Bagage de phylogénie moléculaire

Nous discutons brièvement des concepts de base de la phylogénie moléculaire. Les bases de données biologiques qui mettent à disposition les séquences nucléotidiques, et notamment la base de données sur le VIH du laboratoire national de Los Alamos, sont présentées. Puis nous discutons de l'étape d'alignement, essentielle à toutes analyses de phylogénie moléculaire, qui consiste à mettre en regard les nucléotides de chaque séquence homologue, dérivant d'un même nucléotide ancestral. Nous présentons ensuite, les principaux modèles d'évolution nucléotidique (GTR, HKY, F84, etc.) permettant d'estimer la distance évolutive qui sépare les séquences depuis leur divergence de leur ancêtre commun. Les méthodes de distances d'inférence phylogénétique, comme UPGMA, NJ ou FastME sont rapidement exposées, tout comme les méthodes de parcimonie et les méthodes probabilistes (PhyML, MrBayes). Enfin, nous présentons les algorithmes de parcimonie ACCTRAN, DELTRAN et DOWNPASS qui permettent d'inférer les annotations ancestrales d'une phylogénie enracinée à partir d'annotations aux feuilles. La méthode du shuffling *qui permet d'en dégager la significativité statistique est aussi présentée.*

Sommaire

1.1 Introduction

La phylogénie est une discipline scientifique qui étudie les « parentés entre différents êtres vivants en vue de comprendre l'évolution des organismes vivants »[1]. Les premières phylogénies (Charles DARWIN, 1809-1882 ; Ernest HAECKEL, 1834-1919) se basaient sur des caractères morphologiques, anatomiques et/ou physiologiques afin de comparer les organismes vivants et d'étudier leur parenté. Mais lorsqu'il s'agit de comparer des organismes bactériens ou viraux ces critères de comparaison atteignent leur limite.

Depuis le développement de la biologie moléculaire et la découverte de l'ADN (acide désoxyribonucléique) comme support de l'hérédité dans les années cinquante, de nouveaux caractères sont utilisés comme source d'information pour l'inférence de phylogénies : les séquences de macromolécules (ADN, ARN et protéines). Les premières études phylogénétiques essentiellement basées sur des séquences protéiques remontent au début des années soixante et donnent ainsi naissance à une nouvelle branche de la phylogénie : la phylogénie moléculaire. Mais ce n'est que vers la fin des années soixante-dix, avec le développement de techniques spécifiques permettant de séquencer des fragments d'ADN à grande échelle et à faible coût que la phylogénie moléculaire connait un essor grandissant. En particulier parce que cette discipline est très utilisée en génomique fonctionnelle, science qui étudie le rôle des gènes.

La phylogénie moléculaire est aussi très utilisée par les épidémiologistes car elle permet de mettre en évidence des liens entre différentes souches virales, liens qui reflètent des chaînes de transmission. Un exemple souvent cité car c'est le premier qui utilise des outils de phylogénie moléculaire dans un cadre médico-légal, est celui d'un dentiste de Floride, séropositif, qui est suspecté être la source de contamination de quelques uns de ses patients (Ou *et al*, 1992). Les indices ayant menés à cette hypothèse proviennent d'une patiente atteinte du syndrome de l'immunodéficience acquise (SIDA) mais pour laquelle aucune situation de contamination n'a pu clairement être identifiée, hormis deux interventions chirurgicales venant de son dentiste. Pour confirmer un éventuel lien épidémiologique, des souches virales ont été prélevées chez le dentiste, chez la patiente, ainsi que chez six autres patients qui ont séroconverti pendant l'enquête ; par ailleurs, trente-cinq souches virales provenant d'individus locaux ont été rajoutées comme souches témoins. L'analyse phylogénétique de toutes ces souches virales a révélé que la souche collectée chez le dentiste est phylogénétiquement très proche de celles collectées chez ses patients, confirmant ainsi la source de contamination. Mais le mode de contamination reste indéterminé. De nombreux autres exemples comme celui-là sont disponibles dans la littérature, Leitner et Fitch (1999) en commentent d'autres.

[1] Source Wikipédia.

Dans ce chapitre, nous présentons brièvement les différentes méthodes d'inférence phylogénétique. Mais avant cela, nous présentons les bases de données biologiques, véritables sources d'information pour les études moléculaires, puis l'étape d'alignement, fondamentale à toute analyse phylogénétique. Enfin, nous terminerons ce chapitre par l'exposé de quelques méthodes de parcimonie permettant de reconstruire les annotations ancestrales (par exemple des régions géographiques) à partir d'une phylogénie et des annotations associées aux feuilles de cette phylogénie qui représentent les souches virales de l'alignement. Des compléments d'information peuvent être trouvés dans les ouvrages de Lemey *et al*. (2009b) ou celui de Felsenstein (2003).

1.2 Bases de données biologiques

Les études de phylogénie moléculaire sont souvent basées sur des séquences nucléotidiques. Pour être facilement accessibles, et pour faciliter le traitement de l'information, les séquences nucléotidiques obtenues par les biologistes sont stockées dans des bases de données. Ces bases de données fournissent aussi une pléthore d'outils pour manipuler ou analyser les séquences, mais aussi des informations supplémentaires sur chacune d'elles. Ces informations, ou annotations, sont très utiles car elles renseignent sur l'organisme de collecte, les propriétés de la séquence, les auteurs, etc., permettant ainsi de cibler les recherches dans ces bases.

Il existe de nombreuses bases de données biologiques mais la plupart sont spécifiques à un organisme, une fonction, etc. Toutefois, il existe trois bases de données principales :

- EMBL-Bank (*European Molecular Biology Laboratory*), maintenue par EMBL-EBI (*European Bioinformatics Institute*) à Hinxton au Royaume-Uni ;
- GenBank, maintenue par NCBI (*National Center for Biotechnology Information*) à Bethesda aux États-Unis ;
- DDBJ (*DNA Data Bank of Japan*), maintenue par NIG/CIB (*National Institute of Genetics*, *Center for Information Biology*) à Mishima au Japon.

Ces trois bases de données collaborent ensemble afin de partager les nouvelles soumissions ou les éventuelles mises à jour. L'ensemble des séquences nucléotidiques publiées y est donc accessible. Chaque séquence soumise se voit attribuer un numéro d'accession unique (qui reste le même quelle que soit la base de données) et qui permet de désigner, sans ambiguïté, les séquences dans la littérature. Par convention, les séquences nucléotidiques sont stockées sous le format de l'ADN, mais les bases de données contiennent aussi des séquences d'ARN (acide ribonucléique). Dans ce cas, ces dernières sont codées avec un « T », qui signifie la thymine, à la place d'un « U », pour designer l'uracile.

Dans nos études, nous utilisons la base de données spécifique au VIH maintenue par le laboratoire national de Los Alamos : *HIV Databases* (www.hiv.lanl.gov). Elle met à disposition un grand nombre de séquences nucléotidiques du VIH de type 1 (VIH-1), du VIH de type 2 (VIH-2) et même du SIV (*simian immunodeficency virus*), virus analogue au VIH mais infectant naturellement les singes d'Afrique. Mise à jour périodiquement, elle contient toutes les séquences soumises dans GenBank, avec un décalage de quelques mois sur les dernières entrées de GenBank. En revanche, les séquences sont annotées avec plus d'informations que celles disponibles via GenBank, comme l'origine géographique de collecte, l'année d'isolation, le sous-type d'appartenance, le groupe à risque de l'individu chez lequel elle est prélevée, etc. Ces informations sont récupérées dans les publications correspondantes aux séquences par les gestionnaires de la base de données. De plus, le site internet propose une interface de recherche conviviale, ergonomique et adaptée aux particularités du VIH et du SIV. Il est ainsi très facile d'obtenir des séquences sur une région précise du génome, provenant d'un même pays ou d'un même continent, isolées chez un patient avec un facteur à risque particulier, etc. Des outils sont aussi mis à disposition et permettent le traitement spécifique de séquences du VIH/SIV, comme, par exemple, *Sequence Locator* qui permet de retrouver les coordonnées de début et de fin d'une séquence sur le génome de référence (HXB2 pour le VIH et SIVmm239 pour le SIV).

