

Introduction à la phylogénie moléculaire à l'usage des malacologues amateurs

Johann DELCOURT

Le malacologue amateur est souvent intéressé par comprendre le lien qui existe entre les espèces, et c'est d'autant plus vrai si celui possède des collections de coquilles, que ce soit pour raisons scientifiques ou simplement pour le plaisir. On peut même caricaturer certaines situations rencontrées par le collectionneur : un changement de taxonomie peut induire un petit traumatisme car il implique éventuellement une correction des étiquettes de la collection. Une modification de l'ordre faussement établi des familles par la création de nouvelles familles, voire transfert d'espèces d'une famille à une autre, et c'est le drame. Le collectionneur soigneux, ayant un tiroir par famille se retrouve à devoir changer toutes la structure de sa collection, voire de meuble. Le malacologue se retrouve alors fatalement à s'intéresser à pourquoi telle espèce est rapprochée d'une autre, et pourquoi il y a régulièrement des remises en questions voire des modifications dans les classifications. La faute vient des scientifiques qui ne se contentent pas que de classer les êtres vivants ; ces derniers veulent en plus que cette classification traduise l'histoire évolutive des espèces. Hors l'état de leur connaissance progresse en permanence, et les modifications régulières dans les classifications trahissent ce travail perpétuel de reconstitution. Pour le malacologue amateur qui veut en savoir plus, celui-ci se retrouve rapidement né à né avec des articles scientifiques présentant des arbres phylogénétiques. N'ayant pas toujours reçu une formation de génétique, et qui plus est, récente car il s'agit d'un domaine en constant progrès, il est facile à imaginer que ce malacologue soit un peu perdu face à ces d'arbres. L'aider à comprendre les bases de cette discipline qu'est la phylogénie moléculaire est l'objectif de cet article.

Les origines de la phylogénie

Quelques moments clés :

- ◆ Charles Darwin (1809-1882): **théorie de l'évolution** qui établit le lien de parenté entre les espèces, première esquisse d'un arbre phylogénétique en 1837 (First Notebook on Transmutation of Species).
- ◆ Ernst Haeckel (1834-1919) : popularise et développe l'image de **l'arbre de la vie**.
- ◆ Rosalind Franklin (1920-1958), James Watson (1928-), Francis Crick (1916-2004) découvre la structure de l'**ADN**, le support de transmission d'informations à la descendance: début de la biologie moléculaire
- ◆ 1965 : Emile Zuckerkandl et Linus Pauling : l'**ADN** fournit en plus une information sur **notre histoire évolutive**! Ils développent aussi le concept d'horloge moléculaire.

Qu'est-ce-que la phylogénie

La **phylogénie** est l'étude des relations de parenté entre les organismes ou les taxons. Les taxons sont des rangs taxonomiques identifiés quelques soit son niveau (espèce, genre, famille, ordre, ...). La phylogénie implique l'**hypothèse de l'origine unique de la vie** où tous les organismes sont plus ou moins fortement apparentés. **LUCA** (*Last Universal Common Ancestor*) représente le plus ancien ancêtre commun de l'ensemble des êtres vivants. Les relations entre les organismes sont représentées par des **arbres phylogénétiques** ou arbres de parenté. Ces arbres sont des outils mathématiques qui tentent de reconstituer l'histoire évolutive des espèces sur base d'une série d'informations caractérisant ces espèces. On remarquera aussi que de nombreux gènes ont aussi un gène ancestral commun ; et un arbre phylogénétique d'une famille de gènes peut aussi être réalisé.

A partir de quelles informations ?

Avant la découverte de l'ADN, l'essentiel des informations permettant d'identifier le degré approximatif de parenté entre les espèces était surtout basé sur la comparaison morphologique, surtout les **caractères homologues** (ayant une même origine mais pouvant avoir des fonctions similaires ou différentes) et non analogues (même fonction mais origine différente). Avec la découverte de l'**ADN, de l'ARN et des protéines (chaînes d'acides aminés)**, une foule d'informations quantitatives sont devenues accessibles. Pour l'ADN et l'ARN, le code est quaternaires (4 bases) mais pour les protéines, les acides aminés sont au nombre de 22, à quelques exception près. Cependant, concernant les espèces fossiles, seules l'information morphologique est accessible ; l'ADN fossile est très rare, et le plus vieux ADN fossile obtenu actuellement remonte à seulement 700.000 ans. Les méthodes moléculaires comparent donc pratiquement que les espèces actuelles ou très récemment éteintes.

