# Méthodes de reconstruction des Phylogénies

Olivier Lespinet

*olivier.lespinet@igmors.u-psud.fr*

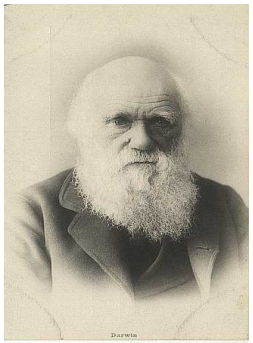
http://olivier-lespinet.info

Equipe Evolution Moléculaire et Bioinformatique des Génomes

Institut de Biologie Intégrative de la Cellule (I2BC) - Bâtiment 400 - Université Paris-Sud

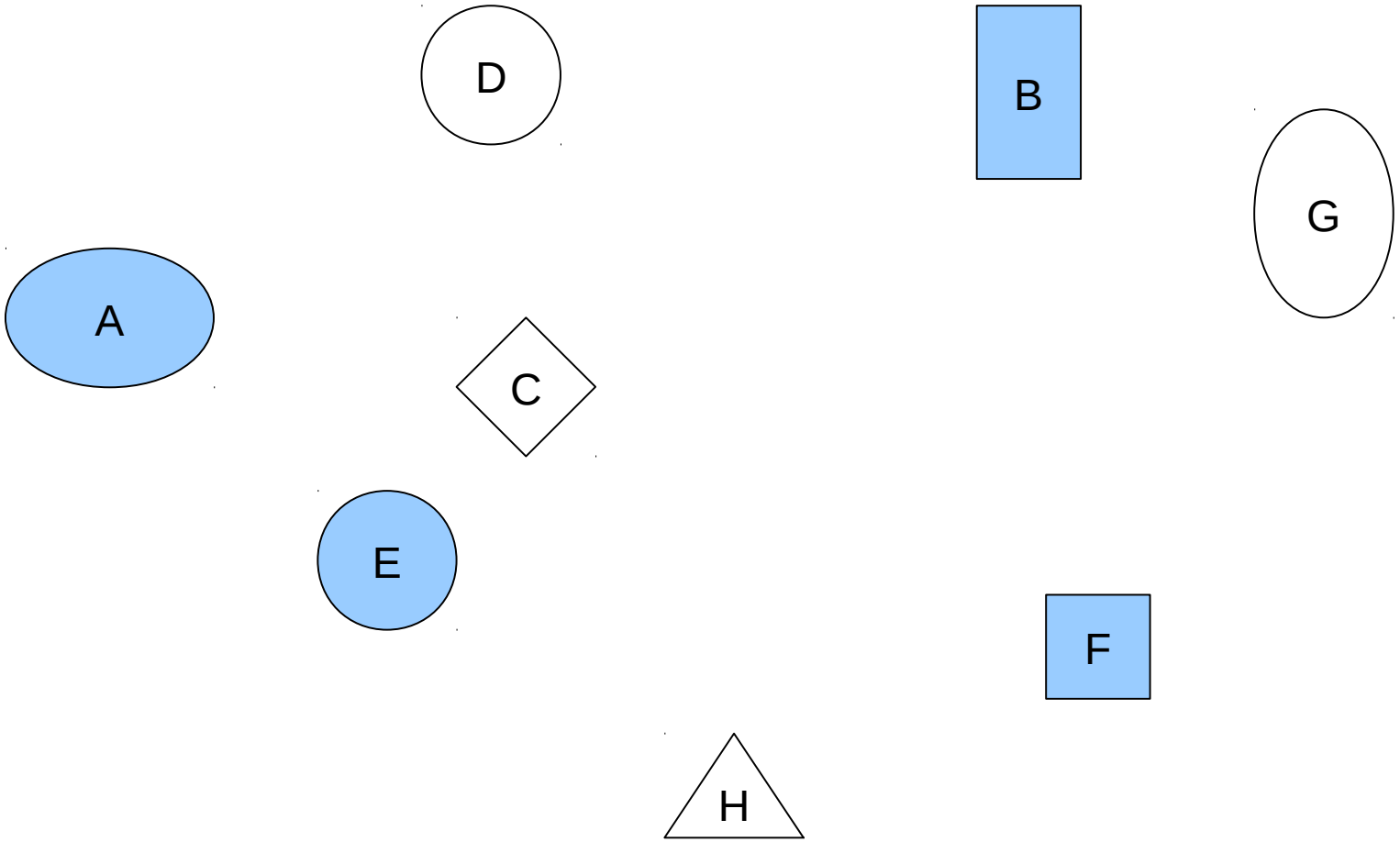
91405 Orsay Cedex

# Comment déterminer les liens de parenté entre plusieurs organismes ?

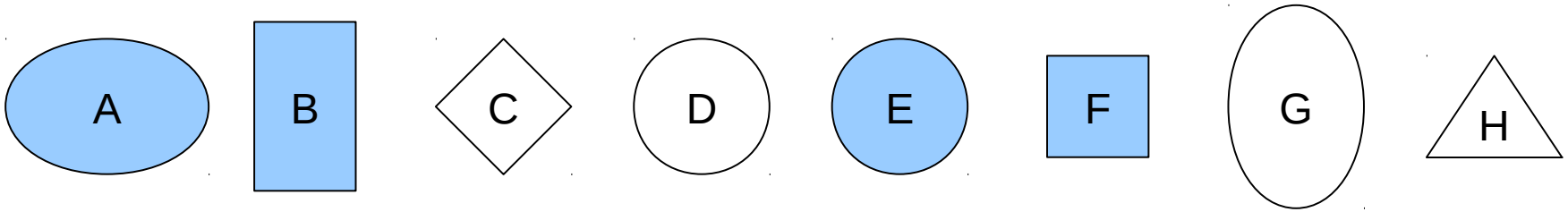


Sont-ils de la même famille ? Du même groupe ? de la même espèce ?

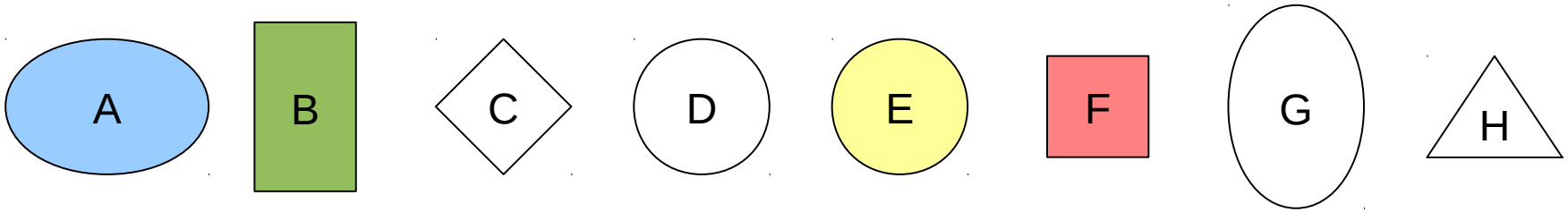
Quelles sont les deux entités les plus ressemblantes ?



# La notion de caractère et d'état possible pour un caractère




| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| Couleur | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 |
| Forme | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 |




| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| Couleur | 1 | 2 | 0 | 0 | 3 | 4 | 0 | 0 |
| Forme | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 |

Comment proposer un scénario évolutif à partir de l'observation de caractères ?

Comment représenter les relations évolutives entre plusieurs espèces ?

Comment représenter les relations évolutives
entre plusieurs espèces ?

Jean-Baptiste Pierre Antoine de Monet
Chevalier de Lamarck
(Bazentin, 1744 – Paris, 1829)

# PHILOSOPHIE
## ZOOLOGIQUE,

OU

## EXPOSITION

Des Considérations relatives à l'histoire naturelle des Animaux ; à la diversité de leur organisation et des facultés qu'ils en obtiennent ; aux causes physiques qui maintiennent en eux la vie et donnent lieu aux mouvemens qu'ils exécutent ; enfin, à celles qui produisent, les unes le sentiment, et les autres l'intelligence de ceux qui en sont doués ;

PAR J.-B.-P.-A. LAMARCK,

Professeur de Zoologie au Muséum d'Histoire Naturelle, Membre de l'Institut de France et de la Légion d'Honneur, de la Société Philomatique de Paris, de celle des Naturalistes de Moscou, Membre correspondant de l'Académie Royale des Sciences de Munich, de la Société des Amis de la Nature de Berlin, de la Société Médicale d'Emulation de Bordeaux, de celle d'Agriculture, Sciences et Arts de Strasbourg, de celle d'Agriculture du département de l'Oise, de celle d'Agriculture de Lyon, Associé libre de la Société des Pharmaciens de Paris, etc.

TOME PREMIER.

A PARIS,

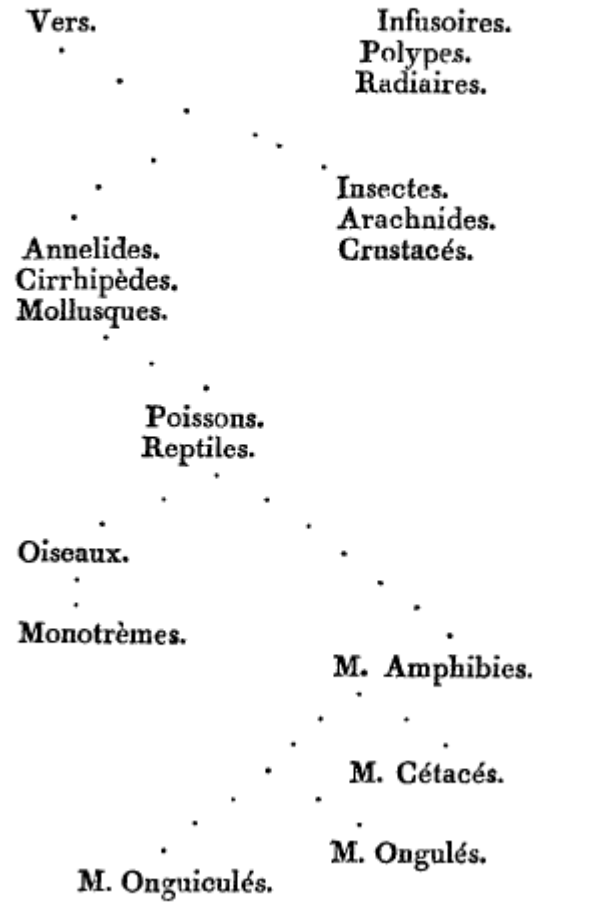
Chez { DENTU, Libraire, rue du Pont de Lodi, N°. 3;
L'AUTEUR, au Muséum d'Histoire Naturelle (Jardin des Plantes).

