

Bioinformatique évolutive

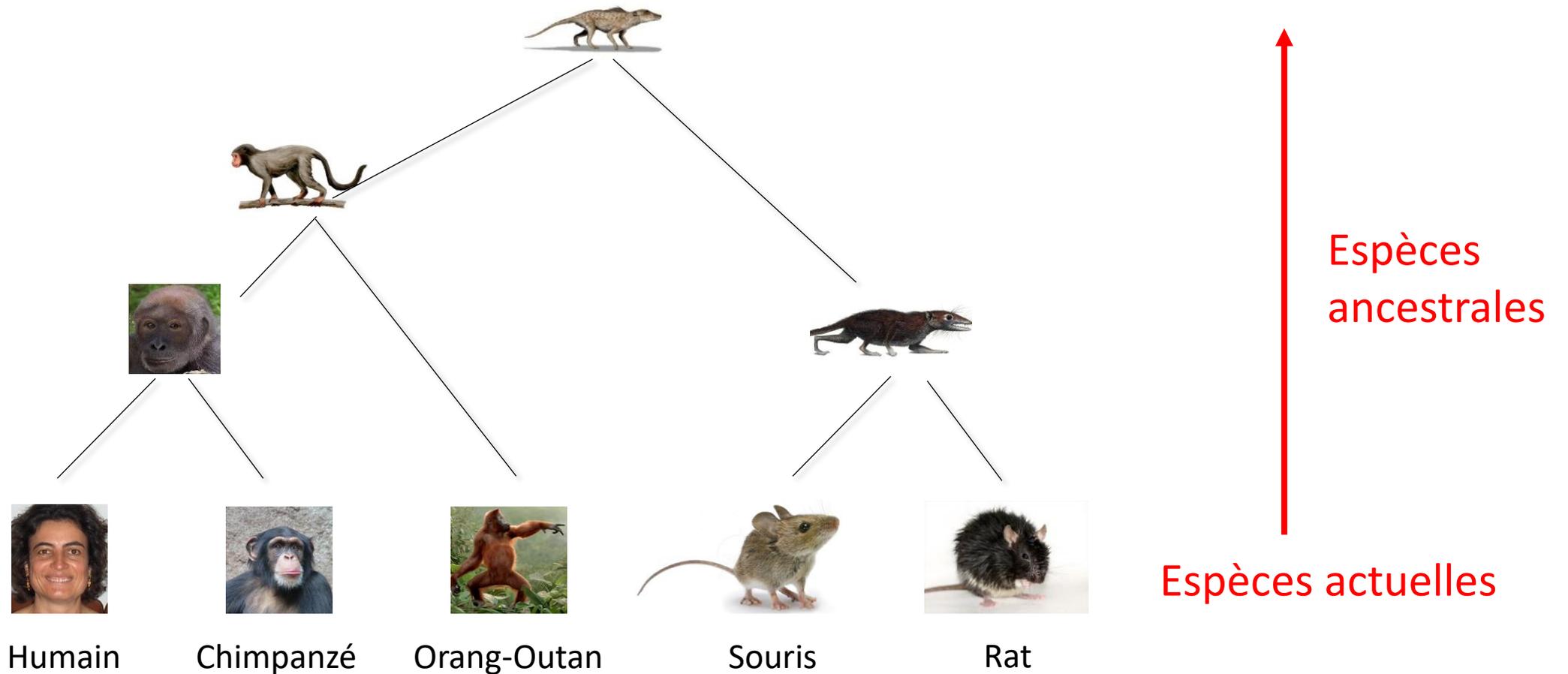
Nadia El-Mabrouk

 DIRO

Université 
de Montréal



Inférer les relations évolutives entre les espèces – Divergence par spéciation



Plan :

- I. Un peu d'histoire et concepts de base
- II. Modèles d'évolution
- III. Définitions formelles sur les arbres
- IV. Dénombrement des arbres
- V. Distance entre arbres

Arbre phylogénétique

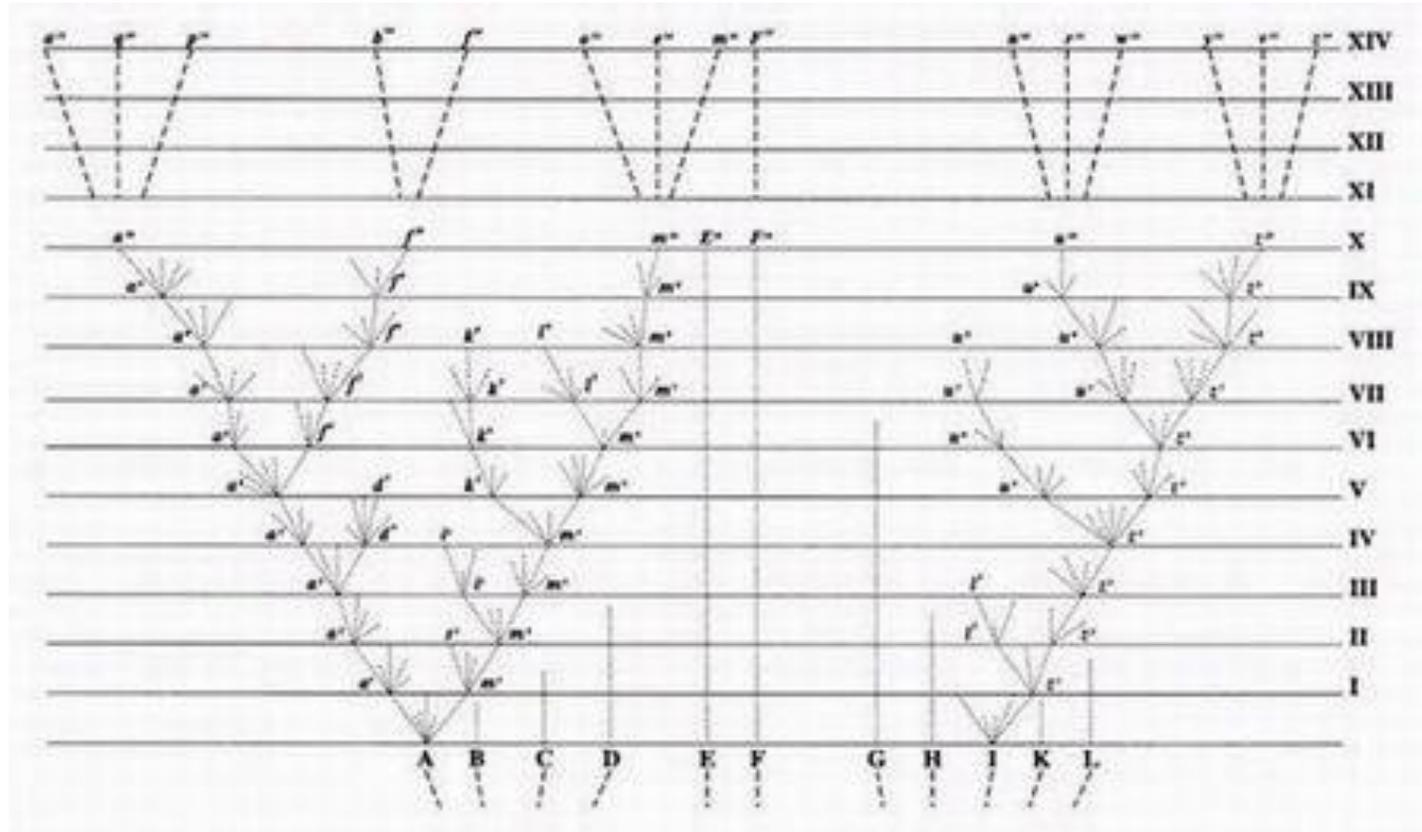


Image de l'arbre de vie apparue dans l'ouvrage de Darwin « On the Origin of Species by Natural Selection », 1859. C'était la seule illustration du livre.

http://commons.wikimedia.org/wiki/File:Darwins_tree_of_life_1859.gif

Spéciation

- Les pinsons de Darwin, à l'origine de la théorie de l'évolution de Darwin

➤ **Spéciation allopatrique**, par isolement géographique. Mode de spéciation de loin le plus fréquent chez les animaux.



I. Un peu d'histoire et concepts de base

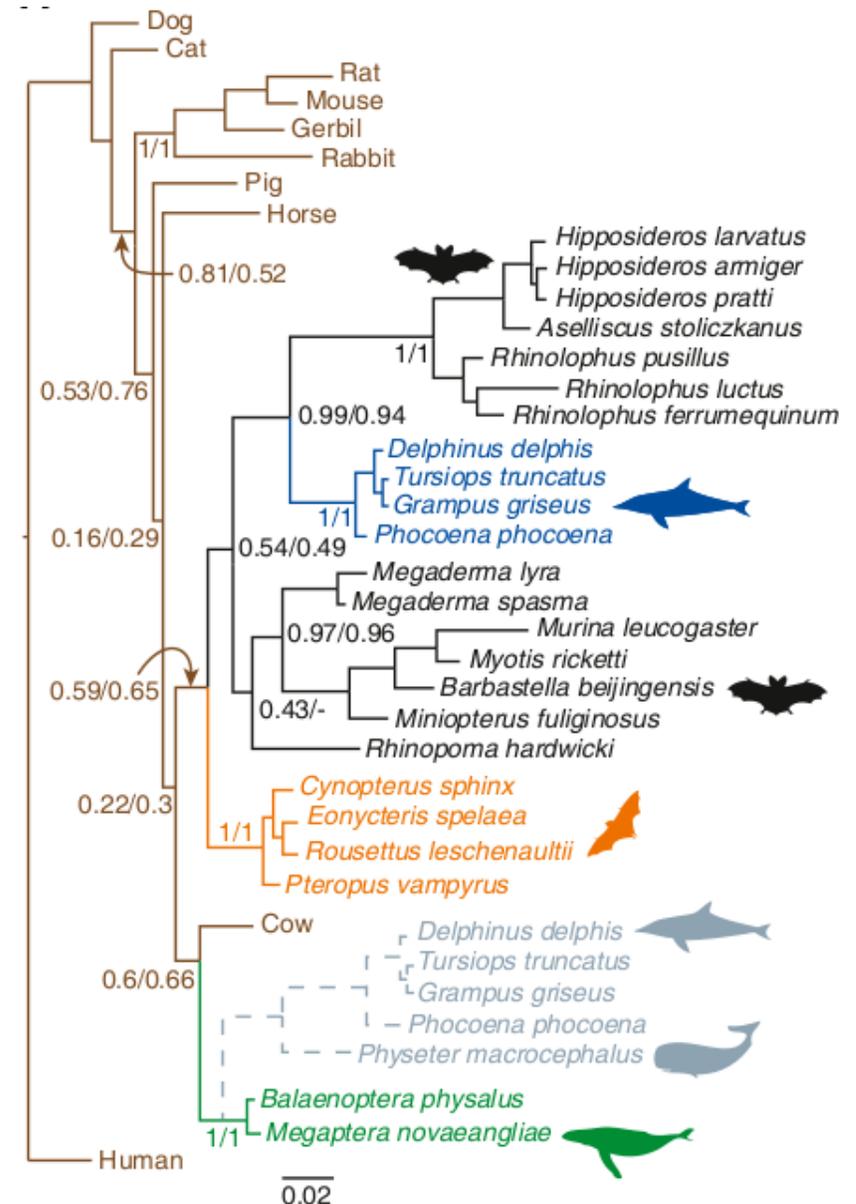
- **Jusqu'au XIX^{ième} siècle**, notion fixiste des espèces: telles que Dieu les avaient créées, immuable et en nombre limité. Selon Linné (1758) : « *il y a autant d'espèces que l'Être suprême a créé dès le début de formes différentes* ».

