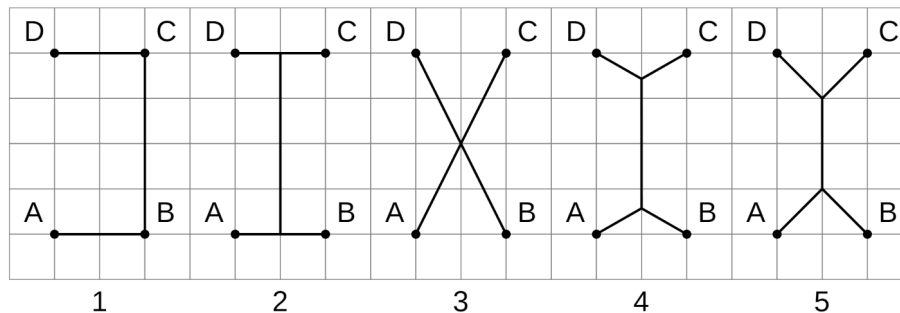


Figure 13 : Une comparaison de cinq projets de réseaux routiers en forme d'arbres reliant quatre villes disposées aux coins d'un rectangle de taille 4×2 . Ces projets ont respectivement des longueurs de 8, 8, 8,94, 7,46 et 7,66.

Conclusion

Les difficultés liées à l'apprentissage des arbres phylogénétiques ont fait l'objet de nombreux travaux, principalement consacrés à leur lecture ; leur construction à partir de matrices de caractères a été beaucoup moins étudiée (Meisel, 2010 ; Tyrrell et Schmiemann, 2024). Dans cet article, nous avons clarifié la nature du problème que les exercices scolaires de phylogénétique prétendent traiter. Cette mise au point a permis de dissiper certains mirages : les recettes scolaires, qui « marchent » parce qu'on choisit uniquement des exemples favorables, ne résolvent pas le bon problème.

Nous avons montré que la méthode des « groupes emboîtés » et les méthodes par « polarisation a priori » violent des propriétés et des invariances essentielles (sensibilité au recodage arbitraire des colonnes 0/1, sensibilité aux colonnes non informatives, sensibilité aux duplications de lignes), mélangent confusément l'inférence de la topologie et son enracinement, et n'utilisent pas la vraie fonction objectif. Il est d'ailleurs révélateur que le critère de parcimonie ne soit jamais mentionné dans ces méthodes et ne soit donc jamais enseigné aux élèves du collège et du lycée. On ne leur apprend jamais non plus à calculer la longueur d'un arbre de topologie fixée (c'est-à-dire à résoudre le problème de « petite parcimonie »).

Le résultat est attendu : des topologies fausses et des longueurs supérieures à l'optimum, sauf à truquer les matrices pour sauver ces procédures.

À l'inverse, deux voies simples peuvent être enseignées pour résoudre le problème posé. D'une part, l'addition pas à pas classique (avec un ordre d'insertion judicieusement choisi, et une évaluation systématique du coût à chaque étape) permet d'atteindre l'optimum en s'appuyant explicitement sur la fonction objectif. D'autre part, l'analyse de compatibilité commence par les caractères qui peuvent être combinés sans conflit afin de construire une première structure de l'arbre, puis utilise le critère de parcimonie pour départager localement les régions non résolues. Ces deux approches parviennent quasiment toujours à trouver l'arbre de longueur minimale dans les cas généraux pour lesquels les méthodes scolaires actuelles se révèlent radicalement inadaptées. Je plaide donc pour un basculement clair de l'enseignement : abandonner les pratiques qui s'opposent à la logique, transgressent les mathématiques, ignorent le consensus scientifique (Meacham, 1984 ; Nixon et Carpenter, 1993, 2013 ; Kitching et al., 1998 ; Aubert, 2023a), et enseigner à la place des notions et des méthodes véritablement rigoureuses et intellectuellement honnêtes.

On ne fait pas de la science si on n'assume pas le problème réel, tel qu'il est : mathématique, exigeant, mais parfaitement accessible et didactisable. L'apprentissage de sa résolution peut se faire très progressivement en mettant en place année après année les notions et les outils d'analyse nécessaires, mais sans prendre de raccourcis incohérents comme c'est le cas actuellement. Entre la lecture simple d'un tableau de caractères et la résolution complète du problème de « grande parcimonie », il existe tout un ensemble d'activités intermédiaires qui permettent d'installer progressivement les concepts utiles si elles étaient implémentées dans les programmes de mathématiques (comme les exemples développés dans la section 7) et de SVT (comme le problème de « petite parcimonie »). Il ne faut donc pas se laisser aveugler par un faux dilemme entre faisabilité didactique et rigueur mathématique. Les nouveaux programmes de mathématiques du cycle 4 insistent d'ailleurs sur la place centrale du raisonnement et sur la rigueur logique (DGESCO, 2026).