M. DCCC. IX.



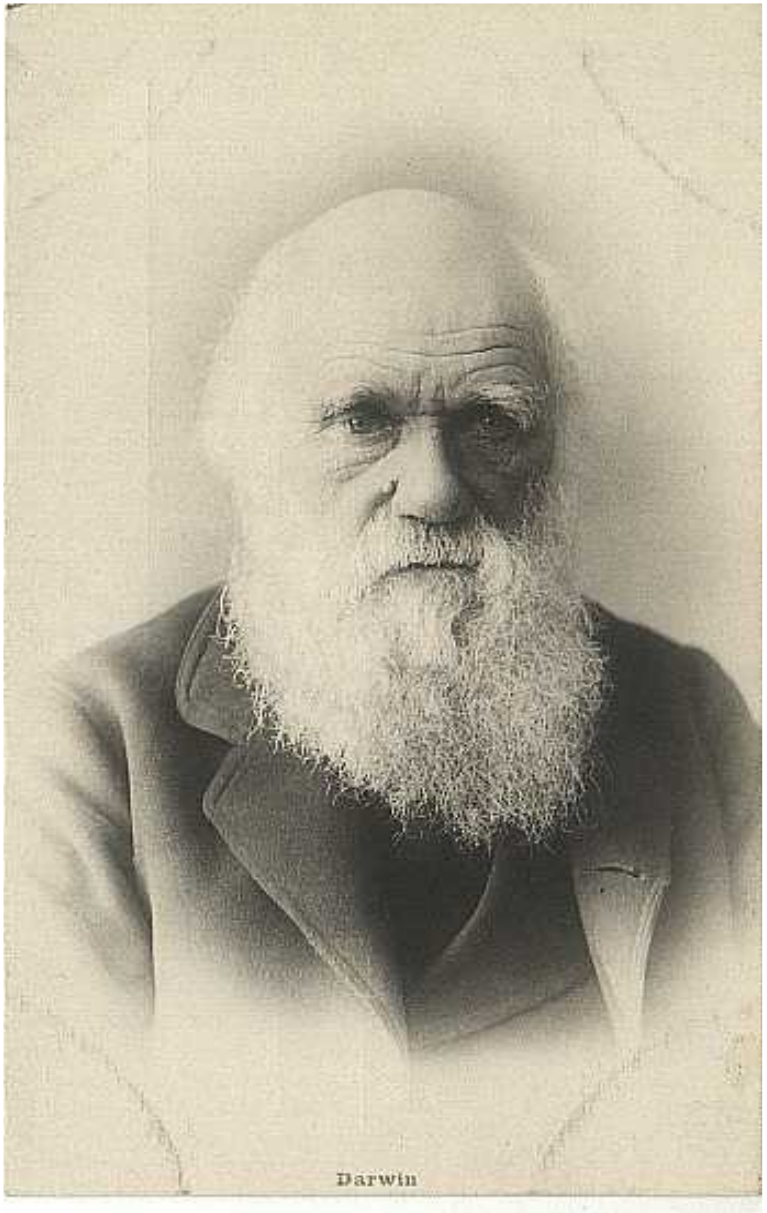
## TABLEAU

*Servant à montrer l'origine des différens animaux.*




Cette série d'animaux commençant par deux

Charles Robert Darwin
(Shrewsbury 1809- Downe 1882)



ON

# THE ORIGIN OF SPECIES

## BY MEANS OF NATURAL SELECTION,

OR THE

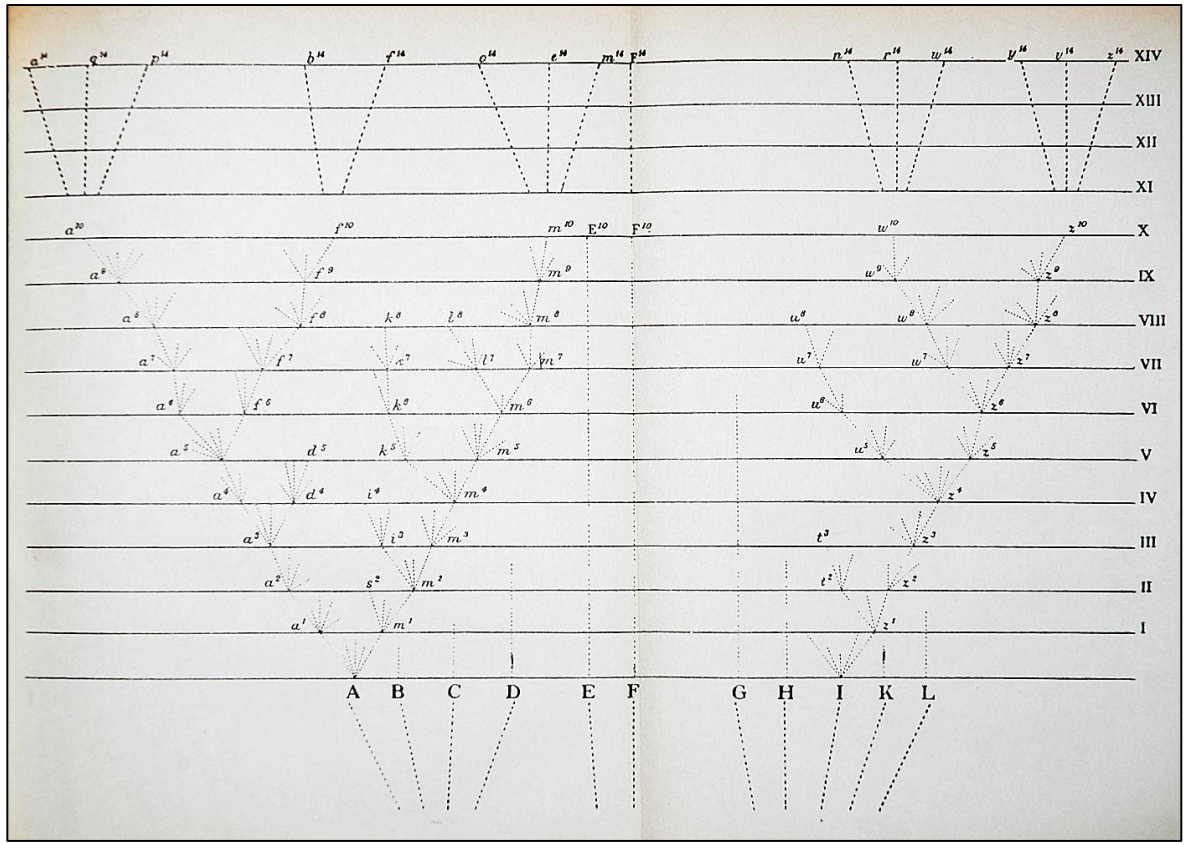
PRESERVATION OF FAVOURED RACES IN THE STRUGGLE FOR LIFE.

By CHARLES DARWIN, M.A.,

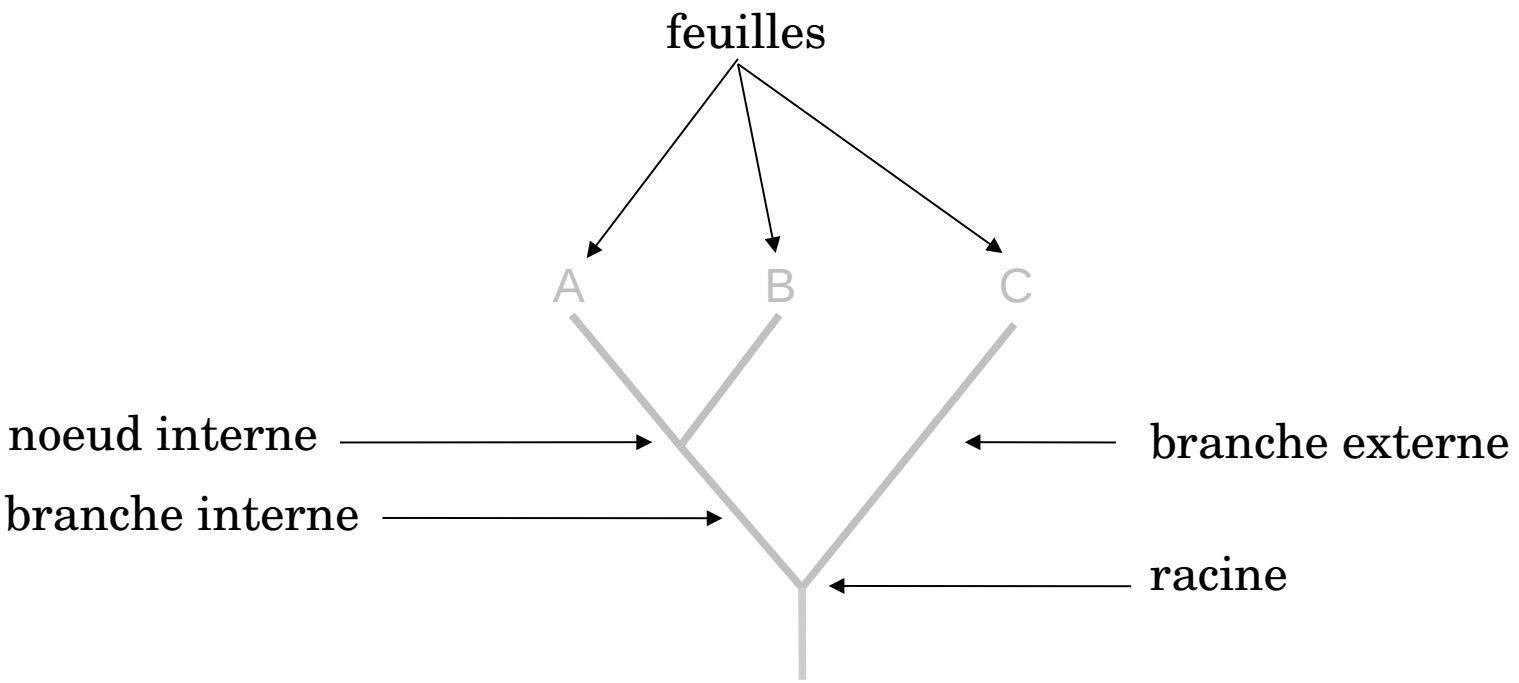
FELLOW OF THE ROYAL, GEOLOGICAL, LINNÆAN, ETC., SOCIETIES;
AUTHOR OF 'JOURNAL OF RESEARCHES DURING H. M. S. BEAGLE'S VOYAGE ROUND THE WORLD.'

LONDON:
JOHN MURRAY, ALBEMARLE STREET.
1859.



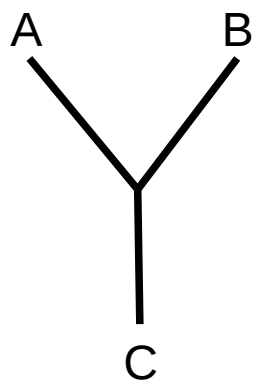


# Quelques définitions

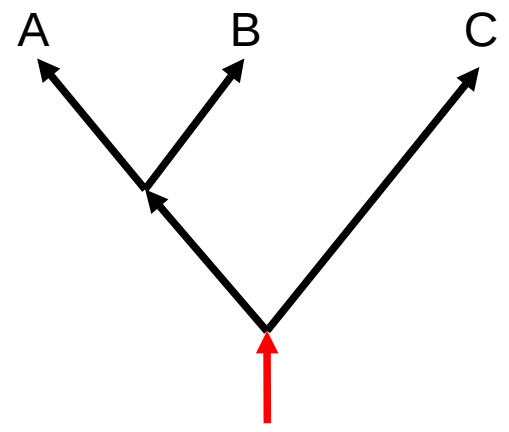



1 arbre = 1 topologie + des longueurs de branches

Différents types d'arbres pour schématiser les relations évolutives entre différentes entités

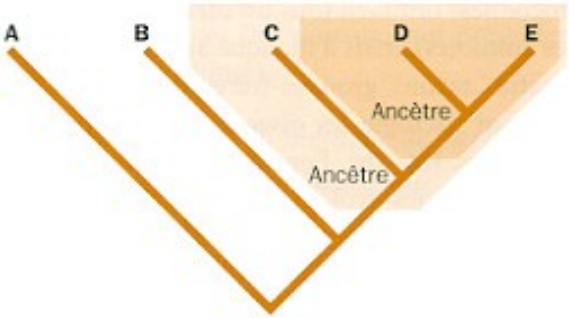



Arbre non raciné
(graphe connexe non cyclique)