But de la taxonomie: inventorier toutes les formes de vies. Spécimens «types» déposés dans des musées, étalon pour des comparaisons.

Différences génétiques versus morphologiques - Convergence évolutive

Sélection naturelle conduisant à l'acquisition du même trait, en parallèle dans deux espèces différentes. Exemples:

- Apparition des ailes chez l'oiseau et la chauve-souris.
- Capacité d'écholocation au moyen de sons à très hautes fréquences. Présent chez les dauphins et les chauves-souris.
- Deux hypothèses:
 - Gène ancestral commun qui a disparu dans les autres branches? Entraîne beaucoup de pertes
 - Convergence évolutive. Plus parcimonieux
- Base génétique de l'écholocation: Gène nommé *Prestin*.
- Phylogénie:
 - Basée sur la séquence nucléotidique du gène: arbre phylogénétique «correct» regroupant les dauphins avec les baleines.
 - Basée sur les séquences d'AA du gène: Phylogénie regroupe dauphins et chauves-souris! Preuve d'une convergence évolutive. Le gène muté procure un avantage sélectif, et est donc fixé dans les deux espèces éloignées. Sélection positive.



Différences génétiques entre espèces ou entre populations

- Différences entre l'homme et le chimpanzé:
 - (Nature 2005) sur la base de la comparaison de 2400 Mb d'ADN, **~ 1% de différence nucléotidique**. En moyenne une différence tous les **100 nucléotides alignés (30 millions de variations ponctuelles)**. Mais seulement pour les parties « alignables ».
 - **~ 4% si on compte insertions et délétions.**
 - **~ 2,7% de différences de CNV** (Copy Number Variation)
 - **~ 6,4% (1418/22000) des gènes humains** n'ont pas leurs orthologues chez le chimpanzés.
- Substitutions nucléotidiques entre notre génome et celui du chimpanzé est **dix fois plus grande** que celle résultant du polymorphisme entre deux êtres humains. Polymorphisme entre deux êtres humains environ **0,1%**.

Espèce, variété, race

- **L'espèce est l'unité de base de la classification des êtres vivants.**
- Les groupes d'individus d'une espèce forment des populations
- **Variétés de plantes et races d'animaux en fonction de regroupements selon des traits particuliers.**

Des variétés et races sont créées par l'homme pour des raisons d'élevage ou d'agriculture, en sélectionnant des individus: Populations homogènes ayant des caractéristiques propres.

Par exemple, plus de 1000 variétés cultivées de maïs. Peuvent se croiser entre elles.

Existe-t-il des races humaines?

- Combien de races? Des dizaines? Des centaines? Tous accordent une place prédominante à la couleur. Mais pas binaire. Grande variabilité des traits physiques.
- « Au début des recherches en génétique, les scientifiques, qui avaient en tête des classifications raciales héritées du siècle dernier, pensaient qu'ils allaient retrouver des gènes des Jaunes, des Noirs, des Blancs... Mais on ne les a pas trouvés. Dans tous les systèmes génétiques humains connus, les répertoires de gènes sont les mêmes. » (André Langaney, 1992).



Existe-t-il des races humaines?

- **Gradation dans la fréquence des allèles.**
- On peut définir des groupes de population selon des fréquences alléliques d'un grand nombre de marqueurs génétiques. On peut même retrouver le continent d'origine d'un individu. On parle de **«groupe géographique»**.
- Permet de retracer les histoires de migration des populations.
- Du point de vue de la médecine (facteurs de risque, réponse aux traitements, etc), la descendance plus importante que le groupe géographique (caucasienne, africaine, asiatique, etc). Populations de proximité plus similaires en terme de gènes.

Existe-t-il des races humaines?

Genetic Structure of Human Populations; N.A. Rosenberg et al., Science, vol. 298, 2002

- La proportion moyenne de différence génétique entre des individus de populations différentes (prédéterminée sur une base culturelle ou géographique) excède de peu celle entre des individus d'une même population.
- Sur les 4199 allèles présent plus d'une fois, seulement 7.4% sont spécifiques à une seule région (Afrique, Europe, Asie du sud, Océanie, etc). De plus ces allèles spécifiques sont rares (~ 1% dans la région d'occurrence)
- On peut cependant identifier des sous-groupes en fonction de la fréquence des allèles.

Existe-t-il des races humaines?

Genetic Structure of Human Populations; N.A. Rosenberg et al., Science, vol. 298, 2002

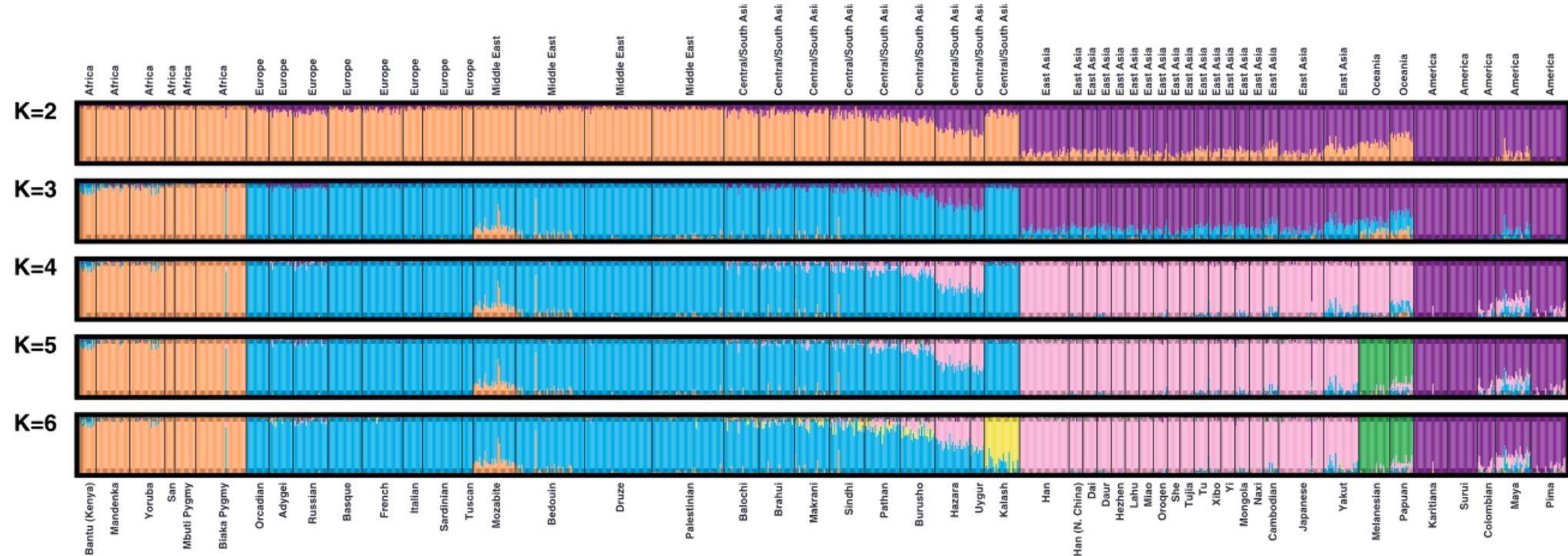


Fig. 1. Estimated population structure. Each individual is represented by a thin vertical line, which is partitioned into K colored segments that represent the individual's estimated membership fractions in K clusters. Black lines separate individuals of different populations. Populations are labeled below the figure, with their regional affiliations above it. Ten *structure* runs at each

K produced nearly identical individual membership coefficients, having pairwise similarity coefficients above 0.97, with the exceptions of comparisons involving four runs at $K = 3$ that separated East Asia instead of Eurasia, and one run at $K = 6$ that separated Karitiana instead of Kalash. The figure shown for a given K is based on the highest probability run at that K .

II. Modèles d'évolution

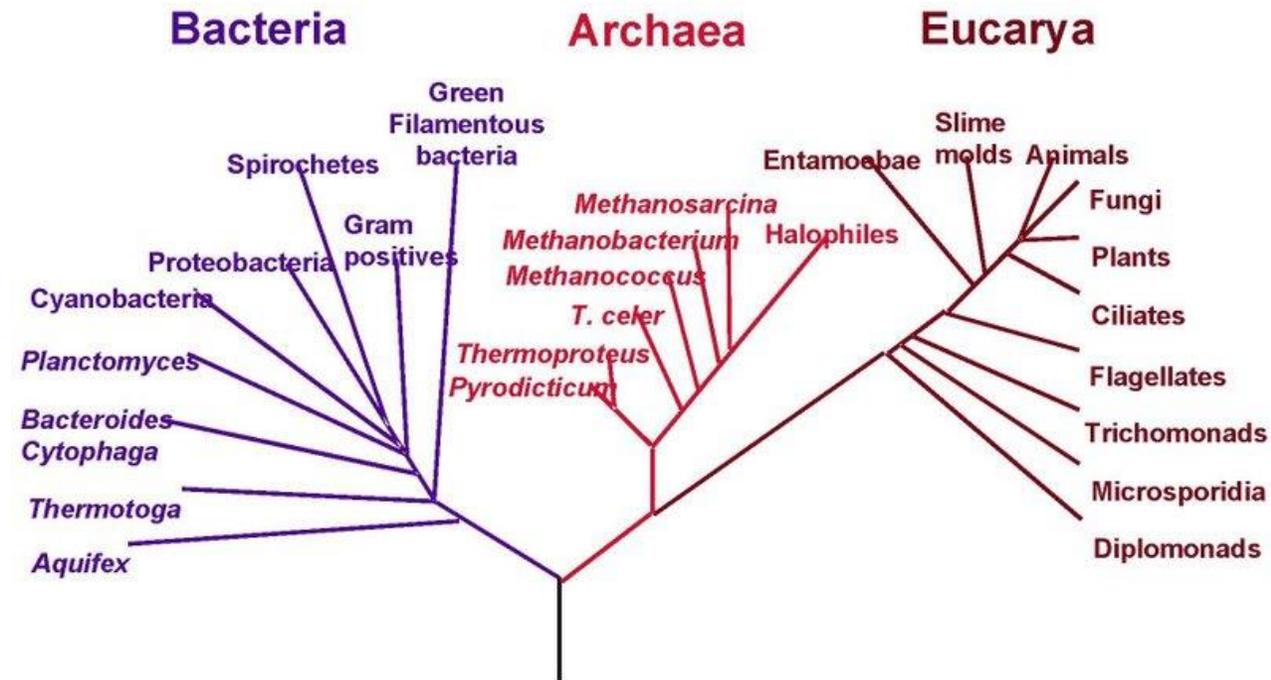
Évolution à partir d'un ancêtre commun

Tout au long de l'évolution, les gènes accumulent des mutations. Lorsqu'elle sont neutres ou bénéfiques à l'organisme elles sont transmises d'une génération à l'autre.