L'analyse ne traite généralement pas l'intégralité d'un génome, mais plutôt des séquences bien choisies, par exemple un gène codant qui existe sur l'ensemble des espèces que l'on étudie. Le choix de ces séquences est crucial pour l'efficacité de l'analyse. Les mutations apparaissant dans une séquence d'un gène important (ex. qui code pour une fonction de base du métabolisme d'un animal) sont généralement plus contre-sélectionnées (car ayant plus facilement un impact négatif voire mortel pour l'individu porteur) qu'un gène codant pour un caractère moins crucial (ex. couleur

des yeux). Donc, certains gènes peuvent accumuler plus de mutations que d'autres. Si on étudie les relations d'un groupe d'espèces dont l'ancêtre commun est très ancien, il est nettement préférable de travailler avec un gène qui n'accumule pas trop vite les mutations. En effet, trop de mutations peuvent faire perdre la piste des liens de parenté. Inversement, trop peu de mutations ne permettront pas de résoudre l'arbre car la comparaison de la séquence sera trop peu informative. Par contre, pour d'autres types d'étude génétique qui implique des périodes plus courtes (par ex. mesurer les déplacements relativement récents de certaines populations), il est intéressant d'étudier des séquences non codantes (ancien gène qui ne sont plus exprimés ou encore des séquences n'ayant jamais eu cette fonction) qui ont pour particularité d'accumuler rapidement à l'échelle évolutive de nombreuses mutations, car leur mutation n'ont que très peu d'impact sur la survie de l'individu porteur. Ajoutons encore que des mutations peuvent être portées que par certains individus d'une espèce, alors que d'autres mutations, généralement plus anciennes, se retrouvent chez l'ensemble des individus d'une population voire de l'espèce.

Description d'un arbre phylogénétique

Un arbre phylogénétique est constitué de **branches** qui réunissent des **nœuds** (=espèces ou unité taxonomique). Les espèces toujours vivantes constituent des unités taxonomique opérationnelle, et forme ce que l'on appelle **les feuilles** de l'arbre (= nœuds terminaux). Les **nœuds internes** correspondent aux unités taxonomiques hypothétiques, supposées être des espèces intermédiaires éteintes, ancêtre commun des espèces représentées par les nœuds plus récents. La forme globale de l'arbre (description des branches et des nœuds) constitue sa **topologie**.

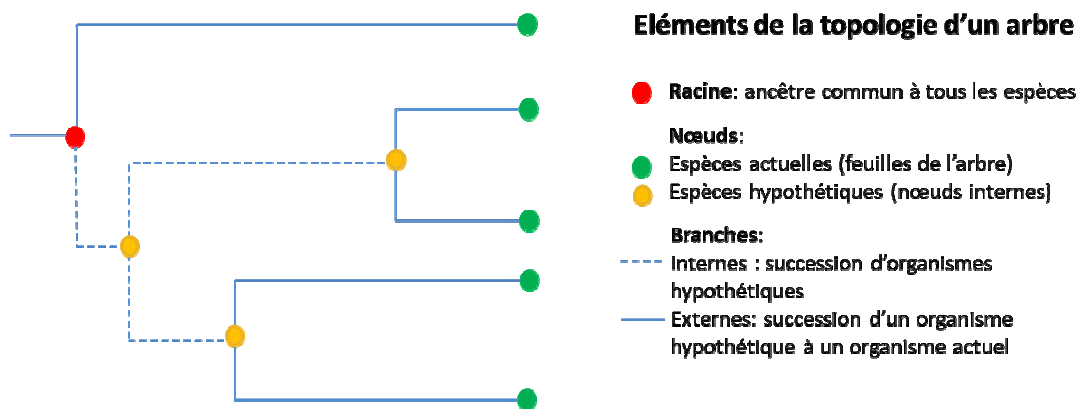


Fig. 1 : Éléments de la topologie d'un arbre phylogénétique

La racine de l'arbre

Les arbres phylogénétiques peuvent ou ne pas présenter de racine, selon que cette arbre tienne compte ou pas d'une dimension temporelle. Cette **racine** représente l'ancêtre commun, et permet de suivre le chemin évolutif séparant chaque espèce de l'ancêtre commun. En absence de racine, l'arbre représente juste le lien de proximité. Cependant, d'un arbre non-(en)raciné peut être calculé la position de la racine, on parle d'**enraciner l'arbre**. Cependant, la racine peut être théoriquement positionnée sur autant de branches que possède l'arbre non-(en)raciné ; hors il n'y a eu qu'une seule histoire évolutive, donc qu'une seule position de la racine. Pour enraciner l'arbre, il existe deux méthodes.