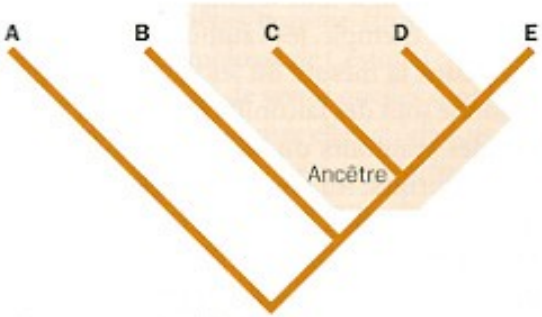

Arbre raciné
(graphe connexe et orienté)

# Monophylie, Paraphylie et Polyphylie



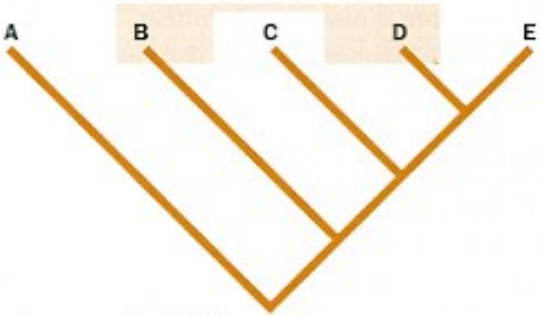

a. Groupes monophylétiques

Un ancêtre commun et tous ses descendants




b. Groupe paraphylétique

Un ancêtre commun et une partie seulerment de ses descendants




c. Groupe polyphylétique

Des membres sans ancêtre commun

# Monophylie, Paraphylie et Polyphylie

Symplésiomorphies

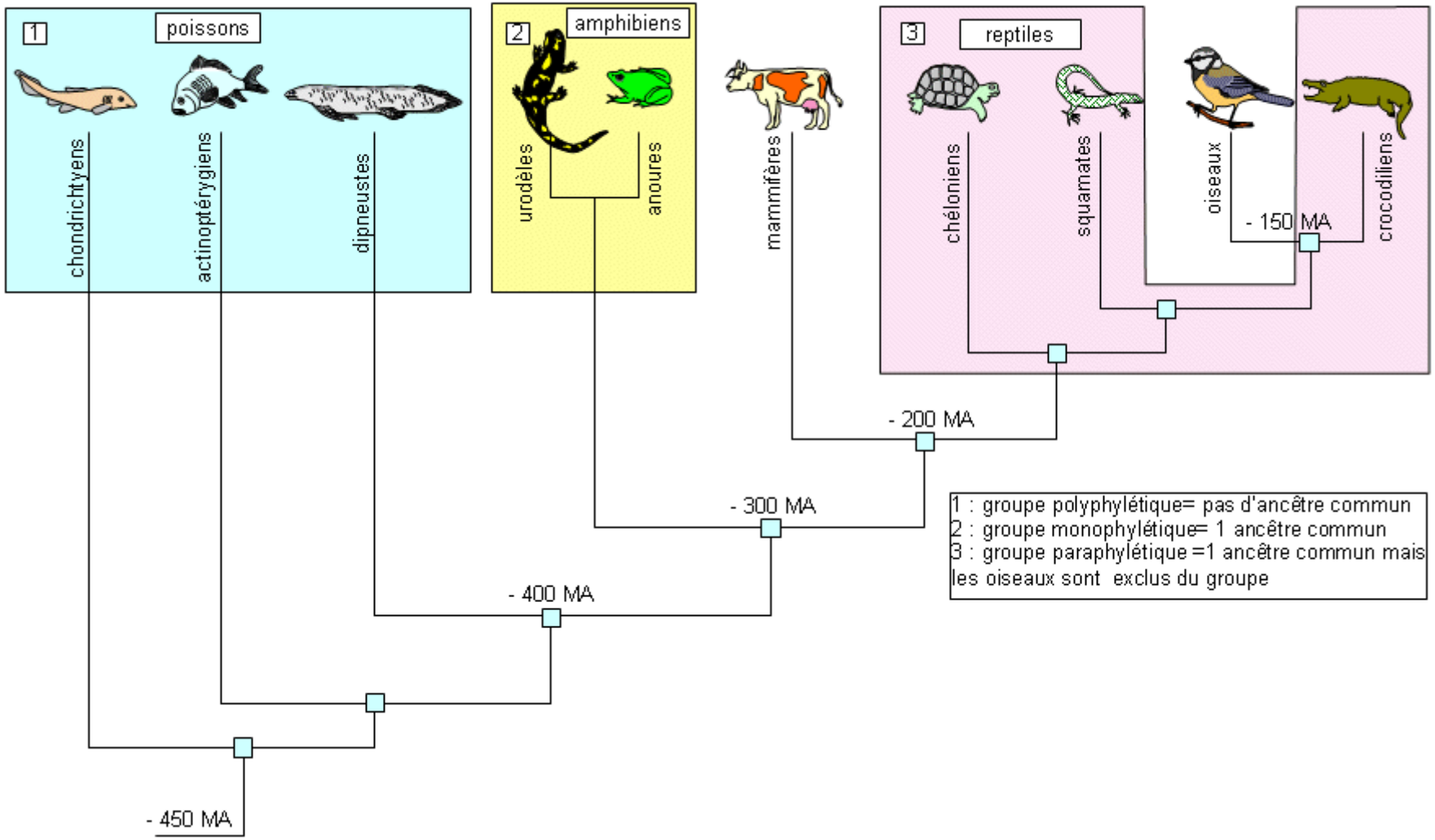
Caractères ancestraux

Synapomorphies

Caractères dérivés partagés

Symplésiomorphies

Caractères ancestraux




d'après Alain Gallien

Comment proposer un scénario évolutif à partir de l'observation de caractères ?

Comment représenter les relations évolutives entre plusieurs espèces ?

Comment proposer un scénario évolutif à partir de l'observation de caractères ?

Etant donnée une liste de caractères associés à un ensemble d'entités, comment construire un arbre retraçant les liens évolutifs entre toutes ces entités ?

Etant donnée une liste de caractères associés à un ensemble d'entités, comment construire un arbre retraçant les liens évolutifs entre toutes ces entités ?

1. Les méthodes de parcimonie
2. Les méthodes phénétiques (de distance)
3. Les méthodes probabilistes (maximum de vraisemblance et Bayesiennes)

# Qu'est ce que la parcimonie ?

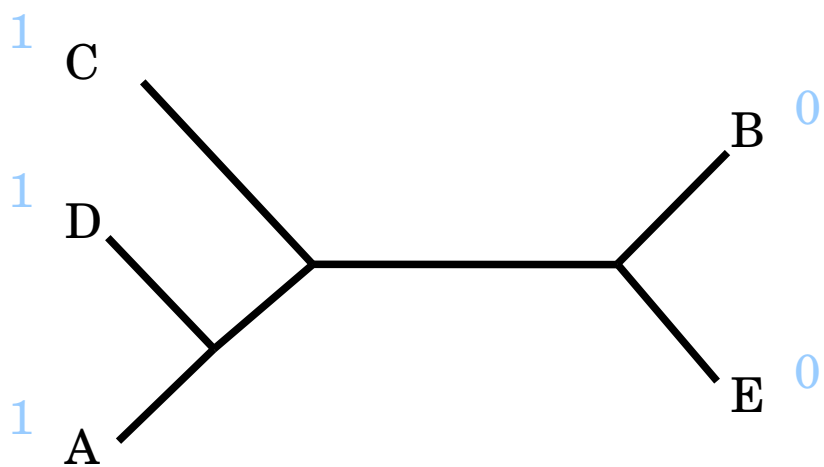
Edwards and Cavalli-Sforza (1963)

Le scénario évolutif proposé nécessite un nombre minimum d'hypothèses.

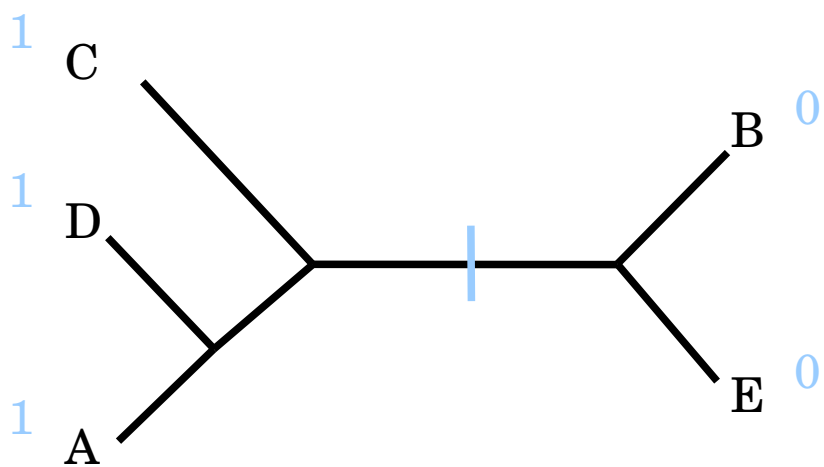
| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| A | 1 | 0 | 0 | 1 | 1 | 0 |
| B | 0 | 0 | 1 | 0 | 0 | 0 |
| C | 1 | 1 | 0 | 0 | 0 | 0 |
| D | 1 | 1 | 0 | 1 | 1 | 1 |
| E | 0 | 0 | 1 | 1 | 1 | 0 |

Cela revient à proposer un scénario ou le nombre de changements évolutifs est minimal

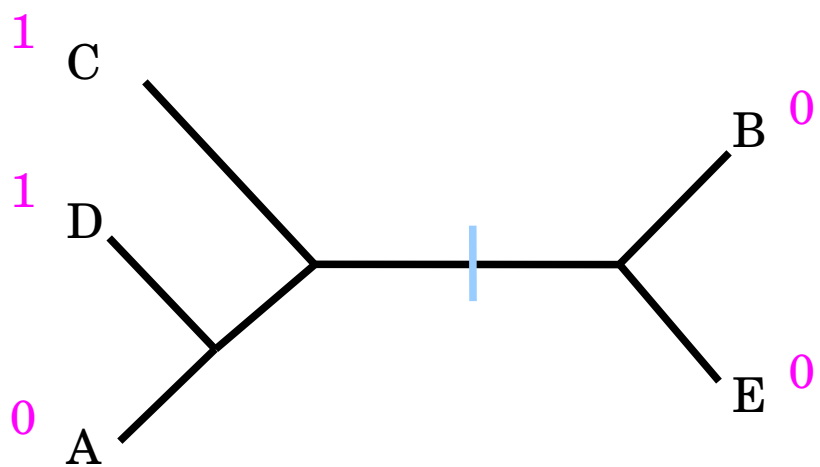
| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| A | 1 | 0 | 0 | 1 | 1 | 0 |
| B | 0 | 0 | 1 | 0 | 0 | 0 |
| C | 1 | 1 | 0 | 0 | 0 | 0 |
| D | 1 | 1 | 0 | 1 | 1 | 1 |
| E | 0 | 0 | 1 | 1 | 1 | 0 |



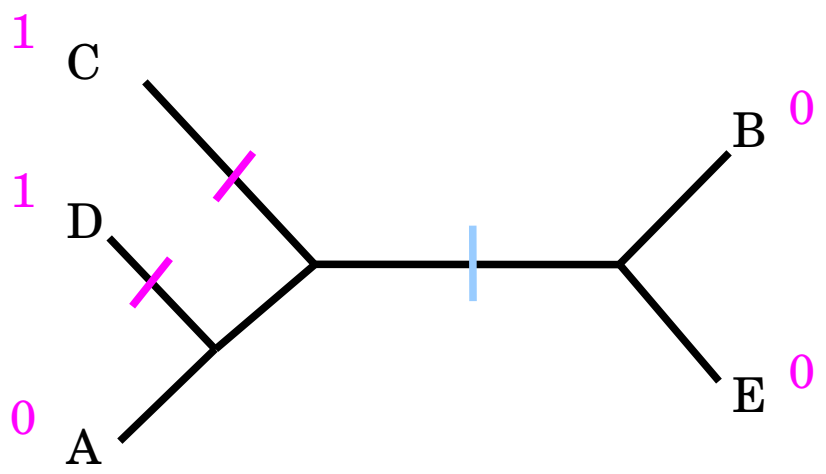
| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| A | 1 | 0 | 0 | 1 | 1 | 0 |
| B | 0 | 0 | 1 | 0 | 0 | 0 |
| C | 1 | 1 | 0 | 0 | 0 | 0 |
| D | 1 | 1 | 0 | 1 | 1 | 1 |
| E | 0 | 0 | 1 | 1 | 1 | 0 |



