Méthode de Parcimonie

Parcimonie

Pour les systématiciens évolutionnistes, la similitude globale ne peut fournir la base de la reconstruction phylogénétique en raison des homoplasies qu'elles englobent et qui ne sont que des coïncidences et non le fruit d'une évolution commune (Simpson 1961, Mayr 1969). Pour eux, seules les homologies permettent la construction phylogénétique.

Pour les cladistes (Hennig 1950 qui en a posé les bases, puis Eldredge et Cracraft 1980, Wiley 1981 ; Nelson et Platnick 1981 ; Schoch 1986 ; Matile et al 1987 ; d'Udekem-Gevers 1990), le concept d'homologie lui-même doit clarifié en distinguant les caractères dans un état ancestral (plésiomorphes) de ceux qui sont dans un état dérivé (apomorphes) et sont les seuls à refléter une origine commune.

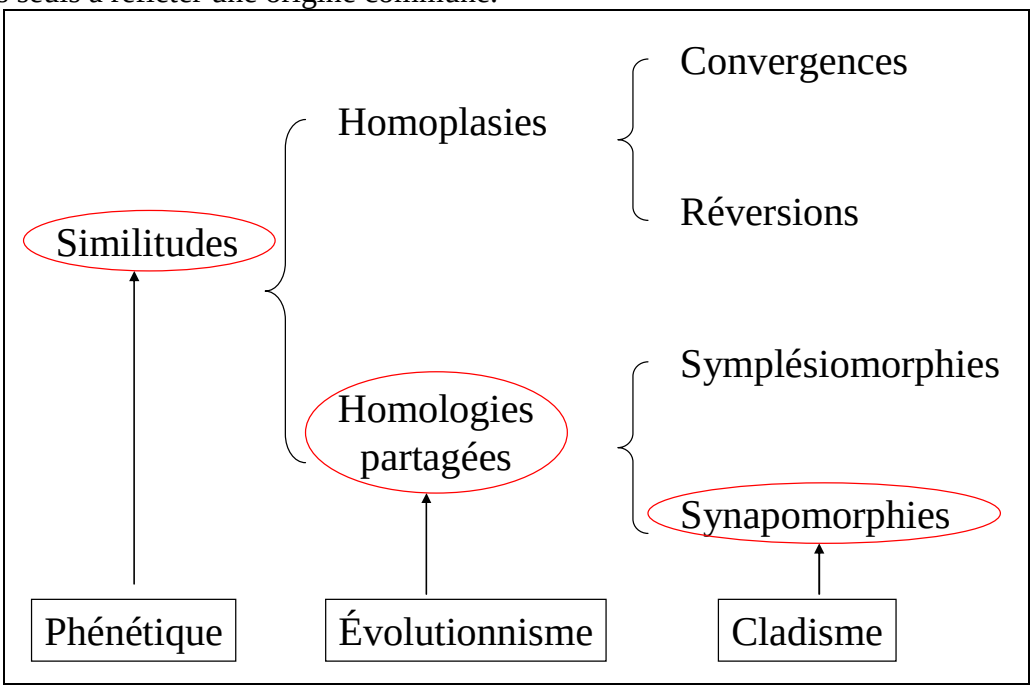
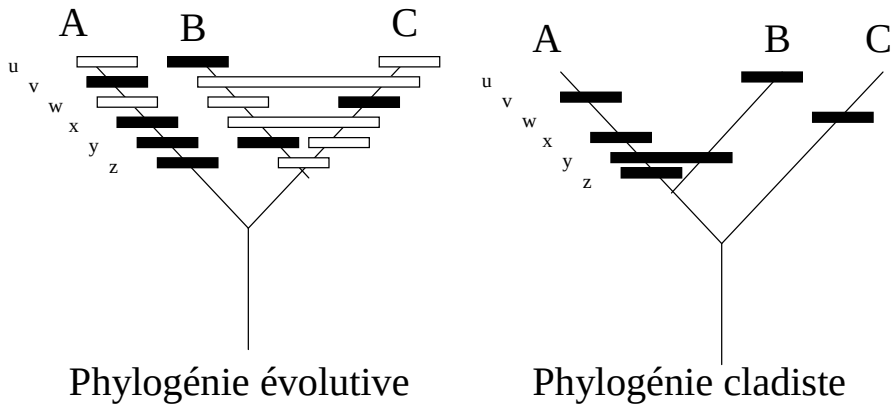


Figure III- 1. Phénétique, évolutionnisme et cladisme.

	A	B	C
u	0	1	0
v	0	0	1
w	1	0	0
x	1	0	0
y	1	1	0
z	1	0	0

	Homologies partagées	Apomorphies partagées
A et B	2	1
B et C	3	0
A et C	1	0



Phylogénie évolutive

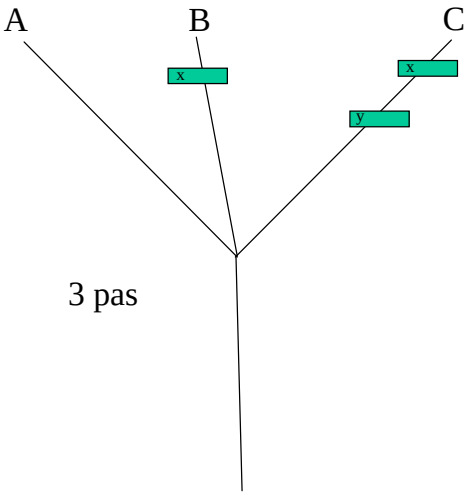
Phylogénie cladiste

Figure III- 2. Utilisation des changements d'état des caractères selon deux écoles.

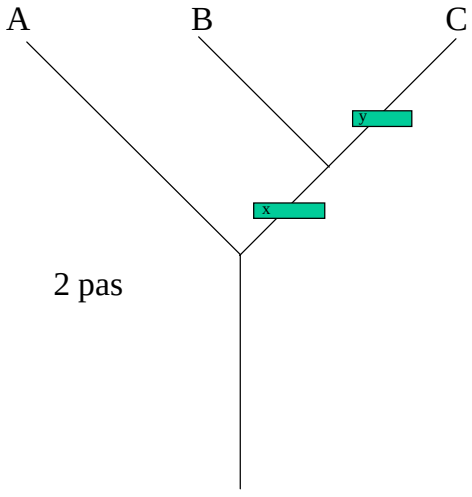
Principe de parcimonie

On construit les arbres vraisemblables et on choisit celui qui a la longueur minimale

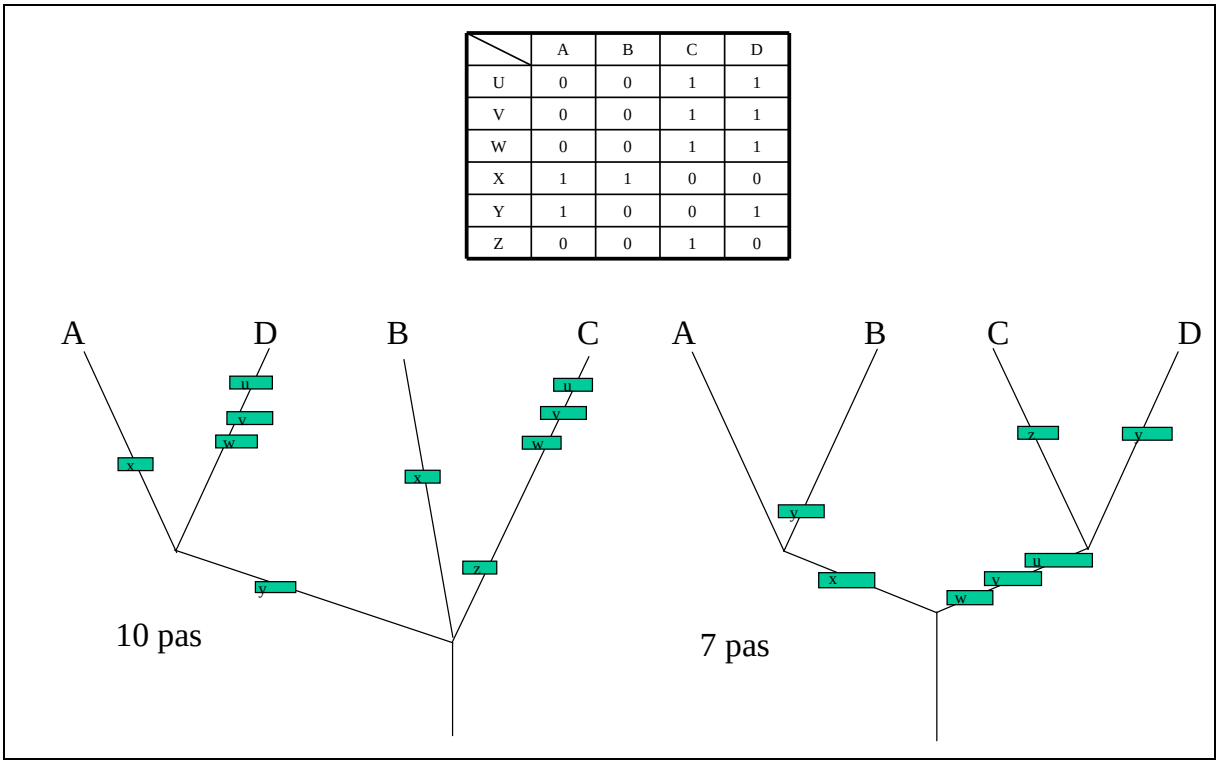
	A	B	C
X	0	1	1
y	0	0	1



3 pas



2 pas



	A	B	C	D
U	0	0	1	1
V	0	0	1	1
W	0	0	1	1
X	1	1	0	0
Y	1	0	0	1
Z	0	0	1	0

Figure III- 3. Dans le premier cas, en vertu du principe de parcimonie l'arbre choisi sera celui de 2 pas et dans le second, celui de 7 pas.

Orientation

Ontogénie

Comment définir les caractères plésiomorphes et apomorphes. Lorsqu'il s'agit de caractères morphologiques, on a des critères ontogéniques : dans la suite du développement ce qui apparaît en premier est plus primitif que ce qui apparaît ensuite.

Chez la limande et la sardine la position des yeux diffère :de part et d'autre de la tête ou du même coté. Si l'on observe le développement embryonnaire on voit chez la limande les yeux d'abord de part et d'autre de la tête puis, au cours du développement l'un passe de l'autre côté. L'état yeux du même côté de la tête est donc un caractère dérivé ou apomorphe.

Il existe des exceptions dont le classique axolotl qui bien que présentant un caractère plésiomorphe (branchies à l'état adulte, phénomène de néoténie) est cependant un proche parent des salamandres terrestres et non un de leurs ancêtres (une analyse de 41 caractères permet d'établir une phylogénie non ambiguë).

Paléontologie

Si dans un groupe monophylétique, l'état d'un caractère est présent chez les fossiles anciens et l'autre état chez des fossiles plus récents, le premier est l'état plésiomorphe et le second , l'état apomorphe. Il faut cependant que les parentés entre fossiles ne soient pas trop lointaines. Tout le problème réside dans la détermination de l'étroitesse des liens.