Malgré le soin apporté au classement et au référencement des séquences, ces bases de données peuvent contenir des informations erronées. Il revient à l'utilisateur de vérifier la justesse des informations.

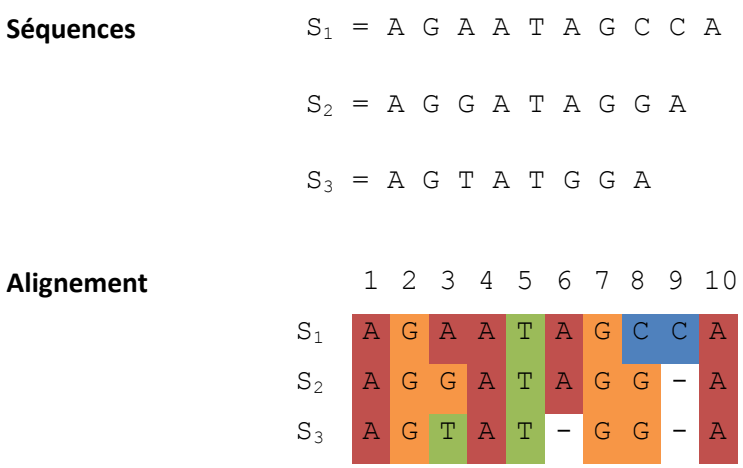
1.3 L'alignement, une étape indispensable

L'alignement de séquences nucléotidiques est une étape clef des études de phylogénie moléculaire. Cette étape ne peut se faire qu'avec des séquences homologues, c'est-à-dire des séquences nucléotidiques partageant un même ancêtre commun, puisqu'elle consiste à identifier, pour chaque séquence, les nucléotides dérivant du même nucléotide ancestral et à les positionner en regard. Le résultat de cette étape est l'obtention d'une matrice, appelée alignement, où chaque ligne correspond à une séquence et où chaque colonne, appelée site, contient les nucléotides dérivés d'un même nucléotide ancestral (Figure 1).

Dans certaines séquences de l'alignement des gaps (ou *indels*) ont pu être introduits. Ils correspondent aux phénomènes biologiques d'insertions (ajout d'un ou de plusieurs nucléotides) ou de délétions (perte d'un ou de plusieurs nucléotides) qui se sont produits au cours de l'évolution. Toutefois, l'utilisation de gaps dans un alignement doit être faite avec parcimonie. Ainsi, un bon alignement est défini comme un alignement qui contient le moins d'évènements de mutation possibles,

avec des pondérations différentes pour les différents évènements mutationnels (substitution, insertion, délétion, ouverture de gap, prolongation de gap, etc.).

Figure 1. Exemple d'alignement de séquences.
L'alignement du bas est un alignement possible résultant des trois séquences du haut. Les positions 1, 2, 4, 5, 7 et 10 ne présentent aucune modification. La position 3 présente deux substituions et la position 8 une substitution pour la séquence S_1. La position 6 présente une délétion pour la séquence S_3 et la position 9 une insertion pour la séquences S_1. D'autres interprétations de l'alignement sont possibles mais elles impliquent davantage d'évènements de mutation.
L'exemple est extrait de Caraux et al. (1995).



Comme l'alignement est la base de toutes méthodes de phylogénie moléculaire, il est indispensable d'avoir un alignement d'une qualité optimale afin d'inférer des phylogénies fiables. Dans le cas contraire, elles peuvent contenir des erreurs ou être aberrantes. C'est pour cela que les biologistes ôtent de l'alignement les sites les plus incertains, comme ceux contenant des gaps ou les parties trop divergentes (souvent en début ou en fin de l'alignement).

Des méthodes automatisées existent pour résoudre des alignements. La plus simple concerne l'alignement entre deux séquences en se basant sur la distance d'édition (ou distance de Levenshtein). Cette distance mesure la similarité entre deux mots. Pour cela, elle calcule le nombre minimum de remplacements (ou substitutions), de délétions ou d'insertions nécessaires pour transformer un mot en l'autre. Rappelons que les séquences nucléotidiques peuvent être vues comme des mots sur l'alphabet génétique $\mathcal{A} = \{A, C, G, T\}$. Un algorithme simple de programmation quadratique permet de calculer la distance d'édition en $O(n \times m)$, où n et m sont les longueurs respectives des deux séquences. Néanmoins cet algorithme calcule uniquement la distance (ou le score) de l'alignement optimal. Un algorithme supplémentaire est nécessaire afin d'en déduire l'alignement, il se fait en $O(n + m)$ en réutilisant le tableau construit lors du calcul de la distance d'édition. Lorsque l'on souhaite aligner plus de deux séquences simultanément, le problème devient très vite complexe. Il est bien sûr possible d'adapter l'algorithme précédent dans le cas de plusieurs séquences, mais la complexité devient alors exponentielle sur le nombre de séquences, et l'application sur plus de

quatre ou cinq séquences est inenvisageable. Pour contrer ce problème, des heuristiques sont proposées mais elles ne permettent pas de résoudre avec exactitude le problème de l'alignement. Les biologistes utilisent donc ces heuristiques afin d'obtenir une base convenable de l'alignement, puis le modifient manuellement avec des logiciels d'éditions.

De nombreux programmes sont disponibles pour résoudre le problème d'alignement multiple de séquences. Une liste exhaustive est trouvée dans Lemey *et al*. (2009b). Dans nos études, seul le logiciel MAFFT (Katoh *et al*, 2005) est utilisé car il a été démontré qu'il est l'un des plus performants (Thompson *et al*, 2011).

1.4 Modèles d'évolution moléculaire

La distance évolutive entre deux séquences nucléotidiques est définie comme « le nombre moyen de substitutions par site s'étant produites depuis que ces séquences ont divergé de leur ancêtre commun » (Perrière & Brochier-Armanet, 2010). Pour calculer la distance évolutive qui sépare deux séquences dans l'alignement, une approche simpliste consisterait à compter le nombre de dissemblances (c'est-à-dire le nombre de sites différents) et de le diviser par la longueur de l'alignement. Cette distance évolutive est appelée p-distance (exprimée en substitutions par site) et correspond à la distance observée entre les deux séquences, et non à la distance évolutive réelle. En effet, imaginons qu'entre deux séquences données, et sur un site donné, les nucléotides A et G sont observés. La p-distance comptabilise une substitution, car c'est ce qui est observé. Mais si la base A est remplacée par la base T, puis par la base G, il y a eu deux évènements de substitutions réelles, mais toujours une substitution observée. Donc la p-distance sous-estime la distance évolutive réelle, puisque de nombreuses substitutions cachées ont pu se produire.

