Analyses cladistiques
Méthode de Parcimonie

Les méthodes du maximum de parcimonie ont été initialement développées pour traiter des données morphologiques. Certaines méthodes ont ensuite été utilisées sur des données moléculaires.

Le concept a été introduit en 1963 par Edwards et Cavalli-Sforza :

L'estimation la plus plausible d'un arbre évolutif est celle qui fait appel à la quantité minimale d'évolution *i.e.*, celui qui implique le moins d'évènements évolutifs

Exemple d'une reconstruction cladistique

Extrait de P. Tassy, Pour la SCIENCE Dossier L'Evolution (Janvier 1997, page 74)
Exemple à partir de données morphologiques

Distribution de cinq caractères crâniens chez les proboscidiens et les siréniens

- 1) remplacement dentaire dit "horizontal"
- 2) orbite antérieure
- 3) forme particulière de l'os tympanique
- 4) configuration du trou auditif externe
- 5) fosses nasales reculées au-dessus ou en arrière des orbites

Les cases contenant un 1 indiquent l'état transformé du caractère.

Caractères / Espèces	1	2	3	4	5
Autres Mammifères	0	0	0	0	0
Lamantin	1	1	1	0	0
Dugong	0	1	1	0	0
Moeritherium	0	1	0	1	0
Phomia	0	0	0	1	1
Eléphants	1	1	0	1	1

Représentants des siréniens :
- Lamantin
- Dugong

Représentants des proboscidiens :
- Moeritherium (fossile 40 millions d'années)
- Phomia (fossile 30 millions d'années)
- Eléphant

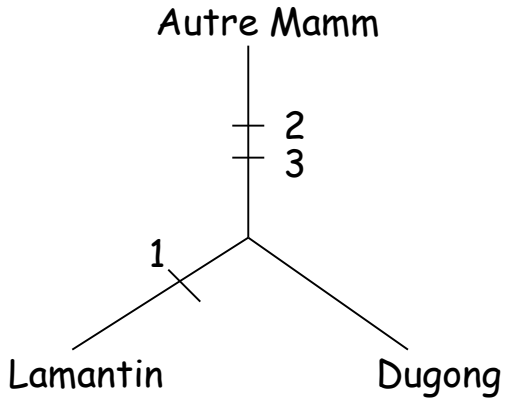
Groupe externe :
- Autres mammifères

Exemple d'une reconstruction cladistique

Caractères / Espèces	1	2	3	4	5
Autres Mammifères	0	0	0	0	0
Lamantin	1	1	1	0	0
Dugong	0	1	1	0	0
Moeritherium	0	1	0	1	0
Phomia	0	0	0	1	1
Eléphants	1	1	0	1	1

A partir de ce tableau on va déterminer les relations de parenté par une méthode de parcimonie.

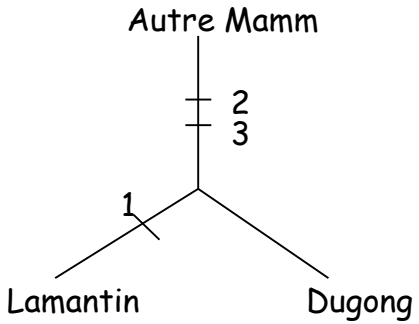
1ère étape : On construit un arbre avec les 3 premières espèces et on reporte sur les branches le numéro du caractère transformé.

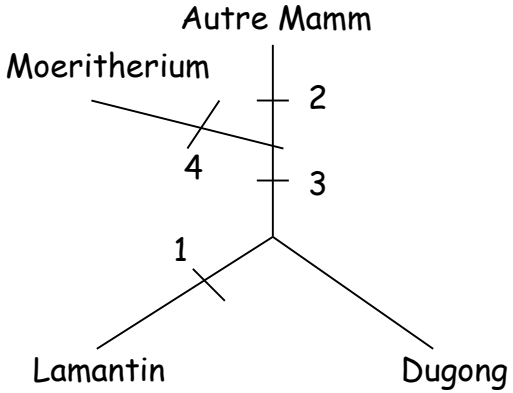


Exemple d'une reconstruction cladistique

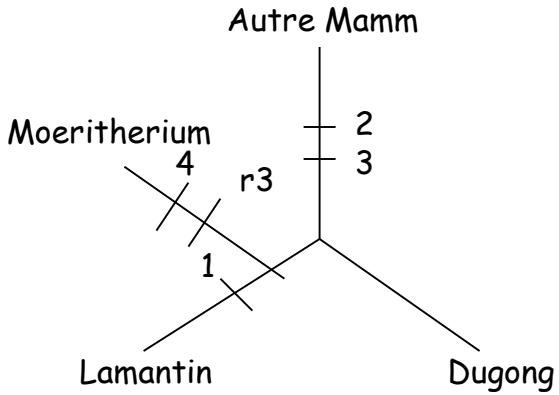
2ème étape : on rajoute la 4ème espèce, 3 possibilités sur chacune des 3 branches.

Espèces \ Caractères	1	2	3	4	5
Autres Mammifères	0	0	0	0	0
Lamantin	1	1	1	0	0
Dugong	0	1	1	0	0
Moeritherium	0	1	0	1	0
Phomia	0	0	0	1	1
Eléphants	1	1	0	1	1





Arbre 1 : 4 changements

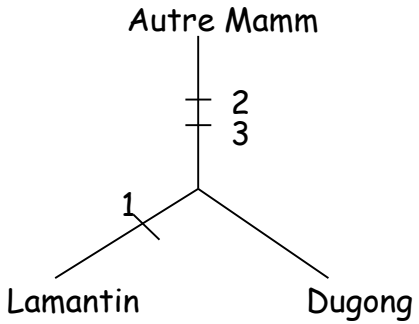


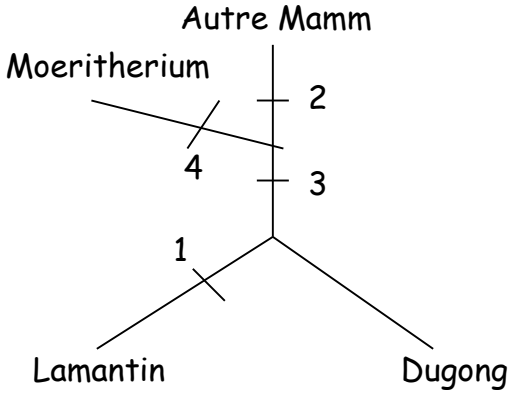
Arbre 2 : 5 changements

Exemple d'une reconstruction cladistique

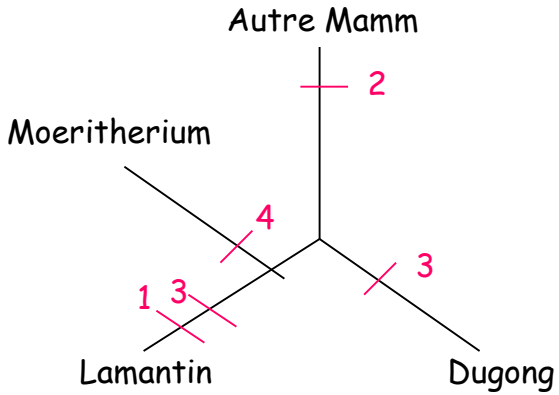
2ème étape : on rajoute la 4ème espèce, 3 possibilités sur chacune des 3 branches.