Sur une période d'au moins 3.8 milliards d'années le premier être vivant sur terre n'a cessé de se séparer en espèces différentes.

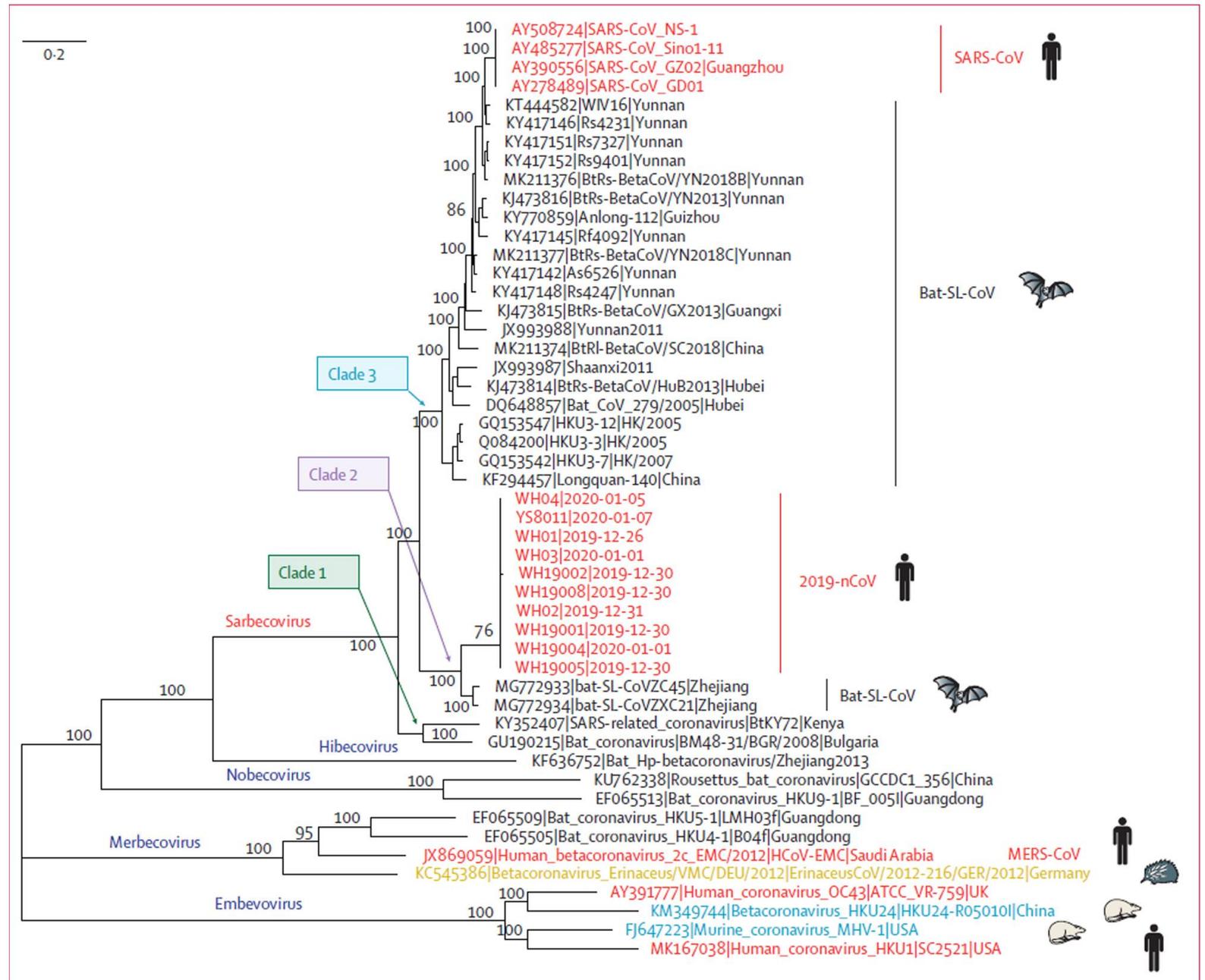
Objectif principal : Reconstruire l'arbre de vie de toutes les espèces vivantes à partir des données génétiques observées.

Phylogenetic Tree of Life



À des fins bio-médicales

- Analyse phylogénétique des génomes complets du virus SARS-CoV-2 et de virus représentatifs du genre Betacoronavirus.
 - **CoV de la chauve-souris et SARS-CoV-2 humain pourraient provenir d'un ancêtre commun**



II. Modèles d'évolution

Caractères utilisés:

- Une région spécifique de l'ADN,
- Une protéine
- Un caractère morphologique
- L'ordre des gènes dans le génome
- ...

Les caractères choisis doivent être **homologues**

Hypothèse généralement considérée: Chaque caractère évolue indépendamment des autres.

II. Modèles d'évolution

- Caractères les plus utilisés pour les études d'évolution: Séquences de nucléotides ou d'AA.
 - Séquences orthologues dans les espèces étudiées
 - Effectuer un alignement multiple des séquences
 - Les **caractères** représentés par les colonnes de l'alignement et les états du caractère sont les nucléotides (ou AA observés)

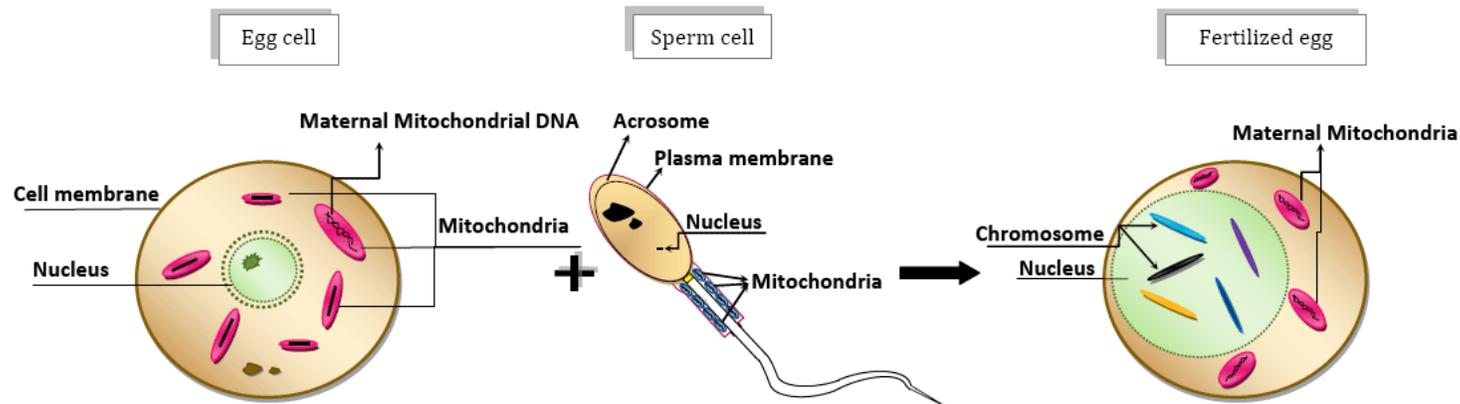
dolphin	ATGACCAACATCCGAAAAACACACCCTCTAATAAAAATCCTC
giant sperm whale	ATGACCAACATCCGAAAATCACACCCATTAATAAAAATCATT
bowhead whale	ATGACCAACATCCGAAAAACACACCCACTAATAAAAATTATT
right whale	ATGACCAACATCCGAAAAACACACCCAGTAATAAAAATTATT
minke whale	ATGACCAACATCCGAAAAACACACCCACTAATAAAAATTATC
fin whale	ATGACCAACATCCGAAAAACACACCCACTAATAAAAATCGTC
blue whale	ATGACCAACATCCGAAAAACACACCCACTAATAAAAATCATC
humpback whale	ATGACCAACATCCGAAAAACACACCCACTAATAAAAATTATC

II. Modèles d'évolution

Choix de marqueurs (séqu. d'ADN)

Comment choisir une région de l'ADN qui « reflète » l'évolution de tout le génome?
Caractéristiques gagnantes:

- **Marqueur « non-recombinant ».** Choisir des marqueurs uni-parentaux, comme les seq. de mitochondries et de chloroplastes: transmission par la mère uniquement.



<http://www2.le.ac.uk/departments/emfpu/genetics/explained/images/mtDNA-egg-and-sperm.gif/view>

II. Modèles d'évolution

Choix de marqueurs (séqu. d'ADN)

- **Marqueur en copie unique**, pour éviter de choisir de mauvais « paralogues » ou:
- Marqueurs en copie multiples subissant une « **évolution concertée** » permettant d'uniformiser toutes les copies.
- **ARNr**: Marqueurs très utilisés pour les études phylogénétiques:
 - Régions répétées de l'ADN subissant une évolution concertée
 - Parmi les familles de gènes les plus conservées dans la cellule
 - Alignements multiples faciles à faire
 - Permet la comparaison d'espèces très éloignées.

II. Modèles d'évolution

Distance évolutive

- **Distance évolutive** d entre deux séquences: nombre moyen de substitutions/site s'étant produites depuis la divergence de ces deux séquences à partir d'un ancêtre commun.
- Estimation des distances évolutives à la base de la plupart des méthodes de reconstructions phylogénétiques.
- Construction d'une **matrice de distance** contenant les distance évolutives entre paire de séquences: Première étape des méthodes phylogénétiques.

Divergence observée

- Calculée directement à partir de la distance d de Levenshtein ou de Hamming (substitutions) entre deux séquences (ADN ou protéines).
 - Taux de divergence = d/n où n est la taille des séquences.
- Pour deux séquences aléatoires d'ADN, le taux de divergence est égal à 0.75
- Divergence observée: seule mesure directement accessible.
- Pas un bon estimateur à part pour les séquences très proches: tendance à sous-estimer la distance évolutive réelle.