Sans discernement et en appliquant ce principe on peut dire que les blattes qui infestent les sous-sols des grandes villes sont plus évoluées que les mammouths qui

vivaient il y a 15 000ans (1981, Nelson et Platnick). Et pourtant… ces deux espèces sont bien monophylétiques appartenant toutes deux aux métazoaires.

Ceci montre bien que ce caractère extrinsèque aux organismes ne peut s'appliquer indépendamment d'un critère principal intrinsèque

Chorologie (distribution géographique)

On admet que lorsqu'une espèce A se divise en deux espèces B et C, l'état transformé (apomorphe) apparaît chez l'espèce la plus éloignée géographiquement de l'espèce initiale. Ici encore il s'agit d'un critère secondaire qui doit être accompagné d'un critère principal.

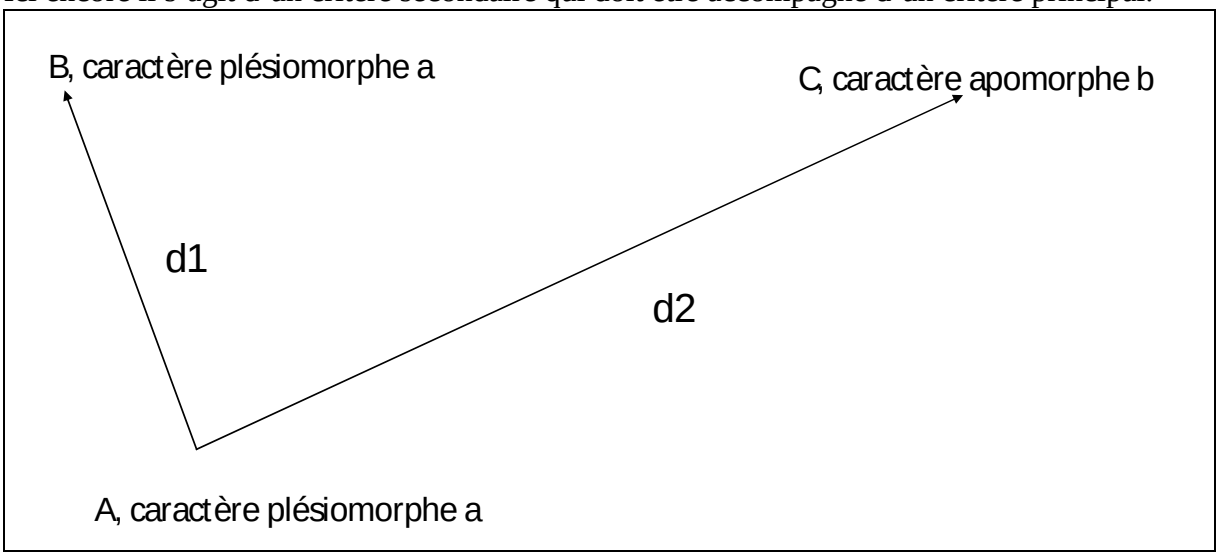


Figure III- 4. Le principe de chorologie détermine l'espèce C comme dérivée de l'espèce A.

Extra-groupe

Si un caractère observé dans le groupe étudié est également présent à l'extérieur de ce groupe, il est plésiomorphe pour le groupe étudié, s'il n'est présent qu'à l'intérieur du groupe, il est apomorphe. Cette définition permet d'étudier le degré d'universalité de l'état du caractère. Cette comparaison ne doit pas se réduire au seul groupe frère du groupe étudié.

Combien faut-il prendre d'extra groupes ?

Avec un seul extra groupe : un caractère dérivé spécifique à l'extra groupe risque d'être confondu avec un caractère plésiomorphe.

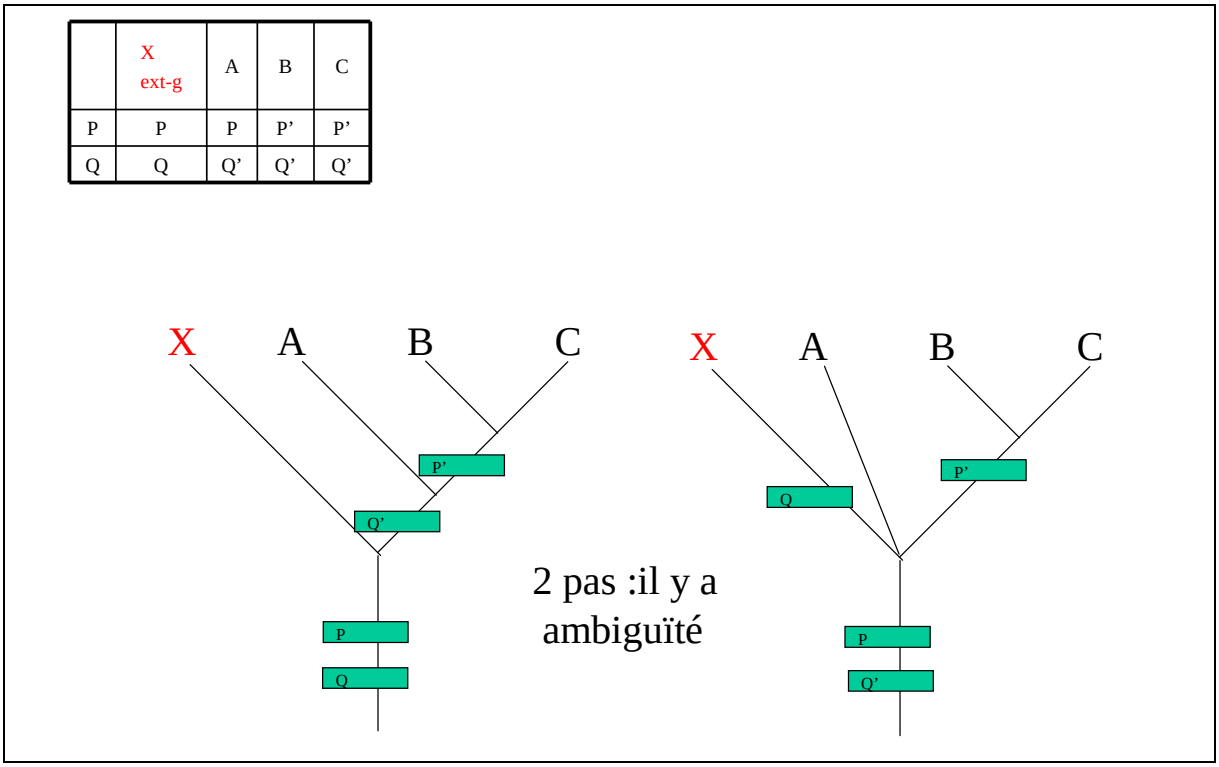


Figure III- 5.Un seul extra groupe ne suffit pas toujours à lever les ambiguïtés.

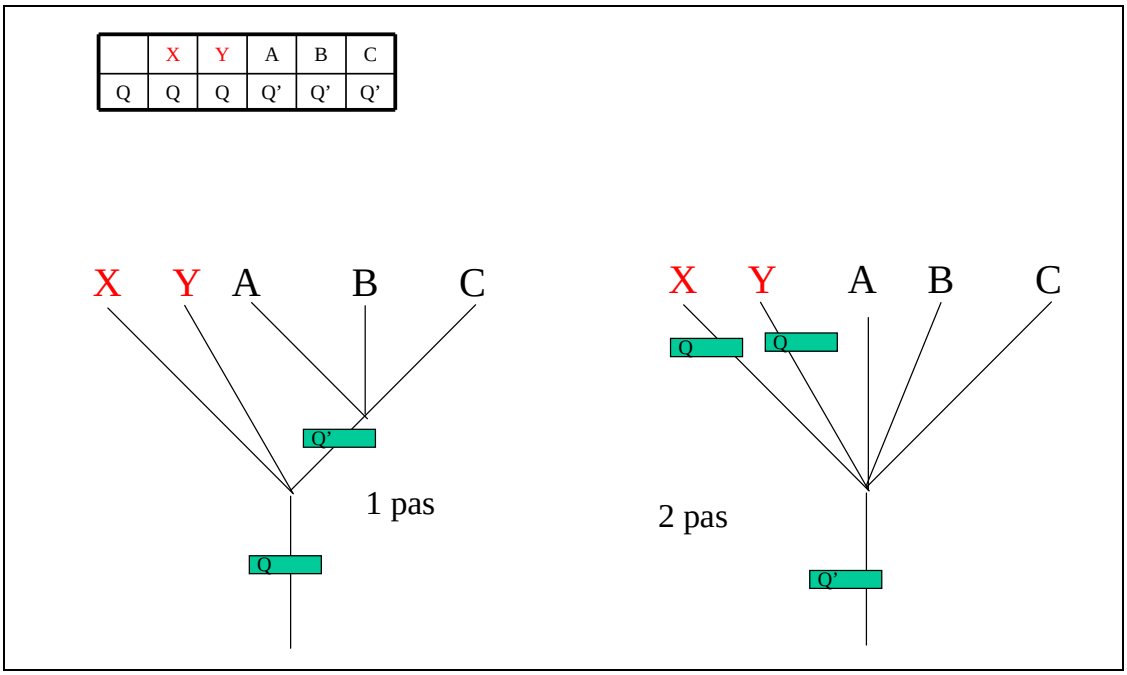


Figure III- 6. Avec deux extra groups l'ambiguïté sur le caractère Q est levée.

Avec les caractères moléculaires seul le critère extra groupe est utilisable. Deux extra groupes non monophylétiques permettent de contourner le problème (parfois plusieurs extra groupes si parmi les deux un a une forme plésiomorphe et l'autre apomorphe :T3 et T4).

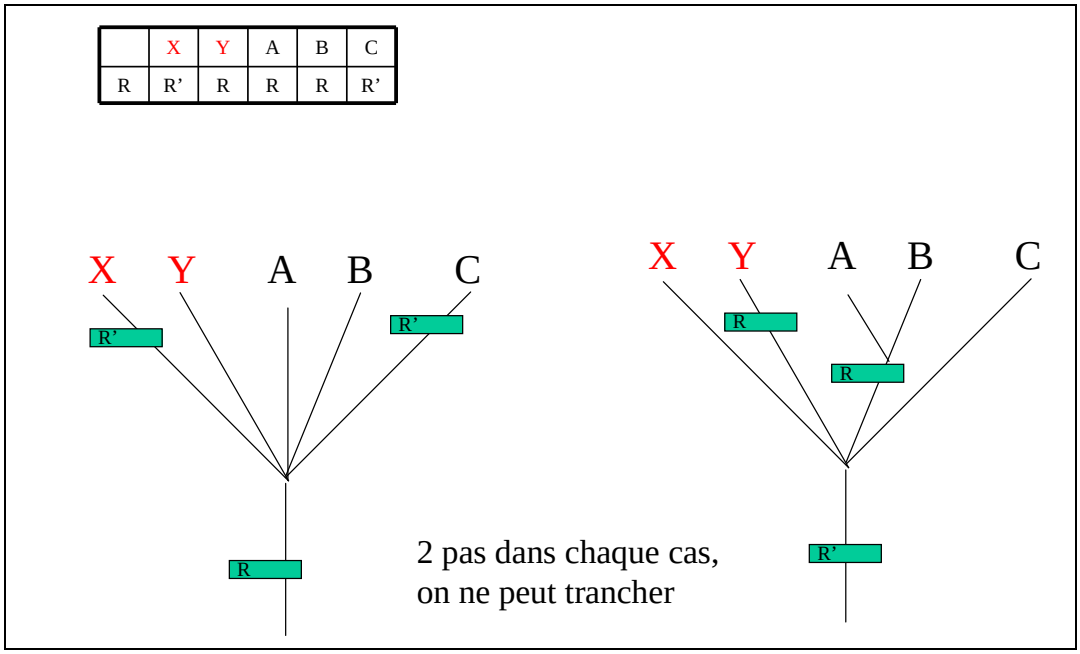


Figure III- 7. Deux extra groups peuvent rester ambigüs.