Espèces \ Caractères	1	2	3	4	5
Autres Mammifères	0	0	0	0	0
Lamantin	1	1	1	0	0
Dugong	0	1	1	0	0
Moeritherium	0	1	0	1	0
Phomia	0	0	0	1	1
Eléphants	1	1	0	1	1

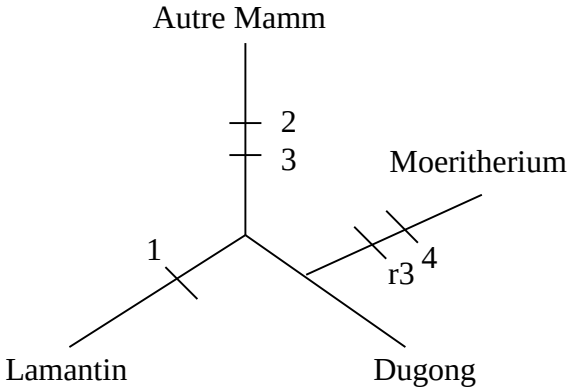




Arbre 1 : 4 changements



Arbre 2 : 5 changements

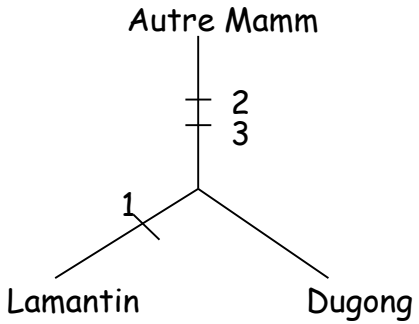


Arbre 3: 5 changements

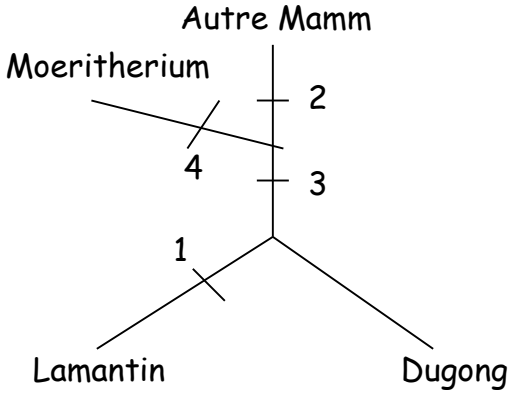
Exemple d'une reconstruction cladistique

2ème étape : on rajoute la 4ème espèce, 3 possibilités sur chacune des 3 branches.

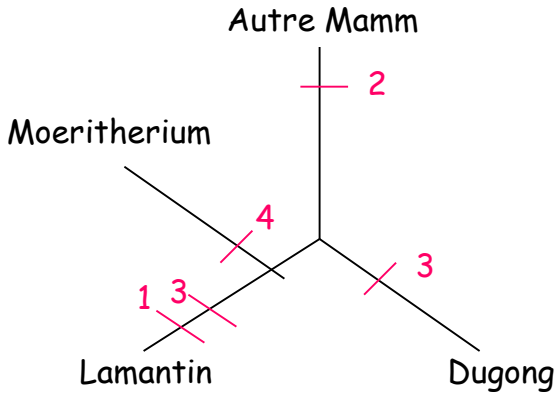
Espèces \ Caractères	1	2	3	4	5
Autres Mammifères	0	0	0	0	0
Lamantin	1	1	1	0	0
Dugong	0	1	1	0	0
Moeritherium	0	1	0	1	0
Phomia	0	0	0	1	1
Eléphants	1	1	0	1	1



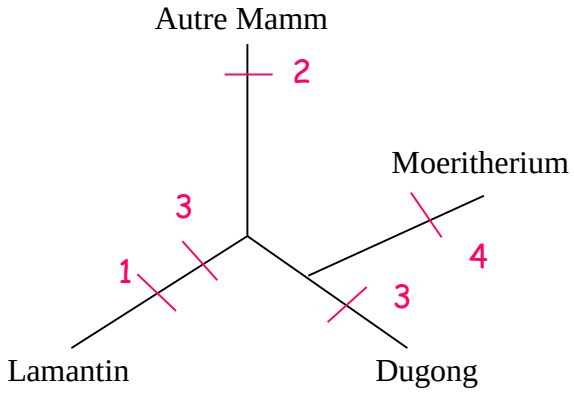
2 explications avec même nb de changements pour la même topologie



Arbre 1 : 4 changements



Arbre 2 : 5 changements

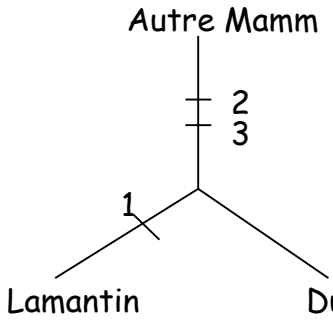


Arbre 3: 5 changements

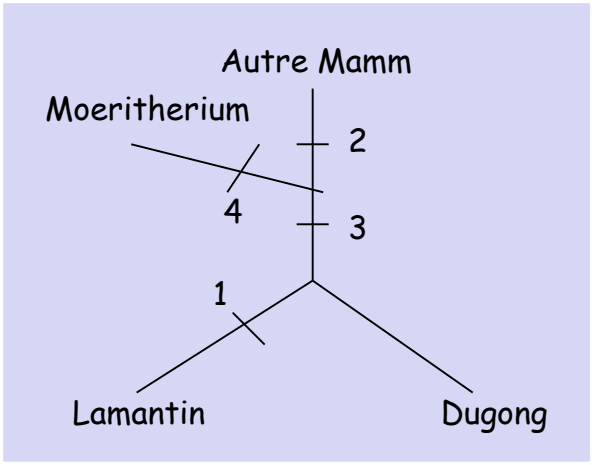
Exemple d'une reconstruction cladistique

2ème étape : on rajoute la 4ème espèce, 3 possibilités sur chacune des 3 branches.