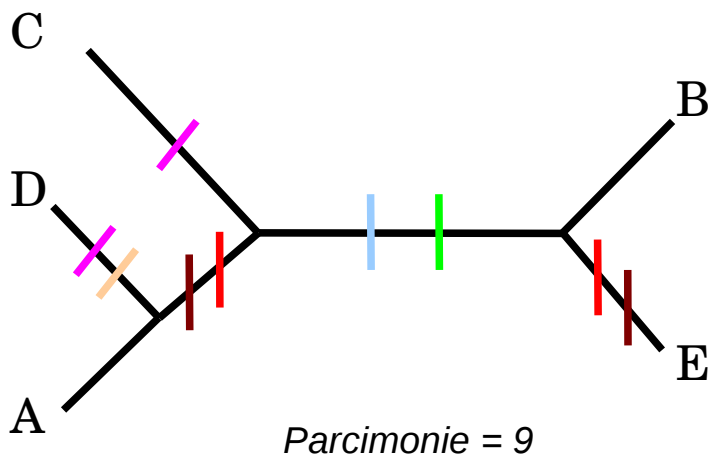
| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| A | 1 | 0 | 0 | 1 | 1 | 0 |
| B | 0 | 0 | 1 | 0 | 0 | 0 |
| C | 1 | 1 | 0 | 0 | 0 | 0 |
| D | 1 | 1 | 0 | 1 | 1 | 1 |
| E | 0 | 0 | 1 | 1 | 1 | 0 |



| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| A | 1 | 0 | 0 | 1 | 1 | 0 |
| B | 0 | 0 | 1 | 0 | 0 | 0 |
| C | 1 | 1 | 0 | 0 | 0 | 0 |
| D | 1 | 1 | 0 | 1 | 1 | 1 |
| E | 0 | 0 | 1 | 1 | 1 | 0 |

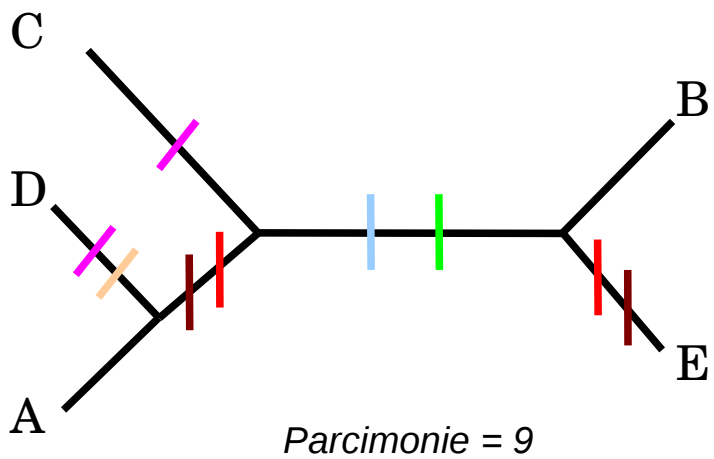


| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| A | 1 | 0 | 0 | 1 | 1 | 0 |
| B | 0 | 0 | 1 | 0 | 0 | 0 |
| C | 1 | 1 | 0 | 0 | 0 | 0 |
| D | 1 | 1 | 0 | 1 | 1 | 1 |
| E | 0 | 0 | 1 | 1 | 1 | 0 |




*Parcimonie = 9*

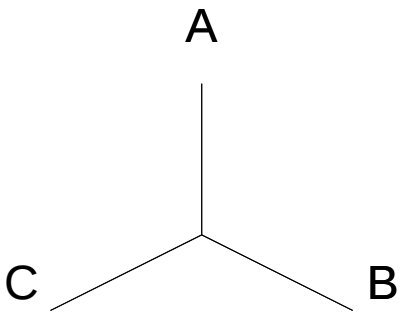
| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| A | 1 | 0 | 0 | 1 | 1 | 0 |
| B | 0 | 0 | 1 | 0 | 0 | 0 |
| C | 1 | 1 | 0 | 0 | 0 | 0 |
| D | 1 | 1 | 0 | 1 | 1 | 1 |
| E | 0 | 0 | 1 | 1 | 1 | 0 |



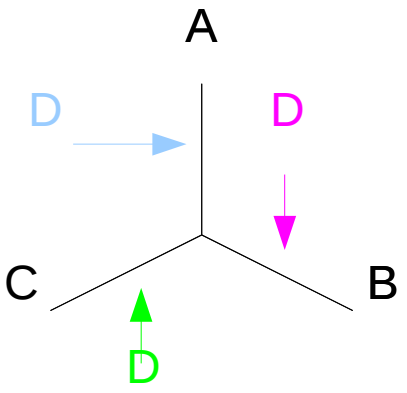

S'agit-il de l'arbre le plus parcimonieux ?

# Combien d'arbres non racinés ?

Il existe un seul arbre avec 3 taxons :




On peut ajouter un *4*ième taxon, sur les branches internes ou externes,



# Combien d'arbres non racinés ?

Un arbre avec ($n$-1) taxons, possède :

($n$-1)-3 branches internes
($n$-1) branches externes

On peut ajouter un $n^{\text{ième}}$ taxon, sur les branches internes ou externes,

soit : ($n$-1)-3+($n$-1) = 2$n$-5 possibilités pour placer le $n^{\text{ième}}$ taxon

Le nombre total d'arbres avec $n$ taxons, $T_n$ est donc défini par :

$$T_n = T_{n-1} * (2n-5)$$

Donc, par récurrence, $T_n = \prod_{k=3}^{n} (2k-5)$

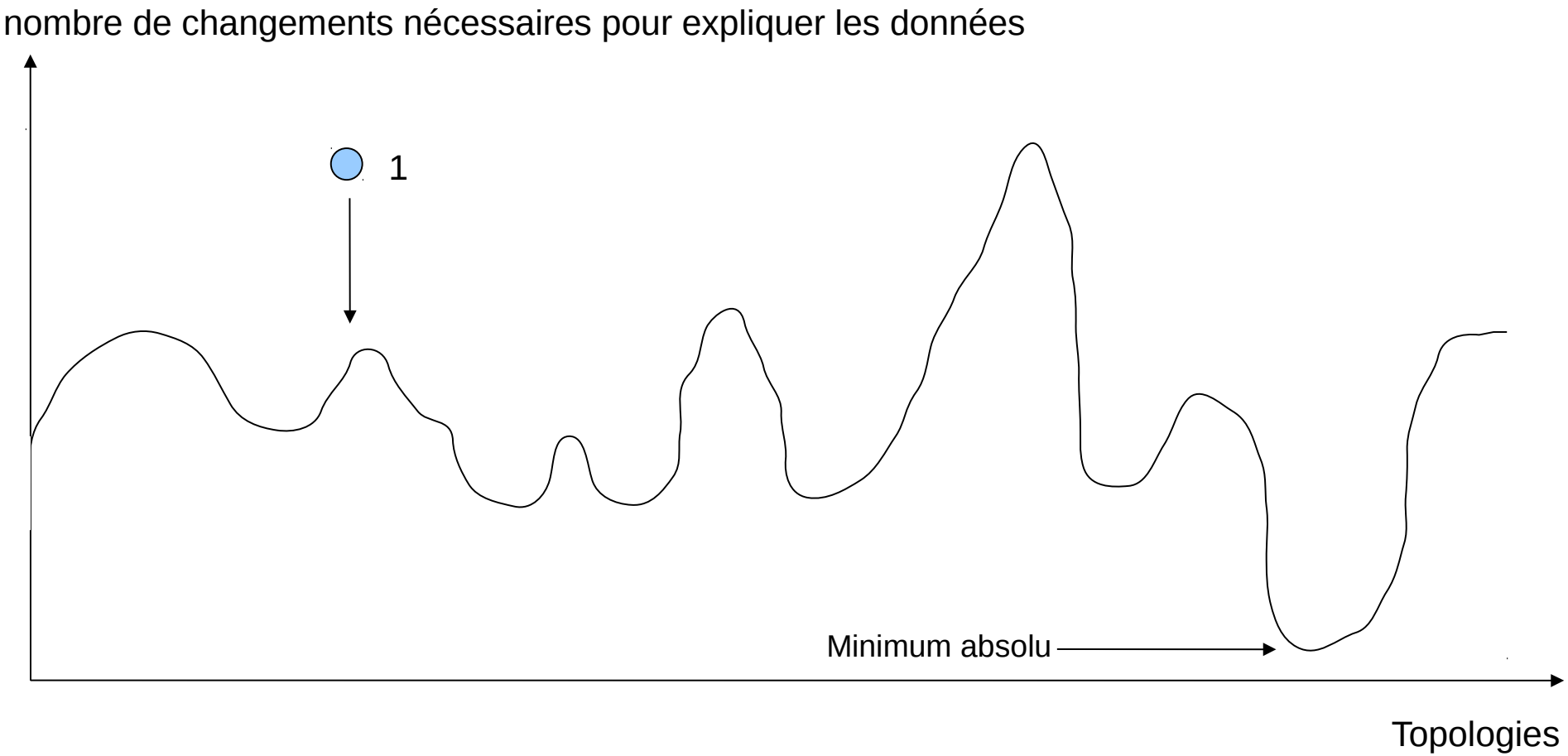
# Combien d'arbres non racinés ?

| Taxons | Nombre d'arbres |
|---|---|
| 3 | 1 |
| 4 | 3 |
| 5 | 15 |
| 6 | 105 |
| 7 | 945 |
| 8 | 10395 |
| 9 | 135135 |
| 10 | $2{,}03.10^{6}$ |
| 20 | $2{,}22.10^{20}$ |
| 30 | $8{,}69.10^{36}$ |
| 40 | $1{,}31.10^{55}$ |
| 50 | $2{,}84.10^{74}$ |
| 100 | $1{,}7.10^{182}$ |