Modèle markovien de l'évolution

- Calcul d'une probabilité de transition d'un état à un autre
- Calcul d'une matrice 4x4:

$$Q = \begin{pmatrix} -\mu_A & \mu_{GA} & \mu_{CA} & \mu_{TA} \\ \mu_{AG} & -\mu_G & \mu_{CG} & \mu_{TG} \\ \mu_{AC} & \mu_{GC} & -\mu_C & \mu_{TC} \\ \mu_{AT} & \mu_{GT} & \mu_{CT} & -\mu_T \end{pmatrix}$$

- μ_{ij} ($i \neq j$) : taux de substitution instantané de l'état i à l'état j .
- $1 - \mu_i$: taux de conservation instantané du nucléotide i .
- Q : matrice des taux du processus de Markov. La somme sur chaque colonne est 0.

Modèle de Jukes et Cantor (JC69)

- Modèle markovien de substitution le plus simple.
- Considère le même taux de substitution instantané pour chacun des changements possible, et un seul taux de conservation global.

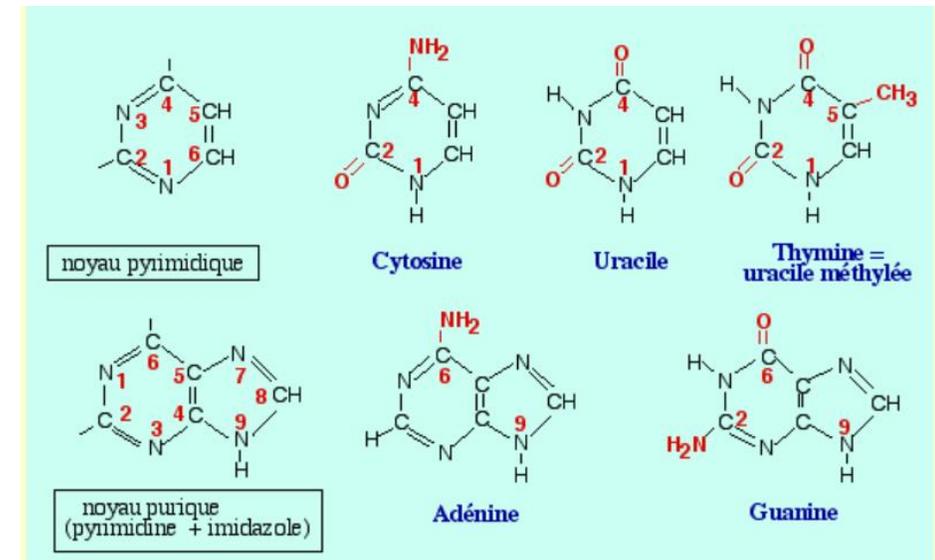
$$Q = \begin{pmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{pmatrix}$$

- $\mu/4$: taux moyen instantané de substitution.

Modèle de Kimura (K80)

- Transitions et transversions ont des taux différents.
 - Transitions:
 - $A \leftrightarrow G, C \leftrightarrow T$
 - Transversions:
 - $A \leftrightarrow T, T \leftrightarrow G, A \leftrightarrow C, C \leftrightarrow G$
- κ rapport entre le taux de transitions et le taux de transversions.

$$Q = \begin{pmatrix} * & \kappa & 1 & 1 \\ \kappa & * & 1 & 1 \\ 1 & 1 & * & \kappa \\ 1 & 1 & \kappa & * \end{pmatrix}$$



Sélection naturelle

- Processus par lequel certaines modifications apparaissant par hasard chez certains individus dans une population sont favorisées et fixées, tandis que d'autres sont défavorisées et perdues.
- Concept initialement formulé par Darwin, basé sur une observation des phénotypes. La sélection naturelle affecte également le génotype.
- Peut mener à la création de nouvelles espèces.

Distance synonyme/non-synonyme pour les séquences codantes

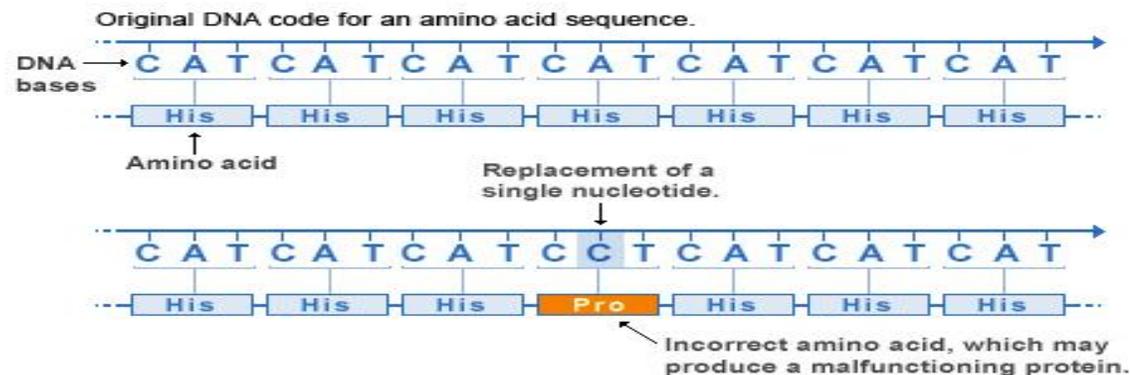
Les gènes sont soumis à plusieurs types de sélection naturelle:

- **Sélection positive**: Processus qui encourage la rétention des mutations qui sont bénéfiques pour un individu.
- **Sélection négative** ou purificatrice: Processus qui tend à faire disparaître des mutations nuisibles.
- **Sélection neutre**: Absence de sélection positive ou négative. Dans le cas de séquences qui ne sont affectées par aucune pression sélective. Peuvent être modifiées sans conséquences sur l'organisme.

Distance synonyme/non-synonyme pour les séquences codantes

- Basée sur la comparaison des substitutions synonymes et non-synonymes (effet sur les codons)
- **Substitution synonyme** (silencieuse): substitution ne provoquant pas la substitution de l'acide aminé initial.
- **Substitution non-synonyme** (non-silencieuse): substitution provoquant la modification d'un acide aminé.

Missense mutation



le code génétique

		Deuxième lettre								
		U		C		A		G		
Première lettre (côté 5')	U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
		UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys	C
		UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop	A
		UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp	G
	C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
		CUC	Leu	CCC	Pro	CAC	His	CGC	Arg	C
		CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	A
		CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	G
	A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
		AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	C
		AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg	A
		AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg	G
	G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
		GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly	C
		GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly	A
		GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly	G
		codon d'initiation				codon de terminaison				

ijk

Troisième lettre (côté 3')

Distance synonyme/non-synonyme pour les séquences codantes

- **Sites synonymes et non-synonymes:**
 - 100% des mutations touchant la 2^{ème} base des codons sont non-synonymes
 - Sous l'hypothèse que les fréquences nucléotidiques sont égales et que les mutations se font au hasard, 95% des mutations touchant la 1^{ère} base et 28% des mutations touchant la 3^{ème} base sont non-synonymes.
- **Distances synonymes et non-synonymes:**
 - d_S distance synonyme entre deux séquences codantes: Nbre de substitutions synonymes s'étant produites par site synonyme.
 - d_N distance non-synonyme: Nbre de subs. non-synonymes par site non-synonyme.

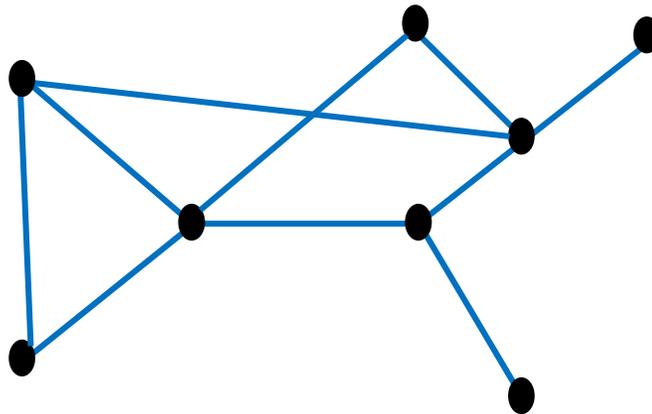
Distance synonyme/non-synonyme pour les séquences codantes

Identification du type de sélection en fonction de d_N/d_S :

- Sélection **neutre**:
→ $d_N/d_S \approx 1$
- Sélection **négative**:
Déficit de substitutions non-synonymes attendu
→ $d_N/d_S < 1$
- Sélection **positive**:
Excès de subst. non-synonymes attendu
→ $d_N/d_S > 1$

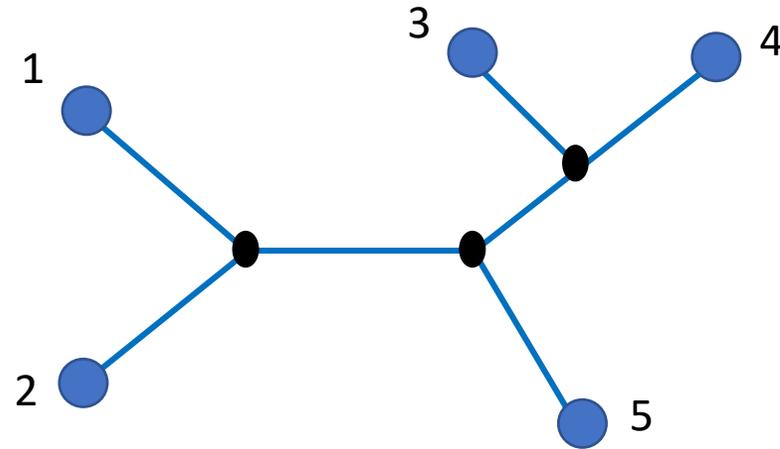
III. Définitions formelles

- **Arbre:** Graphe connexe acyclique; Ensemble de nœuds (ou sommets)
- connectés par des arêtes (ou branches)
- de telle sorte que toute paire de nœuds est reliée par exactement un chemin.



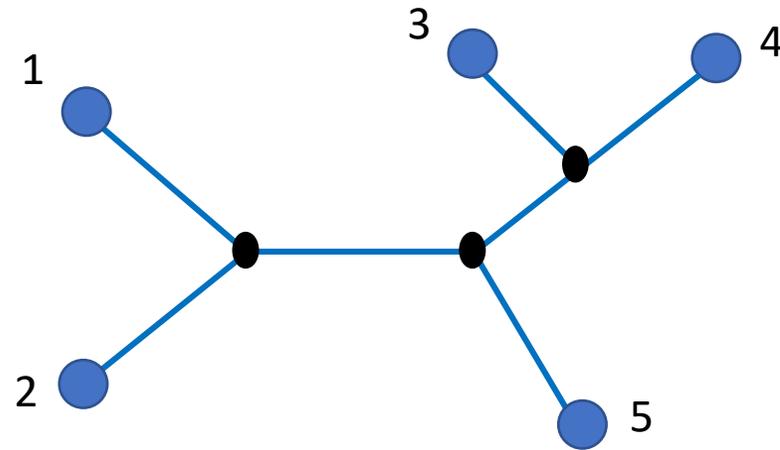
III. Définitions formelles

- Les **feuilles** (nœuds de degré 1) représentent les espèces (ou séquences) actuelles
- Les **nœuds internes** représentent les événements de spéciation (ou duplications dans le cas de familles de gènes).