Des modèles d'évolution sont donc proposés pour estimer au mieux la distance évolutive réelle à partir de la distance évolutive observée. Ces modèles font les hypothèses simplificatrices suivantes :

- les séquences évoluent uniquement avec un processus de substitution nucléotidique, c'est-à-dire que les évènements d'insertion et de délétion ne sont pas pris en compte ;
- les sites de l'alignement sont indépendants les uns des autres, c'est-à-dire que les évènements évolutifs d'un site n'ont aucune influence et ne sont pas influencés par les évènements évolutifs des autres sites de l'alignement ;
- le processus d'évolution est markovien d'ordre 1, c'est-à-dire que l'état futur d'un site ne dépend que de son état actuel et non des états passés précédents ;
- le processus d'évolution est identiquement distribué, c'est-à-dire qu'il est le même quel que soit le site de l'alignement ;

- le processus d'évolution est homogène, c'est-à-dire qu'il ne varie pas au cours du temps, il est donc applicable à toutes les branches de la phylogénie ;
- le processus d'évolution est stationnaire, c'est-à-dire que les probabilités d'observer une base particulière sont celles attendues à l'état d'équilibre (atteint lorsque les séquences ont évolué après un temps infini). Ces probabilités sont donc les mêmes pour toutes les séquences de l'alignement, puisqu'après un temps infini les séquences sont supposées avoir une composition en base identique ;
- au plus une mutation peut se produire dans un temps infinitésimal, c'est-à-dire qu'il ne peut y avoir plus d'une substitution simultanément.

Même si ces hypothèses simplifient fortement les modèles d'évolution, les résultats obtenus sont jugés largement acceptables par les biologistes.

Plusieurs modèles sont proposés afin de simuler le processus d'évolution. Chacun fait des hypothèses différentes en ce qui concerne les fréquences d'apparition des nucléotides et les taux de substitution (probabilité de passer d'un nucléotide i vers un nucléotide j). Le modèle le plus général est le modèle *general time reversible* (GTR) (Lanave *et al*, 1984) qui suppose une fréquence d'apparition différente pour chacun des quatre nucléotides et un taux de substitution relatif différent pour chacune des deux transitions (A ↔ T et C ↔ G) et des quatre transversions possibles (A ↔ C, A ↔ G, T ↔ C et T ↔ G) (dans les modèles réversibles comme GTR, le taux de substitution relatif [ou échangeabilité] de $i \rightarrow j$ est supposé le même que celui de $j \rightarrow i$). Comme il existe deux relations linéaires, une entre les fréquences d'apparition (somme à 1) des nucléotides et l'autre entre les taux de substitution (facteur de normalisation), le modèle GTR a huit paramètres libres (4 fréquences + 6 taux symétriques – 2 relations linéaires). Le modèle de Jukes et Cantor (1969), abrégé JC69, quant à lui, est le moins général. Il suppose que les fréquences d'apparition sont toutes égales et que les taux de substitution sont tous identiques. Il n'a donc aucun paramètre libre. La Figure 2 liste les modèles d'évolution les plus utilisés et les classe suivant leurs paramètres libres.

Les modèles d'évolution décrits ci-dessus supposent que tous les sites évoluent de façon identique, c'est-à-dire que les taux de substitution sont identiques quels que soit les sites de l'alignement. C'est une hypothèse fausse qui peut amener à des estimations biaisées si elle est fortement contredite par les données étudiées. Une alternative est d'ajouter un paramètre qui permet de faire varier les taux de substitution en fonction des sites suivant une loi gamma (Yang, 1994, 1993; Jin & Nei, 1990). Dans la littérature, l'ajout d'une loi gamma au modèle d'évolution considéré est noté +G ou +Γ. Généralement on n'est pas capable, pour des raisons mathématiques, d'utiliser la loi gamma continue standard (l'exception est le calcul des distances évolutives entre paires de sé-

quences, où cette loi est utilisable directement) ; on utilise alors une discrétisation de la loi gamma, souvent à 4, 6 ou 8 catégories, notée +Γ4, +Γ6 ou +Γ8.

Figure 2. Liste des modèles d'évolution.
Tableau récapitulant les différents modèles d'évolution généralement employés, organisés en fonction de leur supposition sur les différents paramètres. Les chiffres entre parenthèse indiquent le nombre de paramètres libres du modèle. Le modèle le plus général est le modèle GTR (Lanave *et al*, 1984) et le plus restreint JC69 (Jukes & Cantor, 1969). Les autres modèles (TN93 (Tamura & Nei, 1993), K2P (Kimura, 1980), HKY85 (Hasegawa *et al*, 1985) et F81 (Felsenstein, 1981)) sont des modèles intermédiaires. Le modèle K2P est parfois appelé K80. Le modèle HKY85 est aussi décrit par Felsenstein (1993) mais avec une formulation différente (d'où l'emploi des deux noms).

		Fréquence des nucléotides	
		Identique	**Différente**
Taux relatif (symétrique) de substitution	6 taux relatifs différents	pas utilisé	GTR (8)
	3 taux relatifs différents (transversions et les deux transitions)	pas utilisé	TN93 (5)
	2 taux différents (transitions contre transversions)	K2P (1)	HKY85 et F84 (4)
	1 taux (tous les taux sont égaux)	JC69 (0)	F81 (3)

Ces modèles d'évolutions supposent aussi que tous les sites de l'alignement sont variables, même lorsque le même nucléotide est présent sur chacune des séquences. Ces sites sont peut-être très conservés et évoluent donc à une vitesse nettement différente que celle des sites voisins. Pour supposer qu'une fraction de sites peut être invariable ou varier avec un taux de substitution très nettement inférieur à ceux des autres sites, un paramètre supplémentaire, qui mesure la proportion de sites invariants, est ajouté aux modèles d'évolution (Waddell & Penny, 1996; Gu *et al*, 1995). Dans les articles, les modèles rajoutant cette catégorie (potentielle) de sites invariants sont notés +I.

Pour les séquences codantes, il est parfois préférable d'utiliser des modèles d'évolution protéique (donc, à partir de séquences protéiques) plutôt que des modèles d'évolution nucléotidique. Les séquences protéiques sont plus conservées que les séquences nucléotidiques puisqu'elles ne prennent en compte que les substitutions non synonymes, et de ce fait, elles sont plus adaptées à la comparaison de séquences très divergentes. Les modèles JTT (Jones *et al*, 1992) et WAG (Whelan & Goldman, 2001) sont parmi les modèles d'évolution protéique les plus connus et les plus utilisés, mais il existe une variété de modèle d'évolution protéique, en raison des différentes pressions de sélection subies sur telle ou telle protéine. À cet effet, Dimmic *et al*. (2002) proposent le modèle rtREV qui est adapté à l'évolution et aux pressions de sélection de la transcriptase inverse des rétrovirus. Les modèles d'évolution protéique ne sont pas utilisés dans nos travaux puisque la région génomique utilisée (*pol*) reste très conservée à l'intérieur d'un même sous-type.

1.5 Méthodes d'inférence phylogénétique

Les méthodes d'inférence phylogénétique peuvent être divisées en deux catégories : les méthodes de distances et les méthodes de caractères qui comprennent les méthodes basées sur des modèles probabilistes d'évolution, cités ci-dessus. Les méthodes de distances ont pour base de calcul, non pas un alignement, mais une matrice contenant les distances évolutives entre paires de séquences, tandis que les méthodes de caractères se réfèrent constamment à l'alignement. Remarquons qu'en phylogénie moléculaire, les distances entre paires de séquences sont calculées à partir de l'alignement et en utilisant un modèle d'évolution. Mais ces distances peuvent très bien provenir d'une autre source d'information, comme, par exemple, des données morphologiques. Les méthodes de distances sont bien connues pour être rapides en temps de calcul, tandis que les méthodes probabilistes, plus lentes, sont généralement bien plus précises.

Avant d'aborder ces différentes méthodes, nous exposons la représentation d'une phylogénie.