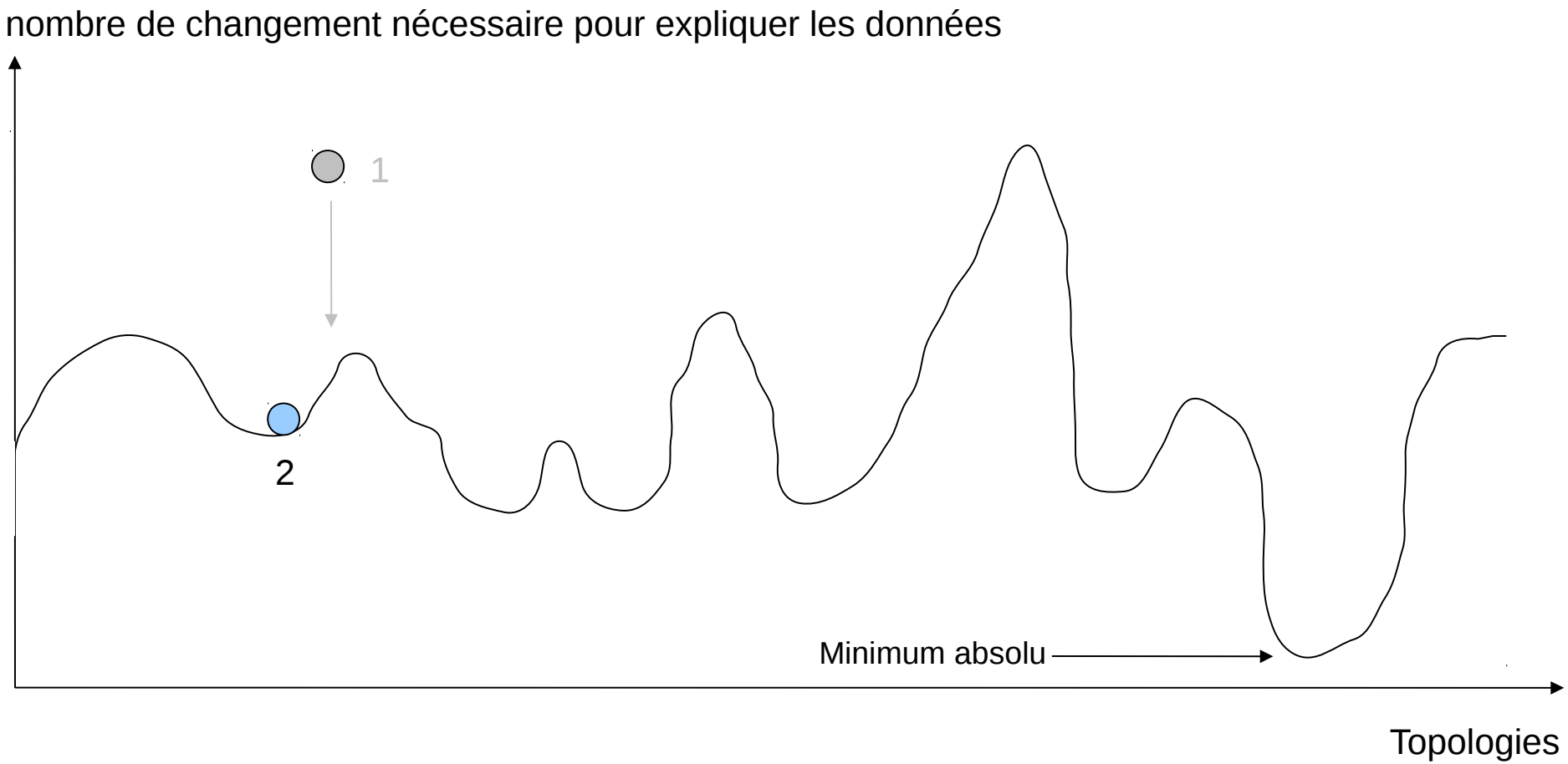
$N_A = 6{,}0221415 \times 10^{23}$

$N_{Atomes} = 10^{80}$

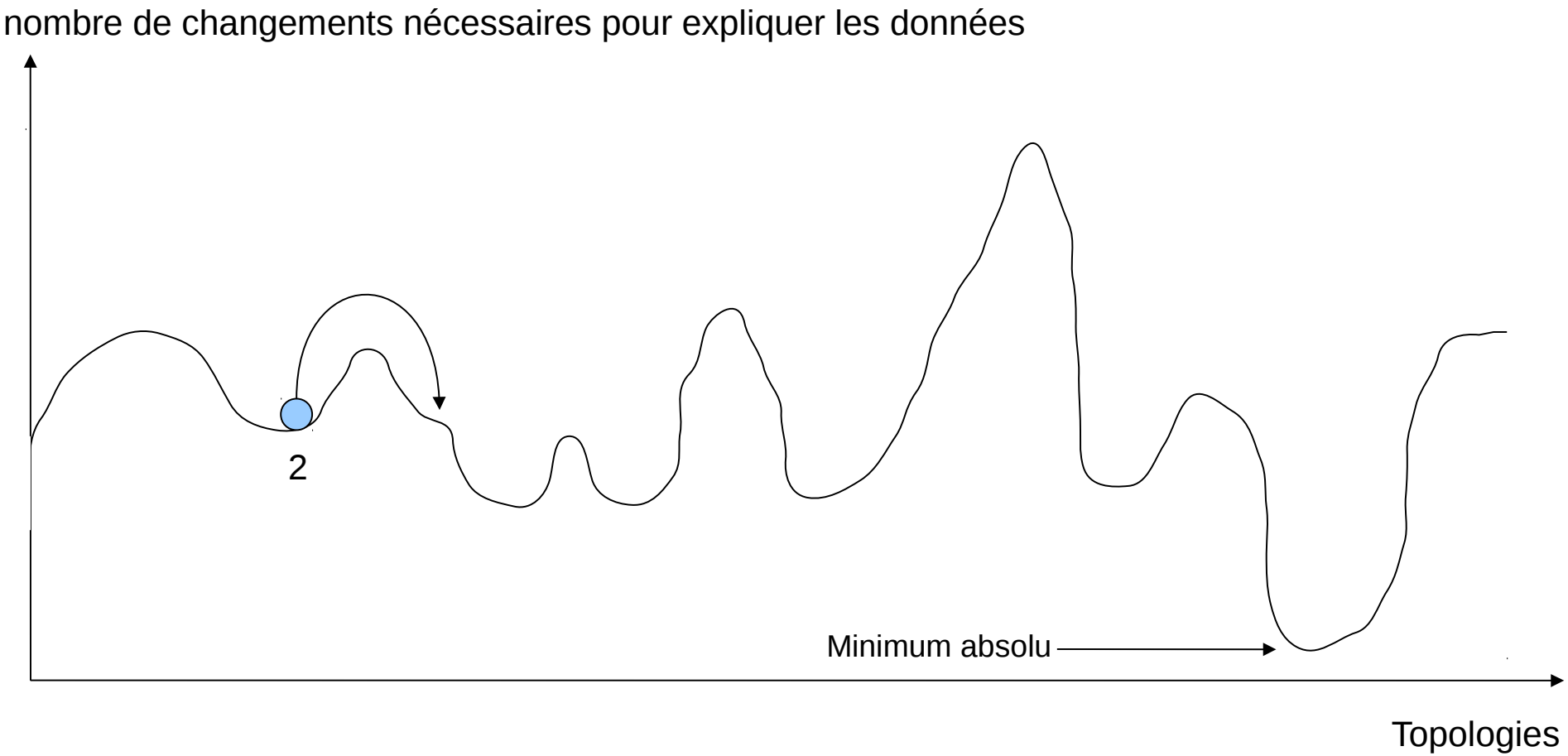
# Rechercher le meilleur arbre par les méthodes heuristiques



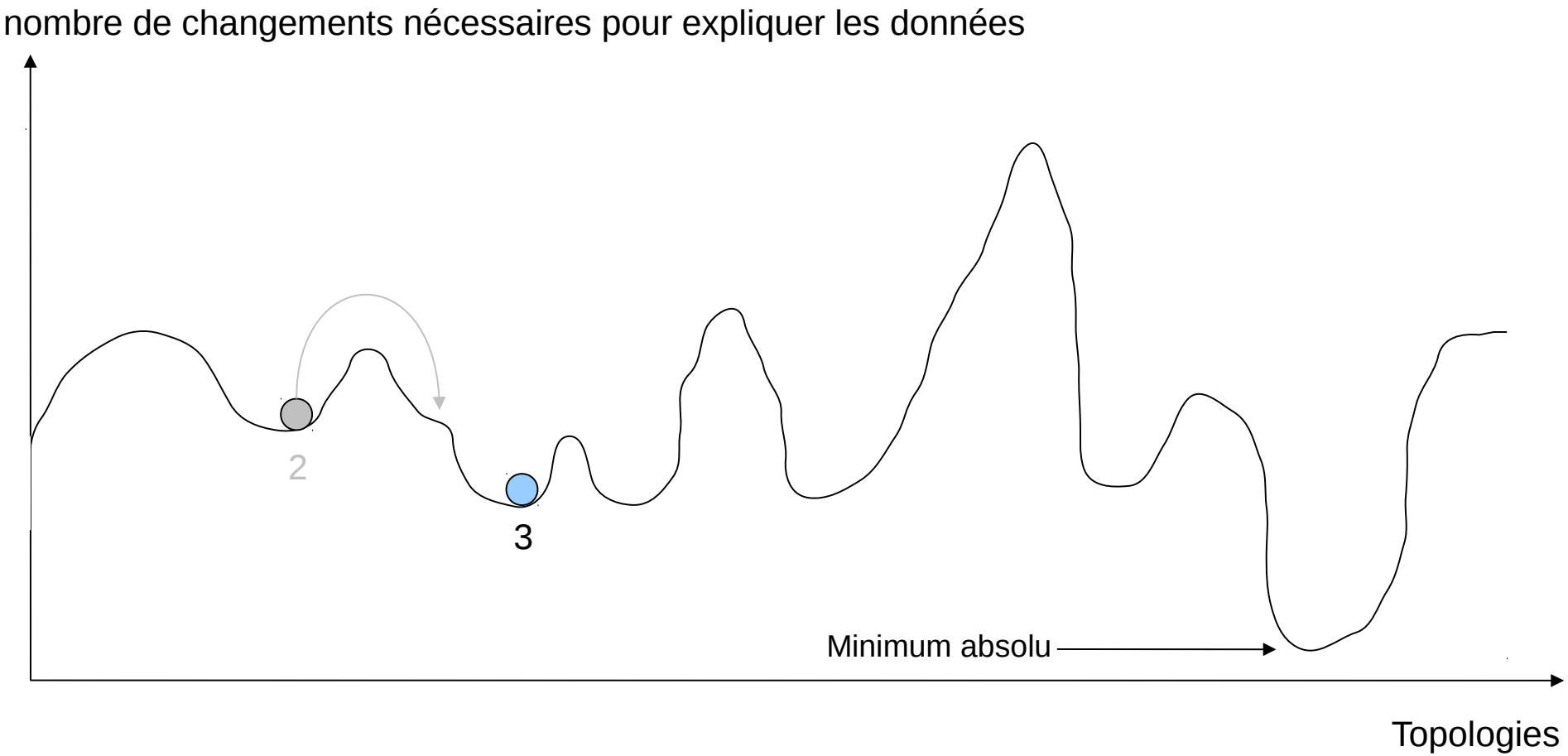
# Rechercher le meilleur arbre par les méthodes heuristiques