La première méthode est basée sur **les poids moyens**, en se reposant sur **l'hypothèse de l'horloge moléculaire**. Cette hypothèse implique que toutes les séquences pour chaque espèce évoluent à la même vitesse, c'est-à-dire accumulent en moyenne le même nombre de mutations par unité de temps. La racine est ici placée au point de l'arbre équidistant (en termes de nombre de mutations) de toutes les feuilles. L'ancêtre commun à tous est distant de toutes les espèces descendantes actuelles par le même nombre de mutation. Si par contre, une lignée d'espèces accumule plus de mutation dans leur séquence génétique que d'autres, cet arbre devient totalement biaisé. Cette méthode est utilisée que pour des groupes d'espèces proches (une même famille par ex.), mais plus pour des espèces assez éloignées (par ex. des phylums ou des classes). Le type d'arbre obtenu est un **chronogramme**.

La deuxième méthode, **l'enracinement par groupe extérieur**, consiste à enraciner l'arbre en incluant quelques espèces dans l'analyse dont on sait a priori qu'elles sont à coup sûr extérieures (plus éloignées) que le groupe étudié (ex. une moule pour étudier les liens de parenté entre des gastropodes Trochidés par ex.). Dans ce cas, l'hypothèse de l'horloge moléculaire n'est pas nécessaire. La racine est définie par le nœud reliant le groupe extérieur aux séquences étudiées.

La longueur des branches

Dans un **cladogramme**, si la topologie des branches indique si une espèce est plus proche d'une autre, la longueur de ces branches est arbitraire et ne reflète pas la distance évolutive séparant les séquences/ espèces. Dans un **chronogramme**, la longueur des branches est proportionnelle au temps, et donc les nœuds indiquent les **temps de divergence** entre deux espèces (ou deux groupes d'espèces). Cet arbre est construit sur base de l'hypothèse moléculaire comme nous l'avons vu au point précédent. Dans le cas des arbres enracinés par un groupe extérieur, la longueur des branches représente le pourcentage de divergence entre les séquences ADN analysées. On parle alors de **phylogramme**. La somme des longueurs des branches d'un arbre permet de mesurer le **degré de similarité** entre les espèces. Cependant, deux espèces plus proches évolutivement, plus **apparentées**, peuvent connaître une similarité plus faible entre elles qu'avec une autre espèce plus éloignée. Ceci s'explique par des vitesses d'accumulation de mutations différentes selon les lignées. Dans ce cas, l'hypothèse de l'horloge moléculaire est alors réfutée dans ce groupe.