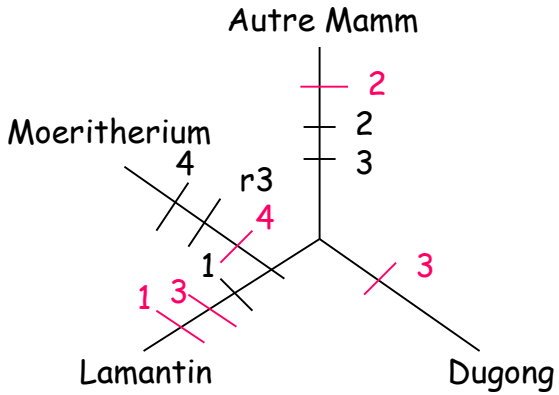
Espèces \ Caractères	1	2	3	4	5
Autres Mammifères	0	0	0	0	0
Lamantin	1	1	1	0	0
Dugong	0	1	1	0	0
Moeritherium	0	1	0	1	0
Phomia	0	0	0	1	1
Eléphants	1	1	0	1	1



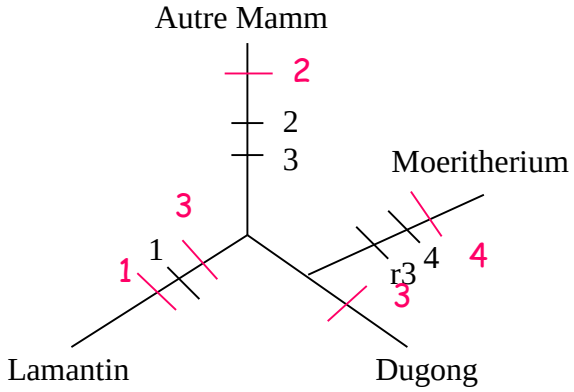
2 explications avec même nb de changements pour la même topologie



Arbre 1 : 4 changements



Arbre 2 : 5 changements



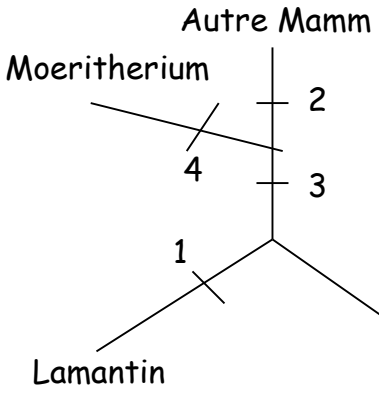
Arbre 3: 5 changements

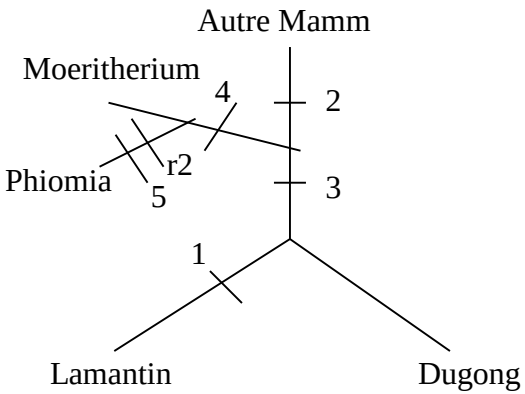
Rouge : caractère 3 convergence; Noire : caractère 3 Réversion

Exemple d'une reconstruction cladistique

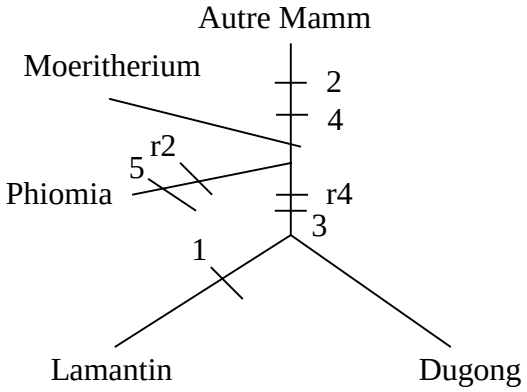
3ème étape : on rajoute la 5ème espèce, 5 possibilités sur chacune des 5 branches.

Caractères / Espèces	1	2	3	4	5
Autres Mammifères	0	0	0	0	0
Lamantin	1	1	1	0	0
Dugong	0	1	1	0	0
Moeritherium	0	1	0	1	0
Phomia	0	0	0	1	1
Eléphants	1	1	0	1	1





Arbre 1 : 6 changements

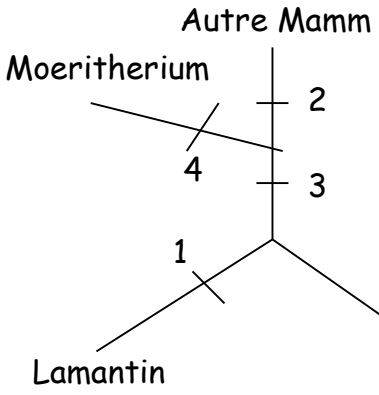


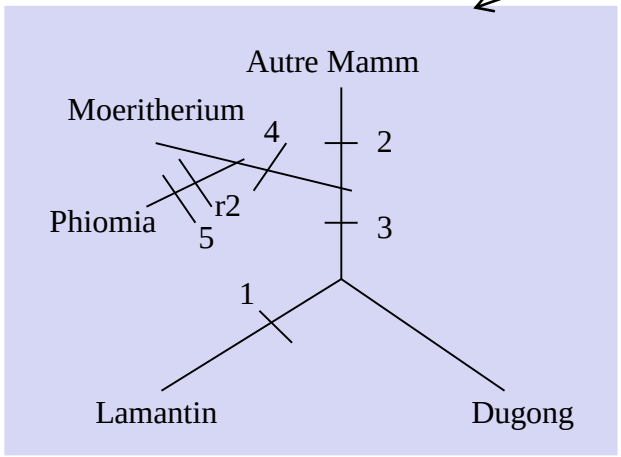
Arbre 2 : 7 changements

Exemple d'une reconstruction cladistique

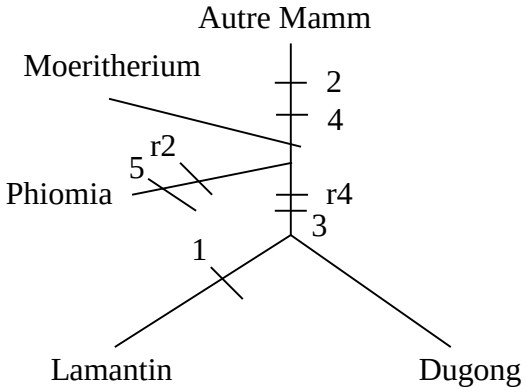
3ème étape : on rajoute la 5ème espèce, 5 possibilités sur chacune des 5 branches.

Caractères / Espèces	1	2	3	4	5
Autres Mammifères	0	0	0	0	0
Lamantin	1	1	1	0	0
Dugong	0	1	1	0	0
Moeritherium	0	1	0	1	0
Phomia	0	0	0	1	1
Eléphants	1	1	0	1	1





Arbre 1 : 6 changements



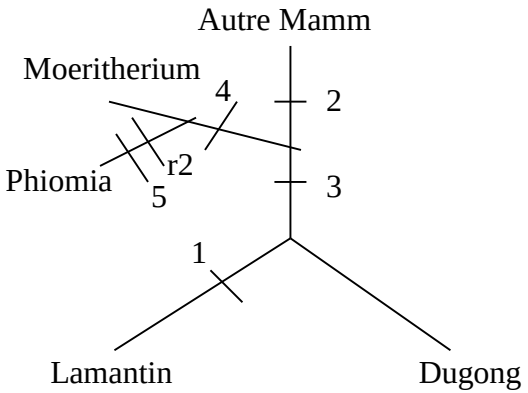
Arbre 2 : 7 changements

Le premier arbre est le plus parcimonieux (vous pouvez tester les autres topologies) et sera utilisé pour ajouter la dernière espèce.