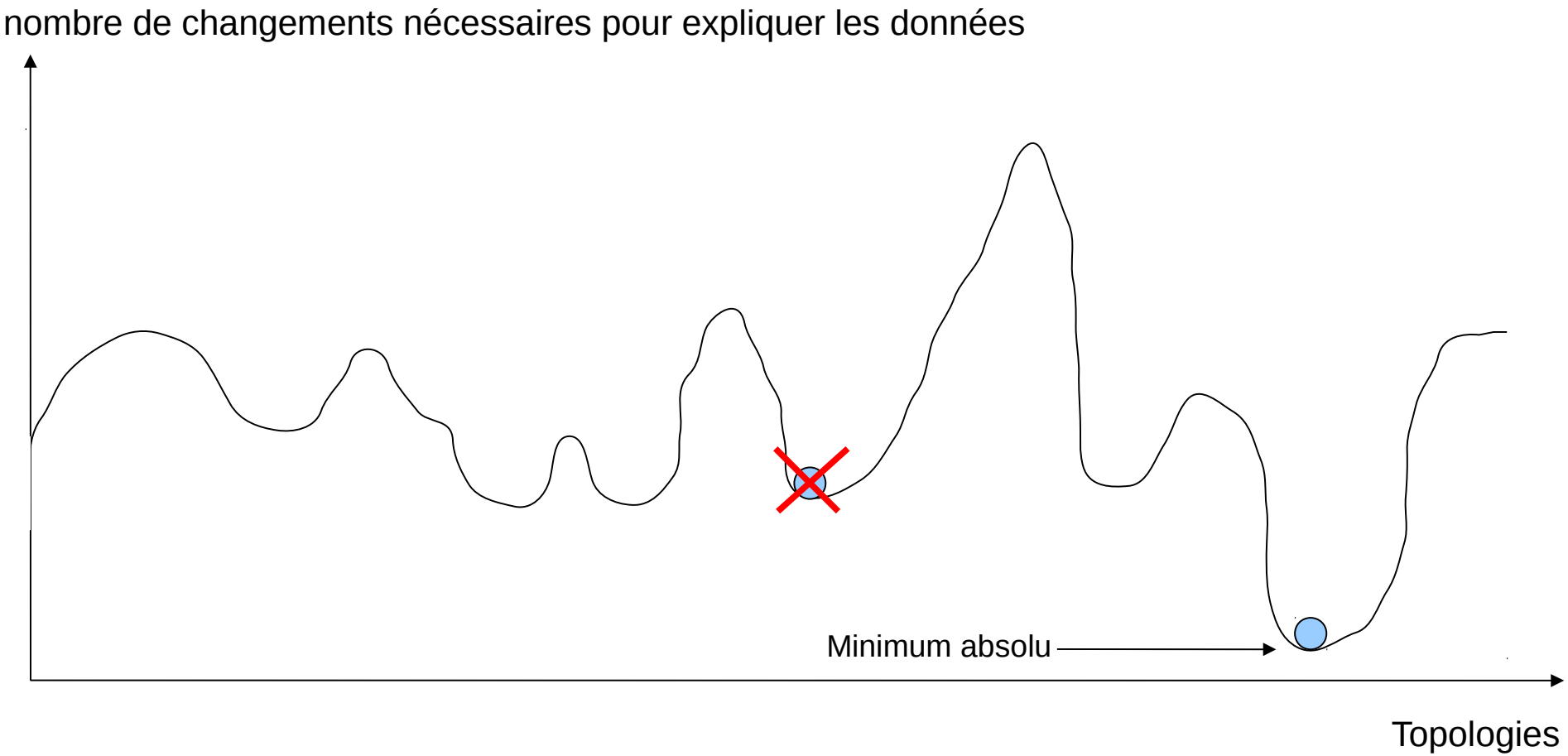
# Rechercher le meilleur arbre par les méthodes heuristiques



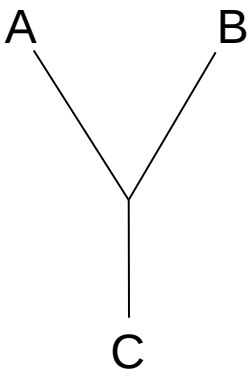
# Rechercher le meilleur arbre par les méthodes heuristiques



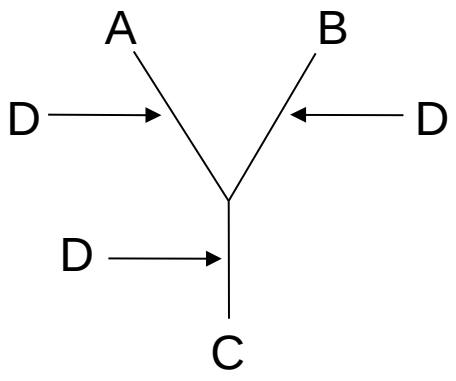
# Rechercher le meilleur arbre par les méthodes heuristiques



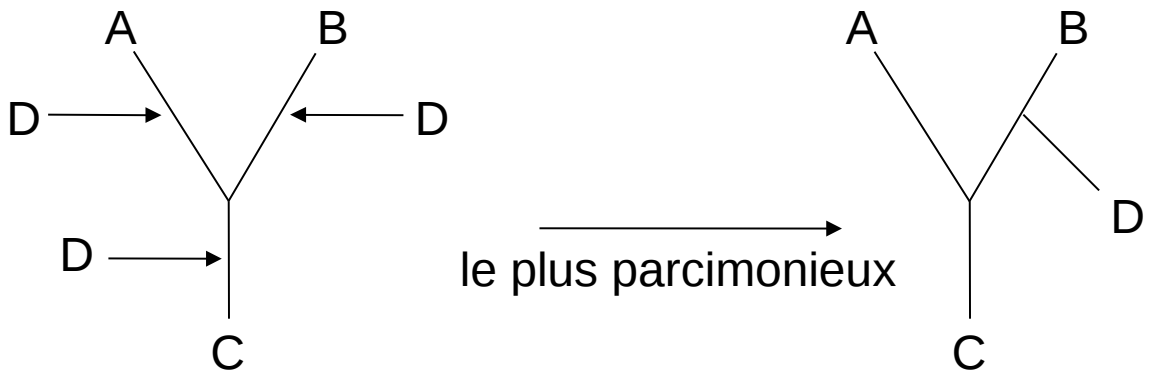
# Construire l'arbre de départ par addition séquentielle



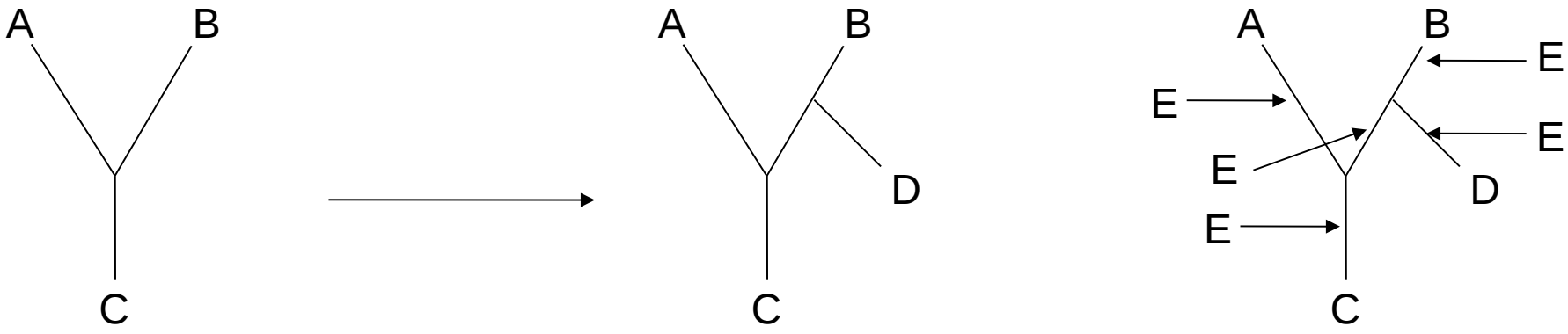
# Construire l'arbre de départ par addition séquentielle



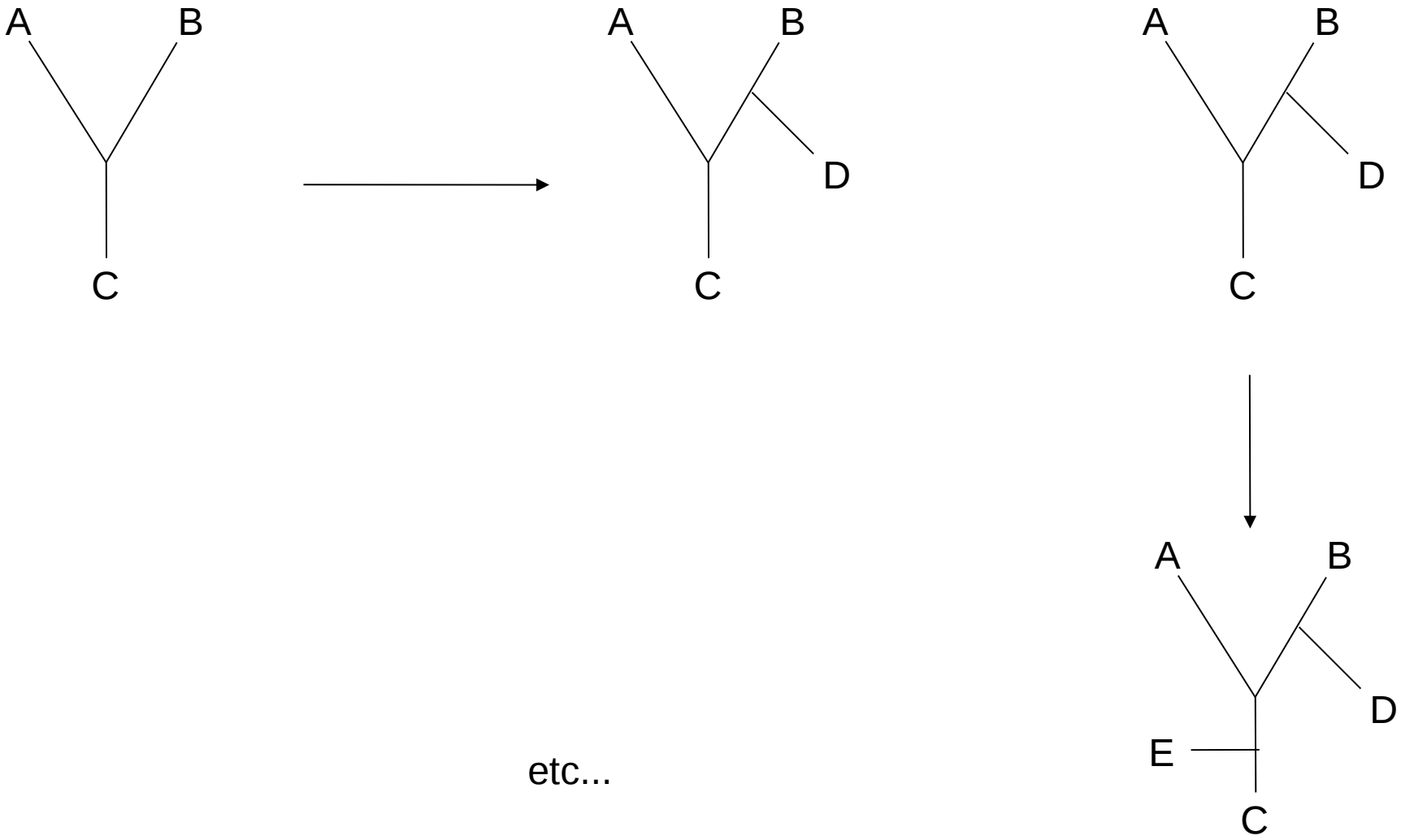
# Construire l'arbre de départ par addition séquentielle



