Introduction à la Phylogénie

Table des matières

- Introduction à la phylogénie: Dogme central, spéciation, phylogénie, taxonomie.
- Arbres phylogénétiques: définitions formelles.
- Les caractères utilisés, modèles d'évolution, sélection naturelle
- Dénombrement des arbres.
- Comparaison d'arbres: « Maximum agreement subtree », Distance Robinson-Foulds, Mouvements NNI, STT, quartets.
- Construction d'arbres: méthodes de distance, méthodes de parcimonie.

I. Introduction - Phylogénie

- HYPOTHÈSE DE BASE: Tous les êtres vivants descendent d'un ancêtre commun.

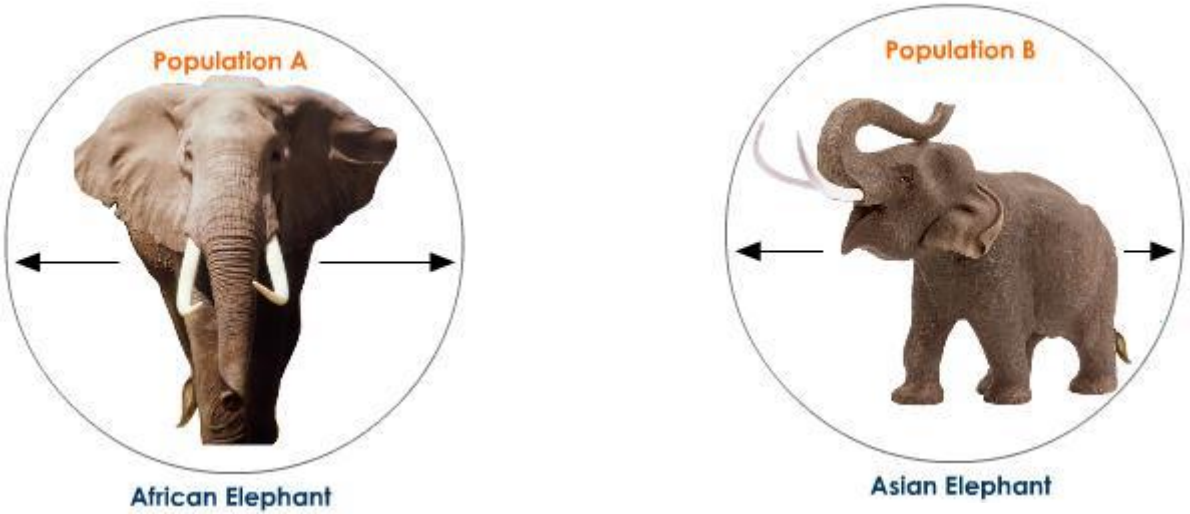
  Sur une période d'au moins 3.8 milliards d'années le premier être vivant sur terre n'a cessé de se séparer en espèces différentes.

- Les êtres vivants évoluent à partir d'un ancêtre commun par une suite de mutations suivies de spéciations.

  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.

Phylogénie

- L'isolement d'une population et l'adaptation à son environnement peut entrainer la création d'une nouvelle espèces

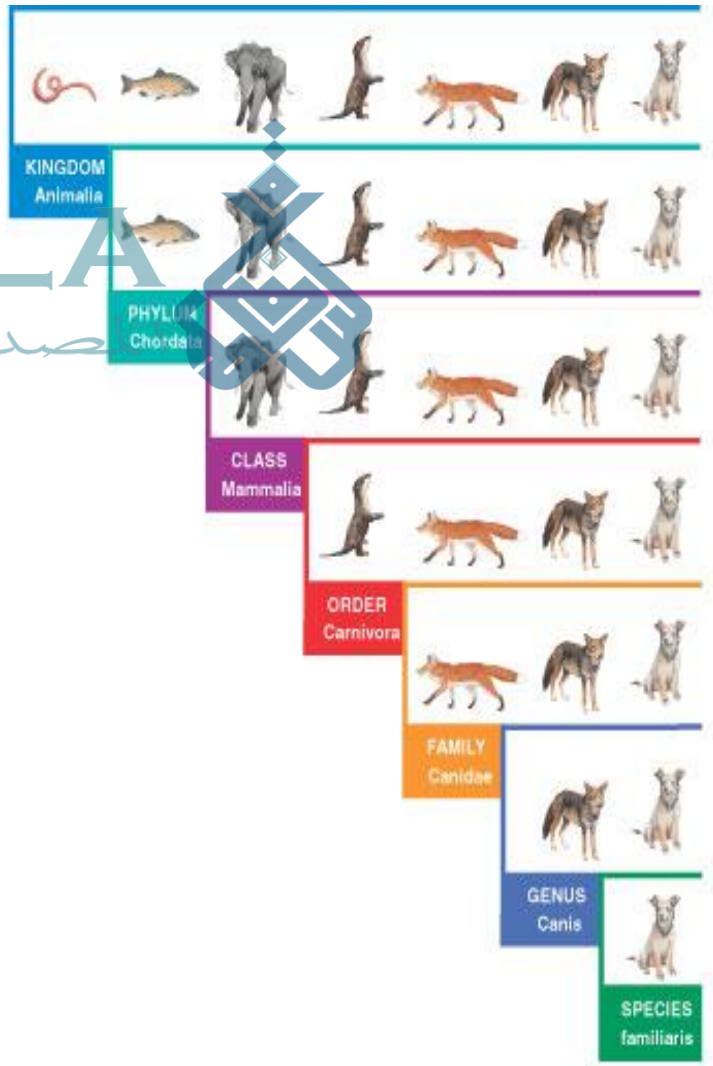


http://www.tutorvista.com/content/biology/biology-iv/biotic-community/speciation.php

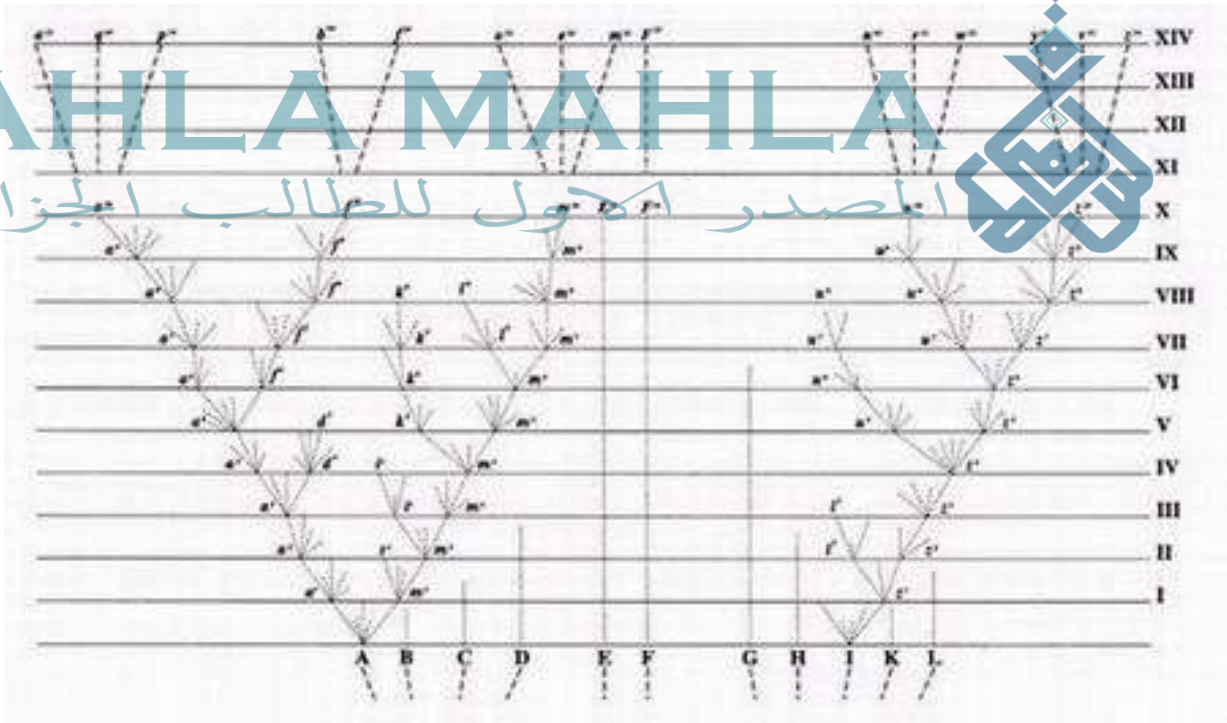
Phylogénie

- Étude des relations d'évolution entre des groupes d'organismes (espèces, populations). Basée sur la notion d' « héritage»
- Taxonomie: Science qui consiste à classifier identifier et nommer les organismes. Basée sur des caractéristiques communes, différentes du reste de la diversité biologique.