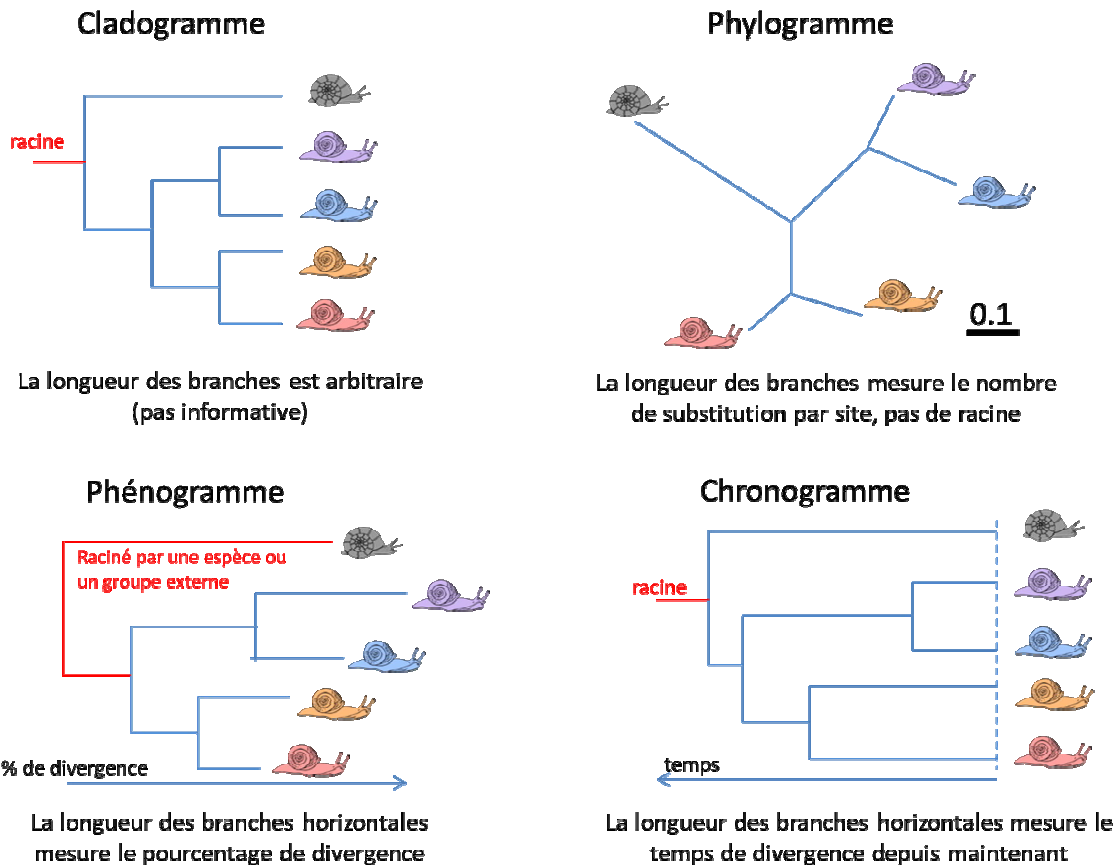


Fig. 2 : Les différents types d'arbres phylogénétiques

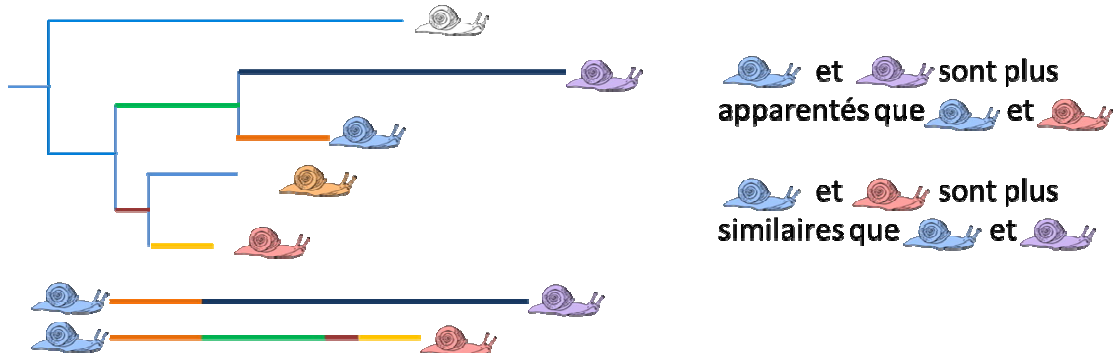


Fig. 3: Notion d'apparentement et de similarité

Quel arbre est le bon ?

Plus il y a d'espèces pris en compte, et plus le nombre d'arbres possible augmente rapidement de manière astronomique, qu'il soit (en)raciné ou pas. Avec 10 espèces, ce chiffre dépasse le million. Mais dans l'histoire évolutive, **un seul arbre est vrai**. Comment calculer la robustesse de notre arbre, ou plutôt la robustesse de certaines branches et nœuds ? La **robustesse** se définit ici comme la particularité d'une partie d'un arbre à conserver sa structure, donc à présenter peu de variation, lorsqu'on modifie quelques peu le jeu de données. Si par exemple, on oublie volontairement une espèce ou bien qu'on en rajoute une nouvelle dans l'analyse, si les regroupements sur l'arbre restent non affectés, on peut dire qu'ils sont robustes, et que ce résultat plaide à ce que statistiquement, la topologie est probablement proche de celle réellement opérée par l'évolution. Evidemment, cela se teste concrètement par beaucoup de petits événements perturbateurs pour quantifier cette stabilité (ou inversement la variation) des regroupements.