III. Définitions formelles

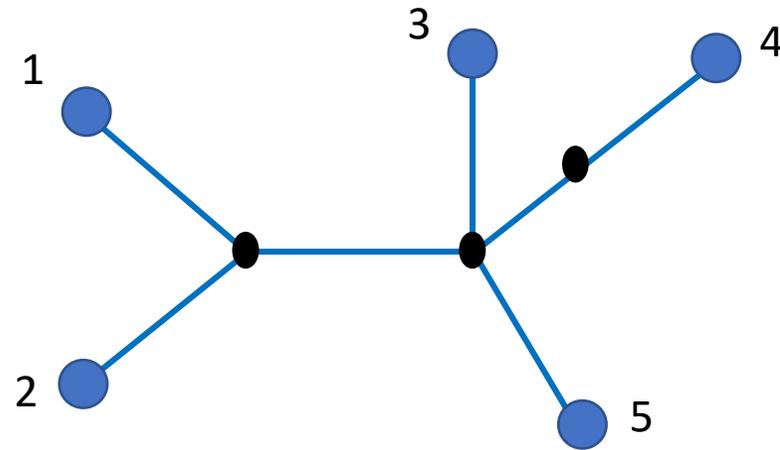
- **Arbre binaire:** Chaque nœud interne est de degré 3



binaire

III. Définitions formelles

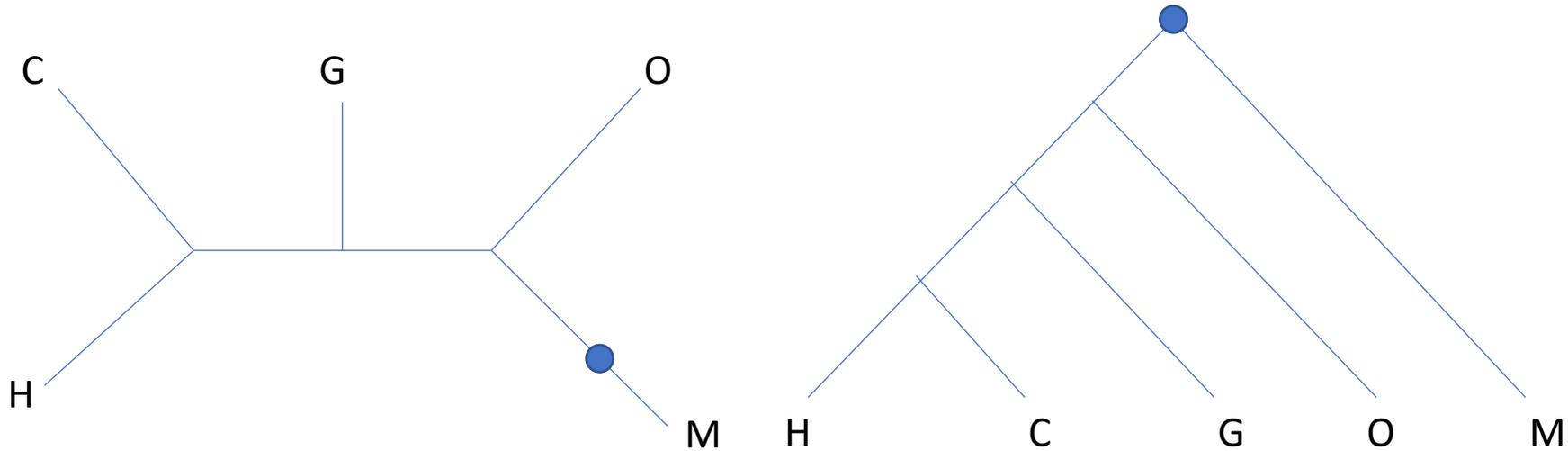
- **Arbre binaire:** Chaque nœud interne est de degré 3



non-binaire

III. Définitions formelles

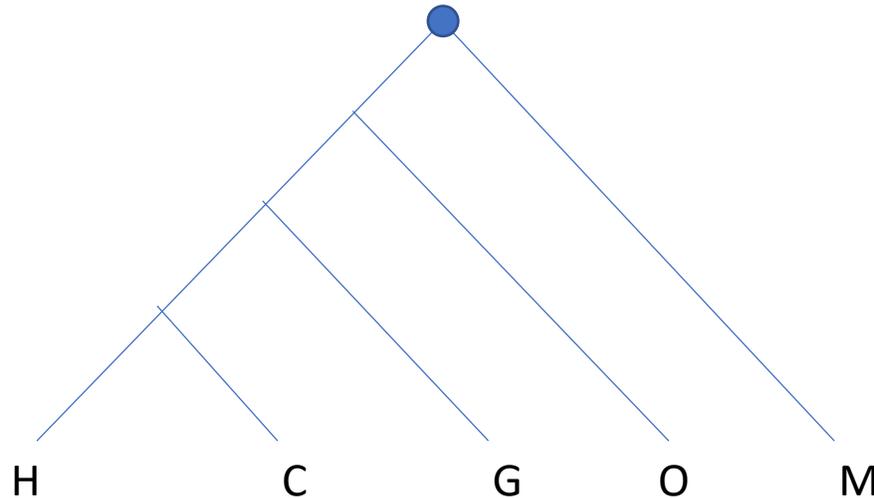
- **Arbre:** Graphe connexe acyclique; Ensemble de nœuds (ou sommets) connectés par des arêtes (ou branches) de telle sorte que toute paire de nœuds est reliée par exactement un chemin.
- **Arbre enraciné:** Un nœud est créé sur une branche et désigné comme étant la racine; permet d'orienter la lecture de l'arbre; le temps s'écoule de la racine vers les feuilles.



III. Définitions formelles

- La **racine** représente l'ancêtre commun
- **Arbre raciné binaire**: Chaque nœud interne a deux fils.

Nœuds internes est de degré 3 à part la racine qui est de degré 2.

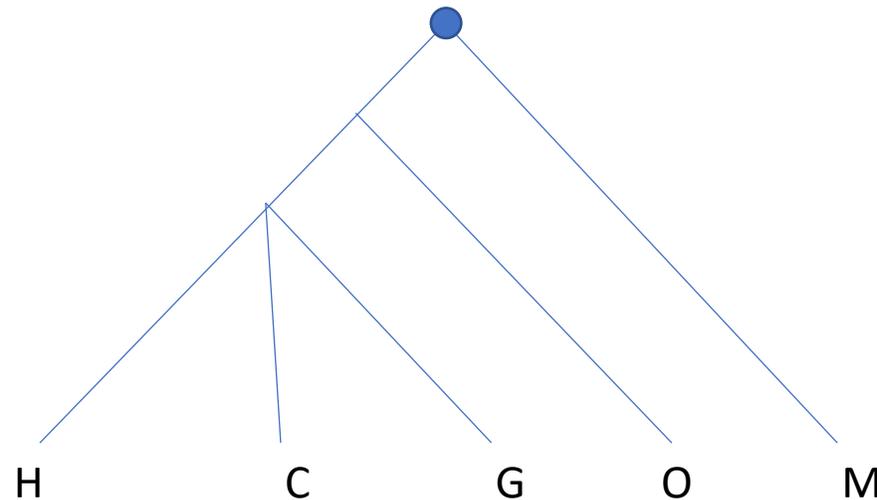


binaire

III. Définitions formelles

- La **racine** représente l'ancêtre commun
- **Arbre raciné binaire**: Chaque nœud interne a deux fils.

Nœuds internes est de degré 3 à part la racine qui est de degré 2.

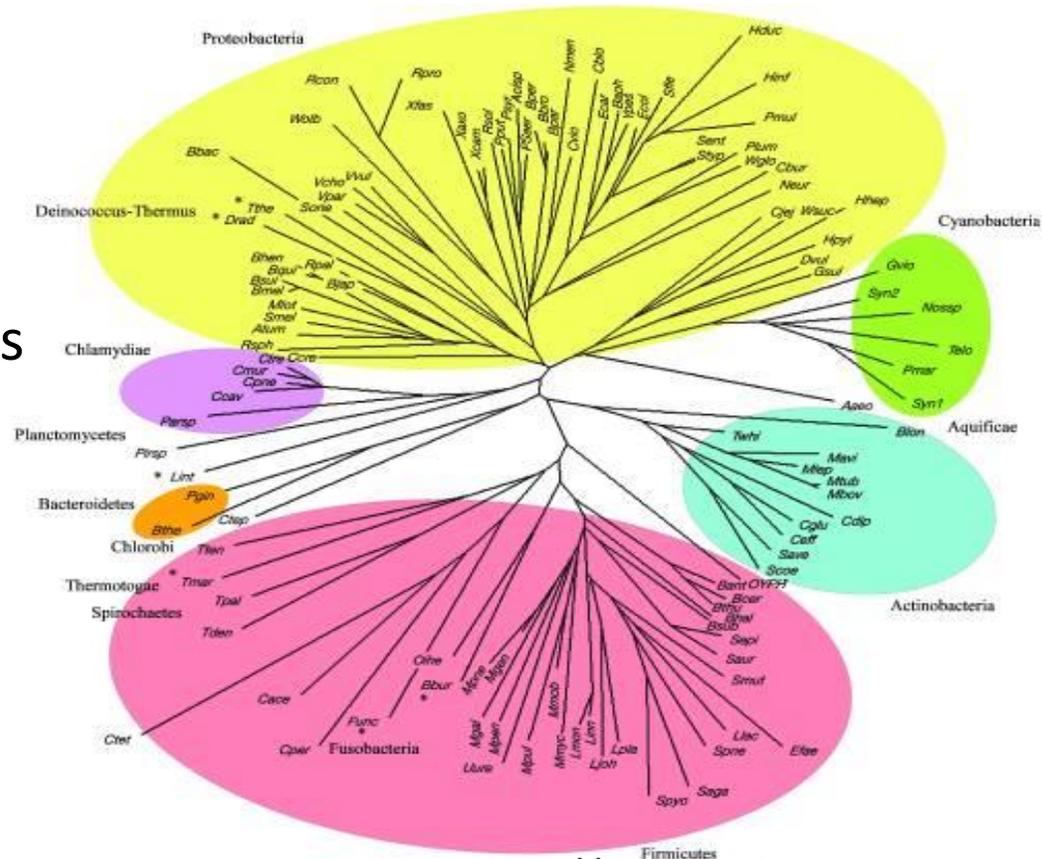


non-binaire

III. Définitions formelles

- Un arbre phylogénétique peut-être **binaire ou non-binaire**.
- Un nœud non-binaire représente généralement un nœud non-résolu de l'arbre