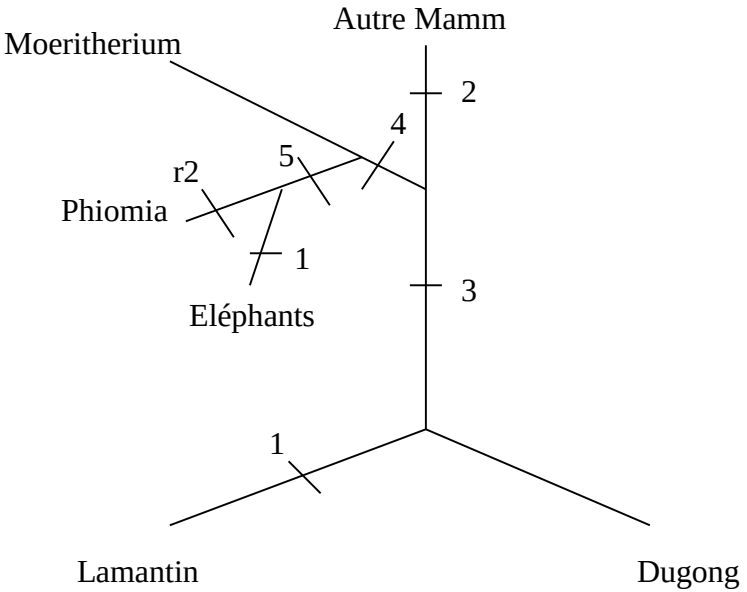
Exemple d'une reconstruction cladistique

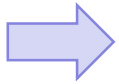
4ème étape : on rajoute la 6ème espèce, 7 possibilités sur chacune des 7 branches.

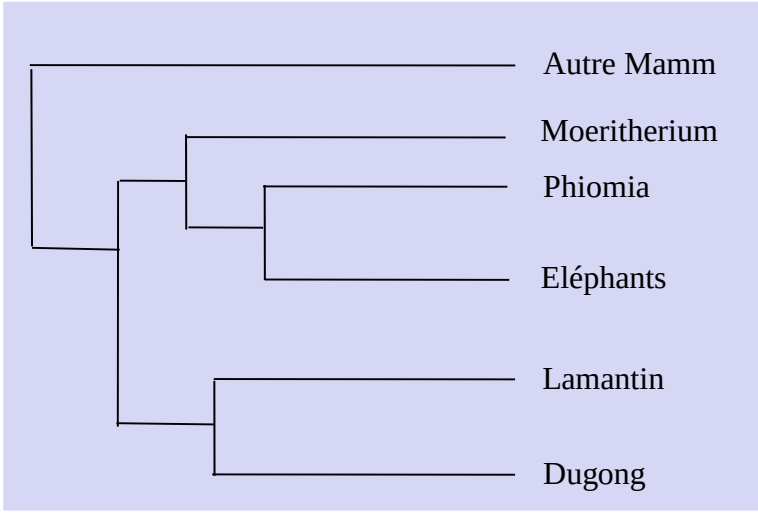
Caractères / Espèces	1	2	3	4	5
Autres Mammifères	0	0	0	0	0
Lamantin	1	1	1	0	0
Dugong	0	1	1	0	0
Moeritherium	0	1	0	1	0
Phomia	0	0	0	1	1
Eléphants	1	1	0	1	1



Arbre le plus parcimonieux : 7 changements







Méthode de parcimonie

Dans le cadre de données séquences :

Hypothèses :

- Les séquences ont évolué à partir d'une séquence ancestrale commune au travers d'un processus de mutation-sélection.
- Les différents sites évoluent indépendamment.
- Les lignées se différencient les unes des autres de façon autonome.
- La vitesse d'évolution est lente et constante au cours du temps

Cette approche :

- ne prend en compte que les sites informatifs, *i. e.*, les sites qui permettent de discriminer une topologie par rapport aux autres. Sont donc exclus les sites invariants (même résidu pour toutes les OTU) et les sites variables impliquant le même nombre de substitutions quelle que soit la topologie.
- ne fait pas de correction pour les substitutions multiples.
- dans la majorité des cas, ne donne aucune information sur la longueur des branches (des solutions ont été proposées, mais rarement utilisées).
- ne fait aucune distinction entre les changements évolutifs ce qui est irréaliste. La parcimonie pondérée utilise une matrice de pondération permettant de données différents poids aux changements (poids plus faible pour les changements les plus fréquents (ex : transitions plus fréquentes que transversions)). Dans ce cas matrice arbitraire, donc choix plus rigoureux des poids par une approche de pondération dynamique (on part d'une matrice pour construire un arbre, à partir de l'arbre on reconstruit une matrice qui est utilisée pour recalculer un arbre. On réitère le processus jusqu'à convergence (topologie obtenue idem à la précédente). Problèmes : temps de calcul pouvant être très long et pas de garantie de la convergence

Méthode de parcimonie

Méthode :

- Chercher toutes les topologies d'arbre possible et conserver celle présentant le minimum de changements. Vite irréalisable.
- Des heuristiques mises en place pour explorer qu'un sous-espace des arbres possibles comme celle présentée avant. Problème : pas de garantie de trouver les meilleurs ou le meilleur arbre possible.

  Solution proposée : on part d'un arbre initial (celui obtenu par la NJ est une bonne approximation) puis

  - petits réarrangements de branches successifs → exploration des arbres voisins.
  - si un des arbres voisins est plus parcimonieux, on le garde.
  - on continue le processus jusqu'à obtention d'un arbre pour lequel aucun réarrangement ne donne un arbre plus parcimonieux.
  - on trouve un optimal local dans l'espace des arbres possibles mais pas de garantie que ce soit l'optimal global, *i.e.,* le « meilleur » arbre.

Méthode de parcimonie

Conclusion :

- On peut avoir plusieurs topologies s'expliquant par le même nombre de changements : famille d'arbres.

- Dépend de l'ordre dans lequel sont ajoutées les séquences pour le construction de l'arbre. On n'obtiendra pas forcément le même arbre si on change l'ordre des séquences. Pour pallier à ce problème, heuristique de réarrangement des branches. En répétant plusieurs fois l'opération, on peut trouver l'arbre le plus parcimonieux.

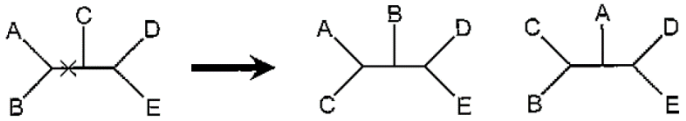
- permet d'inférer l'état des caractères ancêtres.

- critique principale de la parcimonie : méthode non consistante pouvant conduire sous certaines conditions à des résultats erronés (démontré par Felsenstein (1978, Syst. Zool, 27, 401-10)).

