Bioinformatique M1: Lecture 7

P. Derreumaux

PHYLOGENIE

Phylogénie Moléculaire

= branche de la systématique : elle consiste à déterminer l'arbre phylogénétique d'un ensemble de séquences homologues données (ou plus généralement, d'OTU: operationnal taxonomic units, cad la configuration la plus probable pour rendre compte du degré de parenté existant entre ces séquences

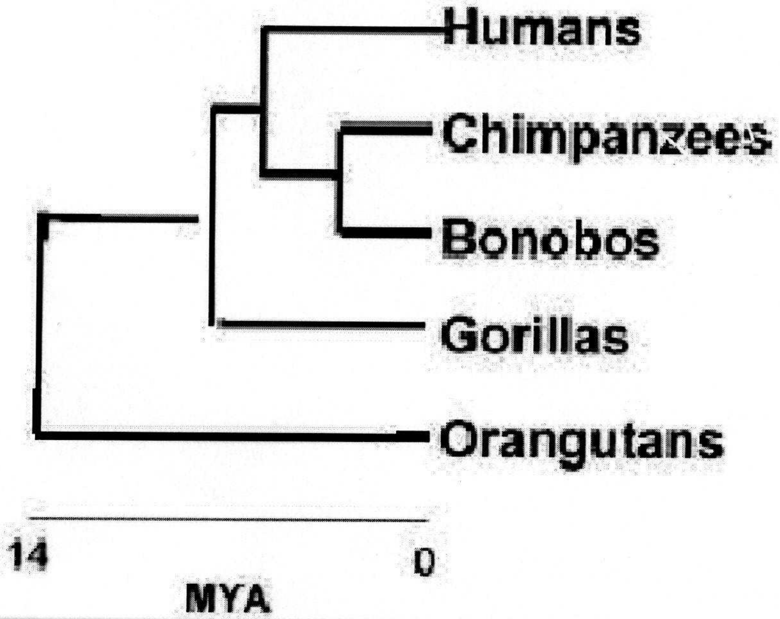
Objectifs

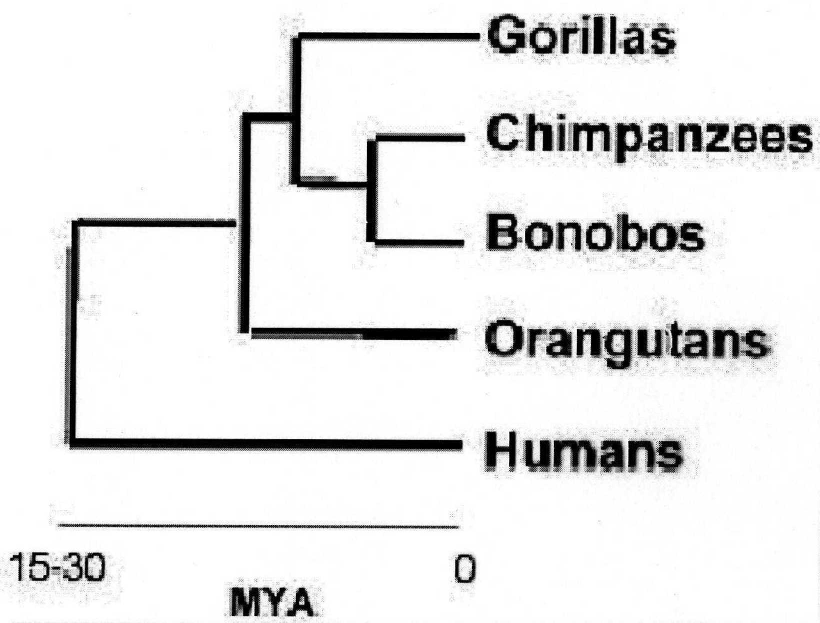
- Mieux comprendre les mécanismes de l'évolution et les mécanismes moléculaires associés.
- Connaître l'arbre de la vie (taxonomie)
- Etudier la biodiversité

Phylogeny Applications

- Tree of life: Analyzing changes that have occurred in evolution of different organisms
- Phylogenetic relationships among genes can help predict which ones might have similar functions (e.g., ortholog detection)
- Follow changes occuring in rapidly changing species (e.g., HIV virus)
- Protein domain databases

QUELLES SONT LES ESPECES PROCHES DE L'HOMME ?





Les données moléculaires comme les gènes de la mitochondrie, la plupart des gènes nucléaires et les hybridations ADN/ADN indiquent un clade : homme, chipanzé, bonono.

Divergence Homme-Singe : 5 Millions

Les données de la paléontologie suggéraient un clade Gorille, Chimpanzé, Bonobo, Orangutan.