1.5.1 Arbre phylogénétique

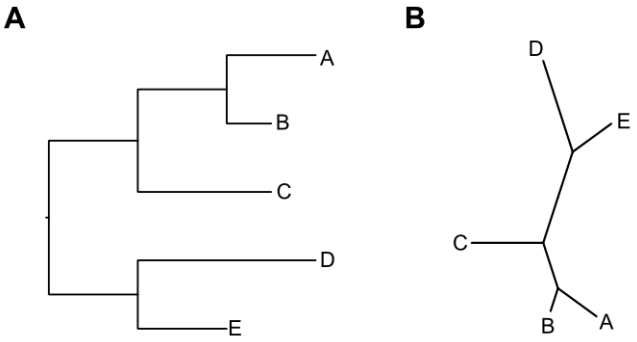
Une phylogénie est représentée graphiquement par un arbre, c'est-à-dire qu'elle est composée de nœuds internes, de nœuds externes (ou feuilles) et de branches. Les nœuds externes représentent les séquences étudiées, parfois appelées OTU (*operational taxonomic unit*), et les nœuds internes représentent des ancêtres communs hypothétiques, parfois appelées HTU (*hypothetical taxonomic unit*). Lorsque chaque nœud interne est adjacent à exactement trois branches, la phylogénie est dite binaire et les nœuds internes des bifurcations. La plupart des phylogénies sont binaires, mais cela n'est pas une généralité. Lorsque la phylogénie n'est pas binaire, les nœuds internes sont appelés des multifurcations. Par la suite, une phylogénie sera toujours considérée binaire.

Un groupe d'OTU est dit monophylétique si aucun autre OTU ne partage leur ancêtre commun, nous disons qu'ils forment un clade dans la phylogénie. Si des OTU additionnels sont inclus dans ce clade, ils sont paraphylétiques. En principe, une phylogénie est valuée, c'est-à-dire que chaque branche a une valeur qui représente le nombre de substitutions par site. Ainsi, en lisant une phylogénie il est possible de connaître la distance évolutive qui sépare deux OTU, elle correspond à la longueur du plus court chemin qui les sépare.

Une phylogénie peut être enracinée ou non (Figure 3). Contrairement à une phylogénie non enracinée, une phylogénie enracinée indique le sens du processus d'évolution, c'est-à-dire le sens de l'écoulement du temps. Plusieurs méthodes existent afin d'enraciner une phylogénie. La plus utilisée est sans doute l'ajout d'un ou de plusieurs OTU, appelés *outgroup*, qui sont connus pour être les OTU

distants et monophylétiques par rapport au group d'intérêt ou *ingroup*. Le nœud racine est alors placé sur la branche qui sépare l'*outgroup* de l'*ingroup*.

Figure 3. Différence entre phylogénie enracinée et non enracinée.
La figure A représente une phylogénie enracinée, tandis que la figure B une phylogénie non enracinée. Les deux phylogénies ont la même topologie, mais la phylogénie de la figure A est obtenue en ayant supposée que les OTU D et E réfèrent à un *outgroup*, le nœud racine a donc pu y être placé.



Dans le cas d'une phylogénie non enracinée, il y a exactement $2n - 3$ branches internes, où n est le nombre d'OTU, et $n - 2$ nœuds internes. Si elle est enracinée, il faut considérer une branche supplémentaire et un nœud supplémentaire.

Le nombre de topologies possibles croît exponentiellement en fonction du nombre d'OTU. Pour n OTU, il existe

$$B(n) = \frac{(2n-3)!}{2^{n-2}(n-2)!}$$

phylogénies enracinées. Le nombre de topologies non enracinées pour n OTU est égal au nombre de topologies enracinées pour $n - 1$ OTU. Avec 6 OTU nous avons donc 954 topologies possibles de phylogénies enracinées et avec 9 OTU les deux millions sont dépassés. Il devient donc vite impossible de parcourir l'ensemble des topologies au fur et à mesure que le nombre d'OTU considéré grandit. Des heuristiques sont proposées comme alternative à ce problème. Citons les algorithmes *nearest-neighbor interchange* (NNI), *subtree pruning and regrafting* (SPR) et *tree bisection and reconnection* (TBR) qui sont largement utilisés pour explorer un sous-espace de l'ensemble des arbres possibles.

1.5.2 Méthodes de distances

Les méthodes de distances essayent de faire correspondre un arbre (une phylogénie) à une matrice de distances. Les distances de la matrice sont généralement obtenues à partir d'un alignement et sont symétriques, c'est-à-dire que pour tout x et y, $d(x, y) = d(y, x)$. Par la suite nous noterons d_{xy} au lieu de $d(x, y)$. Les méthodes de distances peuvent être scindées en deux groupes : les méthodes agglomératives et les méthodes optimisant un critère.

1.5.2.1 Les méthodes agglomératives

Les méthodes agglomératives utilisent un algorithme pour construire pas à pas une phylogénie à partir d'une matrice de distances. Généralement, elles agglomèrent deux OTU x et y, répondant à un critère agglomératif précis, en un nouvel OTU u, puis calculent une nouvelle matrice de distances où les OTU x et y sont remplacés par l'OTU u. Le même procédé est de nouveau répété sur la dernière matrice de distances et ceci jusqu'à l'agglomération de tous les OTU. La phylogénie est ainsi calculée.

La méthode la plus connue et la plus ancienne est UPGMA (*unweighted-pair group method with arithmetic means*) (Sokal & Michener, 1958). Cette méthode est rarement utilisée car elle fait l'hypothèse d'une horloge moléculaire stricte (Zuckerkandl & Pauling, 1965, 1962). L'hypothèse de l'horloge moléculaire stricte stipule que l'évolution est un processus constant et uniforme. Faire cette hypothèse sur une phylogénie, c'est admettre que chaque feuille se situe à égale distance de la racine. C'est une hypothèse très forte qui nécessite souvent une justification lors de son utilisation. En revanche, si la matrice de distances satisfait la condition d'ultramétricité, alors l'algorithme UPGMA construit la phylogénie optimale et, de plus, elle est enracinée. Dans le cas contraire, cet algorithme n'est plus utilisé. Une matrice de distances est ultramétrique si pour tout triplet x, y et z, la condition d'ultramétricité (ou condition des trois points)

$$d_{xy} \leq \max\{d_{xz}, d_{zy}\}$$

est vérifiée. Cela signifie que deux des trois distances d_{xy}, d_{xz} et d_{zy} sont égales et maximales.

Pour surpasser l'hypothèse de l'horloge moléculaire stricte faite par l'algorithme UPGMA, une autre méthode agglomérative est suggérée, il s'agit de *neighbor-joining* (NJ) (Studier & Keppler, 1988; Saitou & Nei, 1987). NJ est l'une des méthodes de distances les plus utilisées et l'est encore aujourd'hui (Ye *et al*, 2011), même si de nombreuses variantes visant à améliorer cet algorithme sont proposées. Citons entre autre BIONJ (Gascuel, 1997), *generalized neighbor-joining* (Pearson *et al*, 1999), *weighted neighbor-joining* (Bruno *et al*, 2000), etc. Cette méthode construit une phylogénie non enracinée et lorsque la matrice de distances est additive, la phylogénie est optimale ou exacte. Une matrice de distances est additive si la condition des quatre points (Buneman, 1971) est satisfaite, c'est-à-dire que pour tout x, y, z et t, nous avons

$$d_{xy} + d_{zt} \leq \max\{d_{xz} + d_{yt}, d_{xt} + d_{yz}\}.$$

Cette condition implique que les deux plus grandes sommes sont égales. Seules les matrices additives peuvent aboutir à une phylogénie non enracinée, telle que la distance fournie en entrée entre deux OTU x et y est strictement égale à la somme des longueurs de branches sur le chemin reliant x et y

dans la phylogénie. La condition d'ultramétricité implique la condition des quatre points, donc si la matrice de distances est ultramétrique (elle est donc aussi additive) alors l'algorithme NJ construit la phylogénie optimale et la racine se trouve sur le point équidistant à chaque feuille.