Techniques de réajustement

Techniques de réajustement les plus courantes :

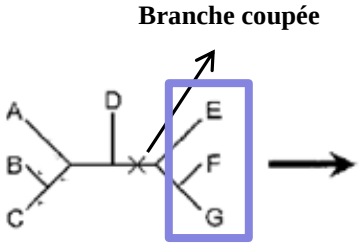
❖ Nearest Neighbor Interchange (NNI) : examiner les arbres qui se trouve à une distance topologique de 2.

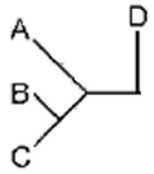


Exemple extrait de Perrière et Brochier-Armanet (2010) Concepts et méthodes en phylogénie moléculaire

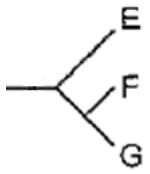
Réarrangement portant sur la branche interne marquée par x.
Seulement deux topologies à une distance topologique de 2

❖ Subtree Pruning et Regrafting (SPR) : l'arbre est coupée au niveau d'une branche ce qui conduit à deux sous-arbres : l'arbre élagué (pruned tree) et l'arbre résiduel. La méthode consiste à rebrancher l'arbre élagué sur chacune des branches de l'arbre résiduel et à calculer pour chaque topologie le critère d'optimisation.





Arbre résiduel

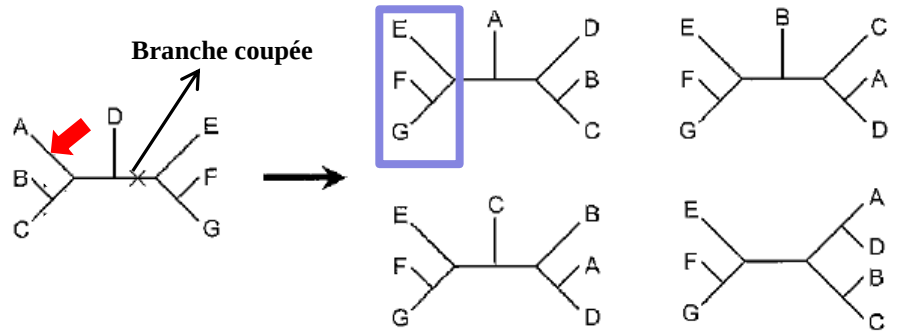


Arbre élagué

Exemple extrait de Perrière et Brochier-Armanet (2010) Concepts et méthodes en phylogénie moléculaire

Techniques de réajustement

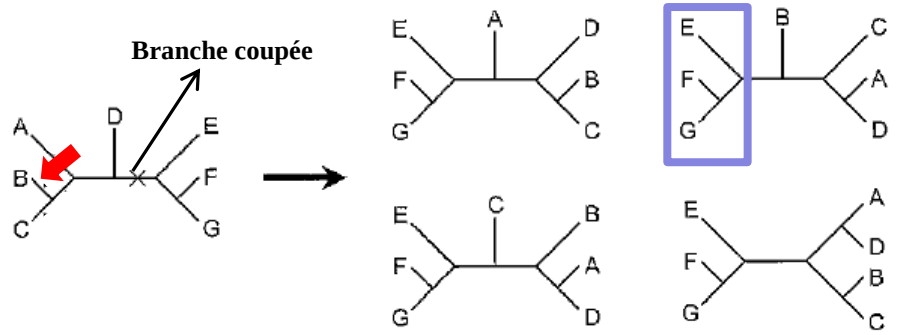
❖ Subtree Pruning et Regrafting (SPR) : l'arbre est coupée au niveau d'une branche ce qui conduit à deux sous-arbres : l'arbre élagué (pruned tree) et l'arbre résiduel. La méthode consiste à rebrancher l'arbre élagué sur chacune des branches de l'arbre résiduel et à calculer pour chaque topologie le critère d'optimisation.



Exemple extrait de Perrière et Brochier-Armanet (2010) Concepts et méthodes en phylogénie moléculaire

Techniques de réajustement

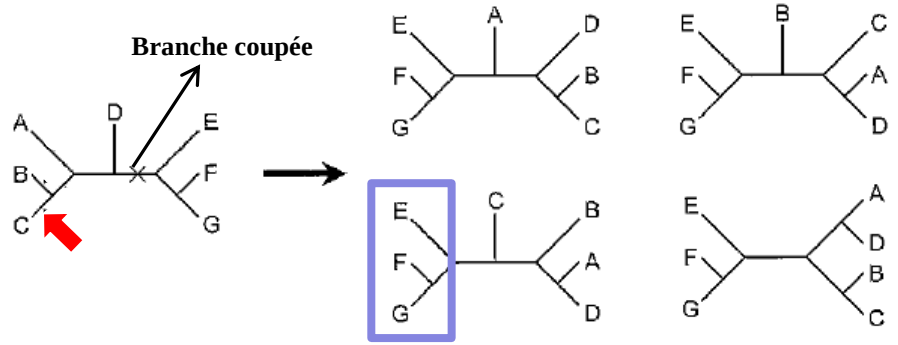
❖ Subtree Pruning et Regrafting (SPR) : l'arbre est coupée au niveau d'une branche ce qui conduit à deux sous-arbres : l'arbre élagué (pruned tree) et l'arbre résiduel. La méthode consiste à rebrancher l'arbre élagué sur chacune des branches de l'arbre résiduel et à calculer pour chaque topologie le critère d'optimisation.



Exemple extrait de Perrière et Brochier-Armanet (2010) Concepts et méthodes en phylogénie moléculaire

Techniques de réajustement

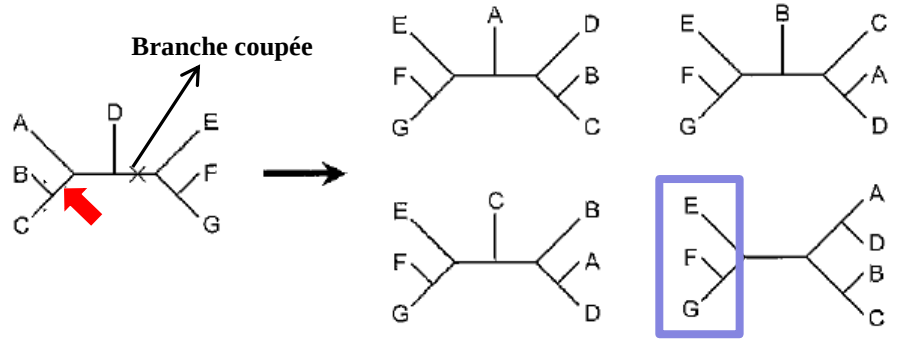
❖ Subtree Pruning et Regrafting (SPR) : l'arbre est coupée au niveau d'une branche ce qui conduit à deux sous-arbres : l'arbre élagué (pruned tree) et l'arbre résiduel. La méthode consiste à rebrancher l'arbre élagué sur chacune des branches de l'arbre résiduel et à calculer pour chaque topologie le critère d'optimisation.