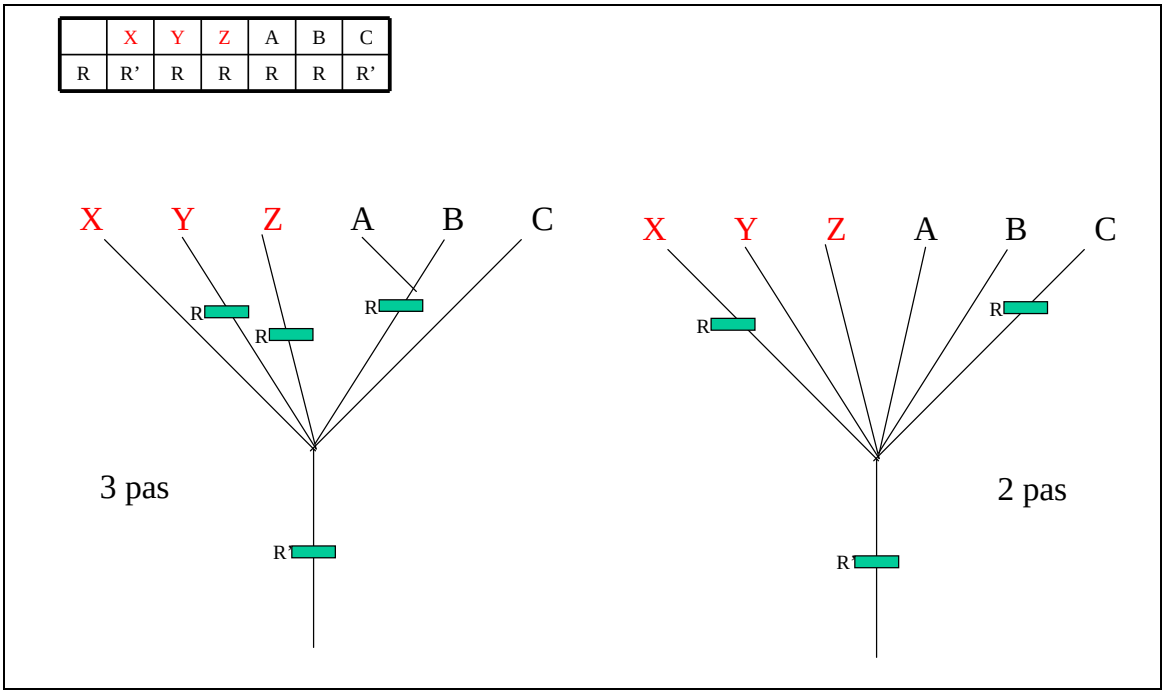


Figure III- 8. Plusieurs extra groupes permettent une meilleure information sur l’état d’un caractère (pléiosomorphe ou apomorphe).

De toute façon, le choix des groupes externes est un a priori qui repose sur d’autres connaissances préétablies. Si on remet en cause cette propriété d’extra groupe, on peut trouver un arbre plus court (T5).

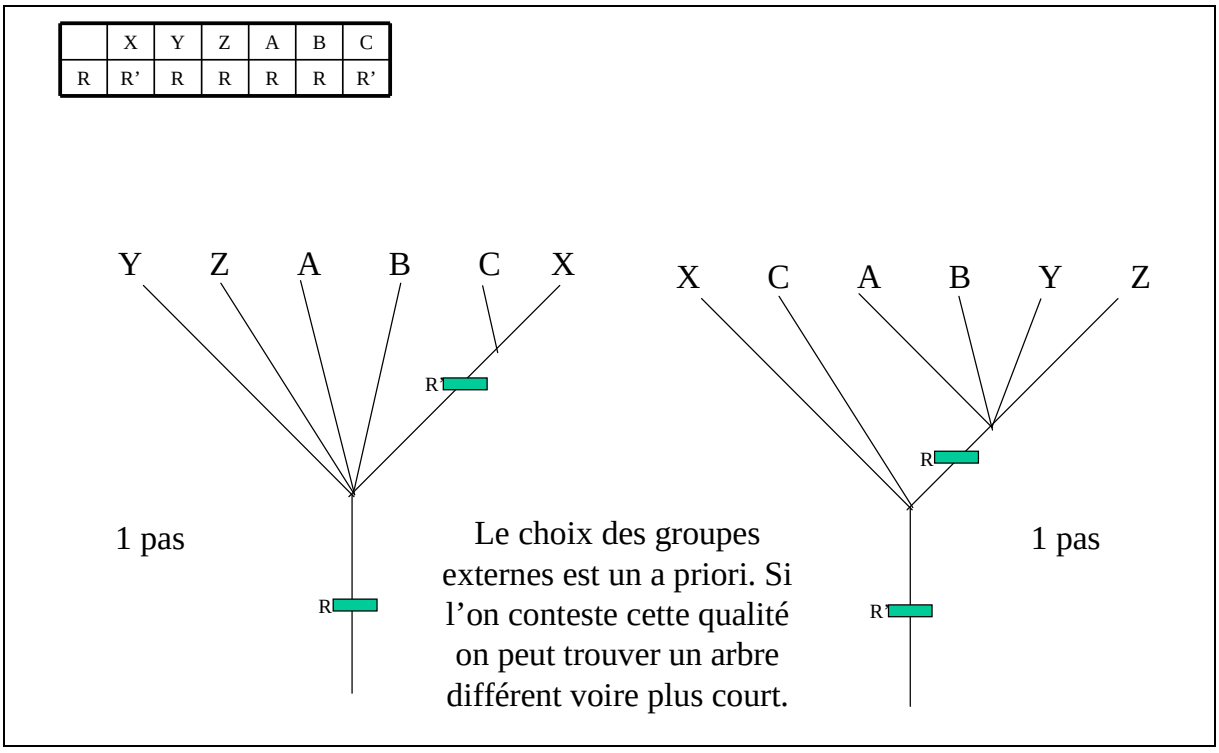


Figure III- 9. Le choix des groupes externes est un à-priori. Si l'on conteste cette qualité on peut trouver un arbre plus court.

Caractères

Un caractère peut être de différents types

- caractères binaires et états multiples (ou polymorphisme)
- caractères réversibles
- caractères irréversibles
- caractères non additifs
- caractères additifs

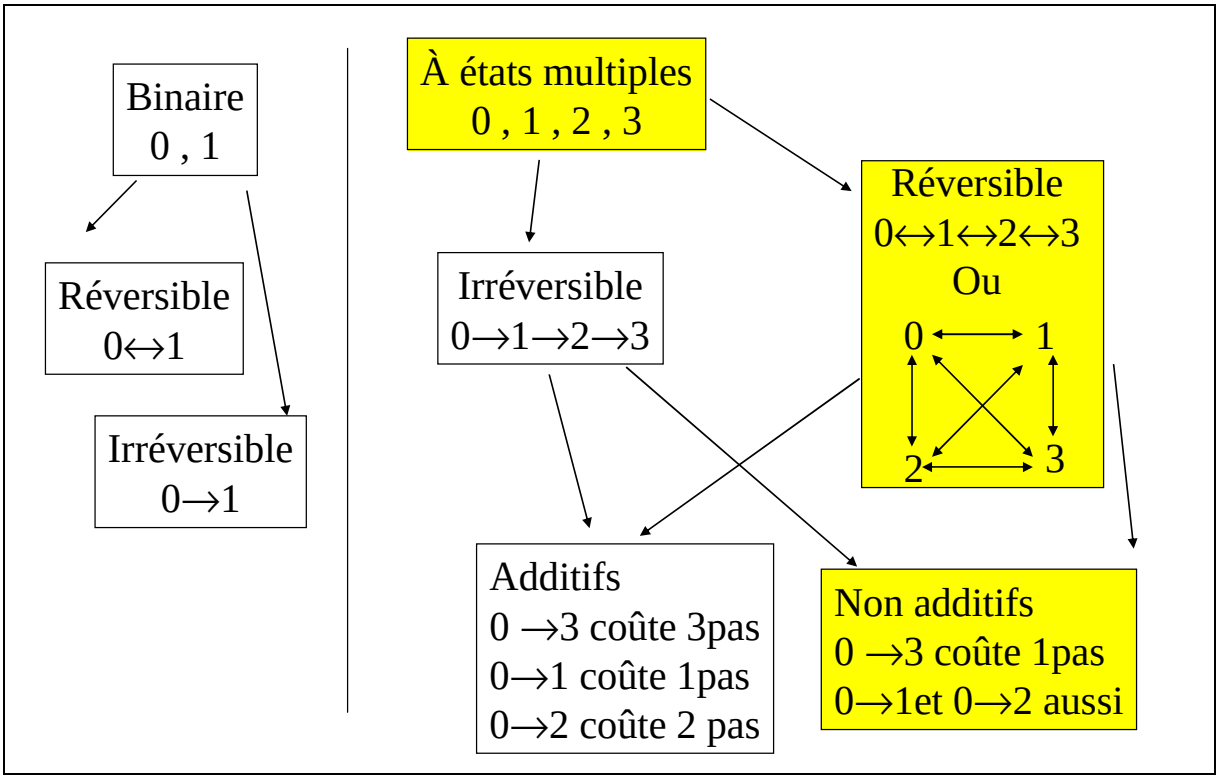


Figure III- 10. Différents types de caractères.

En évolution moléculaire on a en général des caractères à états multiples (4 ou 20), non additifs, non irréversibles.
La parcimonie distingue plusieurs sortes de caractères.

- constant : qui ne présente qu'un seul état parmi tous les taxa étudiés
- variable : qui présente deux états dont un n'est présent qu'une seule fois de telle sorte que quelque soit l'arbre il sera placé sur un segment terminal et comptera pour un pas.
- informatif : qui présente au moins deux états présents chacun au moins deux fois de telle sorte que si l'on trace tous les arbres possibles le nombre de pas dus à ce caractère va changer. Donc en vertu du principe de parcimonie il intervient dans le choix de l'arbre le plus parcimonieux.