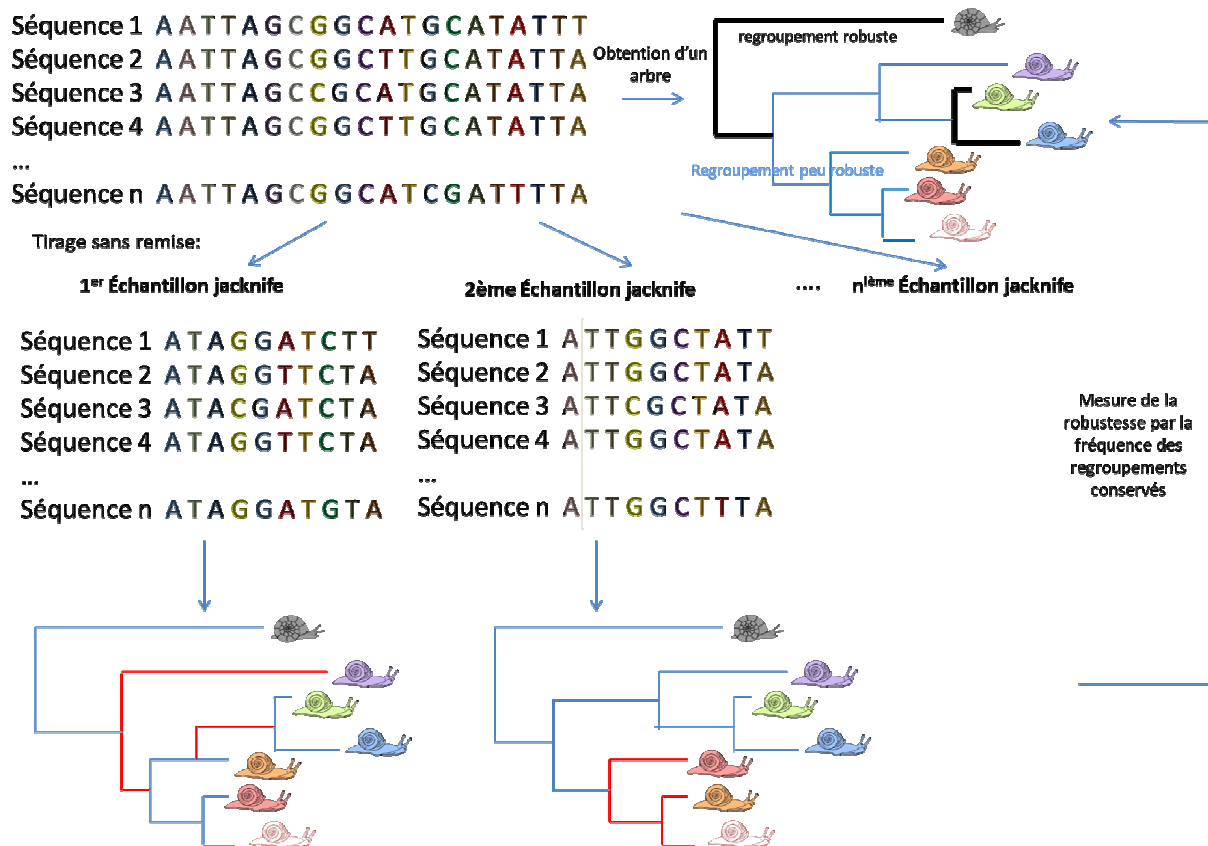


Fig. 4 : Principe du test de la robustesse des regroupements par la technique du Jackknife. Chaque barre couleur correspond à des bases homologues. Les branches rouges sont les éléments instables générés par un échantillonnage. Les branches noires épaisses les branches qui sont les plus conservées et donc les plus robustes.

Deux méthodes principales sont utilisées pour tester cette robustesse. Ces méthodes consistent d'abord à aligner les séquences génétiques homologues. A chaque base homologue correspond une position précise. Ensuite, on échantillonne la séquence. Dans la première méthode, **le jack(k)nife**, on échantillonne (sans remise) une seule fois une partie des positions, et on reconstruit l'arbre. Cette méthode n'est cependant pratiquement plus utilisée, on lui préfère la méthode du **bootstrap**. Ici aussi il y a un échantillonnage des positions, mais une même position peut être ici tirée plusieurs fois, il s'agit techniquement d'un échantillonnage avec remise. Quel que soit la méthode, la construction d'arbre par échantillonnage est réalisée un très grand nombre de fois ; la robustesse s'exprime alors en pourcentage de fois où le regroupement a été conservé.

Cependant, aussi bon soit-il pour expliquer l'histoire évolutive, un arbre restera pour autant une hypothèse susceptible d'être remis en question par tout élément nouveau. On ne parle donc pas d'arbre vrai mais d'**arbre inféré**.