La divergence avec l'homme était estimée à environ 12 Millions années.

continental flux isotopic composition to about 0.71049 [similar to the value proposed in (1)]. Also, the additional global continental Sr flux from groundwater would cause a rise in $^{87}Sr/^{86}Sr$ of 0.0095 over 40 My if left unbalanced. This is higher by a factor of 7 than the observed rise over the past 40 My.

Thus, we conclude that the groundwater data have an enormous effect on the interpretation of the seawater Sr isotope balance. Although we do not claim that the new values presented in Table 2 should be considered as final, these data urge caution about overinterpreting Sr isotope data from a few local watersheds in this area. For example, trying to use the seawater Sr isotope curve to infer the detailed tectonic uplift history of the Himalayas as well as for estimating effects on global climate change still involves considerable uncertainty. Because of the highly variable nature of $^{87}Sr/^{86}Sr$ in the G-B river system, reliable average values are difficult to estimate.

References and Notes

1. S. J. Goldstein, S. B. Jacobsen, *Chem. Geol. Isotope Geosci. Sect.* **66**, 245 (1987).
2. S. B. Jacobsen, *Earth Planet. Sci. Lett.* **90**, 315 (1988).
3. J. M. Edmond, *Science* **258**, 1594 (1992).
4. F. M. Richter, D. B. Rowley, D. J. DePaolo, *Earth Planet. Sci. Lett.* **109**, 11 (1992).
5. S. B. Jacobsen, A. J. Kaufman, *Chem. Geol.* **161**, 37 (1999).
6. S. Krishnaswami, J. R. Trivedi, M. M. Sarin, R. Ramesh, K. K. Sharma, *Earth Planet. Sci. Lett.* 109, 243 (1992).
7. M. R. Palmer, J. M. Edmond, *Geochim. Cosmochim. Acta* **56**, 2099 (1992).
8. M. E. Raymo, W. F. Ruddiman, *Nature* **359**, 117 (1992).
9. A. Galy, C. France-Lanord, L. A. Derry, *Geochim. Cosmochim. Acta* **63**, 1905 (1999).
10. M. Sharma, G. J. Wasserburg, A. W. Hofmann, G. J. Chakrapani, *Geochim. Cosmochim. Acta* **63**, 4005 (1999).
11. L. A. Derry, C. France-Lanord, *Earth Planet. Sci. Lett.* **142**, 59 (1996).
12. N. B. W. Harris, *Geology* **23**, 721 (1995).
13. J. D. Blum, C. A. Gazis, A. D. Jacobsen, C. P. Chamberlain, *Geology* **26**, 411 (1998).
14. J. Quade, L. Roe, P. G. DeCelles, T. P. Ojha, *Science* **276**, 1828 (1997).
15. W. S. Moore, *Nature* **380**, 612 (1996).
16. J. Carroll, K. K. Falkner, E. T. Brown, W. S. Moore, *Geochim. Cosmochim. Acta* **57**, 2981 (1993).

horizontal hydraulic gradient was found to be very low at the sample collection sites (Fig. 1). The re-

8 March 2001; accepted 17 July 2001

Genetic Evidence for Two Species of Elephant in Africa

Alfred L. Roca,[1] Nicholas Georgiadis,[2] Jill Pecon-Slattery,[1] Stephen J. O'Brien[1]*

Elephants from the tropical forests of Africa are morphologically distinct from savannah or bush elephants. Dart-biopsy samples from 195 free-ranging African elephants in 21 populations were examined for DNA sequence variation in four nuclear genes (1732 base pairs). Phylogenetic distinctions between African forest elephant and savannah elephant populations corresponded to 58% of the difference in the same genes between elephant genera *Loxodonta* (African) and *Elephas* (Asian). Large genetic distance, multiple genetically fixed nucleotide site differences, morphological and habitat distinctions, and extremely limited hybridization of gene flow between forest and savannah elephants support the recognition and conservation management of two African species: *Loxodonta africana* and *Loxodonta cyclotis*.

Conservation strategies for African elephants have consistently been based on the consensus that all belong to the single species *Loxodonta africana* (*1–3*). Yet relative to African savannah elephants, the elephants in Africa's tropical forests are smaller, with straighter and thinner tusks, rounded ears, and distinct skull morphology (*2–11*). Although forest elephants are sometimes assigned subspecific status and designated *L. a. cyclotis*, their degree of distinctiveness and of hybridization with savannah elephants has been controversial and often ignored (*2–12*). Recently, a comprehensive morphological comparison of metric skull measurement from 295 elephants (*10*, *11*) and a provocative molecular report limited to a single individual (*13*) noted appreciable distinctions between forest and savannah specimens.