1.5.2.2 Les méthodes optimisant un critère

Les méthodes qui optimisent un critère d'optimalité explorent l'espace des arbres à l'aide d'heuristiques, puis choisissent la meilleure phylogénie suivant le critère d'optimalité.

Pour les méthodes de distances, deux genres de critères d'optimalité sont utilisés. Le premier utilise l'approche standard des moindres carrés (Fitch & Margoliash, 1967). Il choisit la phylogénie qui minimise la somme des différences au carré entre la distance mesurée (celle de la matrice de distances) sur une paire d'OTU et la distance qui sépare ces deux OTU dans la phylogénie. La phylogénie résultante est celle qui contient les distances de chemin entre chaque paire d'OTU les plus proches possibles de celles contenues dans la matrice de distances initiale. Le programme FITCH (Felsenstein, 1989) utilise ce critère pour inférer une phylogénie. Le deuxième critère, bien différent du précédent, consiste à trouver l'arbre d'évolution minimum (*minimum evolution*, ME) (Kidd & Sgaramella-Zonta, 1971), c'est-à-dire celui qui minimise la somme des longueurs de branches, celles-ci étant estimées par moindres carrés à partir de la matrice de distances. Le logiciel FastME (Desper & Gascuel, 2002) utilise ce critère dans sa version « balancée » afin de proposer la meilleure phylogénie possible. Par la suite, il a été montré que l'algorithme NJ minimise également ce même critère d'évolution minimum balancé (Gascuel & Steel, 2006).

Diverses études ont montré que cette approche est remarquablement précise et rapide, avec des algorithmes en $O(n^3)$ ou moins pour construire un arbre initial et le modifier itérativement par mouvements NNI et SPR (Vinh & von Haeseler, 2005; Desper & Gascuel, 2004, 2002).

1.5.3 Méthodes de caractères

Les méthodes de caractères regroupent toutes celles qui se basent sur un alignement pour inférer une phylogénie. Comme les méthodes de distances, elles peuvent être scindées en deux catégories. La première regroupe les méthodes qui ne sont pas basées sur un modèle d'évolution explicite, comme la parcimonie. La seconde regroupe les méthodes basées sur un modèle d'évolution explicite. Ces dernières sont actuellement les plus employées par les biologistes, bien que lourdes en temps de calcul, en raison de leur fiabilité.

Les méthodes de parcimonie parcourent l'espace des phylogénies possibles et choisissent celles qui minimisent la quantité de changement évolutionnaire, c'est-à-dire celles qui expliquent l'alignement avec le moins de substitutions possibles (Fitch, 1971; Farris, 1970; Kluge & Farris, 1969).

En ce sens, elles rappellent le critère d'évolution minimum. Le point de vu philosophique derrière ce critère est que les hypothèses les plus simples sont souvent préférables aux plus compliquées. De ce fait, ces méthodes n'utilisent pas de modèle d'évolution à proprement parler. Aujourd'hui les méthodes de parcimonie sont de moins en moins utilisées pour inférer des phylogénies, mais elles restent utilisées pour inférer des annotations ancestrales à partir d'annotations contemporaines et d'une phylogénie précédemment calculée (cf. section 1.6). Dans ce cadre, où il n'existe souvent pas de modèles bien étudiés et incontestables, elles offrent une grande simplicité associée à des algorithmes rapides.

Les méthodes du maximum de vraisemblance (*maximum likelihood*) sont des méthodes probabilistes qui calculent une probabilité conditionnelle (la vraisemblance) exprimant le fait d'observer l'alignement suivant un modèle d'évolution particulier et une phylogénie particulière. Depuis l'introduction de ce principe en phylogénie en 1981 par Joseph FELSENSTEIN (Felsenstein, 1981), son utilisation est devenue de plus en plus populaire, en particulier grâce aux avancés algorithmiques et technologiques qui réduisent leur temps de calculs. Leur but est de choisir la phylogénie (un parcours, généralement heuristique, de l'espace des topologies est nécessaire) et les paramètres du modèle d'évolution (qui peuvent être estimés en même temps que la recherche de la topologie) qui maximisent la vraisemblance, contrairement aux méthodes de distances où c'est le plus souvent l'utilisateur qui choisit les valeurs des paramètres du modèle d'évolution. Outre le fait de se baser sur un modèle d'évolution, un avantage des méthodes de vraisemblance et de distances par rapport aux méthodes de parcimonie est de converger vers la vraie phylogénie au fur et à mesure que la quantité d'information en entrée augmente, et sous la condition que le modèle évolutif choisi soit le vrai modèle auquel obéissent les données (Felsenstein, 1978). Le logiciel PhyML (Guindon *et al.*, 2010; Guindon & Gascuel, 2003) utilise ce principe pour inférer des phylogénies. C'est l'un des logiciels les plus utilisés dans ce domaine et il le sera aussi dans nos études. D'autres logiciels sont aussi disponibles comme RAxML (Stamatakis, 2006) ou DNAML (Felsenstein, 1989).

Une dernière classe de méthodes de caractères existe. Il s'agit des méthodes bayésiennes, très proches des méthodes de vraisemblance. Ces méthodes sont fondées sur le théorème de BAYES (1702-1761) qui fut publié à titre posthume en 1763. Ce théorème combine la probabilité *a priori* d'un arbre $P(T)$ avec la vraisemblance $P(D|T)$ d'observer les données D (qui incluent l'alignement, les paramètres du modèle d'évolution et les longueurs de branches) sachant la topologie T pour en déduire la probabilité *a posteriori* de T, $P(T|D)$, par

$$P(T|D) = \frac{P(D|T)P(T)}{P(D)}.$$

La probabilité $P(D)$ est une constante de normalisation définie comme la somme des numérateurs pour toutes les topologies possibles : $P(D) = \sum_T P(D|T)P(T)$. L'objectif est alors de maximiser la probabilité *a posteriori* $P(T|D)$. En pratique, le calcul de la probabilité $P(D)$ est trop complexe et on utilise des algorithmes de type MCMC (*Markov Chain Monte Carlo*), et la variante Metropolis-Hastings (Metropolis *et al*, 1953; Hastings, 1970), qui permettent de s'affranchir du calcul de $P(D)$. Ces algorithmes effectuent un parcours au hasard de l'espace des arbres, mais à l'aide d'une marche guidée, jusqu'à un état d'équilibre. Puis un consensus des arbres obtenus est calculé après la suppression des premiers arbres calculés avant l'état d'équilibre (*burn-in*). Plus le nombre d'arbres parcourus est grand, meilleure est l'approximation. Généralement, cet algorithme nécessite au minimum un million de générations (nombre d'arbres parcourus) et reste réservé pour des phylogénies ne dépassant pas quelques centaines d'OTU. La différence fondamentale entre les méthodes bayésiennes par rapport aux méthodes de vraisemblance est que les paramètres suivent une distribution donnée *a priori* par l'utilisateur. Ceci en fait une méthode assez controversée puisque les résultats sont modifiables suivant les choix effectués, sans que ces choix puissent être guidés par des principes rigoureux. Le logiciel MrBayes (Ronquist & Huelsenbeck, 2003), utilisé dans nos études, lorsque cela est possible, permet d'inférer des arbres phylogénétiques sous ce principe. C'est aussi un des programmes de phylogénie les plus utilisés à l'heure actuelle.