Construire l'arbre phylogénétique

Le point le plus important des arbres phylogénétiques est leur construction. Il existe une série de méthodes pour y arriver, nous expliquons ici quelques-unes des plus fréquentes. Avant de commencer à comparer des séquences, il faut d'abord opérer à l'**alignement des séquences homologues (ou alignement séquentiel)**, et ne pas tenir compte des parties non-homologues que l'on appelle des **gaps** (trous). En effet, au milieu d'une séquence homologue peut s'être inséré un morceau ADN chez une espèce et pas chez les autres espèces, ou encore une partie d'une séquence peut avoir disparu chez une espèce (délétion). Ces trous sont appelés **indel** (mot valise pour insertion-délétion). Ne pas tenir compte de cette correction rend l'analyse totalement nulle. Cet alignement se réalise grâce à des programmes informatiques d'analyse qui vont tester toute une série de situation afin de faire ressortir les régions similaires homologues. Cette analyse peut se faire sur des séquences de macromolécules biologiques telles que l'ADN, l'ARN et les protéines.

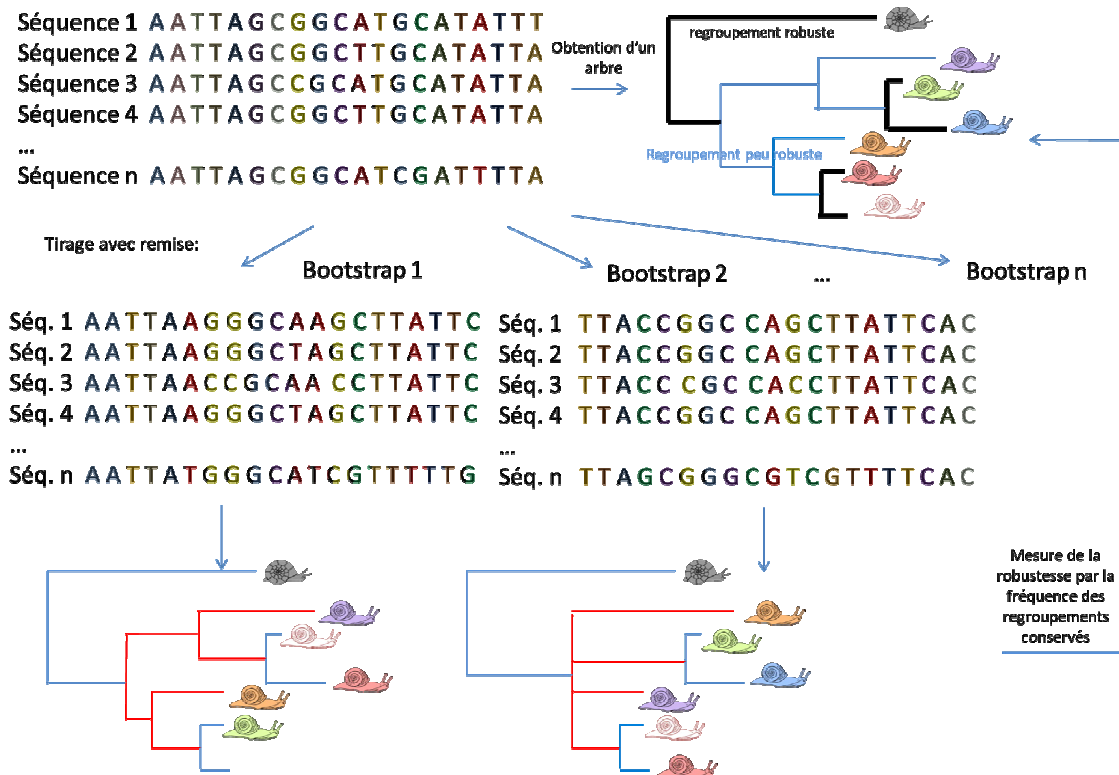


Fig.5 :

Principe du test de la robustesse des regroupements par bootstraps. Chaque barre couleur correspond à des bases homologues. Les branches rouges sont les éléments instables générés par un échantillonnage. Les branches noires épaisses les branches qui sont les plus conservées et donc les plus robustes.



Fig.5 : Alignement des bases homologues et sélection des séquences d'intérêts pour l'analyse.