  → Domain, Kingdom, Phylum, Class, Order, Family, Genus, and Species



Arbre de Phylogénie

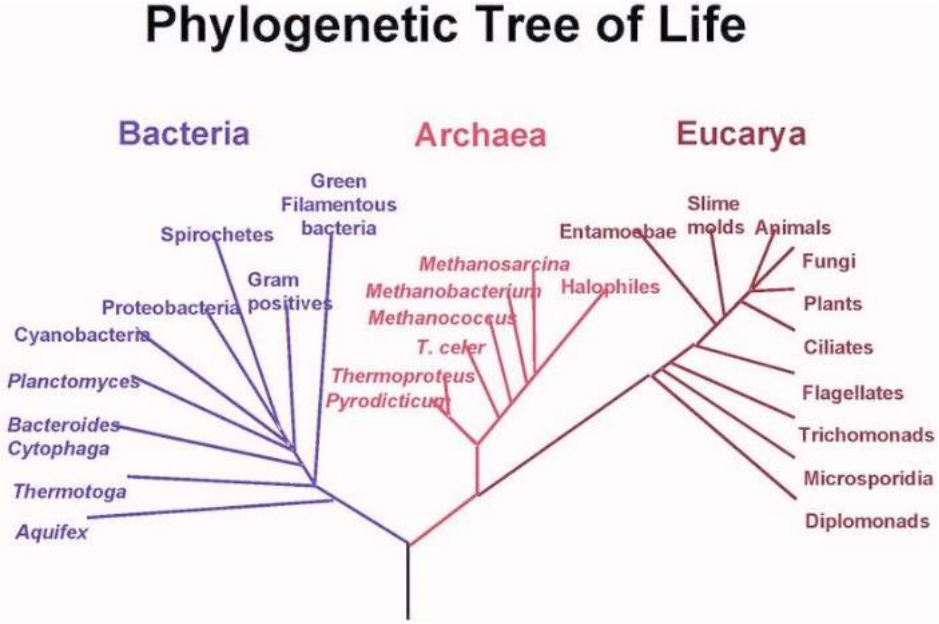


The Tree of Life image that appeared in Darwin's *On the Origin of Species by Natural Selection*, 1859. It was the book's only illustration

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

Arbre de Phylogénie

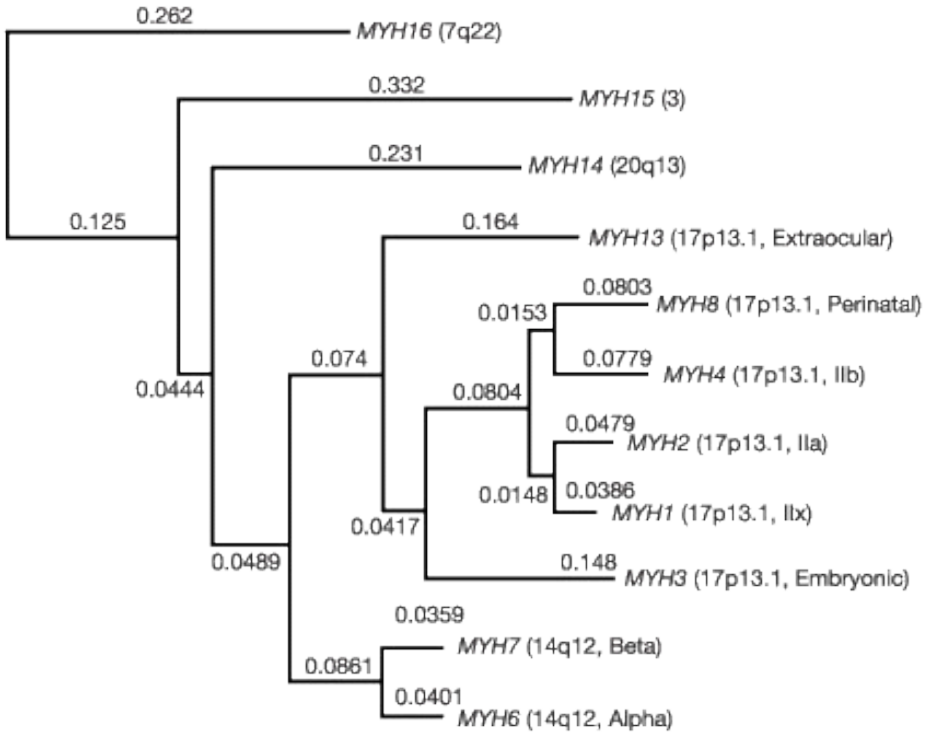
- Premier objectif des études phylogénétiques: Reconstruire l'arbre de vie de toutes les espèces vivantes à partir des données génétiques observées.



NASA:http://www.nasa.gov

Arbre de Phylogénie

- Les arbres de phylogénie sont également utilisés pour représenter l'évolution commune d'une famille de gènes, ou de virus comme le HIV ou l'influenza.

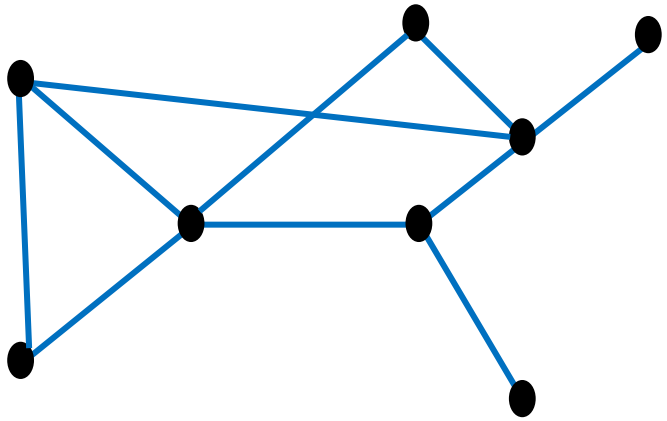


Observation de corrélations entre les mutations du gène Myosin avec certains changements anatomiques dans la lignée humaine. MYH16 chez l'humain très divergeant des autres copies du gène.

http://bio.nyk.ch/Myosin

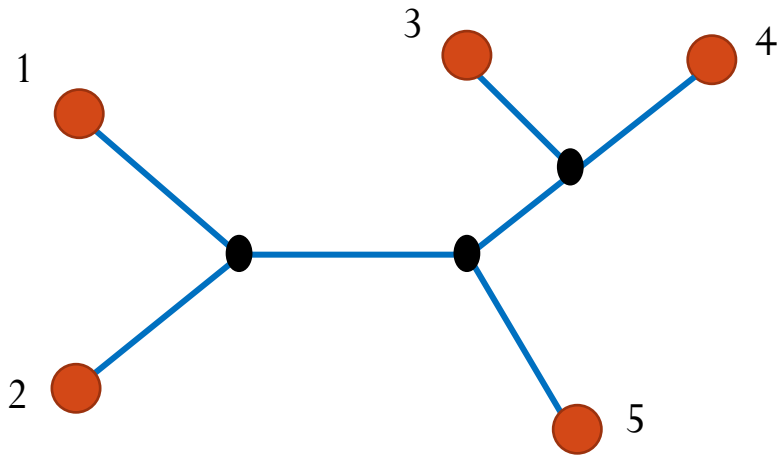
II. Définitions formelles

- 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.