Here we report the patterns and extent of sequence divergence for 1732 nucleotides from four nuclear genes (*14*) among 195 African elephants collected across their range in Africa and from seven Asian elephants (*Elephas maximus*). African elephants were sampled, with biopsy darts (*15*, *16*), throughout the continent, including individuals from 21 populations in 11 of 37 African elephant range nations (Fig. 1). Based on morphology (*2–11*) and habitat (*17*, *18*), three populations were categorized as African forest elephants, whereas 15 populations in southern, eastern, and north-central Africa were categorized as savannah elephants (Fig. 1). DNA sequences from four nuclear genes, including short exon segments (used to establish homology to mammalian genes) and longer introns (which would evolve rapidly enough to be phylogenetically informative), were determined for all elephants (*19*). The genes include *BGN* [646 base pairs (bp)], *CHRNA1* (655 bp),

[1] Laboratory of Genomic Diversity, National Cancer Institute, Frederick, MD 21702, USA. [2] Mpala Research Center, Post Office Box 555, Nanyuki, Kenya.

*To whom correspondence should be addressed. E-mail: obrien@ncifcrf.gov

GBA (100 bp), and *VIM* (331 bp), with sequence from all four genes obtained for 119 individuals. An alignment of variable sites and the composite genotypes are presented in supplemental information (*20*). Among 1732 bp, 73 sites were variable and 52 were phylogenetically informative. These nucleotide variants defined nine unique savannah genotypes among 58 individuals and 24 unique forest genotypes among 24 individuals. We observed nine genetically fixed nucleotide site differences between Asian and African elephants (*BGN* 121, 155, 219, and 513 and *CHRNA1* 011, 079, 274, 301, and 548) and one that approaches fixation (*BGN* 505). There were five fixed site differences between African savannah and forest elephants (*BGN* 304, 485, 508, 514, and 569) and two that were nearly fixed (*CHRNA1* 251 and *GBA* 20) (*20*).

Three methods of phylogenetic analysis (minimum evolution, maximum parsimony, and maximum likelihood) (*21–23*) revealed a concordant deep genetic division between the forest and savannah populations of African elephants (Fig. 2). The forest elephants of Dzanga-Sangha, Lope, and Odzala grouped together, separate from 15 savannah populations, which formed a distinct phylogenetic clade or lineage. An estimated 94% of the observed genetic variation ($F_{ST} = 0.94$, $P < 10^{-5}$) (*24*, *25*) was due to differences between forest and savannah elephants and 6% to intragroup differences. Mantel tests (*26*) revealed only marginal association of genetic versus geographic distance ($r = 0.19$, $P = 0.03$), and that association was attributed completely to forest versus savannah population differences ($P > 0.05$ for forest or savannah populations tested separately).

based on sequences

Although forest and savannah elephants formed two genetically distinct groups, sequences from populations within the two categories could not be distinguished hierarchical analysis of molecular variance (AMOVA) (*24*, *25*). For example, we could not genetically differentiate the forest elephants in Dzanga-Sangha from those of Lope (F_{ST} $P > 0.05$). Despite the extensive geographic distances separating them, the savannah populations in southern, eastern, and north-central Africa were genetically indistinguishable (F_{ST} $P > 0.05$). Forest elephants are genetically more diverse than savannah elephants (Fig. 2). The average number of within-group pairwise differences among 24 forest elephants was 1.74 as compared with a value of 0.06 among 58 savannah elephants (*24*, *25*, *27*). Each forest elephant had a unique composite genotype, whereas the 58 savannah elephants defined only nine distinct genotypes (*20*). Forest elephants displayed larger numbers of heterozygous nucleotide sites than did savannah elephants (an average of 3.54 heterozygous autosomal sites per individual in forest elephants versus 0.39 for savannah elephants) (*20*). These observations suggest a recent founder event in the history of the savannah metapopulation. A potential time venue for the bottleneck is indicated by fossil evidence, which suggests that the savannah elephant's range greatly expanded at the end of the Pleistocene, after *Elephas iolensis*, the predominant African species, became extinct (*3*, *12*).

The genetic and phylogenetic distinctiveness was evident without exception between 36 sampled forest elephants from three populations and 121 savannah elephants collected in 15 populations throughout sub-Saharan Africa. Each savannah population was genetically closer to every other savannah population than to any of the forest populations, even in cases where the forest population was geographically closer. Individuals from two "indeterminate" populations [Mount Kenya and Aberdares (Fig. 1)] contained exclusively savannah elephant genotypes (see Fig. 2, $F_{ST} = 0.88$, $P < 10^{-5}$ in comparing both populations to three forest populations). Genotypes found in the third "indeterminate" population, Garamba, were diverse and predominantly nested within the forest elephant clade in the phylogenetic analyses. The forest populations (including Garamba) were genetically closer to each other than to any savannah populations, including several that