Les méthodes des distances

Ces méthodes comparent deux à deux les différentes séquences, chaque séquence représentant un individu ou une espèce. Le résultat de base est une matrice de distances basées sur le degré de similarité entre ces séquences. Ensuite, tout le processus consiste à reconstruire l'arbre qui traduit le mieux ces relations de distances. Lorsqu'un regroupement est opéré (**clusterisation**), c'est-à-dire que 2 ou plus d'individus/espèces/de séquences s'avèrent plus proches que tous les autres, les méthodes recalculent la distance du nouveau groupe par rapport aux autres. Ceci permet de reconstruire pas à pas les différents nœuds de l'arbre.

La méthode de **UPGMA** (*Unweight Pair Group Method with Arithmetic mean*) s'appuie sur l'hypothèse de l'horloge moléculaire, et permet d'obtenir un chronogramme. Si deux espèces divergent dans leur séquence homologue par 12

mutations, alors statistiquement, depuis leur ancêtre commun, on considère que chaque séquence a accumulé le même nombre de mutation dans sa propre histoire évolutive, indépendante de l'autre ; soit 6 mutations chacune, dont 12 mutations différentes au total. Si une troisième espèce la plus proche est distante de 18 et 20 mutations des autres espèces, elle est distante du groupe/cluster des 2 premières espèces de $(18+20)/2$, soit 19 mutations. Le nombre de mutation qui sépare de l'ancêtre commun aux trois espèces (soit le nœud suivant de l'arbre) est de $19/2=9.5$ de distance. Cependant cette méthode échoue souvent à cause de non-respect de l'hypothèse de l'horloge moléculaire, certaines espèces ayant accumulées des mutations à des vitesses différentes des autres. Elle n'est applicable que pour étudier les relations phylogénétiques d'espèces relativement proches.

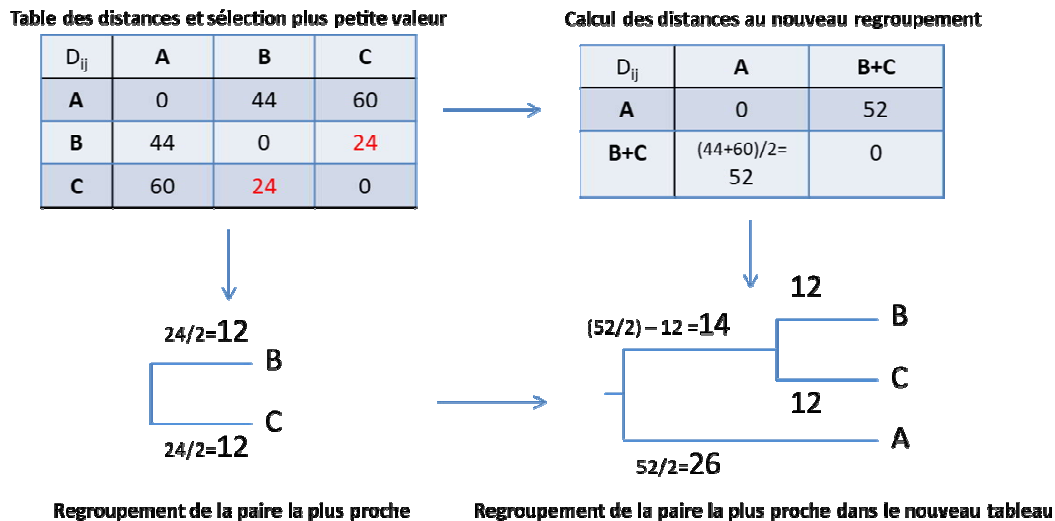


Fig. 6 : Construction d'un arbre phylogénétique (enraciné et de type chronogramme) selon la méthode UPGMA sur base d'une table de distance. A, B et C sont trois espèces différentes.

La méthode qui lui est actuellement préférée est le **neighbour joining (NJ)** qui tolère des vitesses d'accumulation de mutations différentes. Dans cette méthode, l'arbre de départ est une étoile à autant de branche que d'espèce, il n'y a qu'un seul nœud reliant toutes les espèces, et toutes les branches ont la même distance. Ensuite, tout le processus consiste à trouver pas à pas l'arbre qui minimise la somme des distances totales de toutes les branches. Il s'agit d'un arbre non enraciné. A chaque regroupement (formation d'un cluster), les distances sont recalculées, et le processus recommence jusqu'à la résolution de l'ensemble de l'arbre. L'arbre obtenu est de type phénogramme.