# Construire l'arbre de départ par addition séquentielle



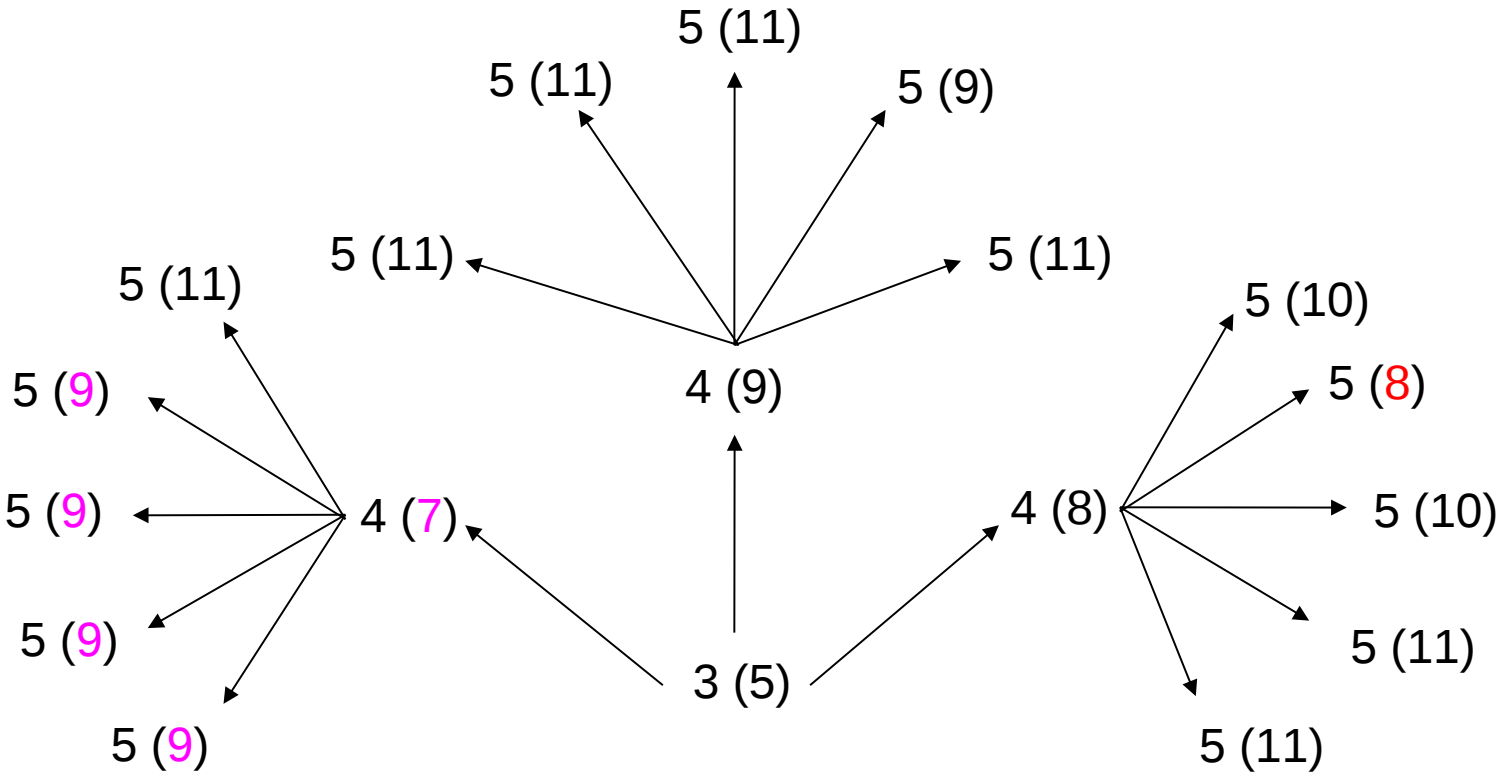
Construire l'arbre de départ par addition séquentielle




etc...

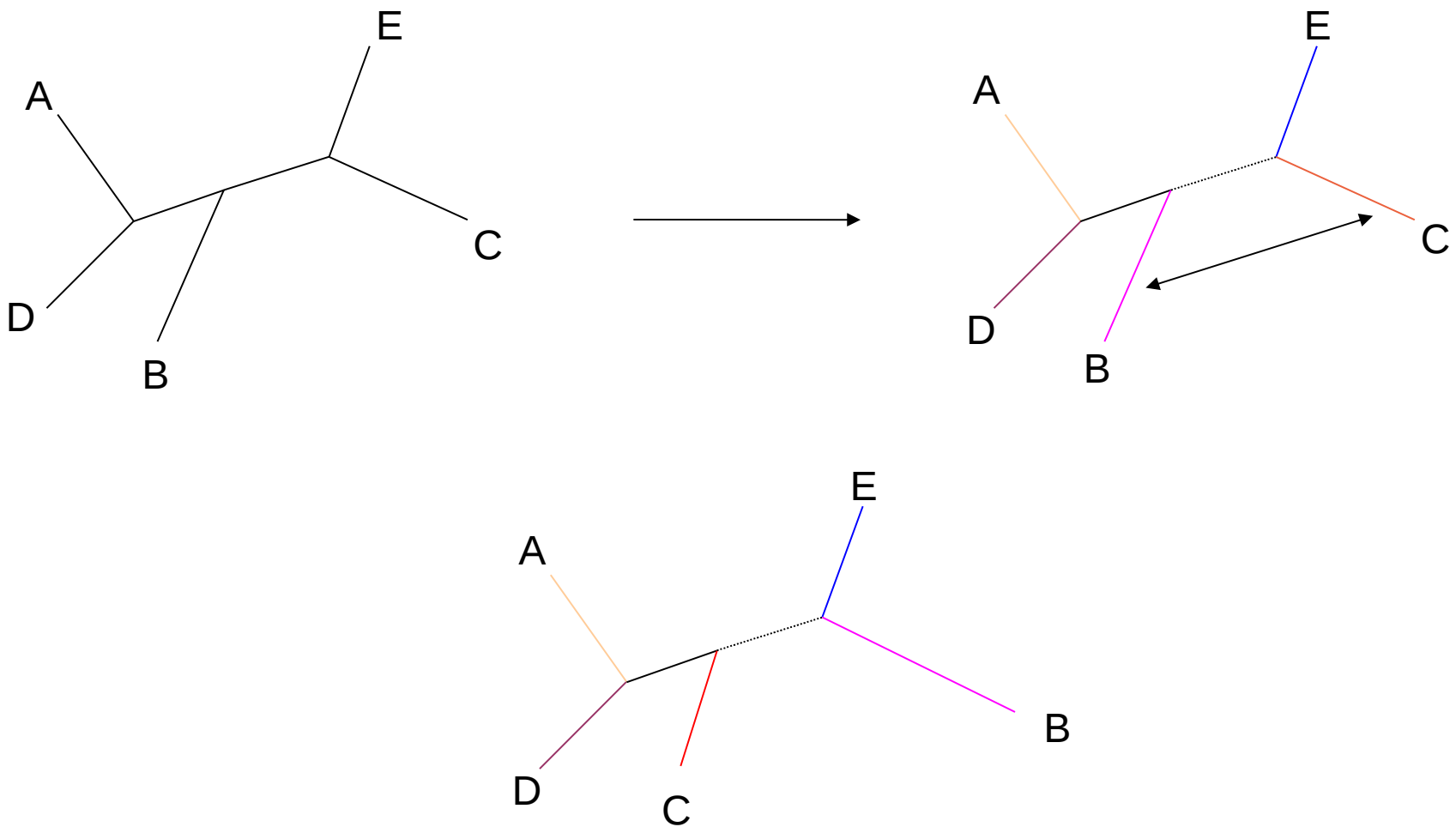
Rechercher le meilleur arbre par les méthodes heuristiques

## Accélerer la recherche par Branch and Bound



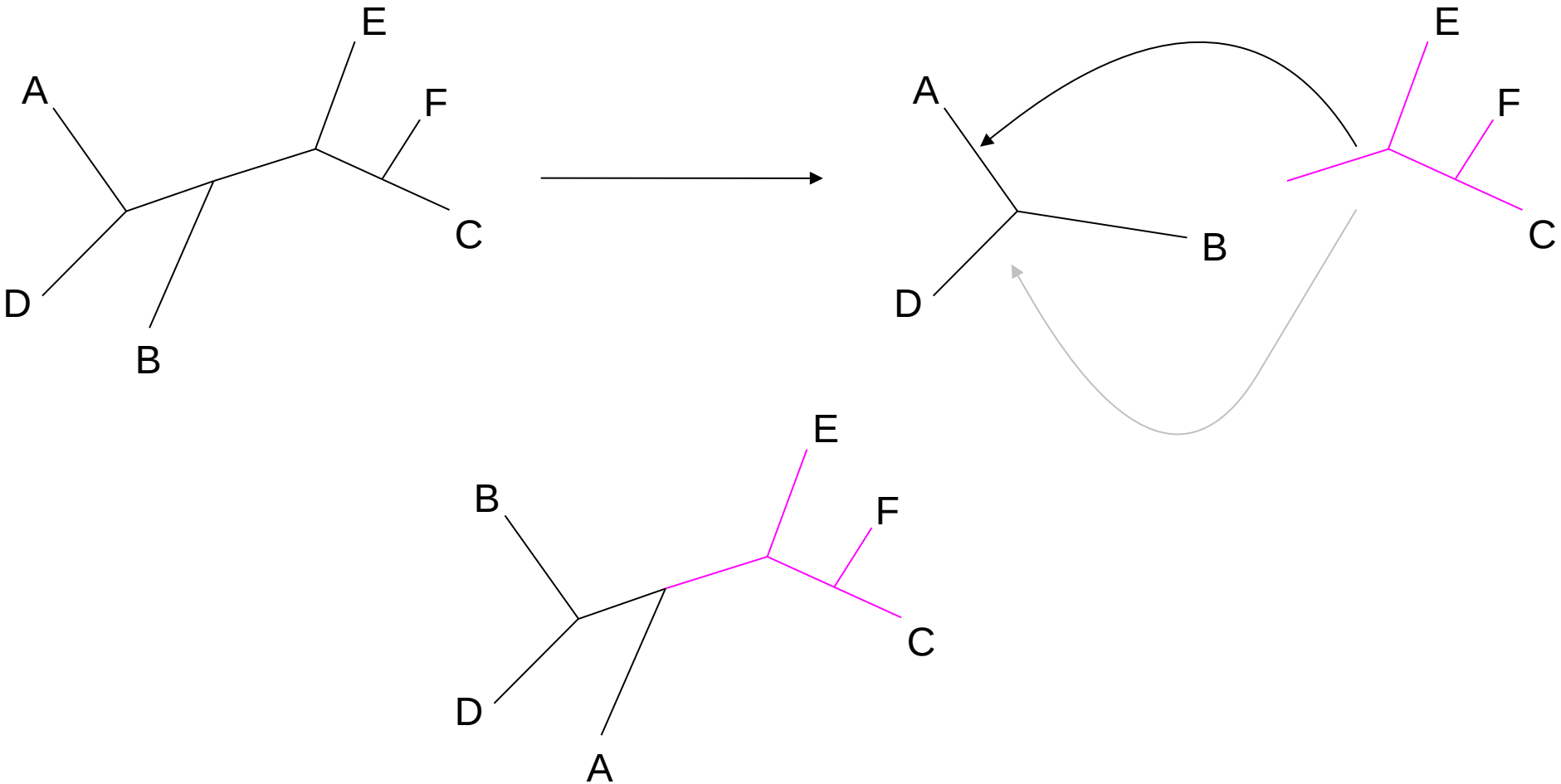
Rechercher le meilleur arbre par les méthodes heuristiques