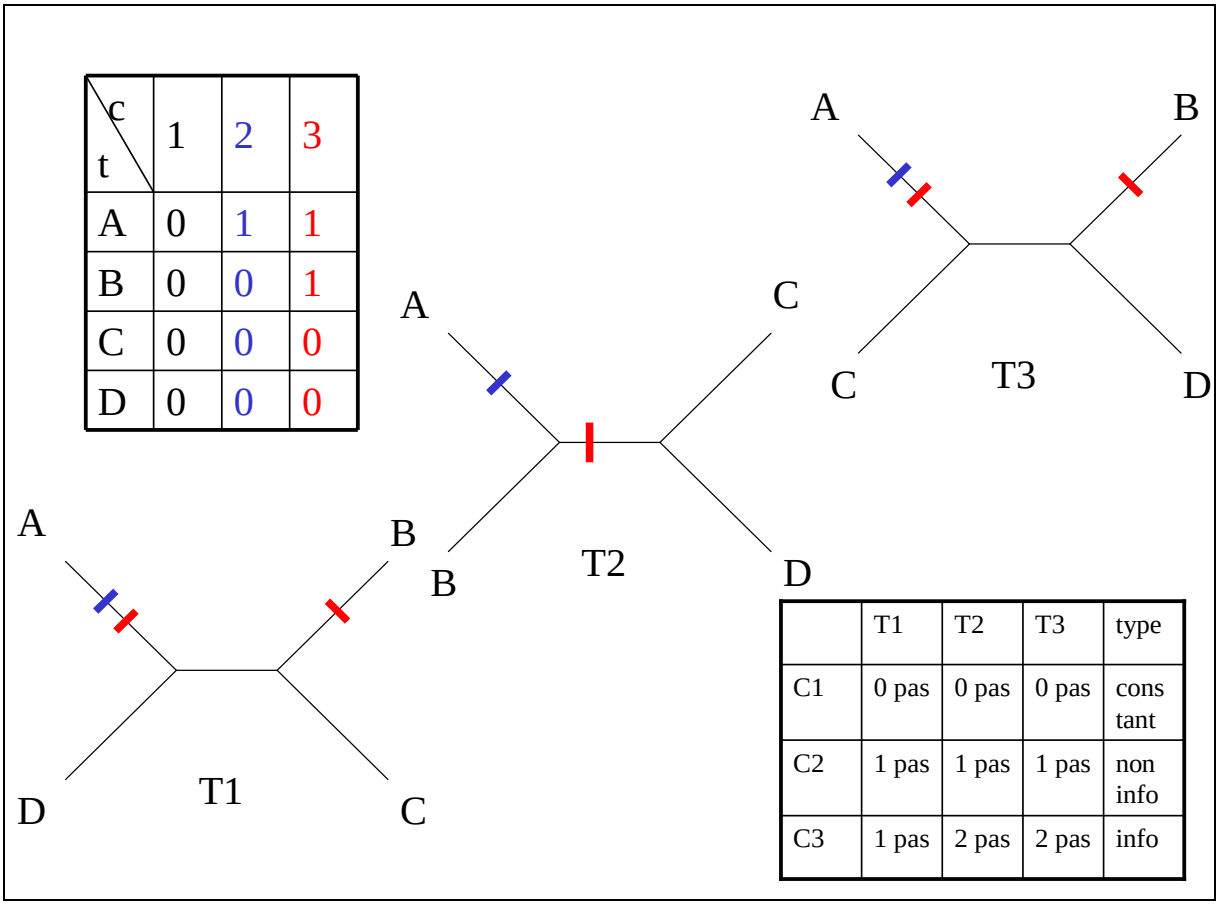


Figure III- 11. Les différentes sortes de caractères . La parcimonie n'utilise vraiment que les caractères informatifs pour la construction d'arbres.

Le changement d'état pour un caractère peut être pondéré.

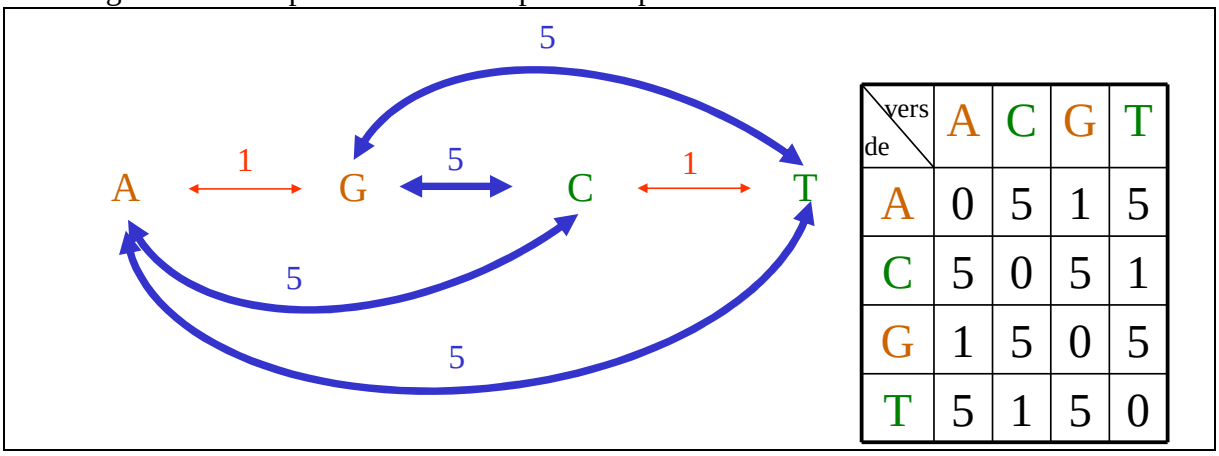


Figure III- 12. Graphe des états d'un caractère et matrice correspondante. Les transversions sont comptées 5 fois plus que les transitions.

Modèle

Lors du calcul d'un arbre, on peut où non imposer des contraintes sur les changements d'états de caractères. On vient de voir que ces états de caractères pouvaient présenter un ordre de succession qui n'était pas indifférent (ce n'est en général pas le cas pour des caractères moléculaires). Même avec un caractère sous deux états l'un (0) est plésiomorphe et l'autre(1)

apomorphe. Pour un caractère donné le passage de 0 → 1 peut se rencontrer sur diverses branches c'est de la convergence, la réversion de 1 → 0 (réversion) peut également se rencontrer. Eviter un de ces deux phénomènes est possible avec quelques pas de plus.

- Le modèle de Wagner admet aussi bien les convergences que les réversions.
- Le modèle de Camin-Sokal n'autorise que les convergences.
- Le modèle de Dollo n'autorise que les réversions (contradiction avec la loi de Dollo qui implique que le retour à l'état ancestral est impossible).

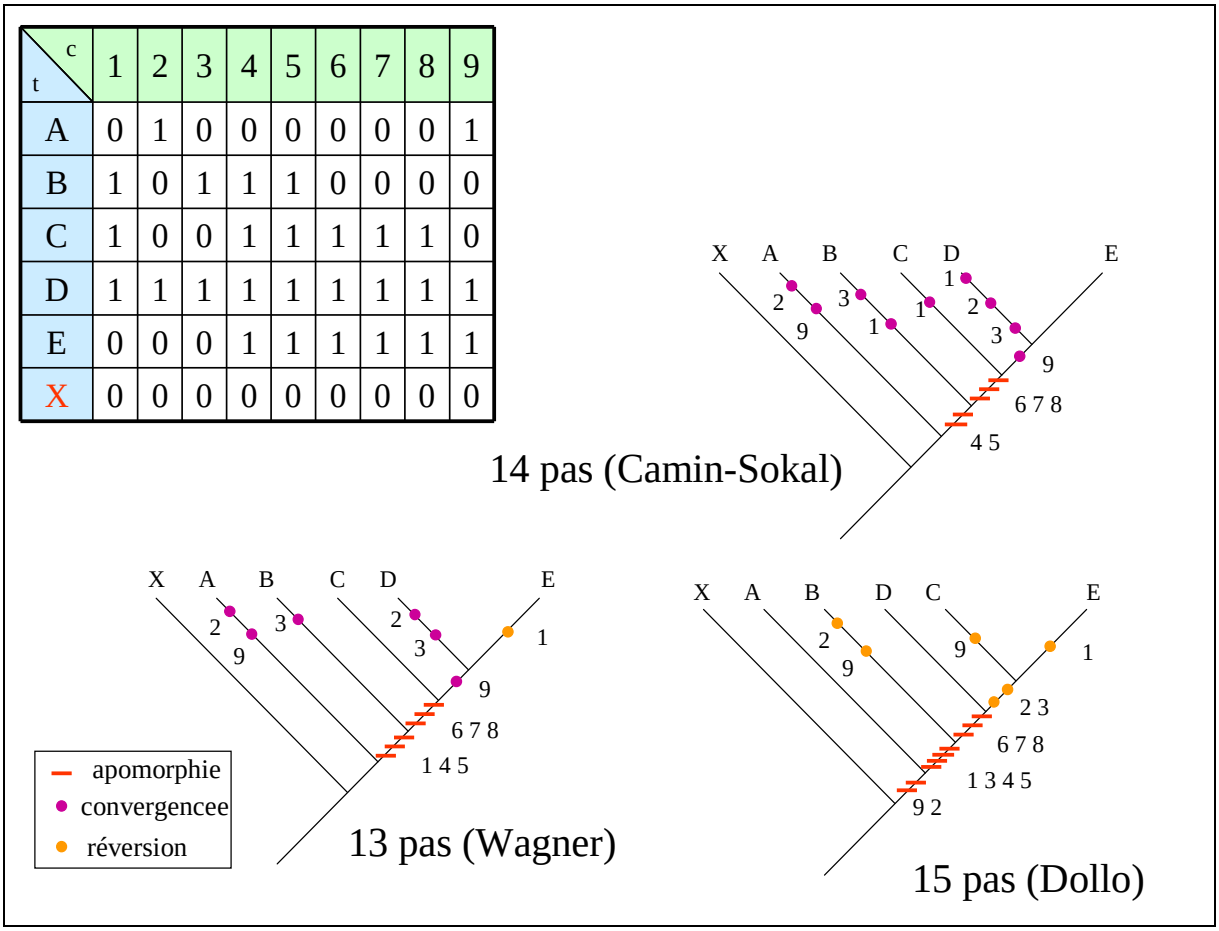


Figure III- 13. Modèles de Wagner, Camin-Sokal, Dollo.

Procédures

Algorithme exact

Exemple du Quagga

Le quagga est un animal éteint en 1883. Il en reste des exemplaires naturalisés dans différents musées. En utilisant des fragments d'os on a pu déterminer la séquence partielle de deux gènes mitochondriaux (cytochrome oxydase et la NADH Déshydrogénase, 229bp en tout)

Figure III- 14. Une espèce disparue: le quagga.

	Cyt Ox				NADH Dase		
	4	**10**	**67**	**103**	**28**	**58**	**71**
Quagga	**A**	**C**	**T**	**T**	**C**	**C**	**T**
Z.pl	**A**	**C**	**T**	**T**	**C**	**C**	**T**
Z.mt	**A**	**T**	**C**	**T**	**T**	**C**	**C**
Cheval	**G**	**T**	**C**	**C**	**C**	**T**	**C**
Vache	**G**	**T**	**C**	**C**	**T**	**T**	**A**

Tableau III- 1. Les positions variables des séquences mitochondriales partielles.