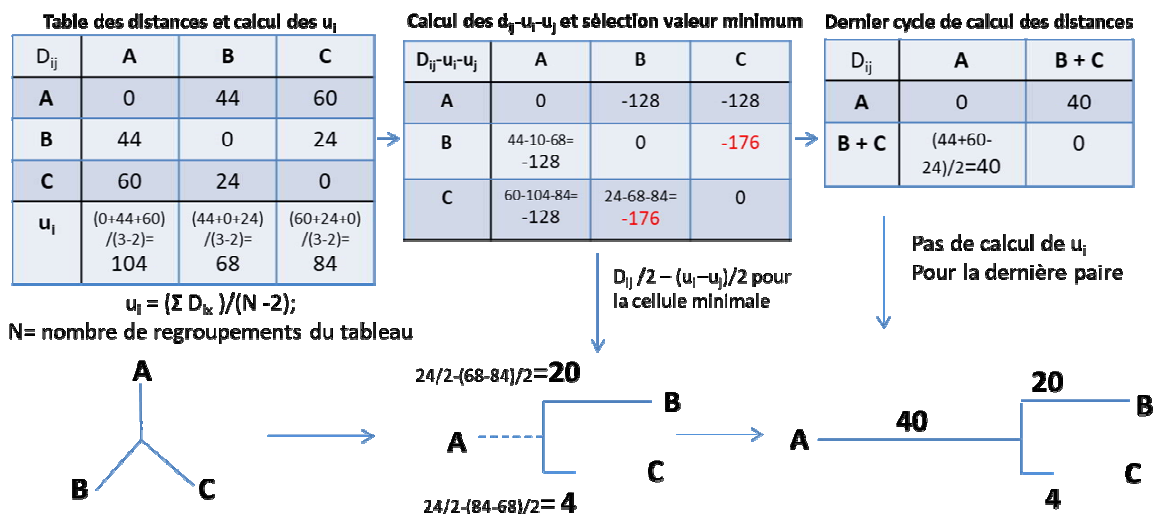


Fig. 7 : Construction d'un arbre phylogénétique selon la méthode NJ (arbre non-enraciné de type phylogramme ou arbre enraciné par le groupe A de type phénogramme) sur base d'une table de distance. A, B et C sont trois espèces différentes.

La méthode cladistique par principe de parcimonie

Dans cette méthode, l'objectif est de former un arbre qui implique le moins d'événements possible pour y arriver (**principe du maximum de parcimonie**), donc impliquant le moins de changements évolutifs que possible. Si d'une séquence ancestrale ont a deux scénarios pour expliquer les changements de séquence chez deux espèces descendantes, si l'une implique moins d'événements de mutation/substitution que l'autre, on considèra que la première est plus probable d'avoir eu lieu. L'arbre se reconstruit ainsi pas à pas, tentant de reconstituer la séquence ancestrale et les regroupements d'espèces descendantes. La méthode crée donc un grand nombre de typologie, et retient que celui qui implique le moins d'événement possible. L'arbre obtenu n'est pas (en)raciné et la longueur des branches n'est pas informative. De plus, il est possible d'obtenir plusieurs arbres différents équivalents car impliquant le même nombre d'événements, l'arbre est alors non-résolu.

La méthode par maximum de vraisemblance

Cette méthode s'appuie sur l'application d'un modèle probabiliste des événements de substitutions a priori dont on connaît la formulation en terme de probabilité, mais dont on ignore les paramètres. On teste alors plein de valeurs de paramètres pour détecter quelles valeurs ont le plus de chance de produire le résultat observé. Le modèle est basé sur des **probabilités conditionnelles**, c'est-à-dire la probabilité si on considère une topologie et des longueurs de bras donnés. On parle aussi d'**analyse Bayésienne**, du nom du théorème de Bayes, à la base des probabilités conditionnelles. Le modèle évalue la vraisemblance de manière indépendante pour chaque site (une base pour le cas d'une séquence ADN) et le résultat est obtenu par multiplication des probabilités pour les n sites. L'arbre qui possède une topologie et des longueurs de bras données qui offrent le maximum de vraisemblances par rapport à tous les autres alternatives possibles est l'arbre retenu.

Pour en savoir plus :

Perrière G. & Brochier-Armanet, C. 2010. Concepts et méthodes en phylogénie moléculaire. Springer. 250p. **Hall B.G.**, 2011. Phylogenetic Trees Made Easy: A How-To Manual (4th Edition). Sinauer Associates, Oxford University Press, 255p.