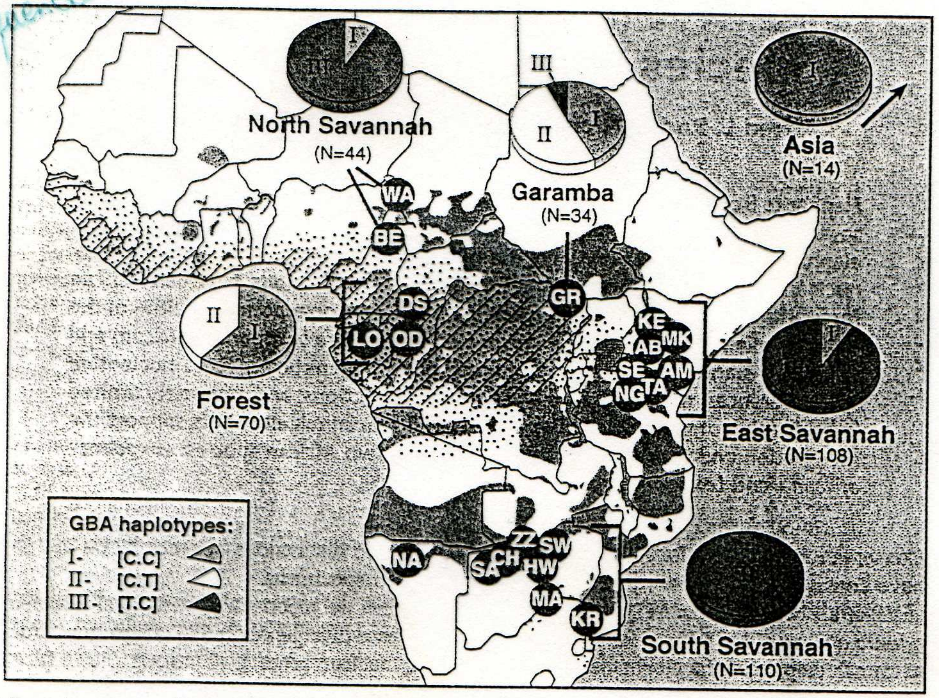


Fig. 1. Locations of sampled African elephant populations. Circles indicate sampling locations and population abbreviations. Green circles are forest populations

million years ago (*12*), the results suggest that forest and savannah elephants diverged approximately 2.63 (±0.94) million years ago (*24, 27, 29*), which is comparable to species-level distinction in other mammalian taxa, including elephants (*12, 30, 31*). This estimate should be considered as a maximum age, however, because the more recent genetic homogenization of the savannah elephants would inflate genetic distance as a consequence of a recent founder event.

Genetic distinctiveness between forest and savannah elephants is also apparent when individual gene variation is examined. For *GBA*, two variable sites in African elephants define three haplotypes ([C.C], [C.T], or [T.C] for nucleotide sites 20 and 79, respectively) that have large forest versus savannah frequency differences (Fig. 1, exact test $P < 10^{-5}$ for forest versus savannah). The predominant haplotype in savannah elephants is [T.C] (frequency = 0.96), whereas alternative [C.C] and [C.T] haplotypes comprise 100% of the forest elephants, suggesting that reproductive isolation exists between the two groups (Fig. 1). For *VIM* and *CHRNA1*, complete and exact haplotypes could not be determined for individuals heterozygous at two or more nucleotide sites, because gametic phase cannot be assessed (for example, for a two-locus genotype, does a double heterozygote G/C,T/A individual contain GT + CA or GA + CT haplotypes?). However, among forest and Garamba elephants, polymorphisms occurred at six nucleotide sites in *VIM* that were genetically monomorphic in savannah elephants (*20*). Similar differences in the occurrence of polymorphic nucleotide sites were apparent within *CHRNA1*: All sites that were variable among forest and Garamba elephants were fixed in savannah populations, whereas the two sites that were variable in savannah elephants were fixed in forest and Garamba elephants (*20*). Likewise, both *CHRNA1* and *VIM* had an insertion/deletion variant limited to forest and Garamba elephants (*20*). The presence of these deletion variants in Dzanga-Sangha, Lope, and Garamba also is consistent with the recent occurrence of gene flow among these forest elephant populations