Exemple extrait de Perrière et Brochier-Armanet (2010) Concepts et méthodes en phylogénie moléculaire

Techniques de réajustement

❖ Subtree Pruning et Regrafting (SPR) : l'arbre est coupée au niveau d'une branche ce qui conduit à deux sous-arbres : l'arbre élagué (pruned tree) et l'arbre résiduel. La méthode consiste à rebrancher l'arbre élagué sur chacune des branches de l'arbre résiduel et à calculer pour chaque topologie le critère d'optimisation.

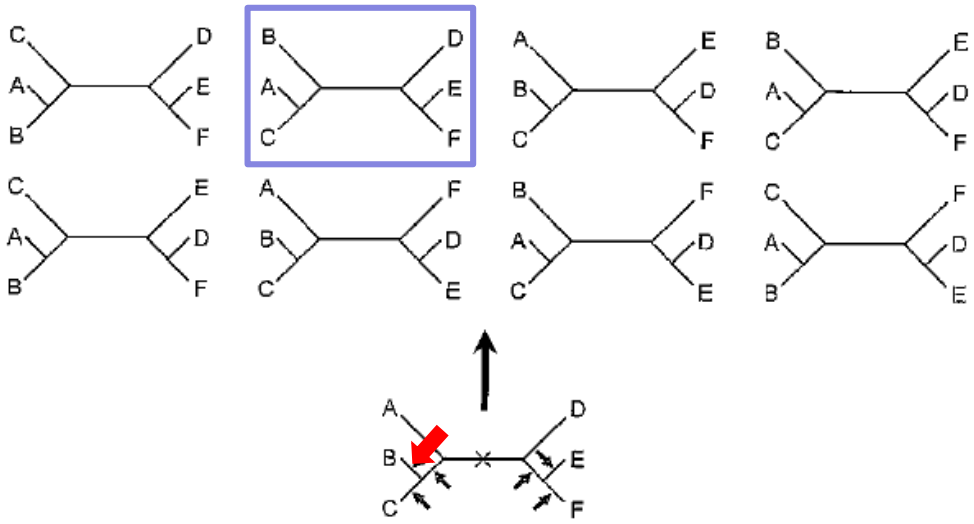


Exemple extrait de Perrière et Brochier-Armanet (2010) Concepts et méthodes en phylogénie moléculaire

Techniques de réajustement

Techniques de réajustement les plus courantes :

❖ Tree Bisection and Reconnection (TBR) : Variante de la SPR. Dans ce cas les deux sous-arbres sont considérés comme indépendants. Toutes les topologies correspondant à toutes les connections possibles entre chacune des branches des deux sous-arbres sont évaluées. Répétée pour l'ensemble des branches internes



Exemple extrait de Perrière et Brochier-Armanet (2010) Concepts et méthodes en phylogénie moléculaire

Méthodes de distance

Introduites en phylogénie dans les années 1960 et se basent donc sur l'utilisation de matrice de distances regroupant les distances calculées entre chaque paire de séquences, ceci au moyen d'un modèle évolutif préalablement choisi.

Méthodes de distance

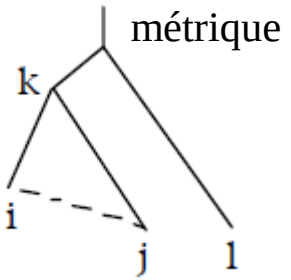
Par définition une distance doit satisfaire les 3 conditions suivantes :
soit 3 OTU i, j et k,

- $d_{ij} > 0$ ($i \neq j$) (positive)
- $d_{ii} = 0$
- $d_{ij} = d_{ji}$ (symétrique)

La distance sera une *métrique* si elle possède la propriété dite « d'inégalité triangulaire »

- $d_{ij} \leq d_{ik} + d_{jk}$

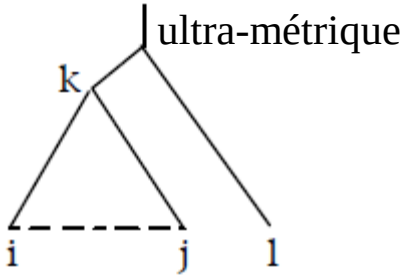
Le chemin direct entre deux point i et j est le plus court (plus court que de passer par k)
L'égalité signifie que la distance est additive $d_{ij}=d_{ik} + d_{jk}$



La distance sera une ultra-*métrique* si de plus elle vérifie

- $d_{ij} \leq \max(d_{ik},d_{jk})$

Cela signifie que les deux plus grandes distances sont égales. Donc $d_{ik} = d_{jk}$, ou $d_{ij} = d_{ik}$ ou $d_{ij} = d_{jk}$
(Conséquence dans la construction de l'arbre, impose l'horloge moléculaire).



Méthodes de distance

L'objectif des méthodes de distance : distances d'arbre (ou patristiques) représentent au mieux les distances présentes dans la matrice de distance.

Deux grands types de méthodes :

- celles basées sur des algorithmes de clustering (UPGMA).
- celles basées sur des critères d'optimisation (moindre carré, neighbor joining).

Méthode de distances actuellement la plus utilisée : la neighbor joining (NJ) et ses variantes.

Méthodes de distance : la NJ

Saitou et Nei (1987) *Mol. Biol. Evol.*, 4, 406-25

Constitue une approximation du minimum d'évolution (critère d'optimisation).

Principe général du minimum d'évolution : Examine toutes les topologies, calcule la somme de la longueur des branches de chacune d'entre-elles et retient celle qui minimise la somme des longueur des branches (arbre de longueur minimum).

NJ : algorithme qui à chaque étape sélection la paire d'OTU qui une fois agglomérée produit l'arbre minimum (algorithme glouton).

Point de départ : une matrice de distances

	1	2	3	4	5
1	0				
2	d_{12}	0			
3	d_{13}	d_{23}	0		
4	d_{14}	d_{24}	d_{34}	0	
5	d_{15}	d_{25}	d_{35}	d_{45}	0

Rappel :Principe

Alignement de séquences

```
CAAACAGCGTT---GGCTCTCTA
AAAATAACACCaacATGCAAATG
AAAACAGCACCaacGTGCAAATG
AAAACAGCACCaacGTGCAAATG
```

Choix d'un modèle évolutif
Calcul d'une matrice de distances

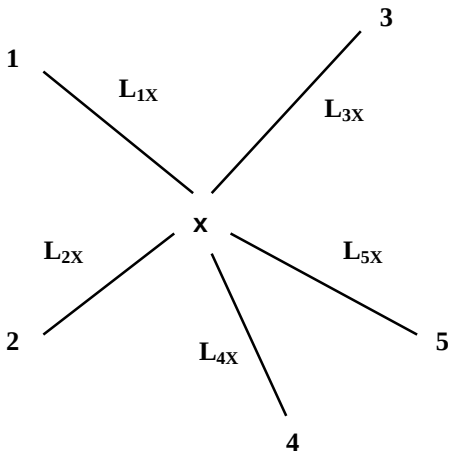
Matrice des distances évolutives entre paires de séquences