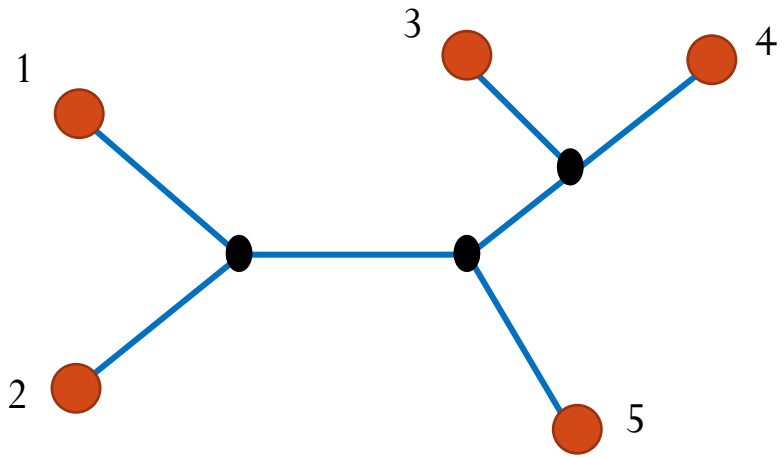
II. Définitions formelles

- Les **feuilles** (nœuds de degré 1) représentent les espèces (ou séquences) actuelles
- Les **nœuds internes** représentent les événements de spéciation



II. Définitions formelles

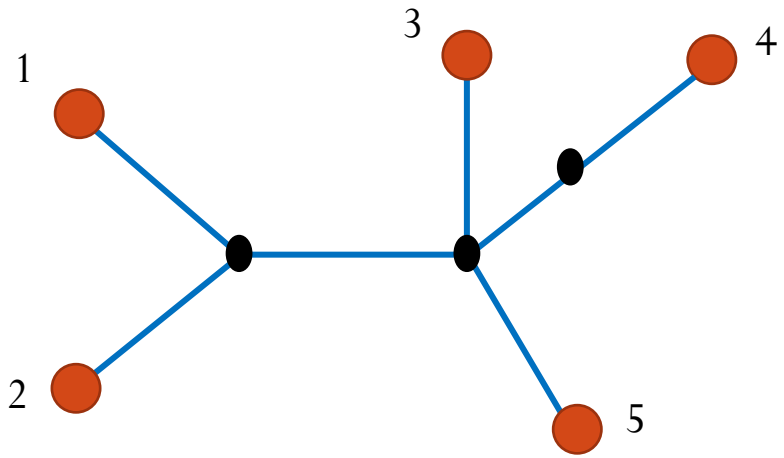
- **Arbre binaire**: Chaque nœud interne de degré 3



binaire

II. Définitions formelles

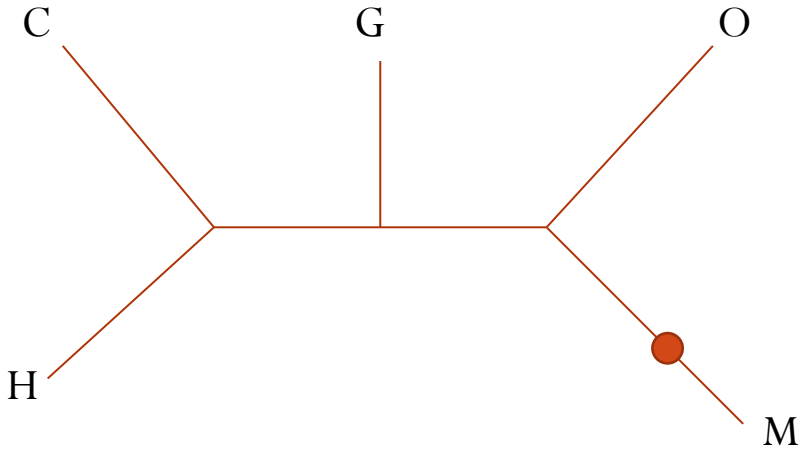
- **Arbre binaire**: Chaque nœud interne de degré 3

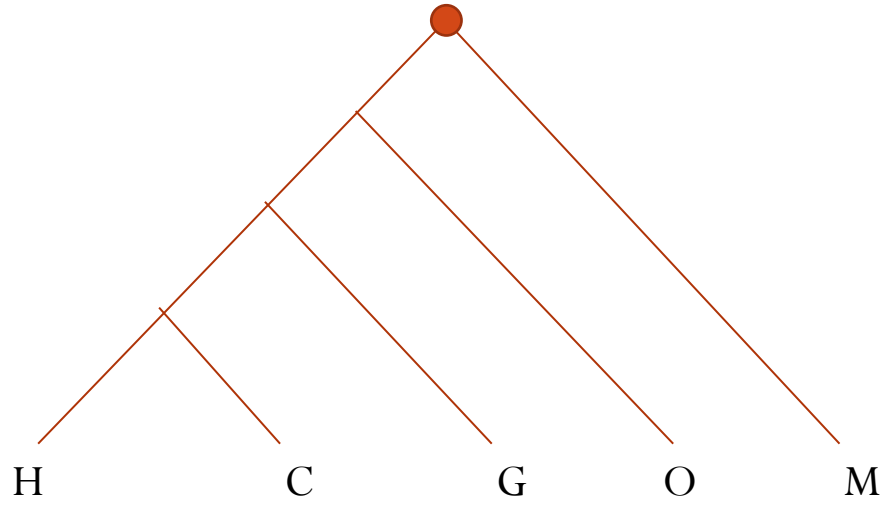


non-binaire

II. 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 raciné**: 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.