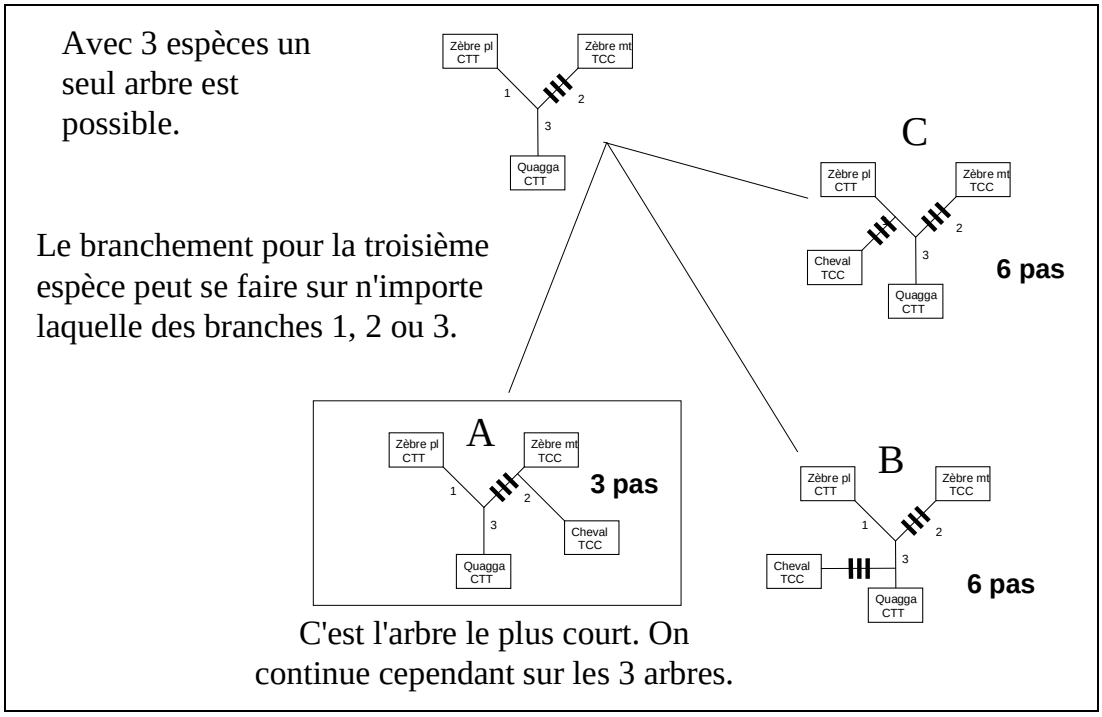


Figure III- 15. Pour construire ces arbres on commence par trois taxons puis on en ajoute un (trois possibilités).

On va calculer tous les arbres possibles pour ces cinq taxons à partir des données moléculaires du tableau 3. On trace d'abord l'arbre étoile pour trois taxons. Il comporte trois branches. Le quatrième taxon peut se placer sur chacune de ces branches. Avec ces quatre taxons il n'y a que trois caractères informatifs. Ajouter la vache augmente le nombre des caractères informatifs à 7. On doit indiquer pour chacun leur état aux nœuds.

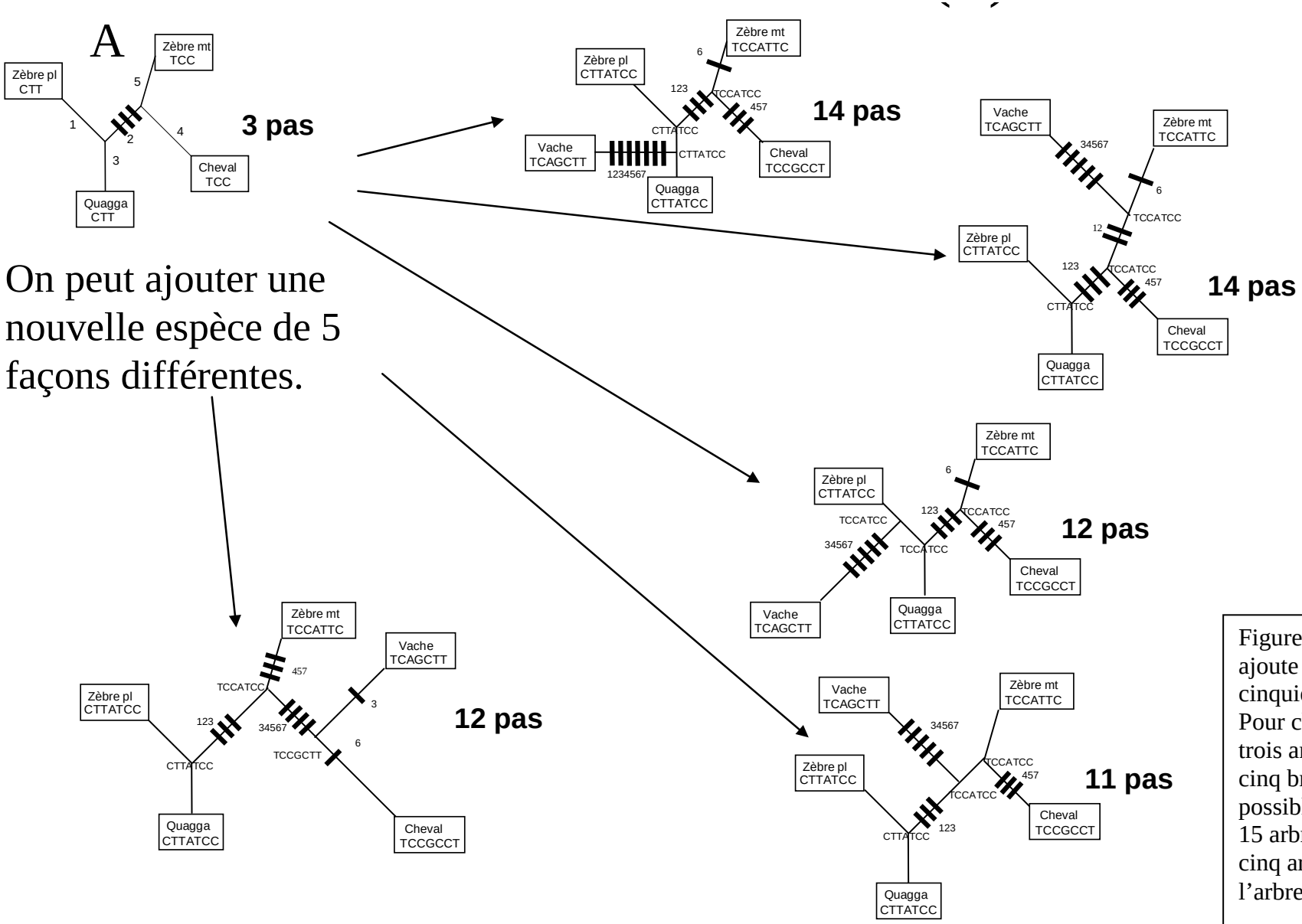


Figure III- 16. On ajoute ensuite le cinquième taxon. Pour chacun de ces trois arbres il y a cinq branches possibles soit en tout 15 arbres. Ici les cinq arbres issus de l'arbre A .

Quinze arbres sont obtenus de longueurs variées.

Arbre	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Pas	15	12	15	15	14	15	15	12	15	14	14	14	12	9	11

C'est l'arbre 14 qui est le plus parcimonieux.

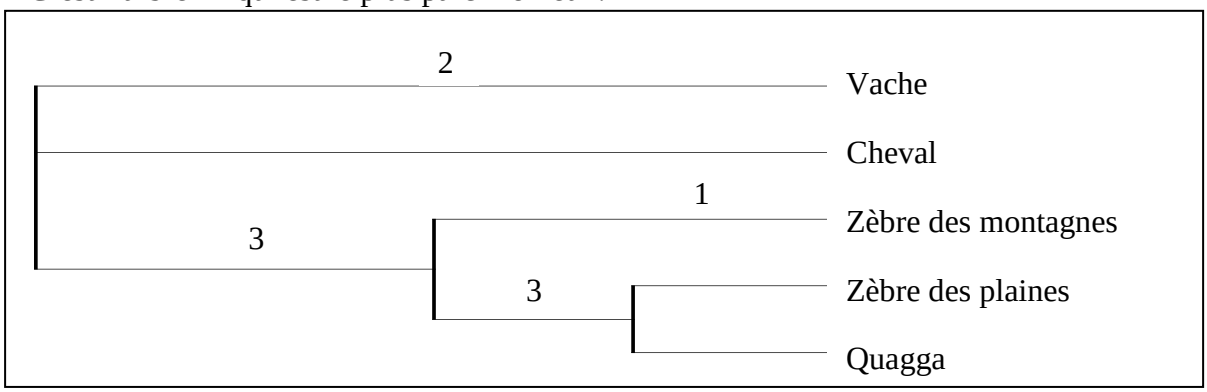


Figure III- 17. Arbre le plus parcimonieux (le nombre de pas est indiqué au dessus de chaque branche).

Cette méthode exhaustive examine tous les arbres possibles en ajoutant les taxons un par un. Le nombre d'arbres possibles augmente rapidement.

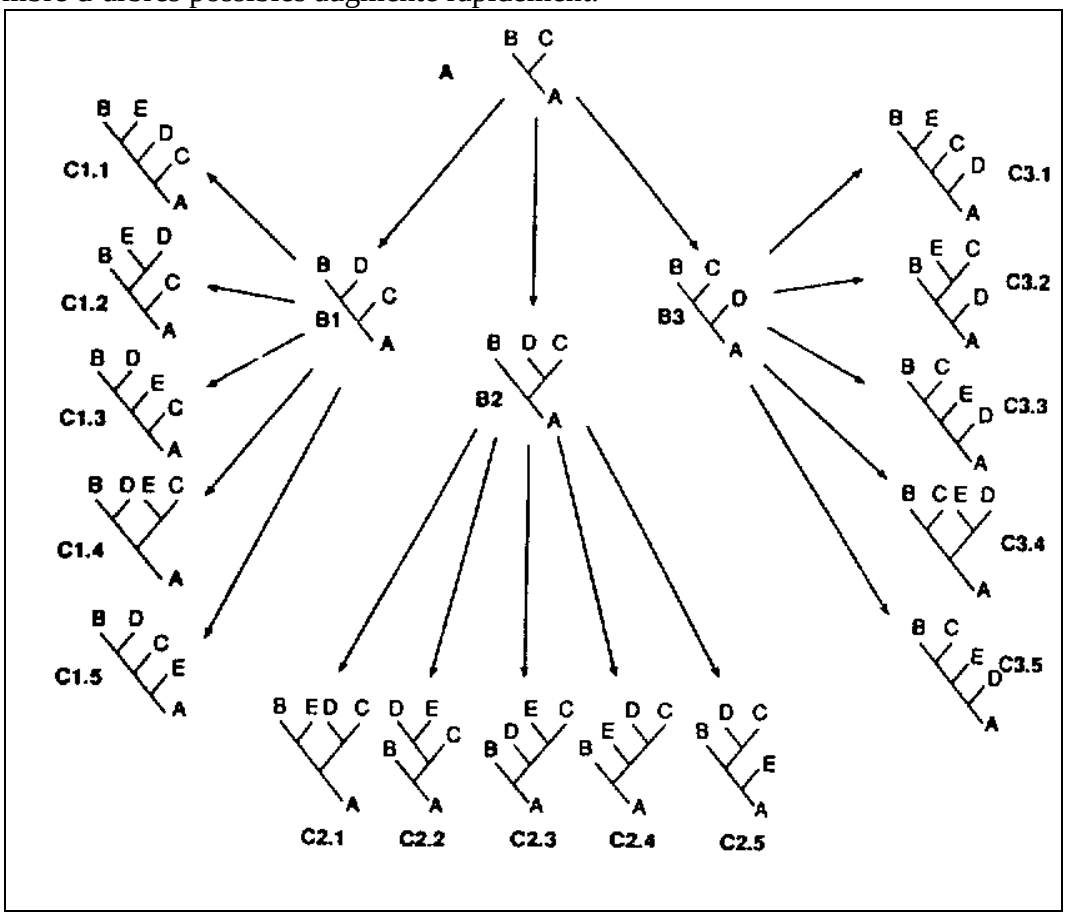


Figure III- 18. Méthode exhaustive : après avoir évalué tous les arbres on choisit le ou les plus courts.