	A	B	C
A	**0**		
B	**3**	**0**	
C	**4**	**3**	**0**

distance matrix

Méthodes de distance : la NJ

Première étape : construction d'un arbre en étoile



Calcul du score d'un arbre en étoile :

Le score S_0 de cet arbre est donné par la somme des longueurs des branches des feuilles *i* au centre de l'étoile X ou L_{ix}, soit :

$$S_0 = \Sigma L_{ix}$$

Or, nous ne connaissons que les distances entre les unités évolutives (OTU, ici les séquences), *i.e.*, les d_{ij}. Cependant, il y a une relation entre ces deux valeurs:

$d_{ij} = L_{ix} + L_{jx}$ (par exemple $d_{12} = L_{1x} + L_{2x}$)

Soit *n* le nombre d'UE, on peut facilement démontrer que :

$$S_0 = \frac{1}{n-1}\sum_{i<j} d_{ij}$$

Exemple : avec n = 5

$$\sum_{i<j} d_{ij} = d_{12} + d_{13} + d_{14} + d_{15} + d_{23} + d_{24} + d_{25} + d_{34} + d_{35} + d_{45}$$

$$\sum_{i<j} d_{ij} = L_{1x} + L_{2x} + L_{1x} + L_{3x} + L_{1x} + L_{4x} + L_{1x} + L_{5x} + L_{2x} + L_{3x} + L_{2x} + L_{4x} + L_{2x} + L_{5x} + L_{3x} + L_{4x} + L_{3x} + L_{5x} + L_{4x} + L_{5x}$$

$$\sum_{i<j} d_{ij} = 4\left(L_{1x} + L_{2x} + L_{3x} + L_{4x} + L_{5x}\right) = 4S_0$$

$$S_0 = \frac{1}{4}\sum_{i<j} d_{ij} = \frac{1}{n-1}\sum_{i<j} d_{ij}$$

Méthodes de distance : la NJ

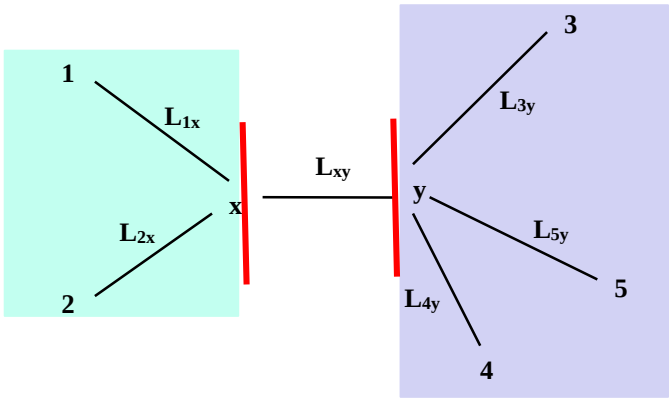
Construction de l'arbre minimum pas à pas

Deuxième étape: recherche de la paire d'OTU qui une fois agglomérée minimise la longueur de l'arbre résultant

Les n (n-1)/2 paires (i,j) sont testées et les S_{ij} des arbres correspondants sont calculés.
Exemple : on regroupe les deux OTU 1 et 2. Ceci nécessite la création des deux nœuds internes x et y, ainsi que celle d'une branche interne reliant x à y.

Le score S_{12} de cette topologie est :

$$S_{12} = L_{1x} + L_{2x} + L_{xy} + \sum_{i=3}^{n} L_{iy}$$



Cette topologie peut être découpée en deux sous-arbres : un arbre avec les OTU 1 et 2 reliées à x et un arbre en étoile avec les (n-2) OTU reliées à y. On sait calculer leurs sommes des longueurs des branches.

$$L_{1x} + L_{2x} = d_{12}$$

et

$$\sum_{i=3}^{n} L_{iy} = \frac{1}{n-3} \sum_{i<j} d_{ij}$$

donc

$$S_{12} = d_{12} + L_{xy} + \frac{1}{n-3} \sum_{i<j} d_{ij}$$

Méthodes de distance : la NJ

$$S_{12} = d_{12} + L_{xy} + \frac{1}{n-3}\sum_{i<j} d_{ij}$$ **(1)**

Il nous reste donc à déterminer L_{xy}, or

$$\sum_{k\neq i,j}\left(d_{ik}+d_{jk}\right) = \sum_{k\neq i,j}\left(d_{1k}+d_{2k}\right) = d_{13}+d_{14}+d_{15}+d_{23}+d_{24}+d_{25}$$

$$\sum_{k\neq i,j}\left(d_{1k}+d_{2k}\right) = L_{1x}+L_{xy}+L_{y3}+L_{1x}+L_{xy}+L_{y4}+L_{1x}+L_{xy}+L_{y5}+L_{2x}+L_{xy}+L_{y3}+L_{2x}+L_{xy}+L_{y4}+L_{2x}+L_{xy}+L_{y5}$$

$$\sum_{k\neq i,j}\left(d_{1k}+d_{2k}\right) = 3\left(L_{1x}+L_{2x}\right)+6L_{xy}+2\left(L_{y3}+L_{y4}+L_{y5}\right) = \left(n-2\right)d_{12}+2\left(n-2\right)L_{xy}+2S_k$$

Car on a n = 5

Soit dans le cas général :

$$L_{xy} = \frac{1}{2(n-2)}\left[\sum_{k\neq i,j}\left(d_{ik}+d_{jk}\right)-\left(n-2\right)d_{ij}-2S_k\right]$$

En remplaçant dans (1) :

$$S_{ij} = d_{ij} + \frac{1}{2(n-2)}\left[\sum_{k\neq i,j}(d_{ik}+d_{jk})-(n-2)d_{ij}-2S_k\right]+S_k$$

Dans (1) on avait $i = 1$ et $j = 2$

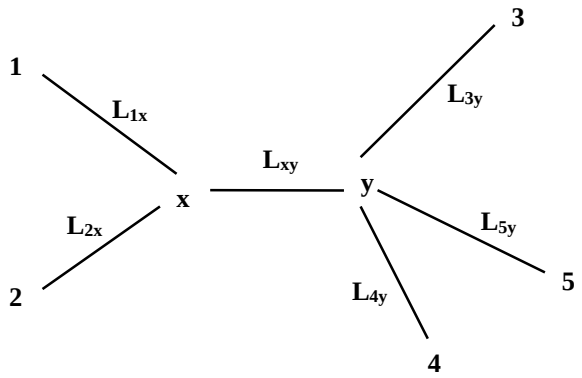
$$S_{ij} = \frac{1}{2}\,d_{ij} + \frac{1}{2(n-2)}\sum_{k\neq i,j}(d_{ik}+d_{jk}) + \frac{n-3}{n-2}\,S_k \qquad \text{or} \qquad S_k = \frac{1}{n-3}\sum_{k\neq i,j;k<l} d_{kl}$$

$$S_{ij} = \frac{1}{2}\,d_{ij} + \frac{1}{2(n-2)}\sum_{k\neq i,j}(d_{ik}+d_{jk}) + \frac{1}{n-2}\sum_{\substack{k\neq i,j\\ k<l}} d_{kl}$$