1.5.4 Fiabilité des phylogénies

Différentes méthodes statistiques permettent de tester la fiabilité des arbres phylogénétiques, en particulier celle des branches internes. Ces techniques sont systématiquement utilisées lors de l'inférence d'arbres phylogénétiques. La plus répandue est celle du *bootstrap* (Felsenstein, 1985). Cette technique utilise le ré-échantillonnage aléatoire pour générer un grand nombre (souvent 1 000) d'alignements bruités. Ces derniers sont construits sur la base de l'alignement de départ. Un alignement bruité est constitué d'une succession de sites (autant que pour l'alignement de base) choisis aléatoirement (avec remise) parmi ceux de l'alignement de départ. Ainsi, certains sites de l'alignement initial peuvent apparaître plusieurs fois, tandis que d'autres jamais. Des phylogénies sont ensuite calculées, sur la base de ces alignements bruités, avec la même méthode et les mêmes paramètres que ceux utilisés pour calculer la phylogénie initiale. Le support statistique attribué à chaque clade de la phylogénie de départ correspond au nombre de fois où ce clade est trouvé dans les réplicas bruités. Plus le signal phylogénétique est fort, plus le support *bootstrap* est élevé. Cette méthode est l'une des plus employée et il est communément admis qu'un support de *bootstrap* supérieur à 80% est statistiquement fiable. Cependant, elle a un inconvénient majeur. Si le temps de calcul nécessaire pour inférer la phylogénie initiale est de x unités de temps, alors $1\,000 \times x$ unités de temps sont nécessaires pour estimer les supports de branches. Par exemple, si l'inférence d'une

phylogénie dure 20 minutes, il faut $20 \times 1\,000$ minutes (soit presque 14 jours) pour calculer les supports de branches associés. En pratique, les supports sont souvent calculés avec 100 réplicas (ce qui revient à un peu plus de un jour de calcul avec l'exemple précédent).

Pour augmenter la vitesse de calcul des méthodes probabilistes, d'autres tests statistiques sont proposés. Les méthodes de vraisemblance utilisent le test *approximate likelihood-ratio test* (aLRT) (Anisimova & Gascuel, 2006) qui estime les supports de chaque branche à l'aide de la seconde meilleure phylogénie parmi les deux différentes phylogénies obtenues par permutation des quatre branches adjacentes à la branche d'intérêt (deux mouvements NNI). En général, les supports sont jugés significatifs à partir de 80-90%. Quant aux méthodes bayésiennes, elles utilisent les arbres générés par la méthode MCMC pour en déduire les supports de branches ou probabilités postérieures. Toutefois, cette dernière méthode a tendance à surestimer les vrais supports de branches (Douady *et al*, 2003), et c'est pour cela que seules des probabilités postérieures supérieures à 95% ou proches de 100% sont jugées statistiquement fiables.

1.6 Reconstruire l'évolution de caractères

Outre le fait d'informer sur les relations de parenté, les phylogénies trouvent de nombreuses autres applications. Certaines sont exposées dans l'introduction de ce chapitre, d'autres dans les chapitres suivants de ce mémoire. Dans cette section, nous introduisons quelques concepts pour reconstruire l'évolution de caractères à partir d'une phylogénie.

Ici, un caractère est un ensemble d'annotations (ou états) qui sont capables d'évoluer de l'une vers l'autre (Maddison & Maddison, 2003). Par exemple, l'annotation « yeux bleus » est cohérente avec les annotations « yeux verts » et « yeux marrons » et il est supposé que l'on peut passer de l'une vers l'autre et vice-versa. Ce caractère, « couleur des yeux », est un caractère discret car la transition d'une annotation à une autre s'opère en une fois et (généralement) non graduellement. Toutefois, les caractères évoluant continuellement (comme la taille ou le poids) peuvent aussi être vus comme des caractères discrets en considérant un intervalle de valeur comme une annotation (Maddison & Maddison, 2003). Les zones géographiques et les pays sont des caractères discrets couramment utilisés, qui feront l'objet d'études dans cette thèse, pour déterminer les flux épidémiques du sous-type C du VIH-1 à la surface du globe.

La reconstruction d'annotations ancestrales cherche à déterminer quelles sont les annotations des HTU à partir : (1) des annotations associées aux OTU et (2) d'une phylogénie enracinée. La phylogénie doit nécessairement être enracinée afin d'orienter le cours de l'évolution. Dans le cas discret, les annotations ancestrales ne peuvent prendre que des valeurs parmi les annotations assignées aux

OTU. Il est donc nécessaire de considérer des OTU pertinents. Ainsi, nous pouvons déjà énoncer trois hypothèses fondamentales sur lesquelles se basent toutes les méthodes de reconstruction de caractères (Omland, 1999) :

- la phylogénie utilisée est la « vraie » phylogénie ;
- tous les OTU pertinents sont dans la phylogénie ;
- les états sont correctement assignés aux OTU, et cela sans erreur possible.

Même si ces trois hypothèses sont parfaitement respectées, les méthodes de reconstruction de caractères ancestraux ne garantissent pas la fiabilité des résultats. Elles suivent des principes divers que nous décrivons maintenant.

Dans le cas de caractères discrets, les méthodes de reconstruction de caractères ancestraux généralement utilisées sont basées sur le principe de parcimonie qui choisit la reconstruction impliquant le moins de changements de caractère le long de l'arbre phylogénétique. Des méthodes plus élaborées existent, certaines sont basées sur le principe du maximum de vraisemblance (Schluter *et al*, 1997) et d'autres sur des approches bayésiennes (Schultz & Churchill, 1999), mais elles nécessitent un modèle probabiliste de transition entre annotations afin d'inférer les annotations ancestrales. L'élaboration d'un tel modèle n'est pas chose aisée et les erreurs de jugement peuvent produire des résultats biaisés. Certaines méthodes récemment développées peuvent auto-estimer les paramètres du modèle probabiliste mais le temps de calcul en est considérablement rallongé (Lemey *et al*, 2009a). Même si les méthodes de parcimonie n'utilisent pas de modèle probabiliste, il est toutefois possible d'attribuer des contraintes ou des poids sur les transitions afin d'en privilégier certaines par rapport à d'autres. On distingue ainsi plusieurs principes de parcimonie (Maddison & Maddison, 2003). Dans le cas de caractères continus, les méthodes de parcimonie sont rarement utilisées car elles n'emploient pas l'information contenue dans les longueurs de branches, information généralement nécessaire pour reconstruire correctement les caractères ancestraux continus. Par la suite, nous nous focaliserons sur les méthodes de parcimonie et leur utilisation sur des caractères discrets, avec en ligne de mire les annotations géographiques qui seront étudiées au Chapitre 6 sur l'épidémie du VIH-1 sous-type C à l'échelle mondiale.

Les méthodes de parcimonie font trois hypothèses supplémentaires à celles émises précédemment (Omland, 1999) :

- les transformations sont identiquement probables quelle que soit la branche, c'est-à-dire que les longueurs de branches ne sont pas prises en compte ;

- les taux d'évolution d'un état vers un autre doivent être relativement lents, et on suppose qu'au plus une transition a lieu sur chaque branche de l'arbre ;
- les coûts de transformations sont symétriques, c'est-à-dire que la probabilité de gagner ou de perdre un état est identique.