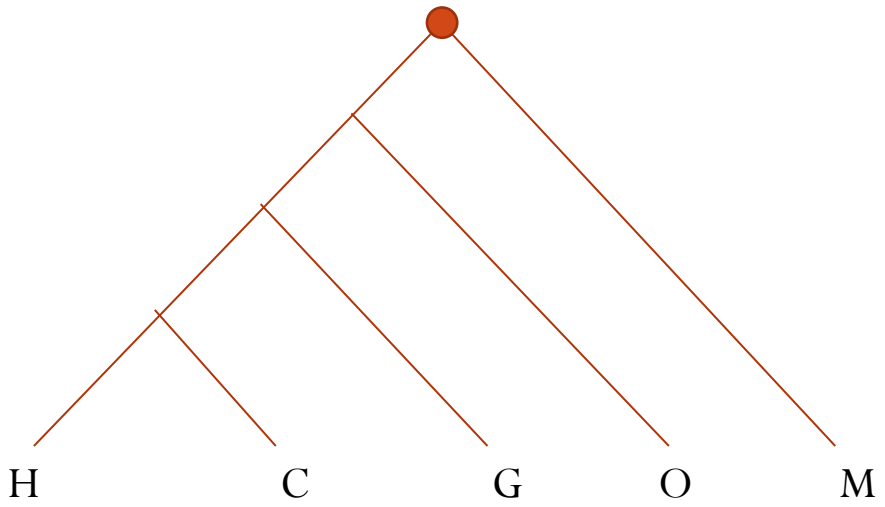




Définitions formelles

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

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

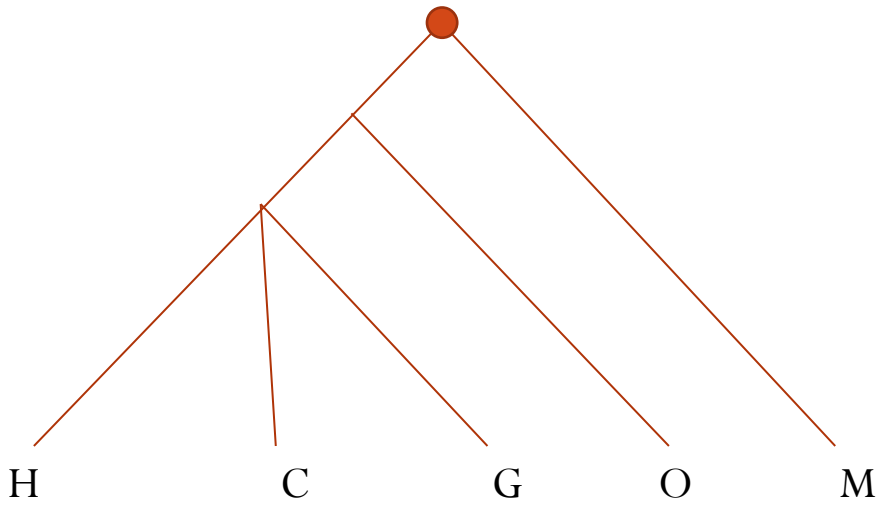


binaire

Définitions formelles

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

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

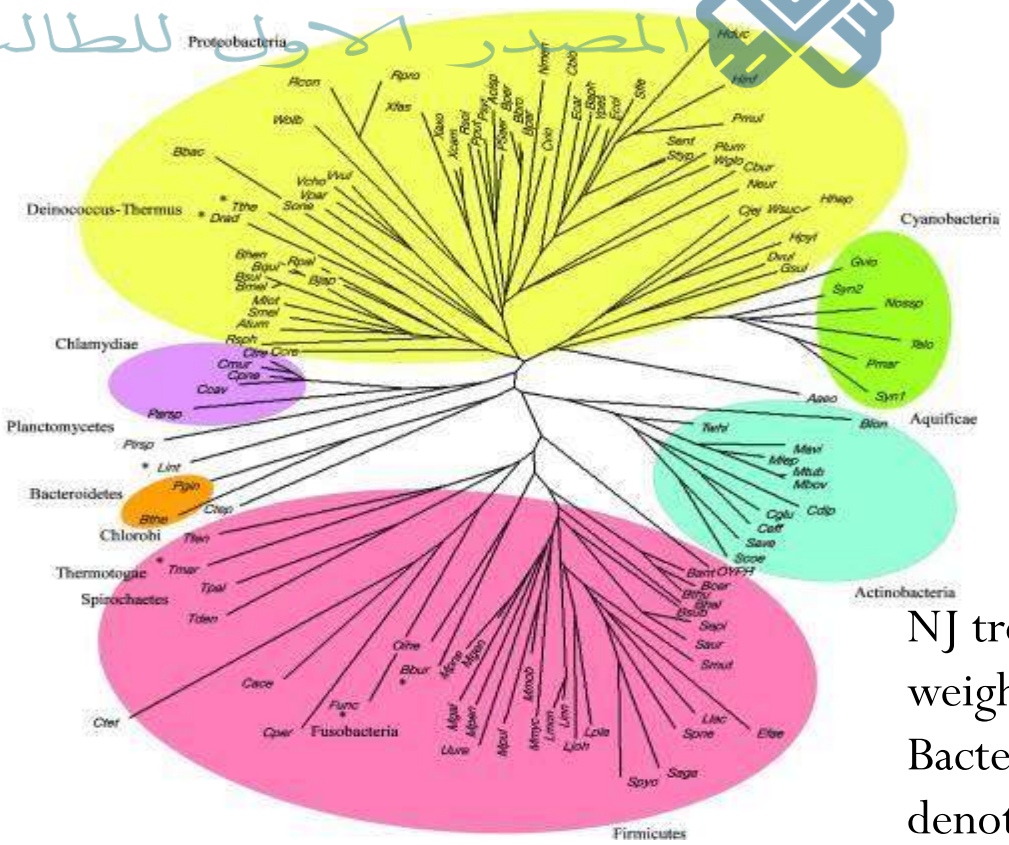


non-binaire

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-scpécifié, les arbres sont considérés binaires

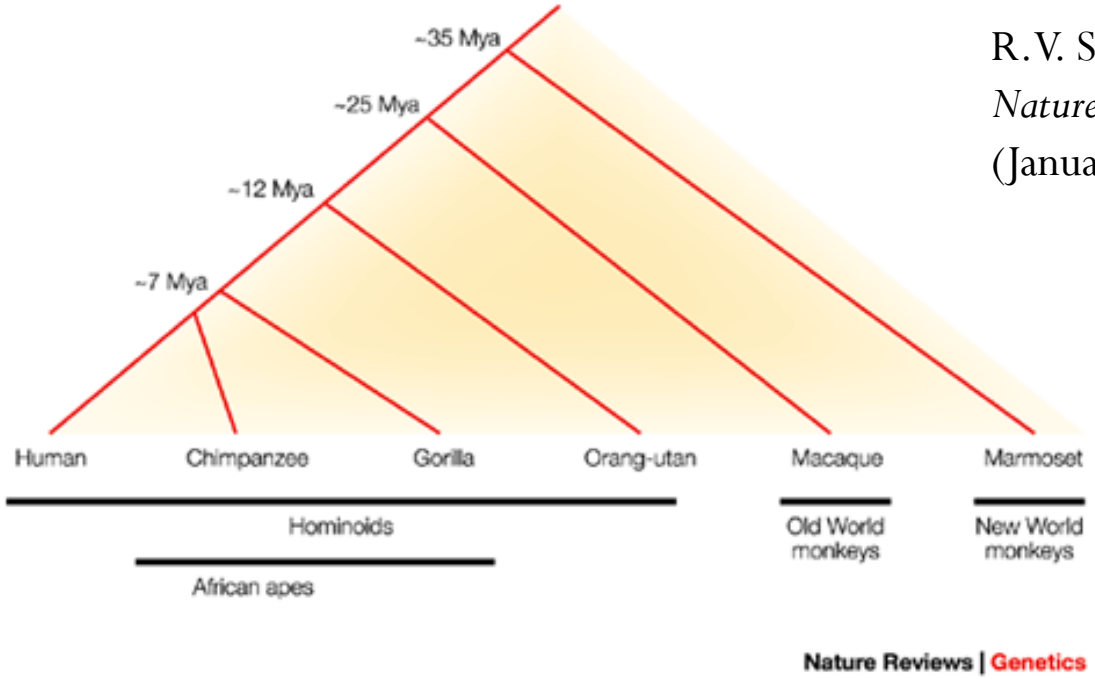


NJ tree (with weighting) of 119 Bacteria. Asterisks denote anomalously positioned taxa.

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC540256/figure/fig3/

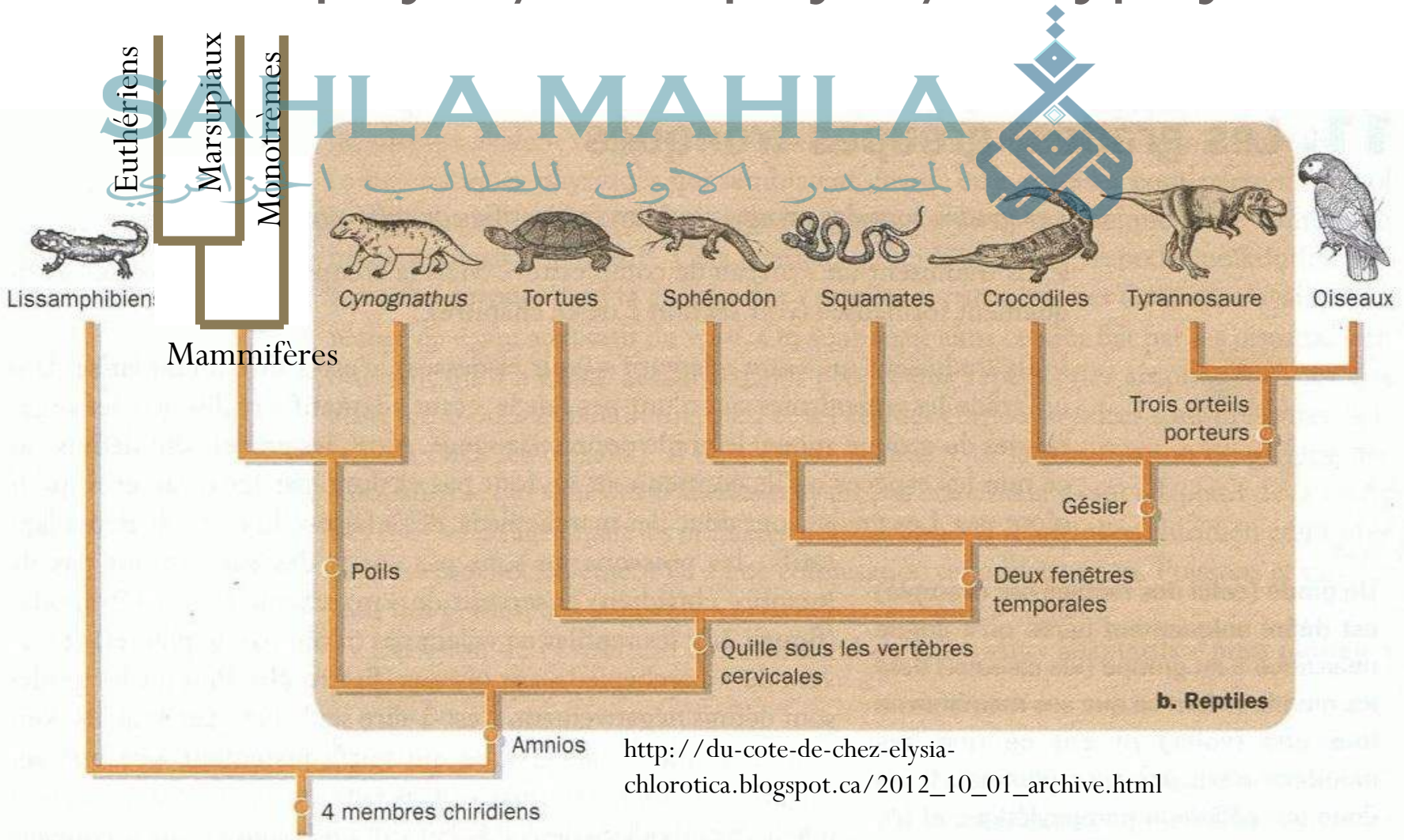
Définition formelle

- 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)

Monophylie/Paraphylie/Polyphylie



http://du-cote-de-chez-elysia-chlorotica.blogspot.ca/2012_10_01_archive.html

Monophylie/Paraphylie/Polyphylie

T: arbre raciné. Soit *M* un groupe d'espèces (actuelles et ancestrales)

- M Groupe Monophylétique si le LCA *e* de *M*, ainsi que tous ses descendants sont dans M. Autrement dit, *M* détermine un sous-arbre de *T*.