Algorithme Branch and bound

Pour aller un peu plus vite sans abandonner la recherche exhaustive la procédure du branch and bound commence par évaluer un arbre au hasard. Ensuite, au cours de la recherche exhaustive des différents arbres, elle va abandonner une voie à partir du moment où le début d'arbre présente déjà une longueur supérieure à celle de l'arbre au hasard. lorsqu'au bout de sa recherche dans une voie elle trouve un arbre plus court que l'arbre au hasard du début, elle garde cette nouvelle valeur comme limite supérieure. Néanmoins, l'algorithme exhaustif seul peut donner tous les arbres suboptimaux.
Un exemple de l'efficacité de cet algorithme : sur un jeu de 11 taxons (34 459 425 arbres binaires en tout)il fallait compter en 1982 55 jours pour trouver tous les arbres possibles. Le branch and bound a permis de réduire ce temps à un peu moins de 5mn.(Testing the theory of descent, Penny, Hendy and Steel (1991) in Phylogenetic analysis of DNA sequences ed Miyamoto and Cracraft).

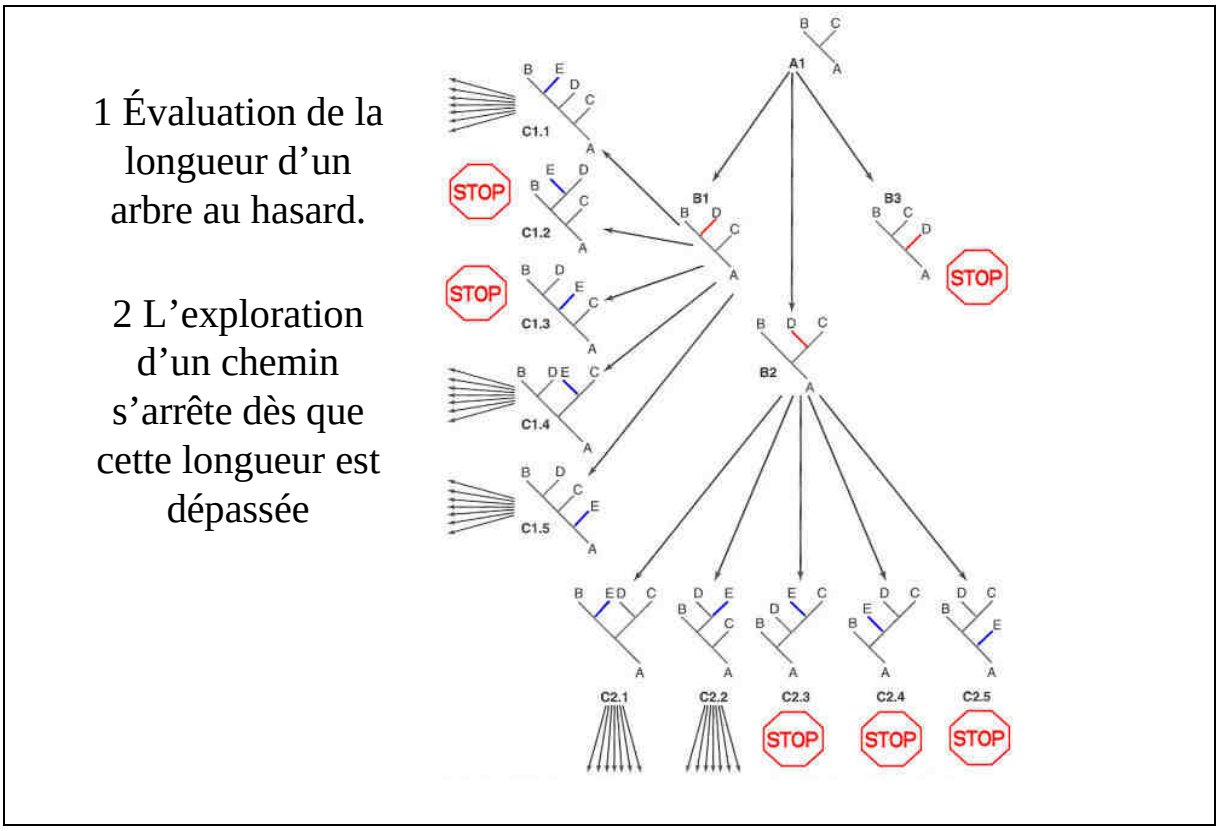


Figure III- 19. Méthode Branch and Bound (BB).

Algorithme heuristique

Les exemples précédents comportaient très peu de taxons et généraient peu d'arbres mais ce nombre croît avec le nombre de taxons. Dans quelles proportions ?
Pour calculer le nombre d'arbres possibles entre plusieurs taxons on considère la façon de déduire les arbres les uns des autres : pour 3 taxons il n'existe qu'un arbre qui présente 1 nœud, 3 segments externes et pas de segments internes. Le 4^{e} taxon peut être ajouté sur chaque segment de cet arbre : 3 arbres possibles qui ont 2 nœuds,1 segment interne et 4 segments externes. Le cinquième taxon peut être inséré sur chacun des segments (interne ou

externe) de chacun des 4 arbres, on obtient ainsi 3*(1+4)=15 arbres. Un arbre à x taxons présente

- x-2 nœuds
- x-3 segments internes
- x segments externes

Il y a Tx arbres possibles. L'insertion du taxon x+1 pourra se faire sur {(x-3)+x}=(2x-3) segments, ce qui générera

Tn=Tn-1 * {2(n-1) – 3} arbres (avec x=n-1)

Par récurrence on montre que

Tn=produits de k=3 à n de l'expression (2k-5)

$$T_n = \prod_{k=3}^{n} (2k-5)$$

En plus ces arbres sont dessinés non racinés. On peut choisir de placer la racine sur n'importe lequel des segments. Le nombre d'arbres racinés est

Tn=produits de k=2 à n de l'expression (2k-3)

$$T_n^{'} = \prod_{k=2}^{n} (2k-3)$$

Nb Taxa	Nombre d'arbres non racinés	Nombre d'arbres racinés
1	1	1
2	1	1
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10 395
8	10 395	135 135
9	135 135	2 027 025
10	2 027 025	34 459 425
11	34 459 425	654 729 075
12	654 729 075	13 749 310 575
13	13 749 310 575,00	316 234 143 225,00
14	316 234 143 225,00	7 905 853 580 625,00
15	7 905 853 580 625,00	213 458 046 676 875,00
16	213 458 046 676 875,00	6 190 283 353 629 370,00
17	6 190 283 353 629 370,00	191 898 783 962 511 000,00
18	191 898 783 962 511 000,00	6 332 659 870 762 850 000,00
19	6 332 659 870 762 850 000,00	221 643 095 476 700 000 000,00
20	221 643 095 476 700 000 000,00	8 200 794 532 637 890 000 000,00

Tableau III- 2. Le nombre d'arbres croît rapidement avec le nombre de taxa.

Même si les arbres à retenir avant le choix sont non racinés, leur nombre augmente considérablement au fur et à mesure du nombre de taxons. Il devient très vite impossible de

considérer tous les arbres possibles. Il va donc falloir choisir parmi quelques uns construits au hasard (procédure heuristique), des essais seront alors nécessaires pour espérer obtenir quand même un arbre parmi les plus court. On utilise des techniques dites d'exploration de collines (« hill-climbing technics ») pour chercher dans un paysage d'arbres de plus en plus grand. La manière d'arriver à une solution optimale locale peut empêcher d'explorer des arbres plus longs (on violerait le principe de parcimonie) pour arriver à une solution optimale globale.

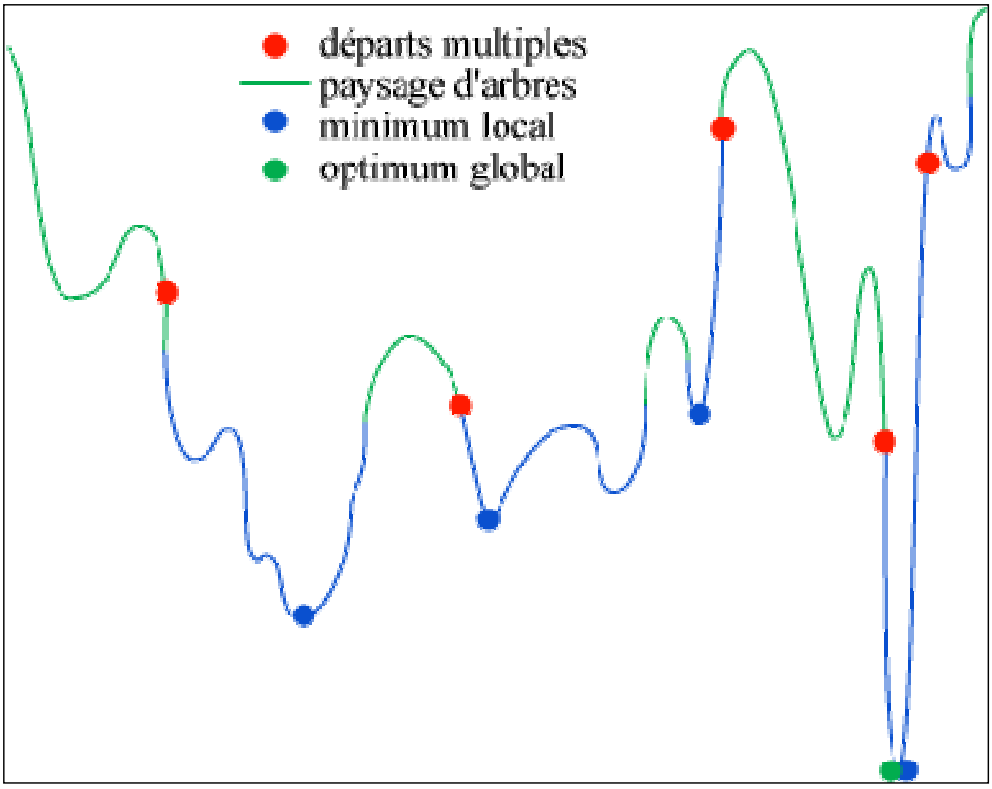


Figure III- 20. Plusieurs points de départ indépendants assurent un meilleur balayage du paysage d'arbres.