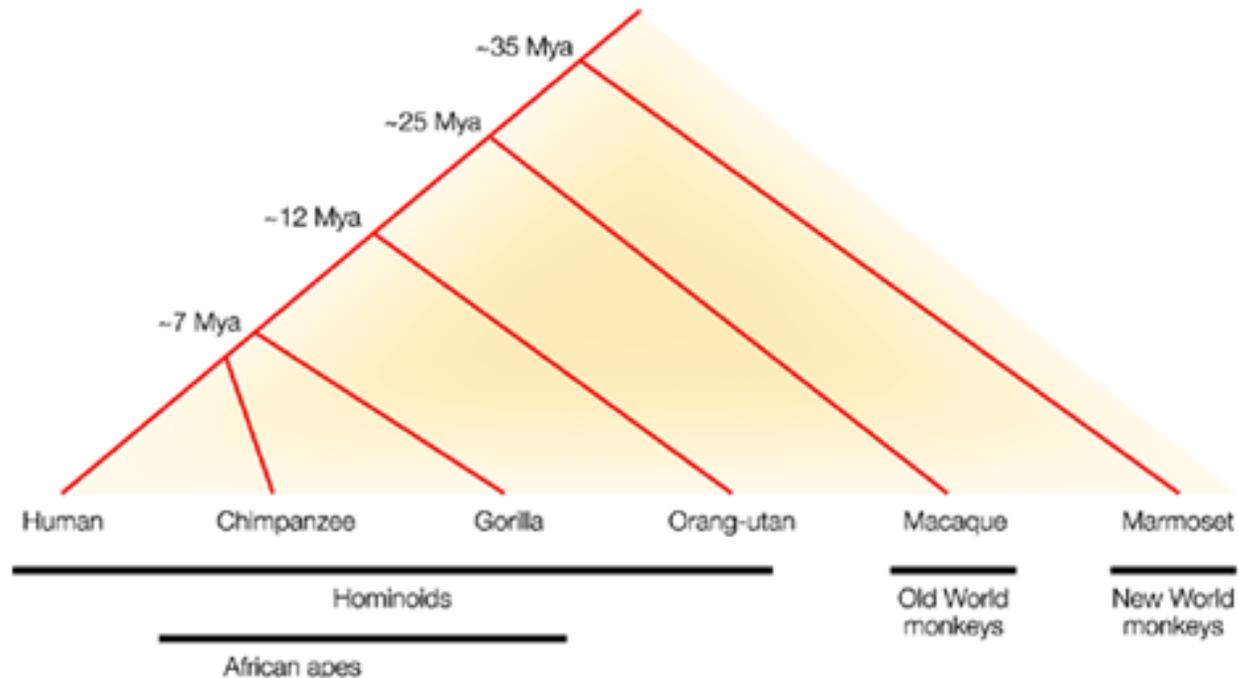
Dans la suite du cours, si non-spécifié, les arbres sont considérés **binaires**



NJ tree (with weighting) of 119 Bacteria. Asterisks denote anomalously positioned taxa.

III. Définitions formelles

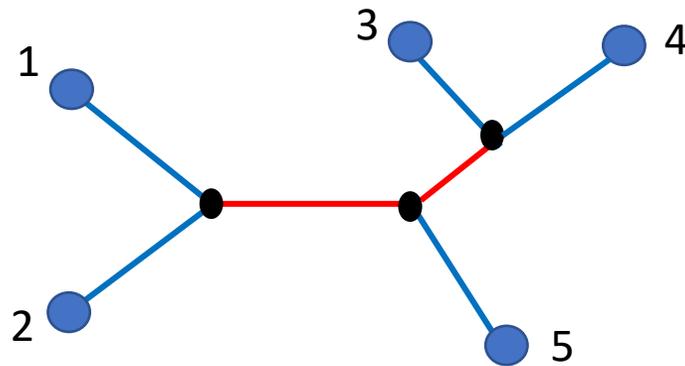
- Les nœuds ou arêtes d'un arbre de phylogénie peuvent être étiquetés. Les étiquettes représentent généralement le taux de mutations survenu, ou la date de spéciation



R.V. Samonte & Evan E. Eichler
Nature Reviews Genetics **3**, 65-72 (January 2002)

IV. L'arbre caché dans la forêt

- Arbre non raciné binaire de **n feuilles**: **n-2** nœuds internes, **n-3 branches internes**, et **2n-3 branches**.

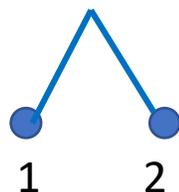


n=5;
3 noeuds internes;
2 branches internes;
7 branches

IV. L'arbre caché dans la forêt

- Arbre non raciné binaire de **n feuilles**: **$n-2$ nœuds internes**, **$n-3$ branches internes**, et **$2n-3$ branches**. Chaque branche définit une *bipartition* de l'ensemble des feuilles. Arbre définit par **$n-3$ bipartitions non-triviales**.

$n=2$:

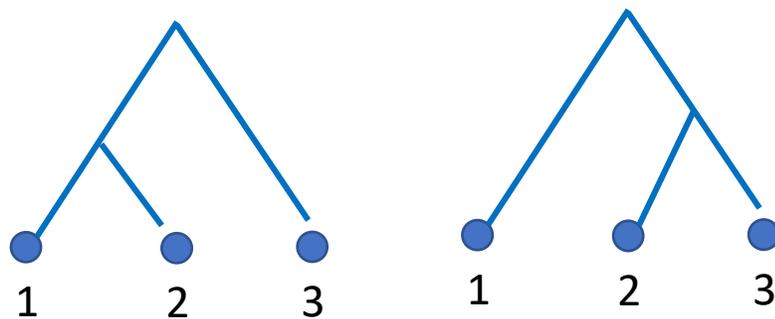
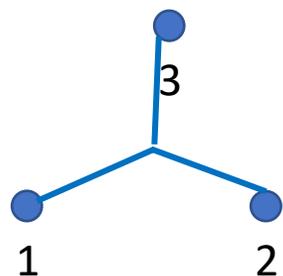


- Arbre non raciné unique
- Arbre raciné unique

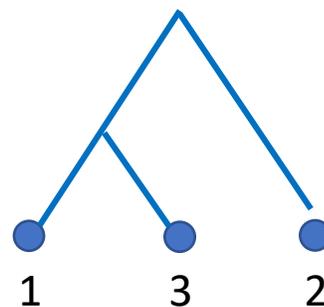
IV. L'arbre caché dans la forêt

- Arbre non raciné binaire de **n feuilles**: **$n-2$ nœuds internes**, **$n-3$ branches internes**, et **$2n-3$ branches**. Chaque branche définit une *bipartition* de l'ensemble des feuilles. Arbre définit par **$n-3$ bipartitions non-triviales**.

$n=3$:



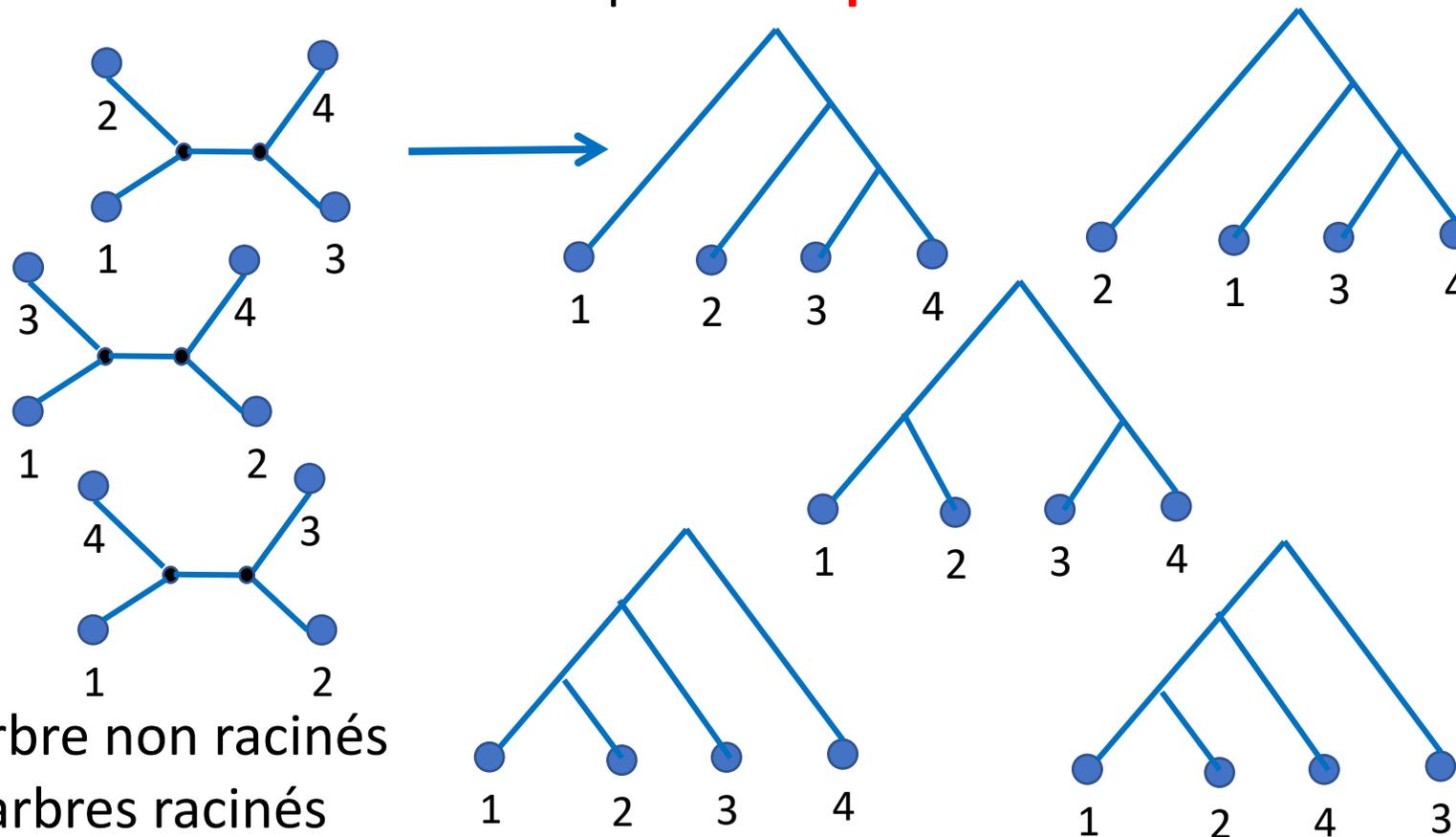
- Arbre non raciné unique
- 3 arbres racinés



IV. L'arbre caché dans la forêt

- Arbre non raciné binaire de **n feuilles**: **n-2 nœuds internes**, **n-3 branches internes**, et **2n-3 branches**. Chaque branche définit une *bipartition* de l'ensemble des feuilles. Arbre défini par **n-3 bipartitions non-triviales**.