  Exemple dans l'arbre des tétrapodes: Mammifères

- M Groupe Paraphylétique si le LCA *e* de *M* est dans *M*, mais que *M* n'est pas complet, i.e. n'inclue pas toutes les espèces du sous-arbres de racine *e*.

  Les Reptiles

- M Groupe Polyphylétique si le LCA de *M* n'est pas dans *M*.

Les tétrapodes à sans chaud ou héméothermes (Mammifères et oiseaux). L'ancêtre des amniotes n'était pas héméotherme.

III. Caractères et 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.

Les caractères ou marqueurs utilisés

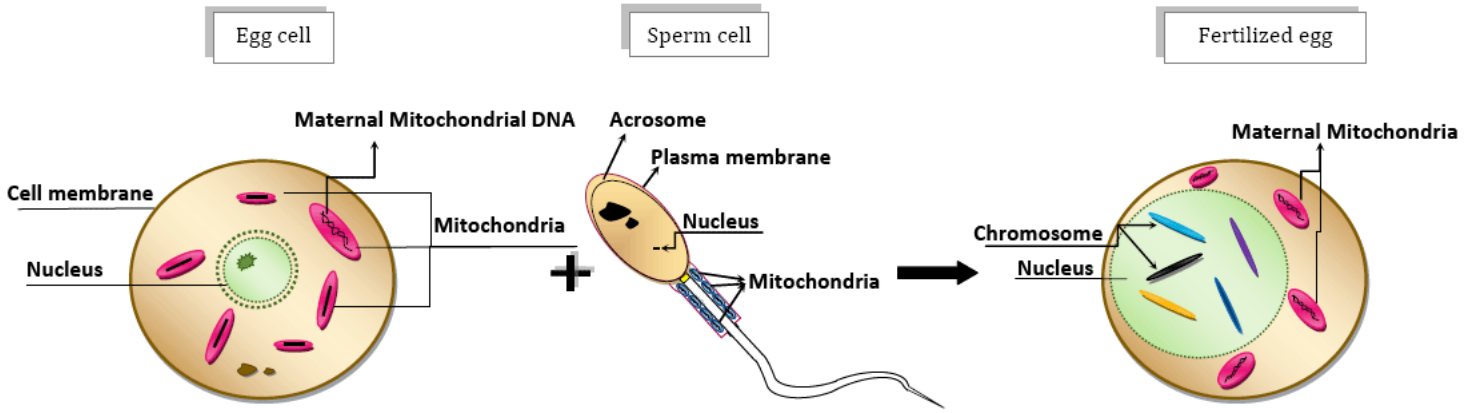
- Caractères les plus utilisés pour les études d'évolution: Séquences de nucléotides ou d'AA.
  - Séquences orthologues dans les espèces étudiées
  - Effectuer un alignement multiple des séquences
  - Les caractères représentés par les colonnes de l'alignement et les états du caractère sont les nucléotides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

Choix de marqueurs (séq. d'ADN)

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

- Marqueur « non-recombinant ». Pour éviter ce problème, 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

Choix de marqueurs (séq. d'ADN)

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

- Marqueur « non-recombinant ». Pour éviter ce problème, choisir des marqueurs uni-parentaux, comme les seq. de mitochondries et de chloroplastes: transmission par la mère uniquement.
- 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.

Modèles d'évolution moléculaire

- 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 0.25
- 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 ←→ G, C ←→ T
  - Transversions: A ←→ T, T ←→ G, A ←→ C, C ←→ G

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

- κ rapport entre le taux de transitions et le taux de transversions.

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.

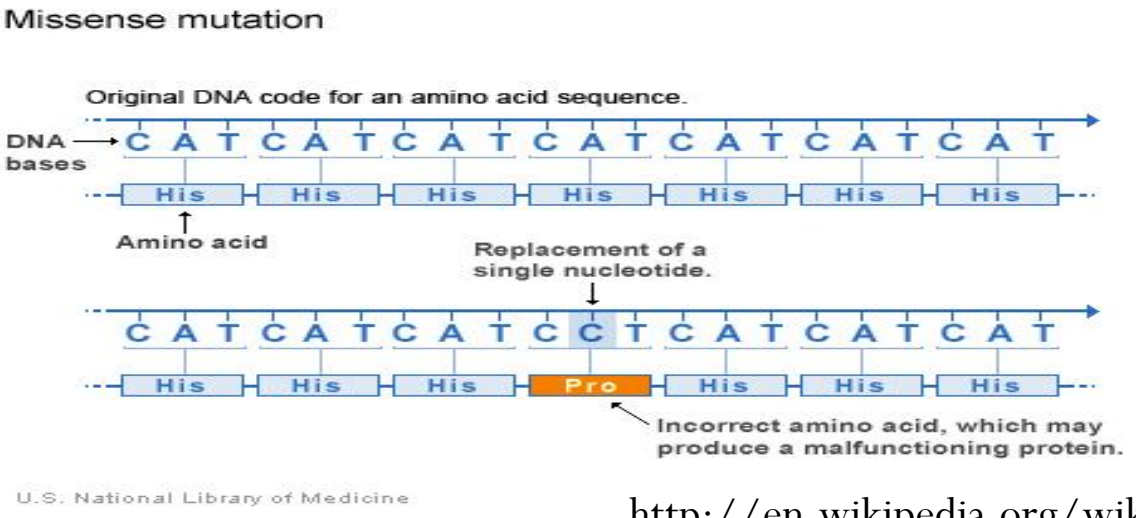
Distance synonyme/non-synonyme pour les séquences codantes

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 non-synonyme (non-silencieuse): substitution provoquant la modification d'un acide aminé.
- Substitution synonyme (silencieuse): substitution ne provoquant pas la substitution de l'acide aminé initial.



http://en.wikipedia.org/wiki/Missense_mutation

le code génétique

Première lettre (côté 5')	Deuxième lettre U		Deuxième lettre C		Deuxième lettre A		Deuxième lettre G		Troisième lettre (côté 3')
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

Distance synonyme/non-synonyme pour les séquences codantes

- Sites synonymes et non-synonymes:
  - 100% des mutations touchant la 2ème 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 1ère base et 28% des mutations touchant la 3ème base sont non-synonymes.
- Distances synonymes et non-synonymes:
  - d_S (aussi notée K_S) distance synonyme entre deux séquences codantes: Nbre de substitution synonymes s'étant produites par site synomyme
  - d_N distance non-synonyme: Nbre de subs. non-synonymes par sire non-synonyme.