Une façon de tourner la difficulté est de recommencer la recherche au hasard d'un point différent. On accroît ainsi la probabilité de trouver l'arbre le plus court de tous, sans pour autant garantir le résultat.

Algorithme de Wagner (1961)

C'est l'ancêtre des algorithmes heuristiques actuels utilisés par les programmes de parcimonie (dont les algorithmes ne sont pas publiés).
Ces algorithmes connectent les unités évolutives entre elles, en construisant un arbre de telle façon que le nombre total de transformations de caractères soit minimal. L'état des caractères est établi pour chaque nœud en maximisant les synapomorphies et en minimisant les homoplasies. Chaque nœud prend valeur d'UEH caractérisée comme les UE. On dresse tout d'abord la liste des distances Manhattan entre tous les couples de taxons. A et B étant les deux taxons.

x_{Ah} = état du caractère h pour l'UE A. Il y a en tout K caractères.

$$d_{AB} = \sum_{h=1}^{K} \left| x_{Ah} - x_{Bh} \right|$$

- On connecte les UE pour lesquelles cette différence est la plus grande (les plus éloignées).
- Une autre UE est ensuite connectée en un nœud Y. Cette troisième UE est choisie de telle sorte que la distance entre elle et Y soit maximale. Cette distance est estimée :

$$d_{CY} = \frac{1}{2}\left(d_{AC} + d_{BC} - d_{AB}\right)$$

- Les états de caractères de l'UEH Y sont définis de la façon suivante : pour un caractère donné c'est la médiane des états de A, B et C(médiane valeur sur l'histogramme du 50^{e} de la population. C'est cette règle qui assure le minimum de transformations.

$$x_{Yh} = médiane\left(x_{Ah}, x_{Bh}, x_{Ch}\right)$$

- Les états de caractères étant connus pour Y il est maintenant possible d'agglomérer une nouvelle UE comme précédemment en prenant comme critère celle qui a une distance maximale du second nœud Y'. Cette distance se calcule en fonction des distances entre UE et non avec les distances entre UE et UEH..

$$d_{DY'} = \frac{1}{2}\left(d_{CD} + d_{YD} - d_{YC}\right)$$

$$d_{YD} = \frac{1}{2}\left(d_{AD} + d_{BD} - d_{AB}\right)$$

$$d_{YC} = \frac{1}{2}\left(d_{AC} + d_{BC} - d_{AB}\right)$$

$$d_{YD} = \frac{1}{2}\left[d_{CD} + \frac{1}{2}\left(d_{AD} + d_{BD}\right) - \frac{1}{2}\left(d_{AC} + d_{BC}\right)\right]$$

Cette méthode qui peut rappeler un processus phénétique s'en distingue totalement par l'attribution d'états de caractères aux nœuds. Cependant il n'est pas assuré d'arriver à l'arbre le plus court et il convient d'effectuer des tests supplémentaires pour s'en assurer.

T \ C	1	2	3	4	5
A	1	0	0	0	0
B	0	1	0	1	0
C	0	0	0	1	1
D	0	1	1	0	0

Distances 2 à 2

AB=3 BC=2 CD=4
AC=3 BD=2
AD=3

1 On connecte C et D (distance la pus grande)

2 Puis on ajoute A (ou B) au nœud Y
AY=1/2(AC+AD-CD)=1/2(3+3-4)=1

BY=1/2(BC+BD-CD)=1/2(2+2-4)=0

3 C'est donc A que l'on ajoute en premier.

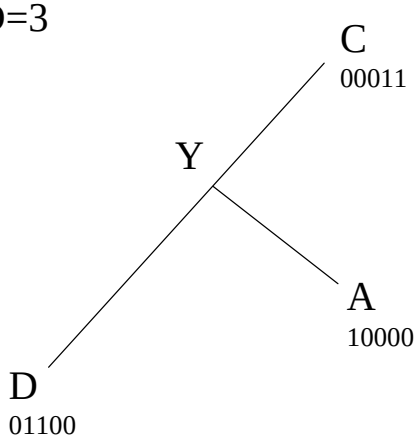


Figure III- 21. Première étape de l'algorithme de Wagner: on prend les deux taxons les plus éloignés et on ajoute celui qui est le plus éloigné du nœud Y. L'état des caractères est alors défini pour ce nœud.

Il reste à placer B sur un des 3 segments YA, YC ou YD.
Sur DY Y'B=1/2(BD+YB-YD) or YB=1/2(AB+CB-AC) et YD=1/2(CD+AD-AC)
Donc Y'B=1/2(BD+1/2(BA+BC)-1/2(CD+AD))= 1/2(2+1/2(3+2)-1/2(4+3)=0,5
Sur AY Y''B=1/2(BA+1/2(BC+BD)-1/2(CA+DA))=1
Sur CY Y'''B=1/2(BC+1/2(BA+BD)-1/2(AC+DC))=0,5
Donc B est mis en Y''

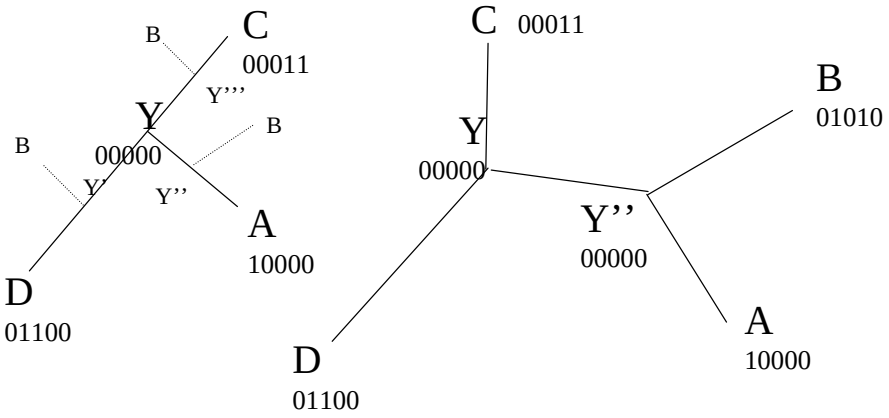


Figure III- 22. L'étape suivante de l'algorithme (Farris, 1971).

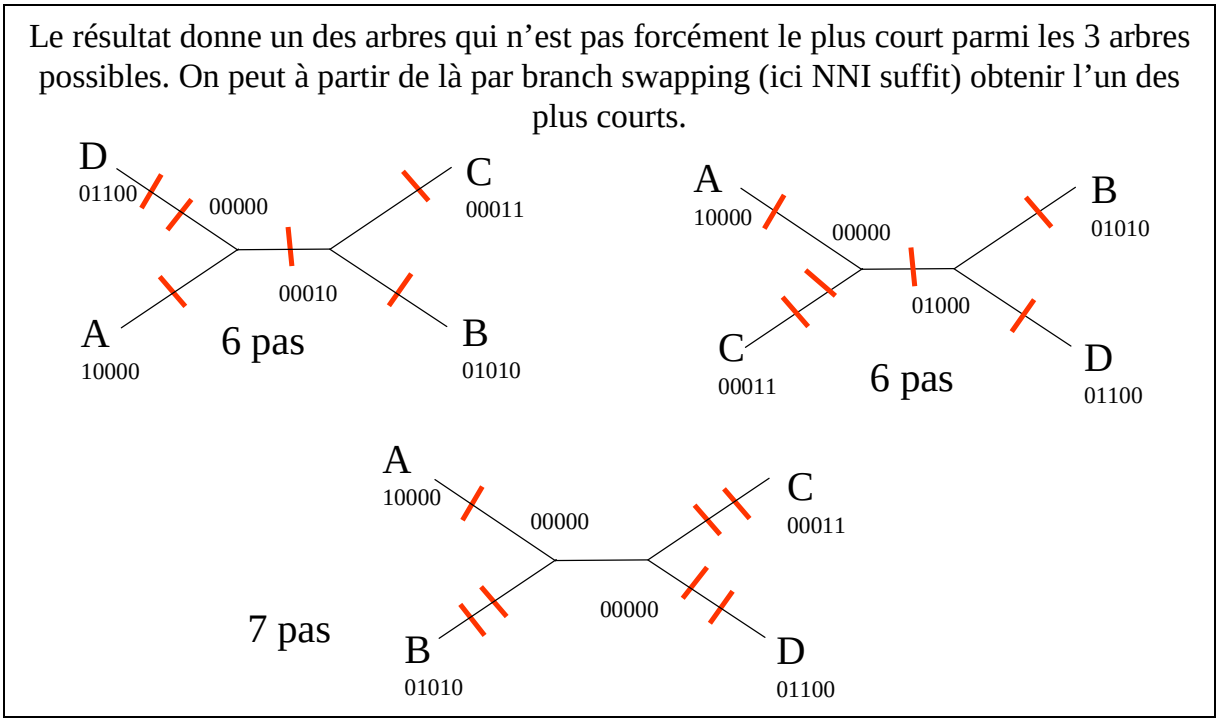


Figure III- 23. L'arbre trouvé par l'algorithme de Wagner, (AB),(CD) n'est pas forcément le plus court.

Branch swapping

En général , les algorithmes d'addition pas à pas ne trouvent pas l'arbre le plus court. Cet arbre est seulement parmi les moins longs. Il faut donc tester l'arbre initialement trouvé en lui faisant subir différents réarrangements qui sont appelés branch swapping (échange des branches). Les procédures pour cela peuvent être locales (Nearest Neighbor Interchanges de PAUP), échange du voisin le plus proche entre 4 taxons, global (Subtree Pruning and Regrafting de PAUP) où chaque sous arbre possible est retioré de l'arbre et réinséré à toutes les positions possibles, et le réarrangement pâr bissection et reconnexion (Tree Bissection Reconnexion de PAUP) où un arbre est coupé en deux le long d'une branche, donnant ainsi deux sous arbres qui sont ensuite reconnectés à toutes les branches possibles de l'arbre.

Comme dans le cas de la procédure du Branch and Bound, si l'obtention d'un arbre minimal nécessite le balayage d'un arbre plus coûteux en pas, l'arbre le plus parcimonieux n'est pas trouvé. Pour pallier à ce problème, il faut que les réarrangements s'appliquent également aux arbres non parcimonieux.