n=4:



- 3 Arbre non racinés
- 15 arbres racinés

IV. L'arbre caché dans la forêt

- Donc le problème d'inférence d'arbres se pose à partir de 3 feuilles pour les arbres racinés, et de 4 feuilles pour les arbres non-racinés.
- Cavalli-Sforza et Edwards (1967) ont montré que le nombre B_r d'arbres racinés à n feuille est:

$$B_r = (2n-3)! / 2^{n-2} (n-2)!$$

- Le nombre B_u d'arbres non racinés à n feuilles est égal au nombre d'arbres racinés à $n-1$ feuilles, donc:

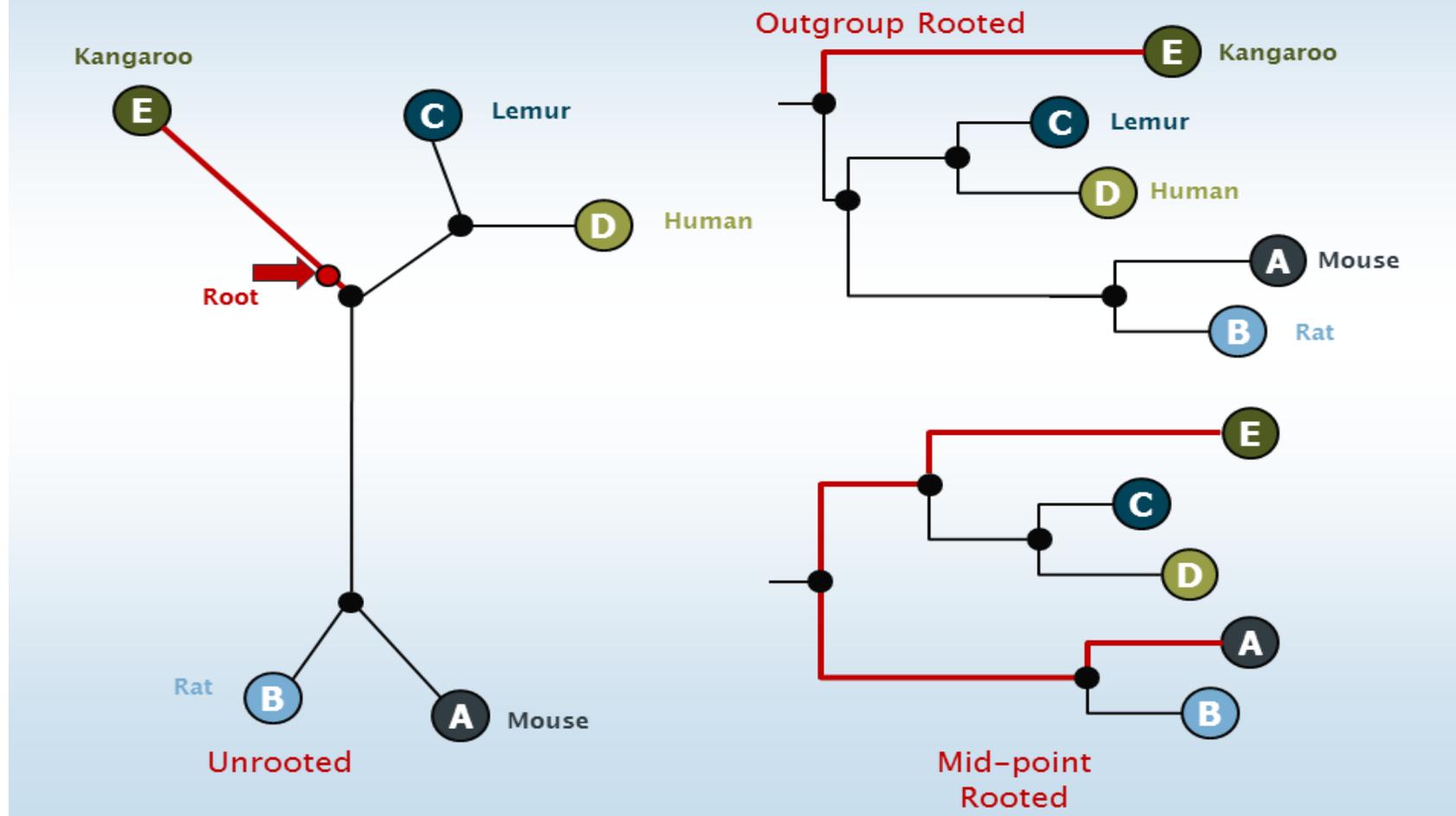
$$B_u = (2n-5)! / 2^{n-3} (n-3)!$$

- Le nombre d'arbres augmente très rapidement avec le nbre de feuilles:
Pour $n=10$, il existe plus de 34 millions d'arbres racinés possibles. Un seul représente la réalité!!

Enracinement

- La plupart des méthodes de reconstruction phylogénétiques produisent des arbres non racinés.
- Pour un arbre non raciné de n feuilles, $2n-3$ enracinements possibles. Plusieurs méthodes existent:
 - **Enracinement au *barycentre***: positionner la racine au milieu du chemin séparant les deux feuilles les plus éloignées. Hypothèse de l'horloge moléculaire. Applicable uniquement aux arbres valués.
 - **Enracinement en utilisant un « outgroup »**. Méthode la plus utilisée. Consiste à rajouter à l'ensemble des séquences des espèces étudiées, une séquence homologue appartenant à une espèce non-apparentée.

Outgroup Rooting



Le kangourou est utilisé comme « outgroup »: Marsupiaux versus mammifères placentaires.

<http://cabbagesofdoom.blogspot.ca/2012/06/how-to-root-phylogenetic-tree.html>

V. Mesures de similarité/dissimilarité entre les arbres

- Plusieurs arbres phylogénétiques peuvent être obtenus pour le même ensemble de taxons.
 - Utilisation de gènes différents ou de parties différentes du génome;
 - Différents modèles d'évolution;
 - Différents algorithmes de reconstruction;
 - Plusieurs arbres statistiquement équivalents
- Comment comparer les arbres?
 - Mesures de distances: **Robinson-Foulds, NNI, STT, quartets.**
 - Mesures de similarité: Structure commune à l'ensemble des arbres. Mesure de similarité populaire: **Maximum Agreement Subtree.**
 - **Consensus d'arbres**

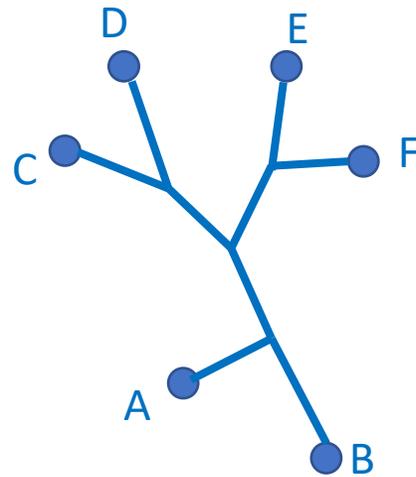
Distance topologique

- Comment comparer deux arbres T_1 , T_2 provenant de données différentes?
- Distance la plus utilisée: **Robinson-Foulds**. Compte le nombre de bipartitions différentes entre T_1 et T_2 .

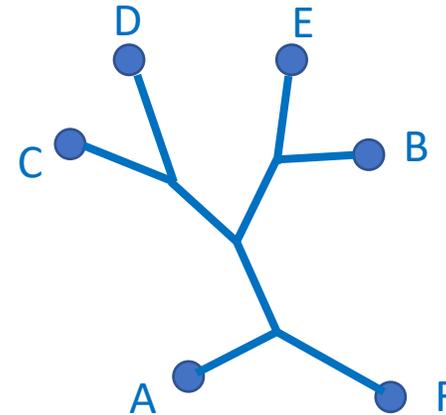
Distance topologique

- Distance la plus utilisée: **Robinson-Foulds**. Compte le nombre de bipartitions (splits) différentes entre T_1 et T_2 .

Bipartitions non-triviales



CD|ABEF EF|ABCD AB|CDEF

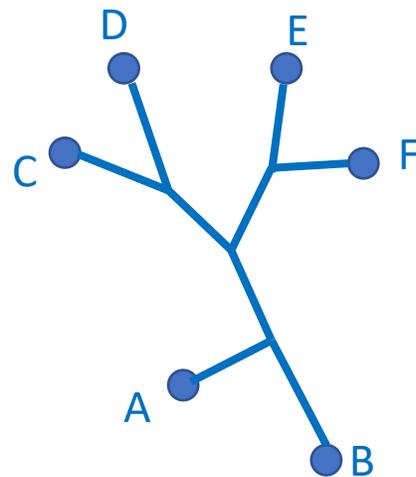


CD|ABEF EB|ACDF AF|BCDE

Distance topologique

- Distance la plus utilisée: **Robinson-Foulds**. Compte le nombre de bipartitions (splits) différentes entre T_1 et T_2 .

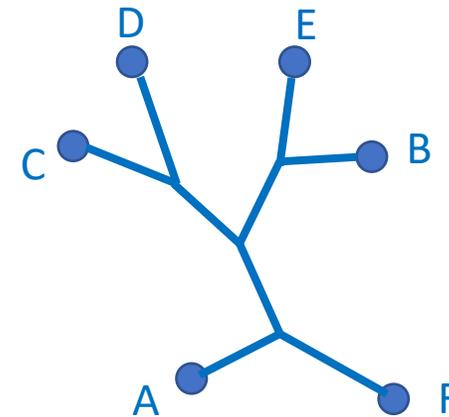
Bipartitions non-triviales



CD|ABEF

EF|ABCD

AB|CDEF



CD|ABEF

EB|ACDF

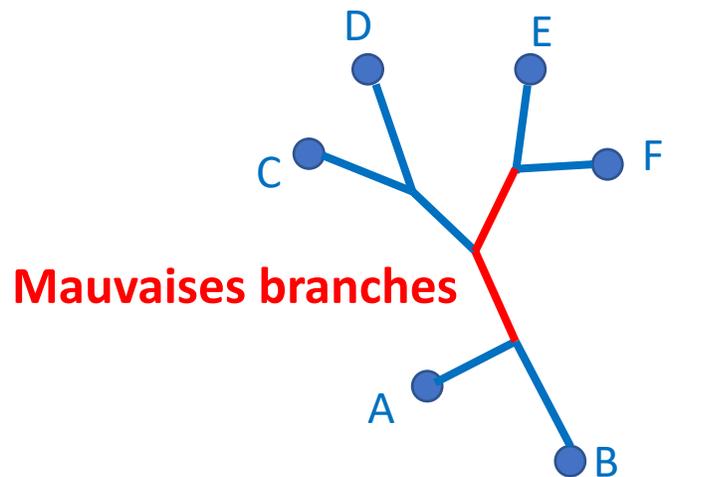
AF|BCDE

Distance topologique $d_T(T_1, T_2) = 4$

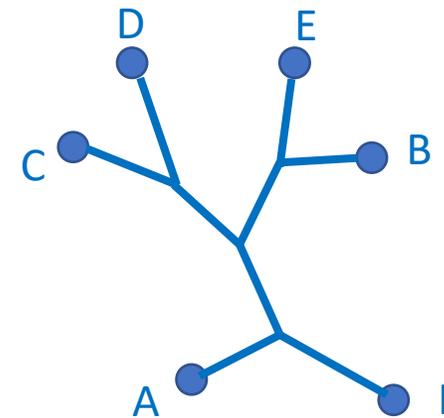
Distance topologique

- Distance la plus utilisée: **Robinson-Foulds**. Compte le nombre de bipartitions (splits) différentes entre T_1 et T_2 .