Distance synonyme/non-synonyme pour les séquences codantes

Identification du type de sélection:

- Sélection négative:

  Déficit de substitutions non-synonymes attendu

  → $d_N/d_S < 1$

- Sélection neutre:

  Aucun déficit en subst. non-synonymes attendu

  → $d_N/d_S \approx 1$

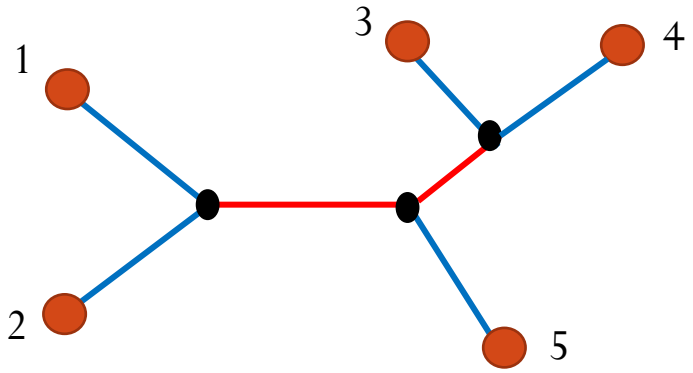
- Sélection positive:

  Excès de subst. non-synonymes attendu

  → $d_N/d_S > 1$

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;

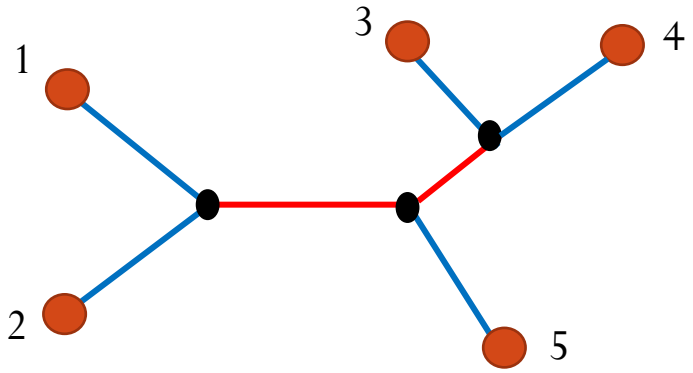
3 noeuds internes;

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**.



1 | 2345
2 | 1345
3 | 1245
4 | 1235
5 | 1234

12 | 345
125 | 34

Bipartitions non-triviales

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**.

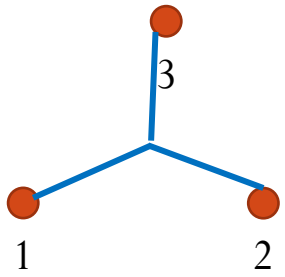
n=2:

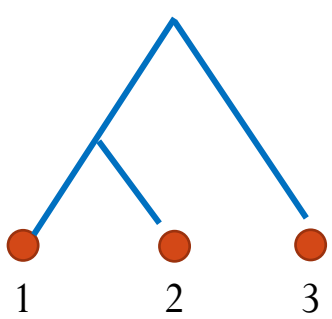
- Arbre non raciné unique
- Arbre raciné unique

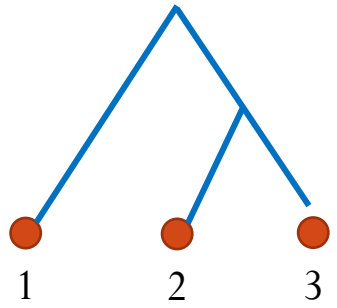
L'arbre caché dans la forêt

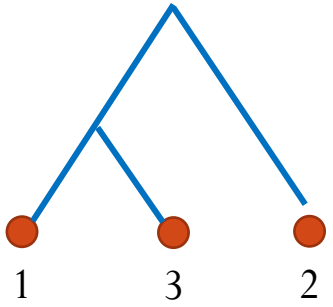
- Arbre non raciné (binaire) de n feuilles: n-2 nœuds internes, n-3 branches internes, et 2n-3 branches. Chaque branche définit une *bipartition* de l'ensemble des feuilles. Arbre définit par n-3 bipartitions non-triviales.

n=3:







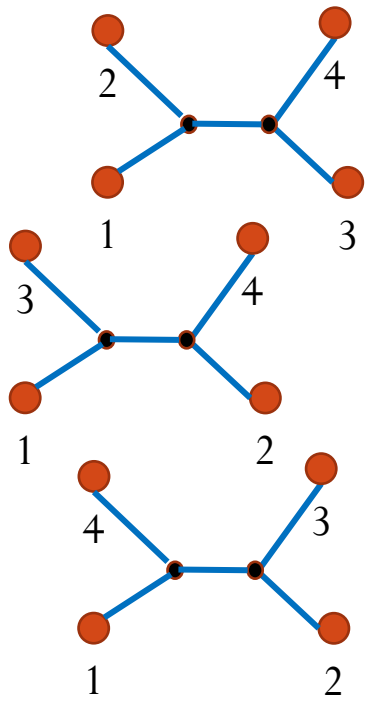


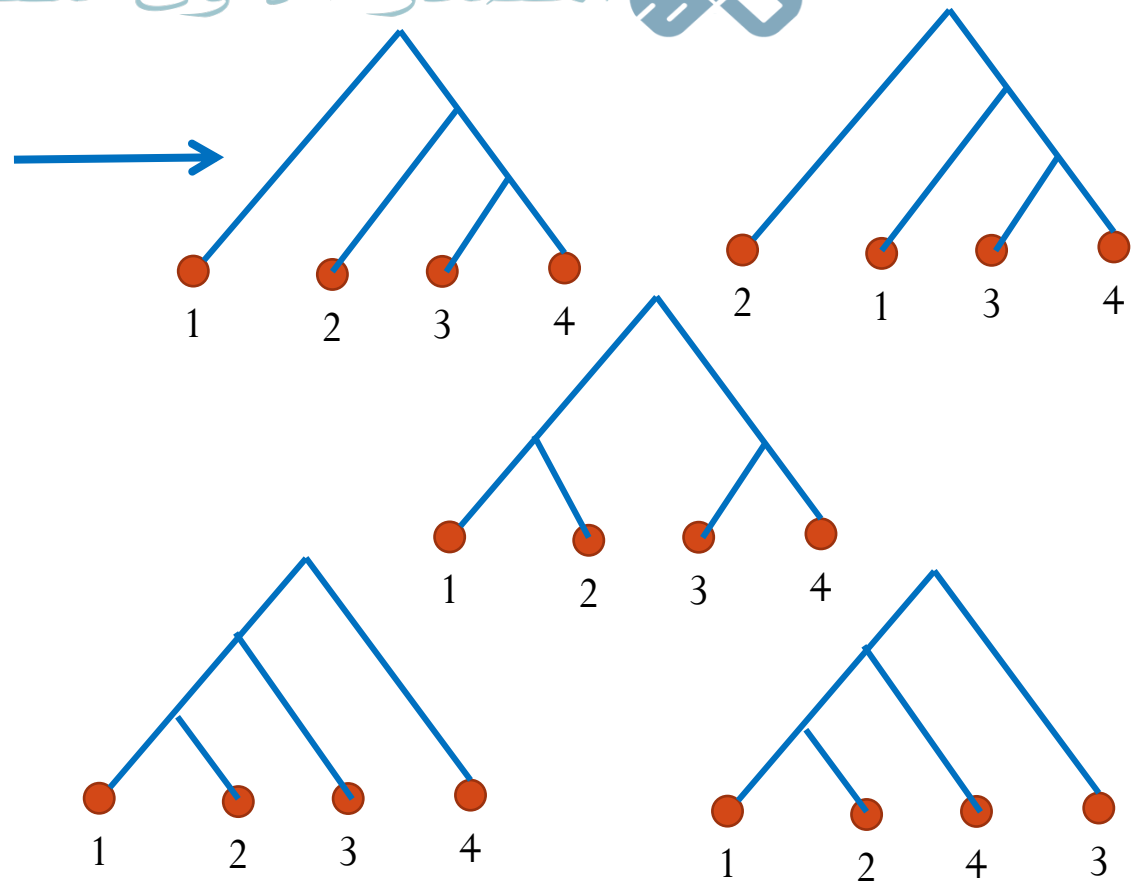
- Arbre non raciné unique
- 3 arbres racinés

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=4:





- 3 Arbre non racinés
- 15 arbres racinés

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 Edwars (1967) ont montré que le nombre B_r d'arbres racinés à *n* feuille est:

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

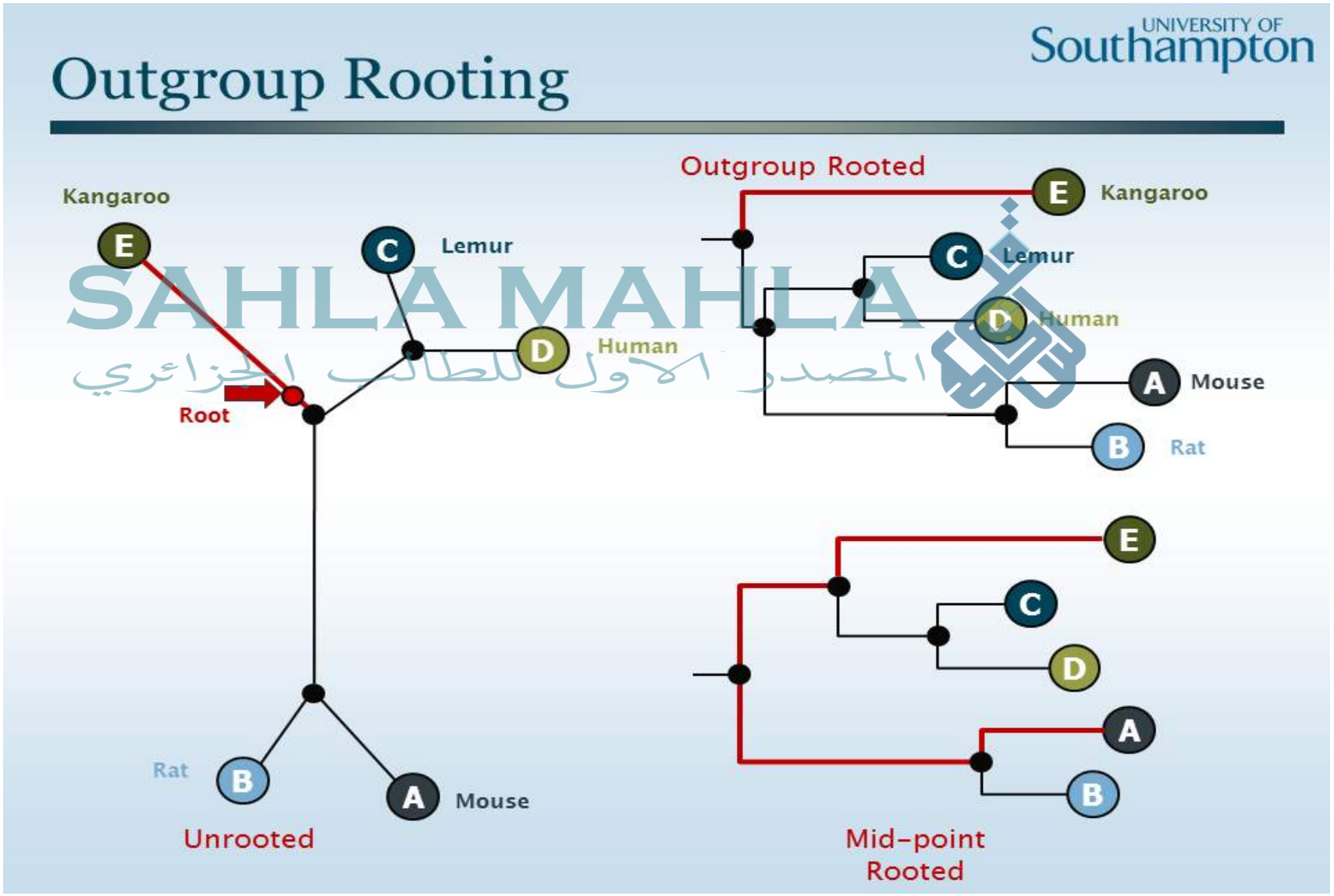
- Le nombre B_u d'arbres non racinés à n feuilles est égal au nombre d'arbres racinés à n-1 feuilles, donc:

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

- 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.



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: MAST.
  - Consensus d'arbres

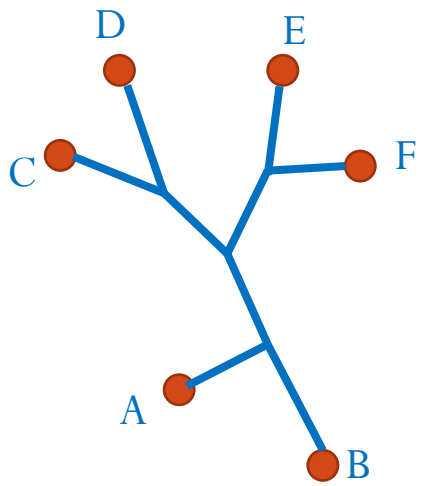
Distance topologique

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

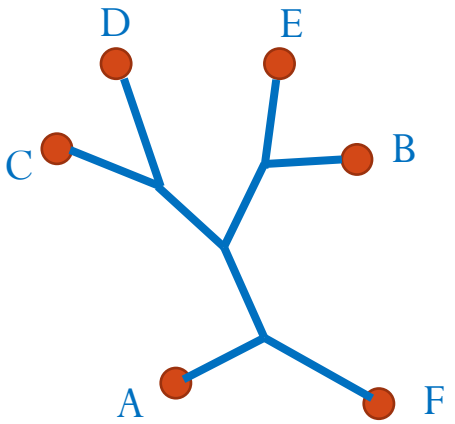
Distance topologique

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

Bipartitions non-triviales



CD | ABEF EF | ABCD AB | CDEF

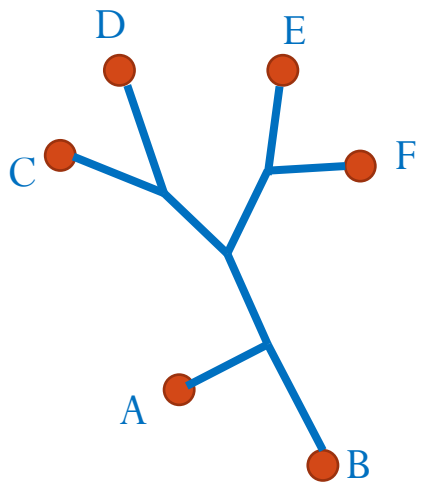


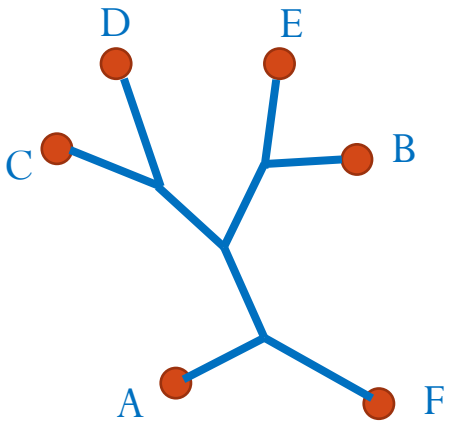
CD | ABEF EB | ACDF AF | BCDE

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**

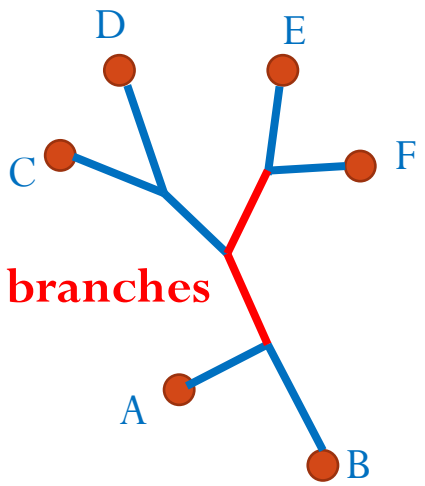
CD | ABEF **EB | ACDF** **AF | BCDE**

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

Distance topologique

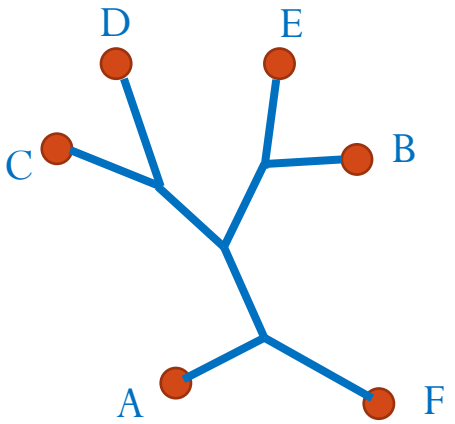
- Distance la plus utilisée: Robinson-Foulds. Compte le nombre de bipartitions (splits) différentes entre T_1 et T_2 .