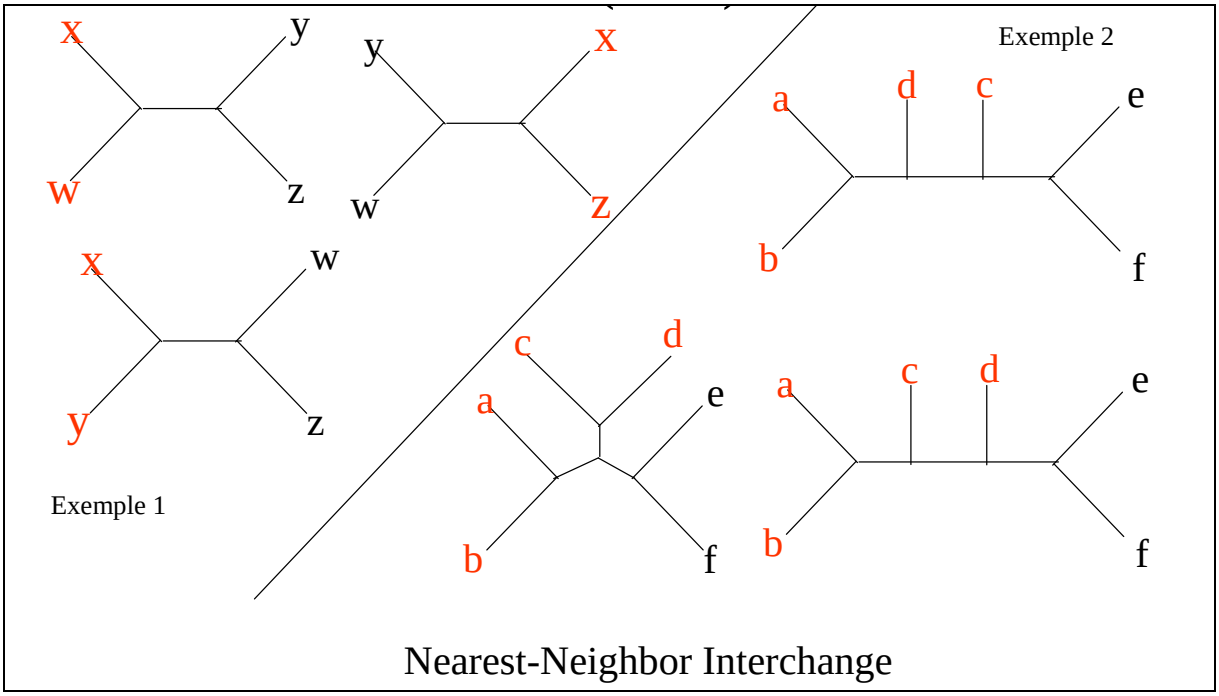


Figure III- 24. Branch swapping:réarrangement local(NNI)

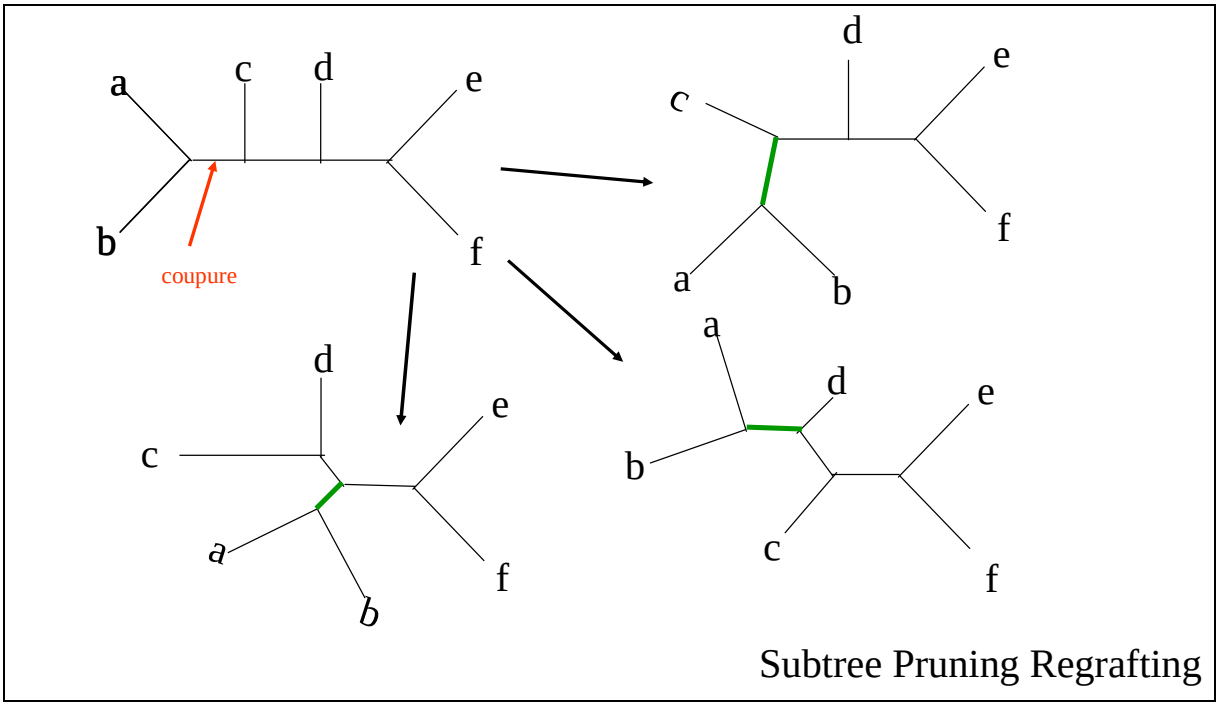


Figure III- 25. Branch swapping:réarrangement global(SPR).

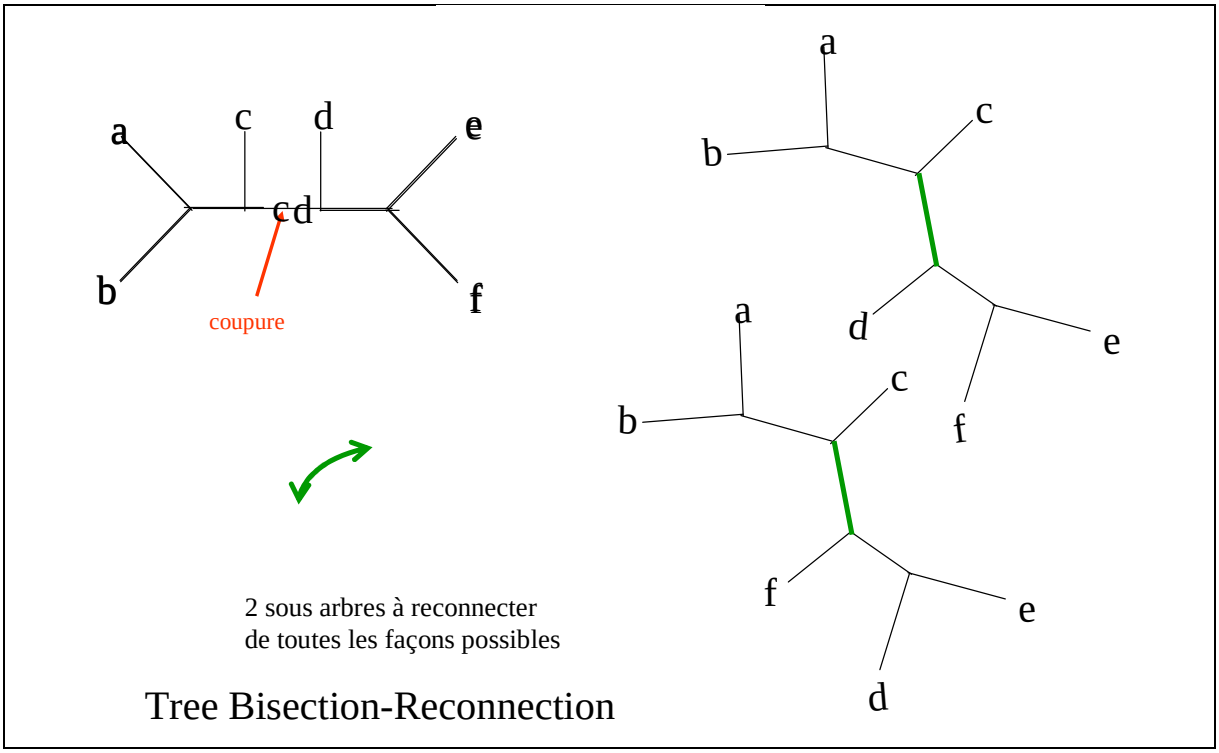


Figure III- 26. Branch swapping:réarrangement global(TBR).

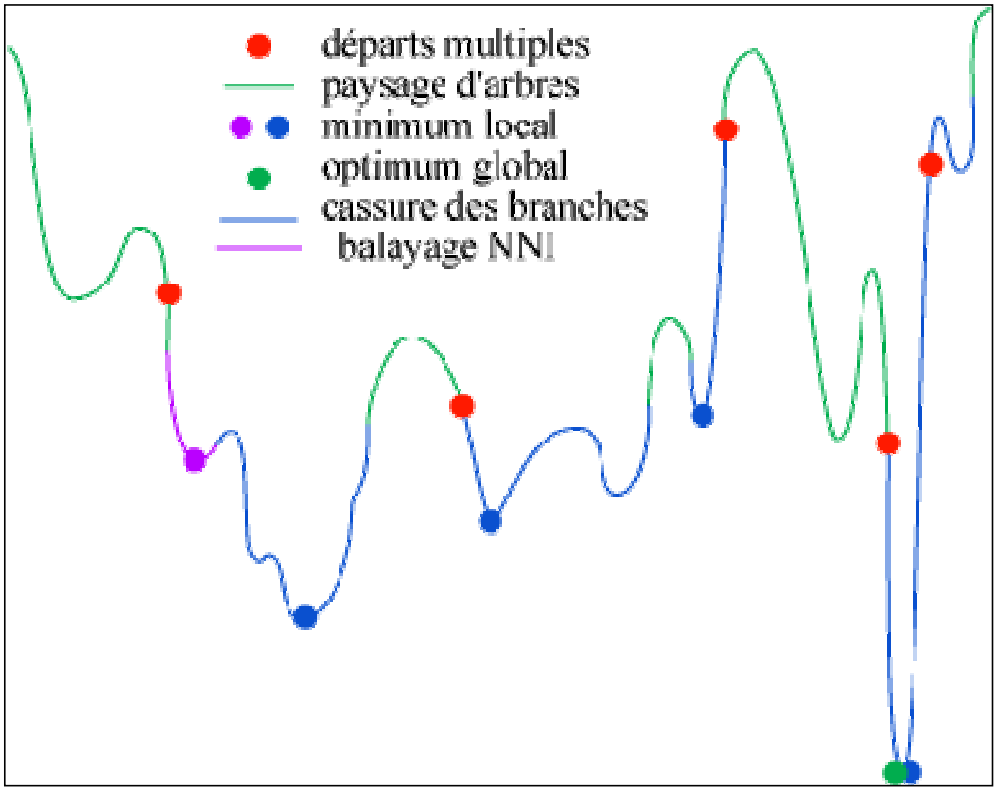


Figure III- 27. L'ensemble du paysage d'arbres. En partant de l'arbre représenté par le point rouge, on explore la zone bleue (clair et foncé représente l'espace exploré par deux méthodes différentes.