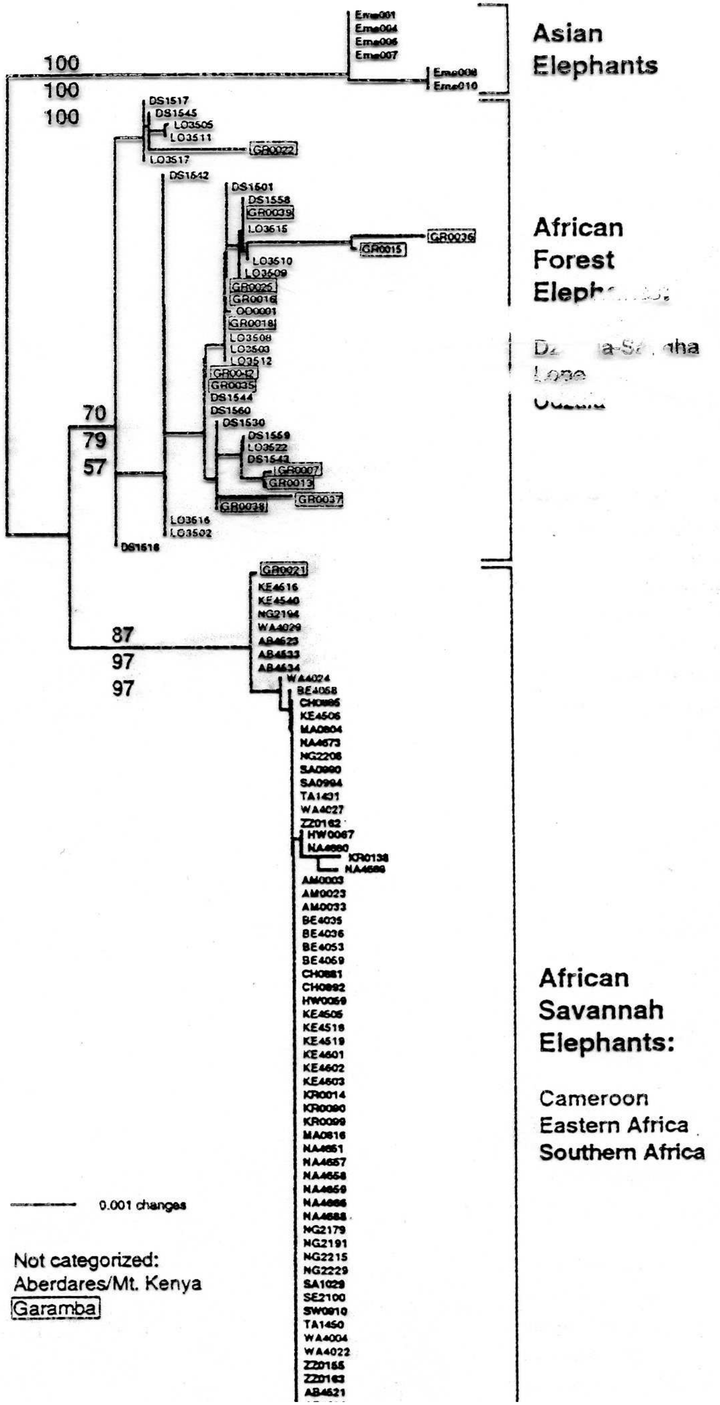
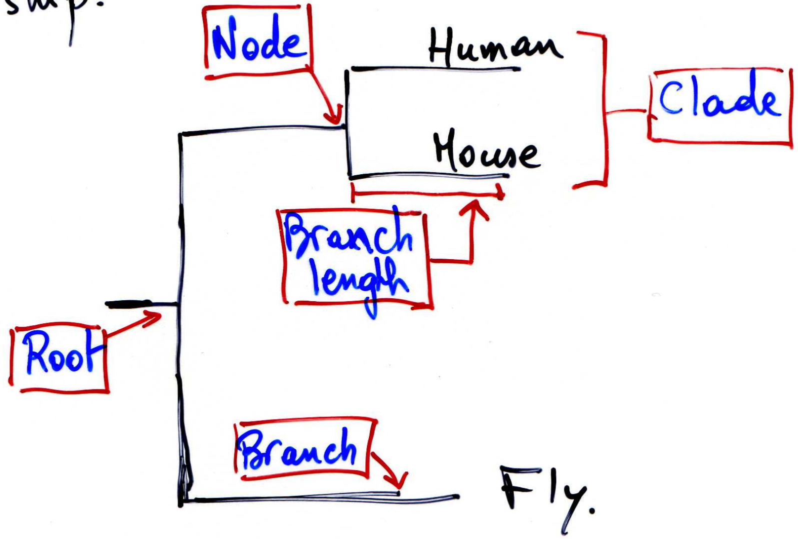


Fig. 2. Phylogenetic relationships for Asian, African forest, and African savannah elephants inferred from comb[illegible] analyses (2[illegible]) [illegible] 1732 bp [illegible], CHRNA1, GBA, and VIM), the two-letter codes for African elephant populations are given in Fig. 1. Asian elephant individuals are coded "Ema." The minimum evolution (NJ) tree is shown. Concordant trees were obtained by MP (tree length was 248 steps; CI = 0.927, RI = 0.934) and ML (-ln [illegible] = 2774.5[illegible]) analyses, which produced the same topology [illegible] defining the three groups. Bootstrap resampling support (100 iterations) is listed on branches for NJ (top), MP (middle), and ML (bottom) analyses for nodes supported by all three methods.