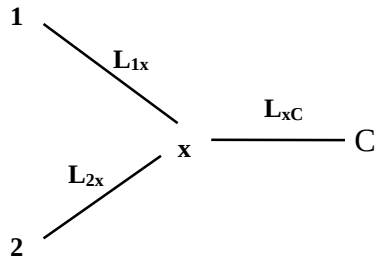
Méthodes de distance : la NJ

Une fois les deux OTU *i* et *j* les plus proches, *i.e.,* donnant la somme des longueurs des branches S_{ij} minimum, ayant été trouvées, il faut estimer les longueur de branches L_{ix} et L_{jx}. Supposons que *i* = 1 et *j* = 2

On applique la méthode de Fitch et Margoliash (méthode des moindres carrés) qui a pour objectif de se ramener systématiquement au cas simple d'un arbre à trois groupes :



devient



Avec C regroupant tous les autres OTU

D'où le système de trois équations :

$$d_{12} = L_{1x} + L_{2x}$$
$$d_{1C} = L_{1x} + L_{xC}$$
$$d_{2C} = L_{2x} + L_{xC}$$

qui une fois résolu donne :

$$L_{1x} = (d_{12} + d_{1C} - d_{2C})/2$$
$$L_{2x} = (d_{12} + d_{2C} - d_{1C})/2$$
$$L_{xC} = (d_{1C} + d_{2C} - d_{12})/2$$

Comme la distance entre 1 et C est donnée par la moyenne des distances séparant 1 de tous les éléments de C (de même pour 2)

On a : $d_{1C} = \frac{1}{n-2}\sum_{k\neq 1,2} d_{1,k}$

$d_{2C} = \frac{1}{n-2}\sum_{k\neq 1,2} d_{2,k}$

Soit dans le cas général :

$$L_{ix} = (d_{ij} + d_{iC} - d_{jC})/2$$
$$L_{jx} = (d_{ij} + d_{jC} - d_{iC})/2$$

$$d_{iC} = \frac{1}{n-2}\sum_{k\neq i,j} d_{i,k}$$

$$d_{jC} = \frac{1}{n-2}\sum_{k\neq i,j} d_{j,k}$$

Méthodes de distance : la NJ

Construction de l'arbre minimum pas à pas

On a trouvé les deux OTU *i* et *j* à regrouper et estimé les longueurs des branches L_{ix} et L_{jx}

Troisième étape: calcul d'une nouvelle matrice

On va recalculer la matrice de distance en tenant compte des deux OTUqui ont été regroupées. La nouvelle matrice aura donc une ligne et une colonne de moins. Dans notre exemple, on est parti d'un tableau à 5 OTU. Le nouveau tableau ne contiendra plus que 4 OTU. Les nouvelles distances sont recalculées en estimant la distance évolutive séparant la nouvelle OTU (*i,j*) de chacune des OTU restantes *k.* Elle est donnée par :

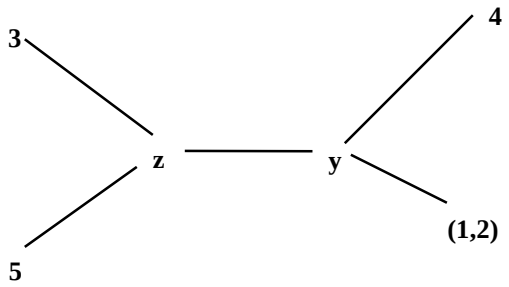
$$d_{i,j,k} = \frac{1}{2}(d_{ik} + d_{jk} - d_{ij})$$

Si on considère qu'à l'étape 2 on a regroupé les OTU 1 et 2, la matrice devient :

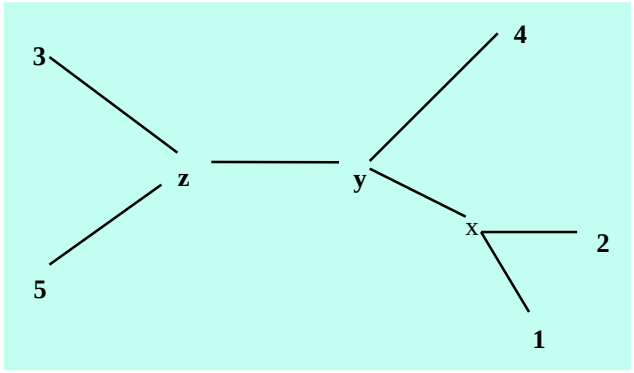
	(1,2)	3	4	5
(1,2)	0			
3	$(d_{13}+d_{23}- d_{12})/2$	0		
4	$(d_{14}+d_{24}- d_{12})/2$	d_{34}	0	
5	$(d_{15}+d_{25}- d_{12})/2$	d_{35}	d_{45}	0

Méthodes de distance : la NJ

On va réitérer les étapes 2 et 3. Donc à partir la nouvelle matrice de distances, on va rechercher les 2 UE les plus proches (3 et 5 par exemple), c'est-à-dire minimisant la somme des longueurs de branche de l'arbre résultant :



Que l'on peut tracer ainsi :



Le cycle :
- regroupement des deux OTU les plus proches par la création d'une arête interne
- calcul des longueurs des branches L_{ix}
- calcul de la nouvelle matrice de distance par regroupement de la paire d'OTU

sera réalisé (*n-3*) fois (recherche des (n-3) branches internes) si on étudie *n* OTU(*n* séquences) car lorsqu'il ne reste plus que trois OTU à placer, il n'y a qu'une seule topologie possible.

Méthodes de distance : la NJ

Des améliorations ont été apportées à l'algorithme d'origine, réduisant sa compléxité et donc le temps de calcul requis (algorithme de Studier et Keppler (1988)). C'est donc cette version qui est implémentée dans la plupart des programmes de phylogénie moléculaire.

Des variantes avec pondérations ont également été développées comme la BIONJ (Gascuel, 1997). La BIONJ apporte de améliorations évidentes surtout quand les séquences sont fortement divergentes et/ou quand elles présentent des vitesses d'évolution différentes.

Conclusion :

- Méthode performante car bon équilibre entre rapidité et efficacité. Elle peut être appliquée sur des très grands jeux de données. Robuste car ne dépend pas de l'ordre des séquences.
- ne fait pas l'hypothèse de l'horloge moléculaire (n'impose pas aux distances estimées d'être ultra-métriques).
- Souvent utilisée pour chercher des arbres qui vont servir de point de départ pour des méthodes plus coûteuses en temps calcul comme la méthode du maximum de vraisemblance (cf. ce chapitre du cours).
- Elle peut être appliquée sur n'importe quel type de distances évolutives.
- Peut conduire à des distances négatives notamment pour les branches terminales mais l'application de la contrainte de non-négativité permet de s'affranchir du problème.
- Problème pouvant être rencontré quand deux paires de voisins différentes donnent des arbres minimums de même longueur. Dans ce cas, tirage au hasard d'une solution. Situation pas fréquemment rencontrée.
- Ne donne pas d'informations sur les états de caractères de l'ancêtre commun.