Bipartitions non-triviales



CD|ABEF EF|ABCD AB|CDEF



CD|ABEF EB|ACDF AF|BCDE

Distance topologique

- Distance la plus utilisée: **Robinson-Foulds**. Compte le nombre de bipartitions différentes entre T_1 et T_2 .
- Un arbre non raciné de n feuilles a $n-3$ branches internes (bi-partitions non-triviales). Donc distance topologique maximale entre deux arbres non racinés est

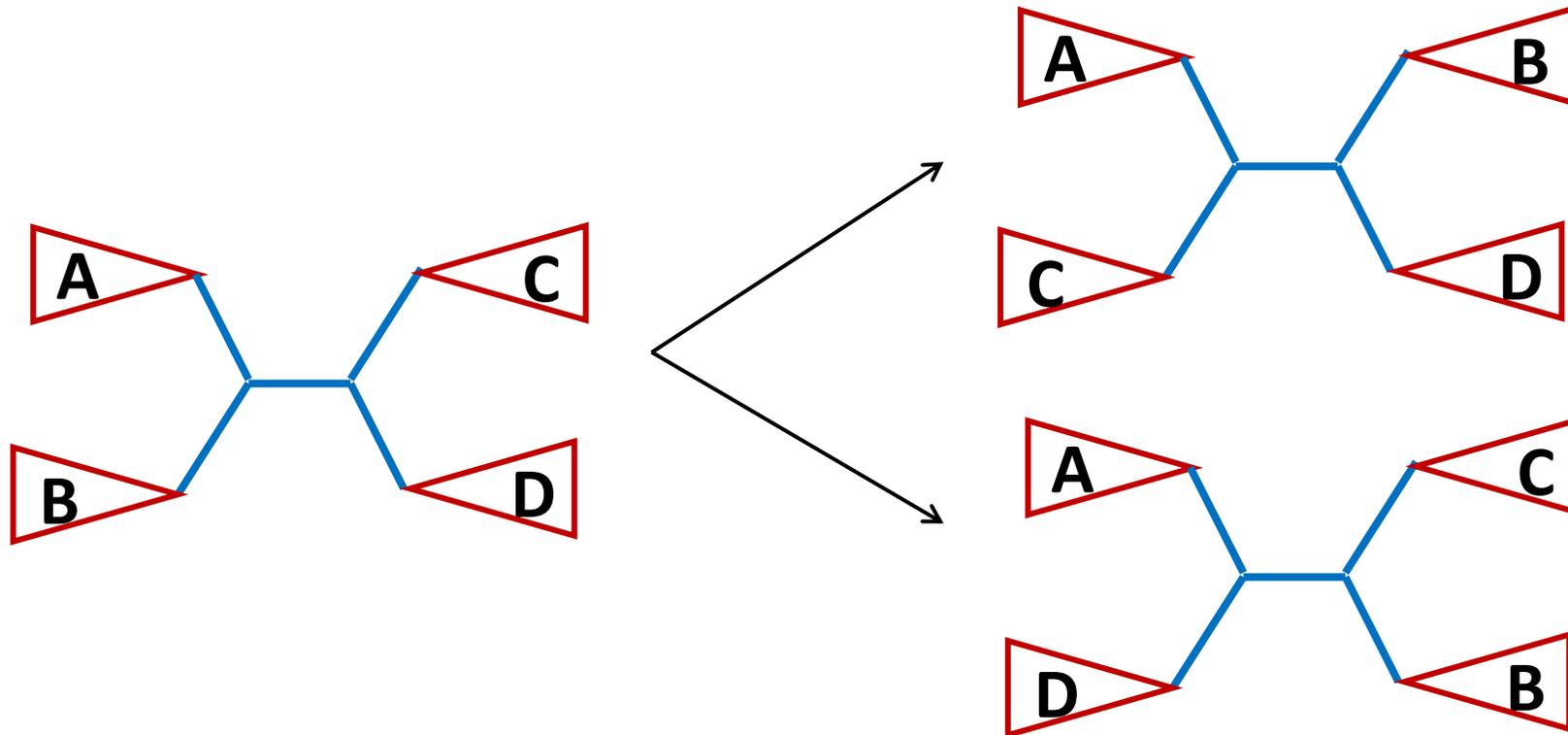
$$d_M(T_1, T_2) = 2(n-3)$$

- *Généralement, la distance topologique est normalisée:*

$$RF(T_1, T_2) = d_T(T_1, T_2) / d_M(T_1, T_2)$$

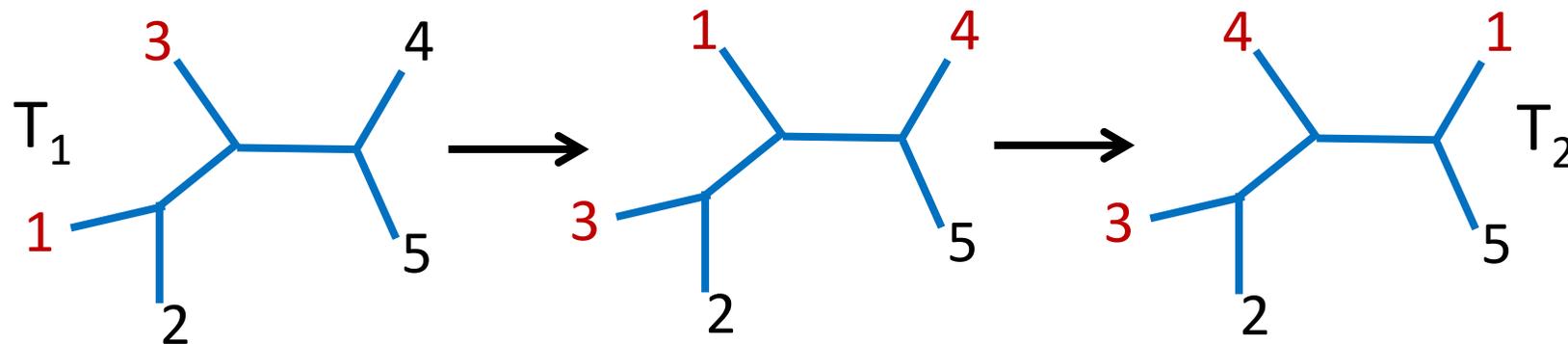
Distance NNI

- NNI “Nearest Neighbor Interchange”: Mouvement permettant d’interchanger deux des sous-arbres incidents à une branche interne. Deux mouvements sont possibles pour chaque branche interne.



Distance NNI

- NNI “Nearest Neighbor Interchange”: Mouvement permettant d’interchanger deux des sous-arbres incidents à une branche interne. Deux mouvements sont possibles pour chaque branche interne.
- Distance NNI entre deux arbres: Nombre minimum de mouvements NNI nécessaire pour transformer un arbre en l’autre.



$$\text{NNI-dist}(T_1, T_2) = 2$$

Distance NNI

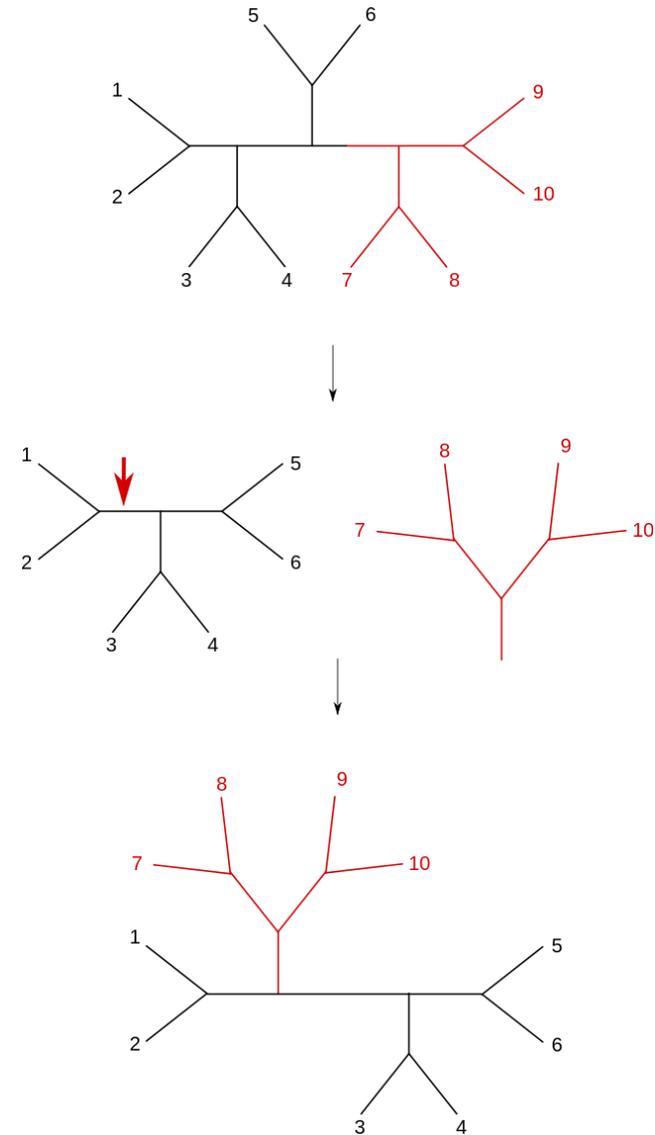
- $\text{NNI-dist}(T_1, T_2) \geq$ nombre de mauvaises branches de T_1 par rapport à T_2 .

En effet, pour supprimer une mauvaise branche, on a besoin d'au moins un NNI.

- Calculer la distance NNI: Problème NP-difficile. Il existe des algorithmes d'approximation.

Autres mouvements

- **Subtree pruning and regrafting (SPR):** Consiste à détacher un sous-arbre et le greffer sur une autre branche de l'arbre.



Autres mouvements

- **Tree bisection and reconnection (TBR):**
Détache un sous-arbre et rebranche une arête de l'arbre initial à une arête de ce sous-arbre.