Phylogenetic Trees: Presenting Evolutionary Relationship.



noeud: Node : represents a taxonomic unit. This can be either an existing species or an ancestor

branche: Branch: defines the relationship between the taxa in terms of descent and ancestry

Branch length: represents the number of changes that have occurred in the branch.

Racine: Root: the common ancestor of all taxa.

feuille: Clade: a group of two or more taxa or DNA sequence that includes both their common ancestor and all of their descendents.

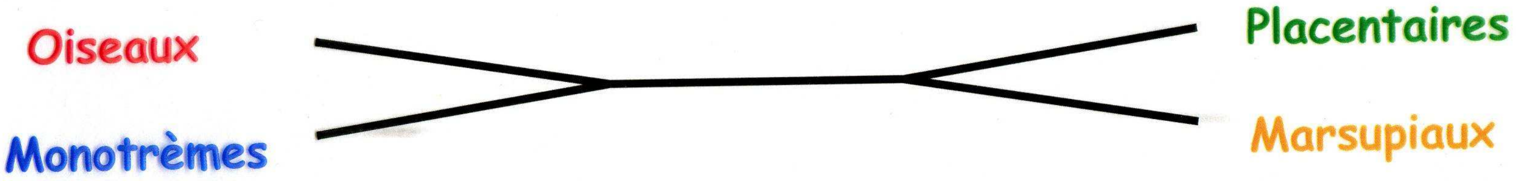
Notions fondamentales (3)

Un arbre phylogénétique est une **structure mathématique** qui est utilisée pour modéliser l'histoire évolutive d'un groupe d'organismes.

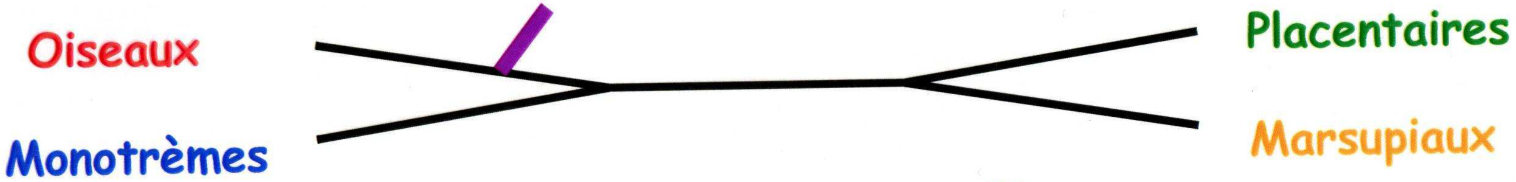
LES PHYLOGENIES SONT DES HYPOTHESES, elles ne peuvent pas être observées, elles ne peuvent être qu'inférées, parce qu'elles reflètent des événements évolutifs passés

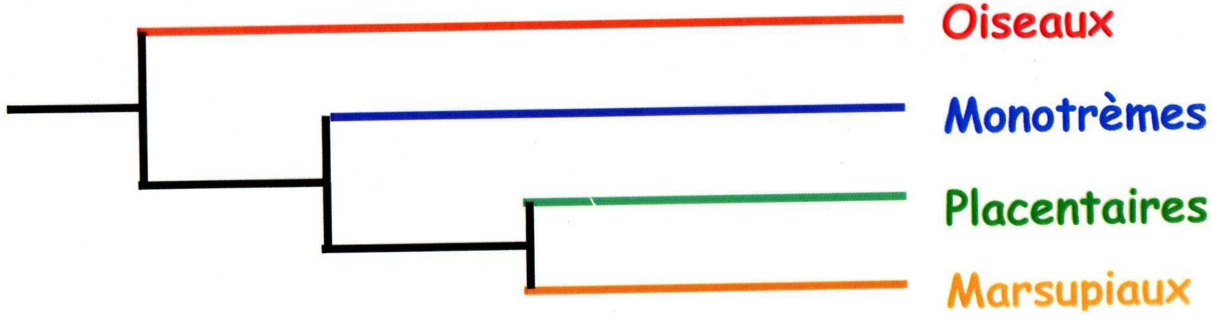
Les arbres les plus courants sont **non racinés**.

⇒ Ces arbres ne représentent pas les relations de parentés entre les organismes car ils n'ont pas de dimension temporelle.