Il existe plusieurs méthodes de parcimonie permettant de reconstruire les annotations ancestrales. La méthode de parcimonie la plus utilisée est la parcimonie non ordonnée (Fitch, 1971; Hartigan, 1973), c'est-à-dire que la transition d'un état vers n'importe quel autre état a le même coût. On parle de parcimonie de Fitch. Pour la calculer et déterminer les états ancestraux on utilise deux étapes successives. Une première étape, appelée UPPASS, parcourt l'arbre des feuilles jusqu'à la racine en assignant à chaque nœud ancestral l'information relative aux seuls nœuds fils. La Figure 4 décrit les étapes de l'algorithme UPPASS. À la fin de cette étape, les états assignés aux nœuds ancestraux ne sont pas forcément les plus parcimonieux, puisqu'ils sont calculés uniquement avec l'information des nœuds inclus dans le clade sous-jacent et ne considère donc pas l'information de toute la phylogénie. Ils sont simplement les plus parcimonieux par rapport au clade sous-jacent. Ainsi, le seul nœud qui y fait exception est le nœud racine, puisque l'information de toute la phylogénie sert à le calculer.

Figure 4. Algorithme UPPASS.
L'algorithme UPPASS est un algorithme récursif de type « *postorder* », dans lequel le calcul à proprement parler se fait après les appels récursifs. Il utilise un paramètre qui est un nœud (initialisé avec la racine de la phylogénie) et calcule les états associées aux nœuds ancestraux. N est le nœud courant, G et D ses fils gauche et droit et P le nœud père. $S(X)$ est l'ensemble des états associés au nœud X.

Entrée : N un nœud
1. **si** N est une feuille **alors**
2. $S(N) =$ état associé à N
3. **sinon**
4. UPPASS(G)
5. UPPASS(D)
6. **si** $S(G) \cap S(D) = \emptyset$ **alors**
7. $S(N) = S(G) \cup S(D)$
9. **sinon**
10. $S(N) = S(G) \cap S(D)$
11. **fin si**
12. **fin si**

Une deuxième étape, appelée DOWNPASS (Maddison & Maddison, 2003), parcourt l'arbre de la racine aux feuilles en considérant pour chaque nœud l'information des nœuds adjacents. Comme la racine contient déjà la valeur la plus parcimonieuse, cette deuxième phase commence avec les nœuds fils du nœud racine. La Figure 5 décrit les étapes de l'algorithme DOWNPASS. Après

l'application de cet algorithme, chaque nœud interne contient les valeurs finales issues de l'information de tous les nœuds et feuilles de la phylogénie. Cette information nous dit essentiellement quels sont les états du nœud qui appartiennent à au moins un scénario optimal de parcimonie. Il faut noter que toutes les combinaisons d'annotations ancestrales ainsi obtenues ne correspondent pas à un tel scénario optimal, loin de là. Nous y reviendrons plus loin.

Figure 5. Algorithme DOWNPASS.
L'algorithme DOWNPASS vient en complément de l'algorithme UPPASS pour intégrer l'information de toute la phylogénie sur chacun des nœuds internes. N est le nœud courant, G et D ses fils gauche et droit et P le nœud père. $S(X)$ est l'ensemble des états associés au nœud X. C'est un algorithme récursif de type « *preorder* », dans lequel le calcul à proprement parler se fait avant les appels récursifs.

Entrée : N un nœud
1. **si** N n'est pas une feuille **alors**
2. **si** N n'est pas la racine **alors**
3. $V = S(P) \cap S(G) \bigcap \mathrm{S(D)}$
4. **si** $V = \emptyset$ **alors**
5. $V = (S(P) \cap S(G)) \cup (S(P) \cap S(D)) \cup (S(G) \cap S(D))$
6. **fin si**
7. **si** $V = \emptyset$ **alors**
8. $V = S(P) \cup S(G) \cup S(D)$
9. **fin si**
10. $S(N) = V$
11. **fin si**
12. DOWNPASS(G)
13. DOWNPASS(D)
14. **fin si**

Après l'utilisation de l'algorithme DOWNPASS, des ambiguïtés peuvent se produire au niveau des nœuds internes, c'est-à-dire qu'un nœud interne peut être associé à plus d'une annotation, signifiant que la parcimonie hésite entre plusieurs solutions. Ces ambiguïtés peuvent être gênantes lors de l'estimation du nombre de transitions (nombre de branches ayant des annotations différentes à ses extrémités), pratique courante dans ce genre d'étude. Il existe plusieurs possibilités afin de diminuer ces ambiguïtés. Les deux plus connues sont les algorithmes ACCTRAN (*accelerated transformation*) (Farris, 1970) et DELTRAN (*delayed transformation*) (Swofford & Maddison, 1987). Ces deux algorithmes font des hypothèses différentes en ce qui concerne le choix final des états de chaque nœud interne. La méthode ACCTRAN force les changements d'états à se produire le plus près possible de la racine et donc à favoriser les transformations reverses, tandis que la méthode DELTRAN force les changements d'états à se produire le plus près possible des feuilles et donc à favoriser les transformations parallèles. La Figure 6 décrit l'algorithme ACCTRAN, proposé conjointement par Farris et Fitch, et la Figure 7 l'algorithme DELTRAN. Il faut bien noter que l'algorithme ACCTRAN remplace

l'algorithme DOWNPASS, alors que l'algorithme DELTRAN vient en complément de l'algorithme DOWNPASS, après son exécution.

Figure 6. Algorithme ACCTRAN.
L'algorithme ACCTRAN remplace l'algorithme DOWNPASS afin de diminuer les ambiguïtés de valeurs au niveau des nœuds internes de la phylogénie. L'annotation d'un nœud privilégie les informations venant des fils plutôt que du père, et va « pousser » les changements vers la racine. N est le nœud courant, G et D ses fils gauche et droit. $S(X)$ est l'ensemble des états associés au nœud X.

Entrée : N un nœud
1. **si** N n'est pas une feuille **alors**
2. **si** $S(N) \cap S(G) \neq \emptyset$ **alors**
3. $S(G) = S(N) \cap S(G)$
4. **sinon**
5. $S(G)$ est inchangé et contient l'information issue de ses fils après UPPASS
6. **fin si**
7. **si** $S(N) \cap S(D) \neq \emptyset$ **alors**
8. $S(D) = S(N) \cap S(D)$
9. **sinon**
10. S(D) est inchangé et contient l'information issue de ses fils après UPPASS
11. **fin si**
12. ACCTRAN(G)
13. ACCTRAN(D)
14. **fin si**