Précisons enfin que notre but n'était pas d'incriminer les professeurs de SVT : chacun, dans sa classe, applique des prescriptions officielles parfois discutables, sans toujours disposer du temps ni de la formation nécessaires pour en mesurer les effets. La responsabilité est d'abord institutionnelle. L'intention de cet article était d'interpeler la communauté des mathématiciens et des professeurs de mathématiques qui sont naturellement concernés par la manière dont les mathématiques sont mobilisées hors de leur discipline. Leur regard peut aider à corriger les dérives, expliciter les objectifs et rigorer les démarches : l'inférence phylogénétique est un problème mathématique, elle doit donc être enseignée dans un cadre mathématique clair.

Damien AUBERT

Professeur de SVT,
Lycée Jeanne d'Arc, Clermont-Ferrand

Références bibliographiques

- Aubert, D. (2023a). Il ne faut pas polariser les caractères ! *Biologie Géologie*, 3-2023(409), 85–103.
- Aubert, D. (2023b). *PhyloGraphe. Application web pour construire des arbres phylogénétiques*.
<https://evoluscope.fr/phylographe/>
 (consulté le 21/10/2025).
- Aubert, D. (2024). *Les Groupes Emboîtés au Collège (SVT)*. Évoluscope, chaîne YouTube.
<https://youtu.be/UK7yCkmBL1w>
 (consulté le 21/10/2025).
- Aubert, D. (2025). *Construire un Arbre Phylogénétique à la Main Facilement (SVT)*. Évoluscope, chaîne YouTube.
<https://www.youtube.com/watch?v=28ka-toylKY>
 (consulté le 21/10/2025).
- Chanet, B., & Lusignan, F. (2010). *Classer les animaux au quotidien : du cycle 2 à la 6^e. 2^e édition*. SCÉRÉN-CRDP de Bretagne, cédérom inclus.
<https://www.reseau-canope.fr/notice/classer-les-animaux-au-quotidien-du-cycle-2-a-la-6e>
 (consulté le 25/08/2026).
- Estabrook, G. F. (2008). Fifty years of character compatibility concepts at work. *Journal of Systematics and Evolution*, 46(2), 109–129.
- Farris, J. S. (1970). Methods for Computing Wagner Trees. *Systematic Zoology*, 19(1), 83–92.
- Felsenstein, J. (2004). *Inferring Phylogenies*. Sinauer Associates.

UNE CLARIFICATION LOGIQUE ET
MATHÉMATIQUE NÉCESSAIRE DE
L'INFÉRENCE PHYLOGÉNÉTIQUE EN SVT

- Fitch, W. M. (1971). Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Zoology*, 20(4), 406–416.
- Foulds, L. R., & Graham, R. L. (1982). The Steiner problem in phylogeny is NP-complete. *Advances in Applied Mathematics*, 3(1), 43–49.
- Hamming, R. W. (1950). Error detecting and error correcting codes. *The Bell System Technical Journal*, 29(2), 147–160.
- Hartigan, J. A. (1973). Minimum Mutation Fits to a Given Tree. *Biometrics*, 29(1), 53–65.
- Hervé, J.-C. (2009). *Phylogène et l'enseignement de l'évolution dans le secondaire*. Dernière mise à jour : 2025-02-15. ENS de Lyon - ACCES.
<https://acces.ens-lyon.fr/acces/thematiques/evolution/logiciels/phylogene/telechargements-enseignants/textes-de-cadrage/enseignement-evolution-secondaire>
(consulté le 21/10/2025).
- Hervé, J.-C. (2010). *Arbres phylogénétiques et compréhension de l'évolution*. Dernière mise à jour : 2024-05-02. ENS de Lyon - ACCES.
<https://acces.ens-lyon.fr/acces/thematiques/evolution/logiciels/phylogene/telechargements-enseignants/textes-de-cadrage/arbres-phylogenetiques>
(consulté le 21/10/2025).
- Hwang, F. K., Richards, D. S., & Winter, P. (1992). *The Steiner Tree Problem*. Elsevier.
- IMT Atlantique (2023). *Le sol et son rôle dans la croissance végétale. Guide enseignant*. Dernière mise à jour : 2023-07-18.
- Collection MERITE.
https://www.imt-atlantique.fr/sites/default/files/formations/innovations-pedagogiques/MERITE/ROBOT%20PEDAGO/LE%20SOL%20ET%20SON%20R%C3%94LE%20DANS%20LA%20CROISSANCE%20V%C3%89G%C3%89TALE_compressed.pdf
(consulté le 25/08/2026).
- Jünger, M., Reinelt, G., & Rinaldi, G. (1995). Chapter 4 The traveling salesman problem. Dans *Network Models. T. 7. Handbooks in Operations Research and Management Science* (p. 225–330). Elsevier.
- Kitching, I. J., Forey, P. L., Humphries, C. J., & Williams, D. M. (1998). *Cladistics: The Theory and Practice of Parsimony Analysis. 2nd edition*. Oxford University Press.
- Kruskal, J. B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, 7(1), 48–50.
- Le Quesne, W. J. (1969). A Method of Selection of Characters in Numerical Taxonomy. *Systematic Biology*, 18(2), 201–205.
- Lecointre, G., & Camus, G. (2007). *La classification du vivant, mode d'emploi*. Planet-Vie, ENS de Lyon.
<https://planet-vie.ens.fr/thematiques/evolution/classification-phylogenie/la-classification-du-vivant-mode-d-emploi>
(consulté le 25/08/2026).
- Lecointre, G. (dir.), Cariou, F., Duco, A., Guillot, G., Lebas, C., Le Louarn, M.-L.,