=> Pour inclure une dimension temporelle il faut placer une racine, c'est-à-dire à positionner l'ancêtre hypothétique commun de tous les organismes étudiés.





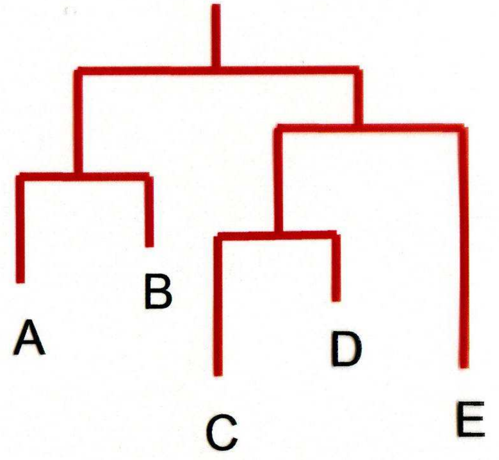
Les arbres

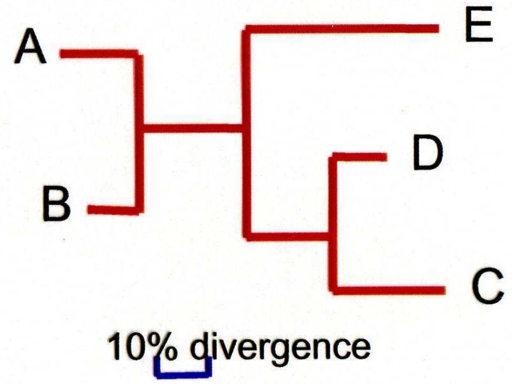
Arbres avec ou sans racine

- Avec racine: montre les relations ancestrales.
- Sans racine: montre les distances.

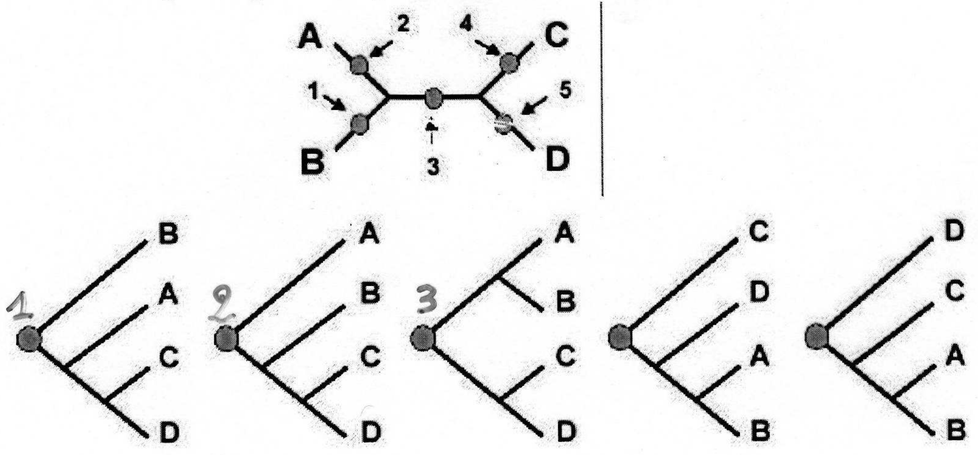
Structure d’un arbre

- Feuilles, branches et noeuds
- Un arbre représente mal les distances entre individus
- Le meilleur arbre est celui qui minimise les distances et dont taille des branches respecte mieux les distances réelles