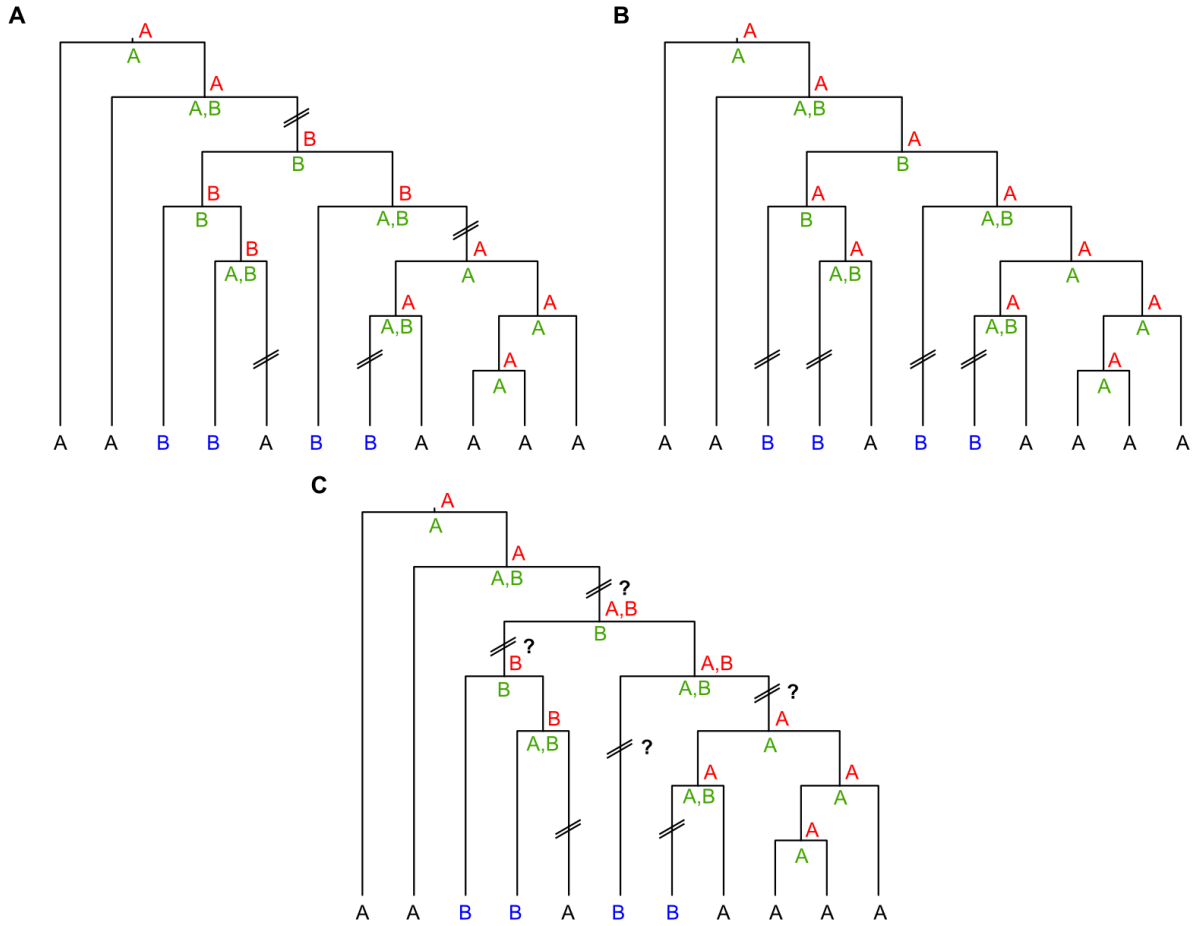
Figure 7. Algorithme DELTRAN.
L'algorithme DELTRAN vient en supplément de l'algorithme DOWNPASS après son exécution afin de diminuer les ambiguïtés de valeurs au niveau des nœuds internes de la phylogénie. L'annotation d'un nœud privilégie ainsi les informations venant du père, et « pousse » les changements vers les feuilles. N est le nœud courant, G et D ses fils gauche et droit et P le nœud père. $S(X)$ est l'ensemble des états associés au nœud X. Comme on se place après DOWNPASS, $S(N)$ contient déjà tous les états les plus parcimonieux, contrairement à ACCTRAN où $S(G)$ et $S(D)$ ne contiennent que les informations issues des clades de racine G et D.

Entrée : N un nœud
1. **si** N n'est pas une feuille **et** N n'est pas la racine **alors**
2. **si** $S(P) \cap S(N) \neq \emptyset$ **alors**
3. $S(N) = S(P) \cap S(N)$
4. **fin si**
5. DELTRAN(G)
6. DELTRAN(D)
7. **fin si**

Les algorithmes ACCTRAN et DELTRAN ne résolvent pas forcément toutes les ambiguïtés des nœuds internes, c'est-à-dire qu'après l'application de l'un ou l'autre, des nœuds internes peuvent encore être ambigus, en particulier lorsque le nœud racine l'est. Dans ce cas, lors de l'estimation du nombre de transitions, certains auteurs ne considèrent pas ces nœuds (Nakano *et al*, 2004) ou alors ils calculent le nombre moyen de transitions parmi toutes les reconstructions les plus parcimonieuses

possibles (Salemi *et al*, 2008). La Figure 8 montre l'application des algorithmes ACCTRAN, DELTRAN et DOWNPASS sur une phylogénie exemple.

Figure 8. Exemple d'application des algorithmes ACCTRAN, DELTRAN et DOWNPASS.
La figure A montre les résultats de l'application d'ACCTRAN, la figure B ceux de DELTRAN et la figure C ceux de DOWNPASS sur une même phylogénie. Les résultats de l'algorithme UPPASS sont indiqués en vert, ceux d'ACCTRAN, DELTRAN ou DOWNPASS en rouge. Les barres obliques indiquent les branches où des transformations ont lieux. Les reconstructions coûtent chacune quatre transitions, mais la reconstruction avec DOWNPASS hésite entre deux scénarios possibles.



Le plus utilisé de ces deux algorithmes semble être ACCTRAN ou, si DELTRAN est choisi, les résultats avec ACCTRAN sont souvent présentés en complément (Agnarsson & Miller, 2008), l'idée étant que ces deux algorithmes constituent deux extrêmes et que la « vérité se situe entre les deux » (cf. ci-après). Cela est du à un commentaire de De Pinna (1991) qui argumente sur le fait que les transformations reverses sont préférables aux transformations parallèles. Mais aucune preuve formelle ne démontre qu'ACCTRAN serait mieux que DELTRAN ou vice-versa. L'utilisation de l'un ou l'autre dépend en réalité largement du caractère étudié. En effet, lorsque l'on considère des caractères morphologiques, on favorise les séquences acquisitions – perte (par exemple d'ailes fonctionnelles) plutôt que l'invention multiple de caractères. En ce sens, la parcimonie ACCTRAN est préférable puisqu'elle force les changements à se produire le plus près possible de la racine et donc défavorise les mutations parallèles par rapport aux évènements reverses. Mais lorsque des caractères épidémiolo-

giques sont considérés, comme par exemple des lieux géographiques, il est plus facile de penser qu'une épidémie s'intensifie dans un lieu donné avant de se diffuser à partir de celui-ci, avec des transmissions multiples. Dans ce cas, l'algorithme DELTRAN est le plus approprié puisqu'il force les changements (de lieux) à se produire le plus près possible des feuilles.

En considérant l'espace de toutes les reconstructions les plus parcimonieuses possibles (*most parsimonious reconstruction*, MPR), c'est-à-dire celles qui minimisent le nombre de changement d'états, ainsi qu'une relation d'ordre sur cet espace, Minaka (1993) a montré que les algorithmes ACCTRAN et DELTRAN sont les deux bornes de cet espace. Ainsi, si les résultats d'ACCTRAN et de DELTRAN sont identiques, il en est de même pour toutes les autres MPR.

Plusieurs méthodes statistiques existent afin de tester la fiabilité des reconstructions de caractères ancestraux par parcimonie. Elles sont toutes basées sur des méthodes de Monte Carlo et comparent la quantité de transitions observées à celle de l'hypothèse nulle ou panmixie, dans laquelle il n'y aurait aucune corrélation entre la phylogénie et les annotations étudiées. Nous présentons ici la méthode de ré-échantillonnage aléatoire, ou *shuffling*, qui mélange les annotations des OTU et estime de nouveau, sous les mêmes conditions, la quantité de transitions. Typiquement ce procédé est répété un grand nombre de fois (1 000 ou 10 000). La quantité de transitions observées est alors comparée à la distribution des quantités obtenues aléatoirement afin de définir sa significativité statistique. Slatkin et Maddison (1989) semblent être les premiers à avoir utilisé cette procédure qui est maintenant un standard dans le domaine. Cette méthode de ré-échantillonnage aléatoire sera utilisée dans notre étude sur l'épidémie mondiale du VIH-1 sous-type C (Chapitre 6) pour établir les significativités statistiques de différents critères. Elle donne une vision plus complète et interprétable que l'approche consistant à soustraire au nombre de transitions observées le nombre de transitions attendues par hasard dans le modèle nul de panmixie (Nakano *et al*, 2004).