- Mardelle, P., & Nicol, É. (2008). *Comprendre et enseigner : La classification du vivant, 2^e édition*. Belin Éducation.
- Collectif d'enseignants (2024). *Enseignement scientifique Terminale, manuel scolaire*. Lelivrescolaire.fr, édition 2024.
<https://www.lelivrescolaire.fr/fiche/enseignement-scientifique-tle-2023>
(consulté le 25/08/2026).
- Meacham, C. A. (1981). A Manual Method for Character Compatibility Analysis. *Taxon*, 30(3), 591–600.
- Meacham, C. A. (1984). The role of hypothesized direction of characters in the estimation of evolutionary history. *Taxon*, 33(1), 26–38.
- Meacham, C. A., & Estabrook, G. F. (1985). Compatibility Methods in Systematics. *Annual Review of Ecology, Evolution, and Systematics*, 16, 431–446.
- Meisel, R. P. (2010). Teaching Tree-Thinking to Undergraduate Biology Students. *Evolution: Education and Outreach*, 3, 621–628.
- Nixon, K. C., & Carpenter, J. M. (1993). On Outgroups. *Cladistics*, 9(4), 413–426.
- Nixon, K. C., & Carpenter, J. M. (2013). More on Absences. *Cladistics*, 29(1), 1–6.
- Sankoff, D. (1975). Minimal Mutation Trees of Sequences. *SIAM Journal on Applied Mathematics*, 28(1), 35–42.
- Swofford, D. L. (2003). *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4*. Sinauer Associates.
- Swofford, D. L., & Maddison, W. P. (1987). Reconstructing ancestral character states under Wagner parsimony. *Mathematical biosciences*, 87(2), 199–229.
- Tyrrell, S., & Schmiemann, P. (2024). Building trees by juggling information and following rules: an expert interview study on tree-building and phylogenetic inference. *Evolution: Education and Outreach*, 17, article 9.

Références institutionnelles

- Direction générale de l'enseignement scolaire (DGESCO) (2020). *Programme officiel*.
<https://eduscol.education.fr/document/621/download> (consulté le 21/10/2025).
- Direction générale de l'enseignement scolaire (DGESCO) (2023a). *Programme d'enseignement du cycle de consolidation (cycle 3) - Annexe 2*.
<https://eduscol.education.fr/document/50990/download> (consulté le 21/10/2025).
- Direction générale de l'enseignement scolaire (DGESCO) (2023b). Programme d'enseignement scientifique de terminale générale. Annexe - Programme d'enseignement scientifique de terminale générale. *Bulletin officiel n° 25 du 22 juin 2023*.
<https://www.education.gouv.fr/sites/default/files/document/Annexe%20%E2%80%93%20Programme%20d%E2%80%99enseignement%20scientifique%20de%20%20terminale%20g%C3%A9n%C3%A9rale-365058.pdf>
(consulté le 25/08/2026).
- Direction générale de l'enseignement scolaire (DGESCO) (2026). *Programme de mathématiques pour le cycle 4 - Annexe 2*.
<https://www.education.gouv.fr/sites/default/files/document/>

UNE CLARIFICATION LOGIQUE ET
MATHÉMATIQUE NÉCESSAIRE DE
L'INFÉRENCE PHYLOGÉNÉTIQUE EN SVT

Annexe%20%20%E2%80%93%20Progr
amme%20de%20math%C3%A9matiques
%20pour%20le%20cycle%204-
480716.pdf
(consulté le 12/03/2026).

Ministère de l'Éducation nationale, de la
Jeunesse et des Sports (MENJS) (2021).
Programmes de sciences de la vie et de la
Terre (BCPST) - Annexe 1. Classes

préparatoires aux grandes écoles - Filière
scientifique. *Annexe de l'arrêté du 16
avril 2021 (NOR ESRS2108111A), publié
au Bulletin officiel n° 26 du 1^{er} juillet
2021.*

[https://cache.media.education.gouv.fr/
file/20/94/8/
ensecsup111_annexes_1407948.pdf](https://cache.media.education.gouv.fr/file/20/94/8/ensecsup111_annexes_1407948.pdf)
(consulté le 21/10/2025).