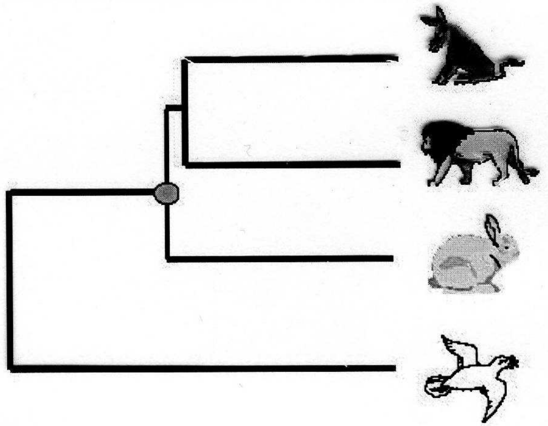


- **Le nombre de places possibles pour la racine :**

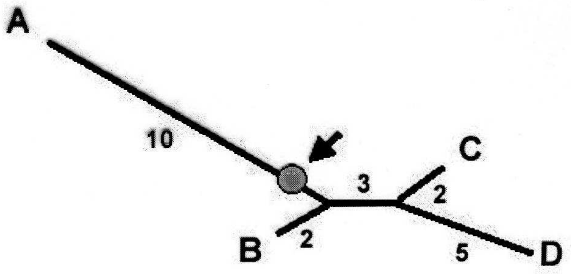


- **Deux mode de racinement pour un arbre.**

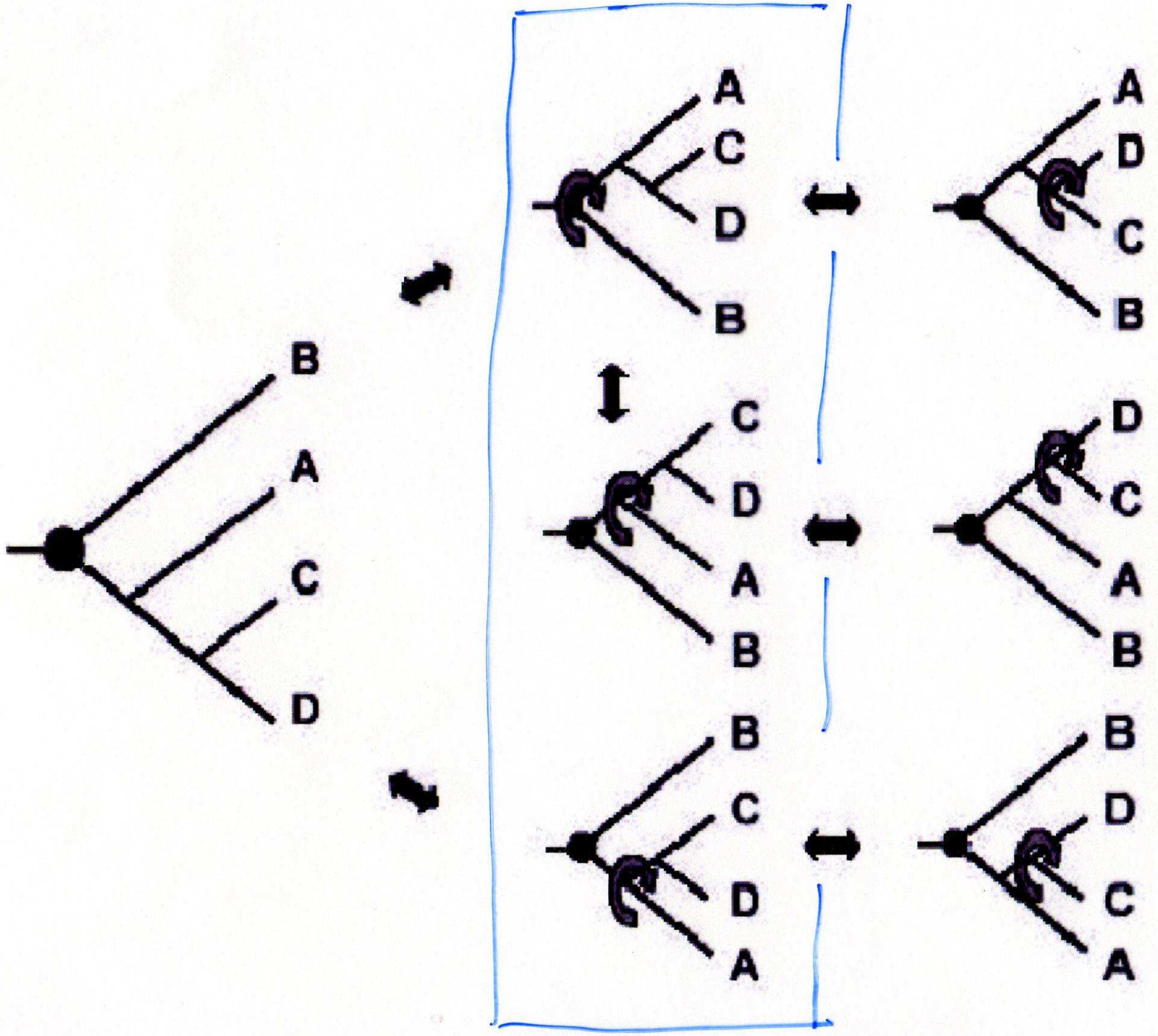
1/ On peut positionner la racine grâce à un groupe externe : sachant (par d'autres données) que les mammifères sont apparus après les oiseaux, l'inclusion d'un oiseau dans la phylogénie permet de mettre en évidence un clade Ane-Lion, avec le Lapin comme taxon externe à ce clade



Quand on n'a aucune possibilité de décider quel taxon peut servir de groupe externe, on place souvent la racine au milieu de l'arbre ; ceci fait implicitement usage de la notion d'horloge moléculaire.



ROTATION DE BRANCHES



NOTE : Contrairement au changement de place de la racine, la rotation de branche (branch swapping) n'a aucune influence sur l'interprétation des résultats !