Bipartitions non-triviales



Mauvaises branches

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 entres deux arbres non racinés est

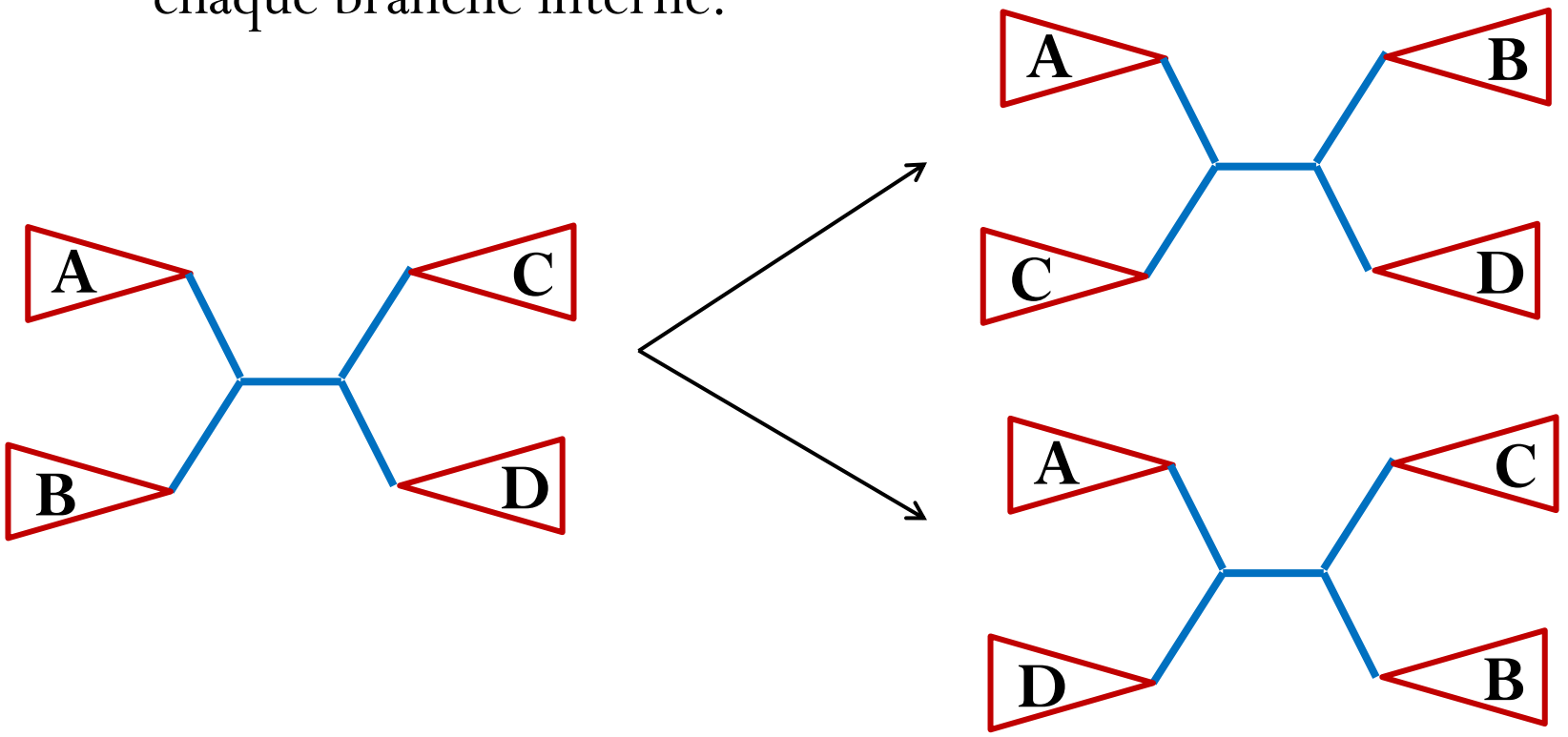
$$d_M(T_1, T_2) = 2(n-3)$$

- *Généralement, la distance tolologique est normalisée:*

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

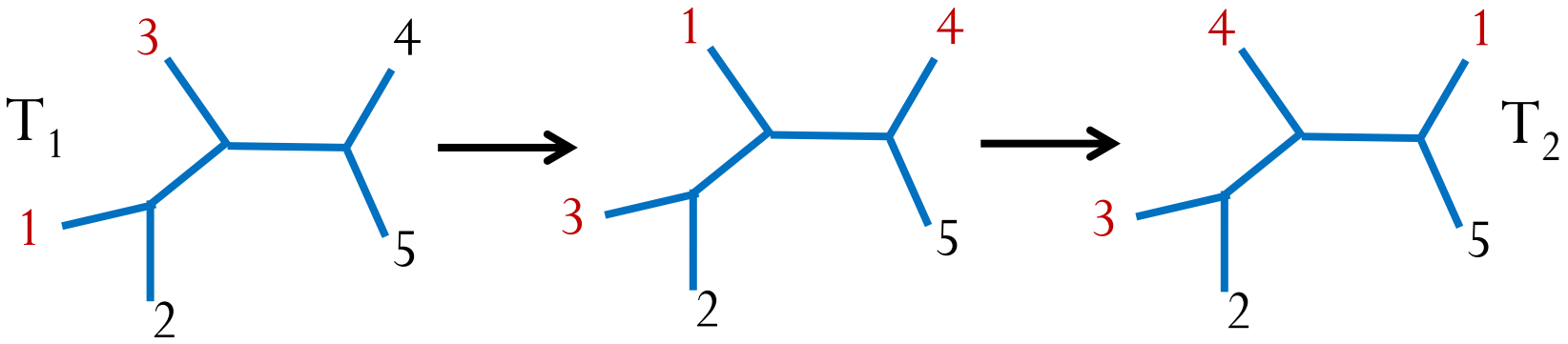
Distance NNI

- 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.



NNI-dist (T_1,T_2) = 2

Distance NNI

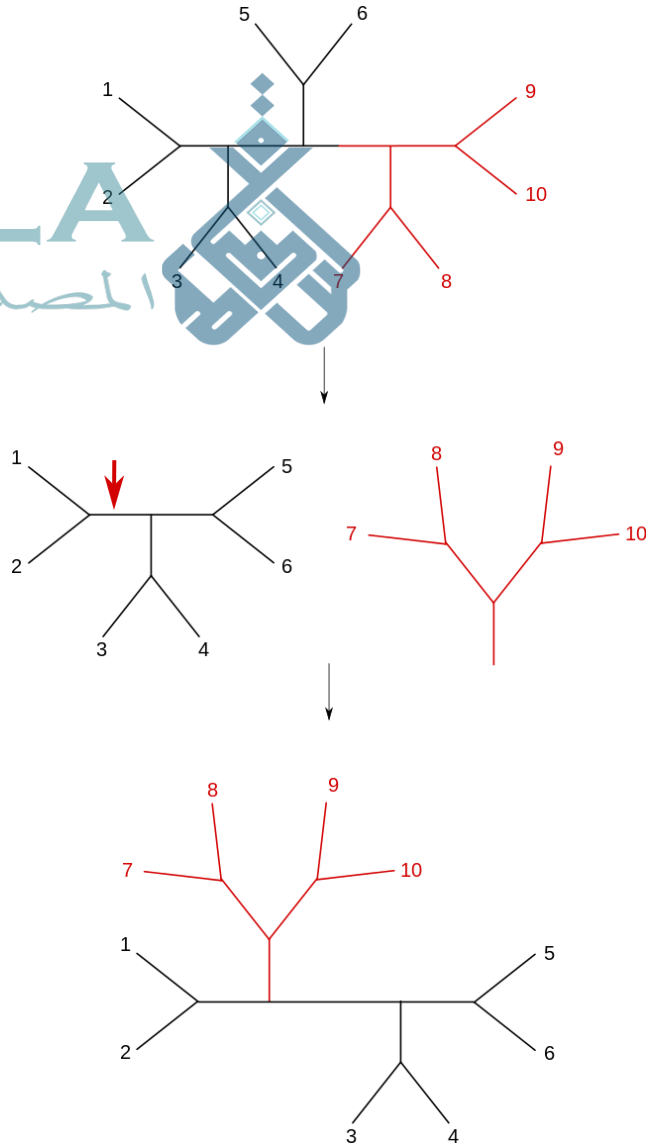
- NNI-dist $(T_1, T_2) \geq$ nombre de mauvaises branches de T_1 par rapport à T_2.

En effet, pour supprimer une mauvaise branche, on a besoin d'au mois 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. le greffer sur une autre branche de l'arbre.