## NNI : Nearest Neighbour Interchange



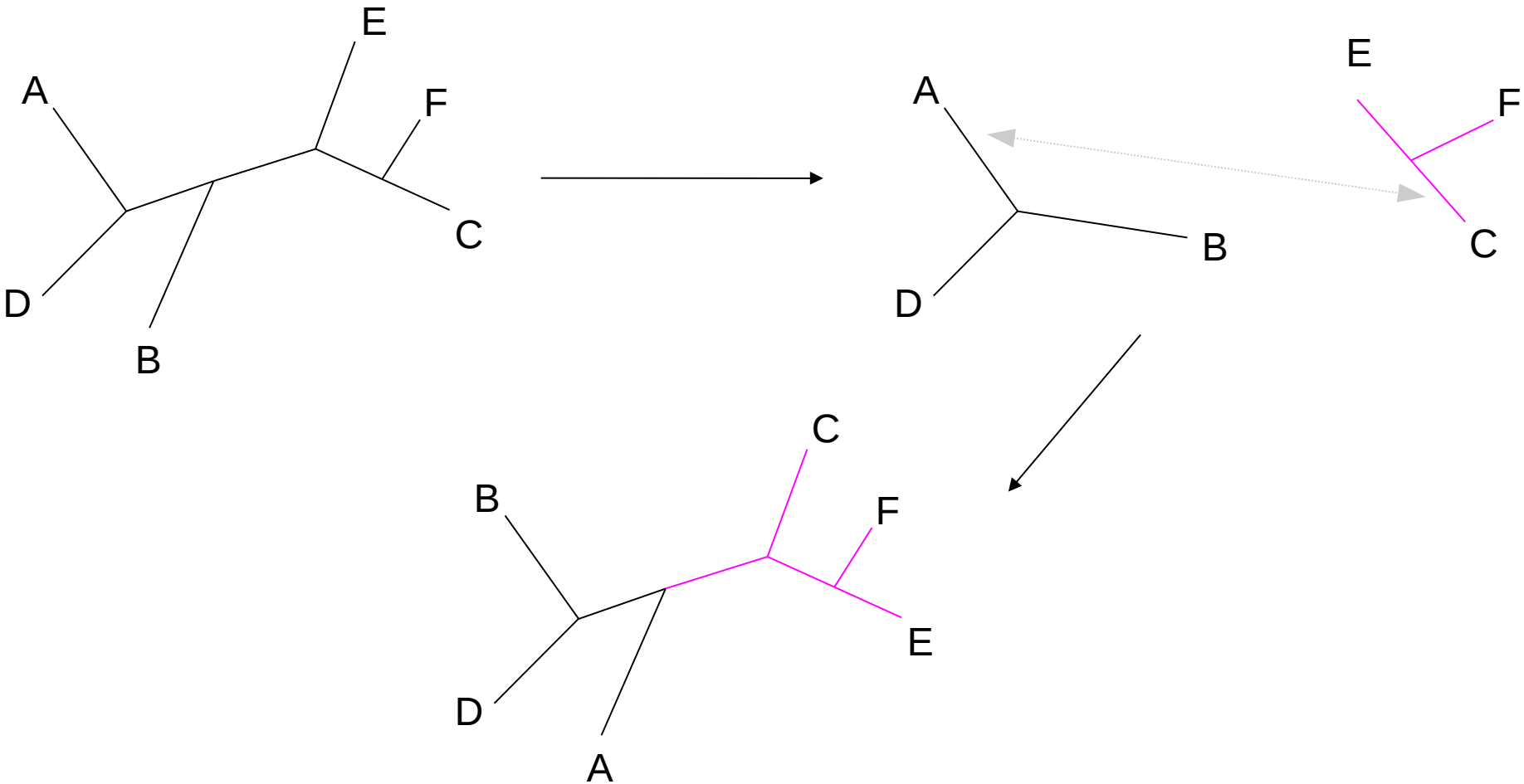
Rechercher le meilleur arbre par les méthodes heuristiques

SPR : Subtree Pruning and Regrafting

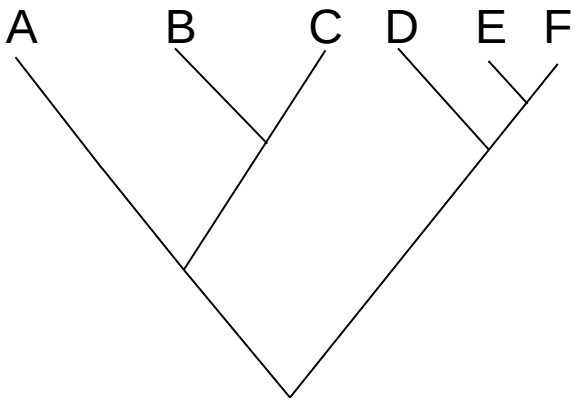


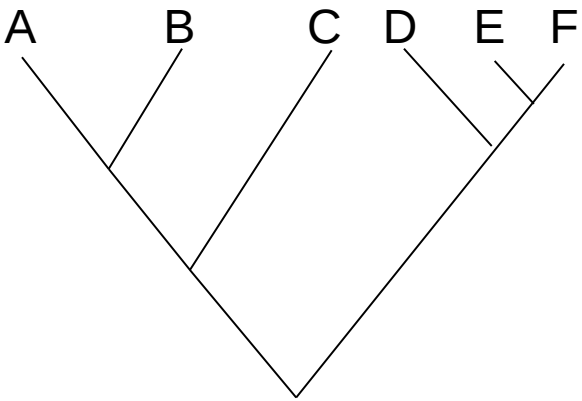
Rechercher le meilleur arbre par les méthodes heuristiques

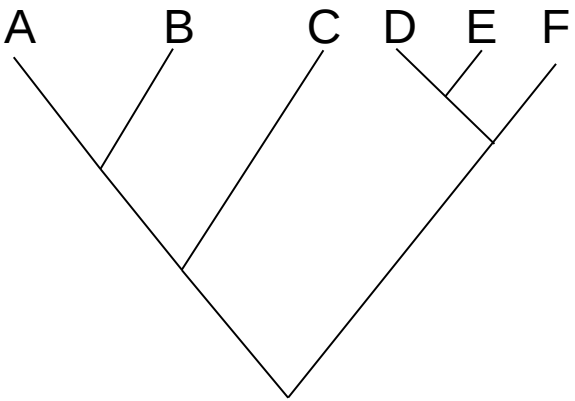
TBR : Tree bisection and reconnection

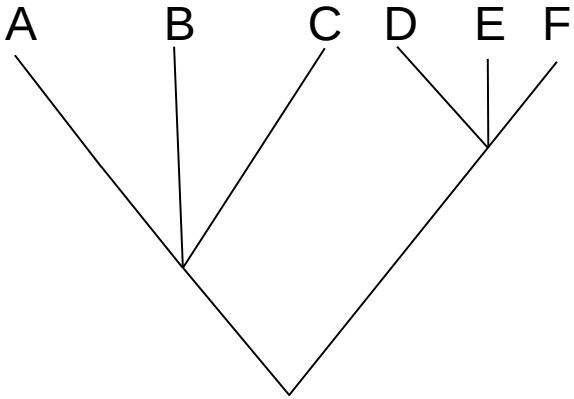


# Arbres de consensus

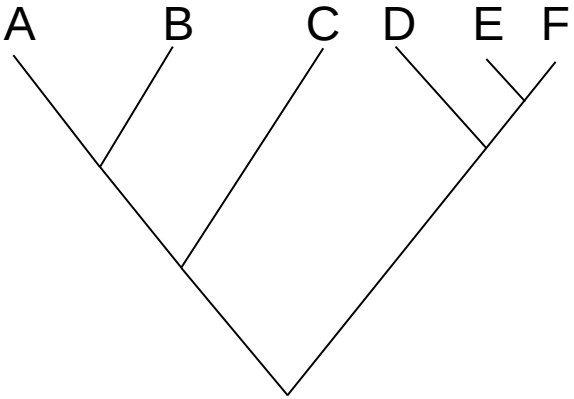



strict




50% majority rule

Camin-Sokal parsimony (1965) : Il y a seulement deux états possibles et les modification ne sont pas réversibles. Elles se font donc toujours de l'état ancestral vers l'état dérivé.

Wagner parsimony (1970) : Les différents états possibles sont ordonnés. Les changements d'états sont possibles dans les deux directions.

Dollo parcimonie (1974) : Il y a seulement deux états possibles, l'état ancestral ne peut évoluer qu'une seule fois et l'état dérivé peu revenir vers l'état initial un grand nombre de fois.

Parcimonie pondérée (1969) : Tous les sites n'ont pas le même poids dans l'analyse

# Parcimonie

Méthode relativement rapide qui  est a correcte à condition :

- que le taux de mutation ne soit pas trop élevé
- qu'il n'y a pas de longues branches

Sinon la méthode devient inconsistante, c'est à dire que l'augmentation du nombre de de caractères ne garanti pas d'obtenir une solution correcte.

# Evaluation des arbres : bootstrap non paramétrique

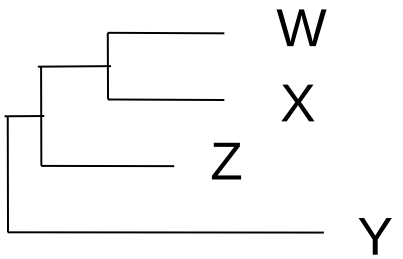
| | 123456789 |
|---|---|
| W | ACTTGACCC |
| X | AGCTGGCCC |
| Y | AGTTGACCA |
| Z | AGCTGGTCC |

Tirage aléatoire

avec remise

de *l* sites

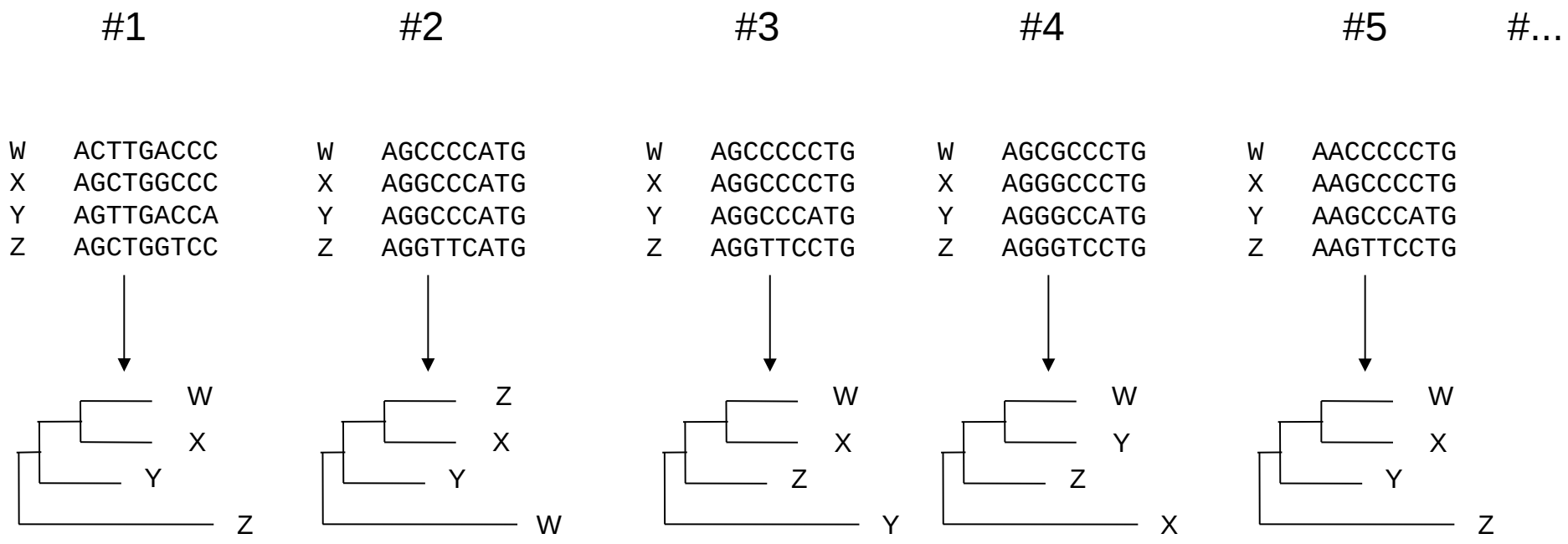
| | 152778945 |
|---|---|
| | 123456789 |
| W | AGCCCCCTG |
| X | AGGCCCCTG |
| Y | AGGCCCATG |
| Z | AGGTTCCTG |




La procédure est répétée un grand nombre de fois

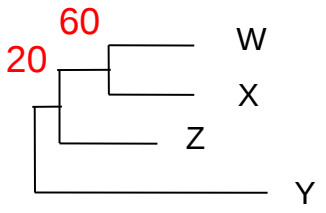
(100 à 1000 fois).

# Evaluation des arbres : bootstrap non paramétrique



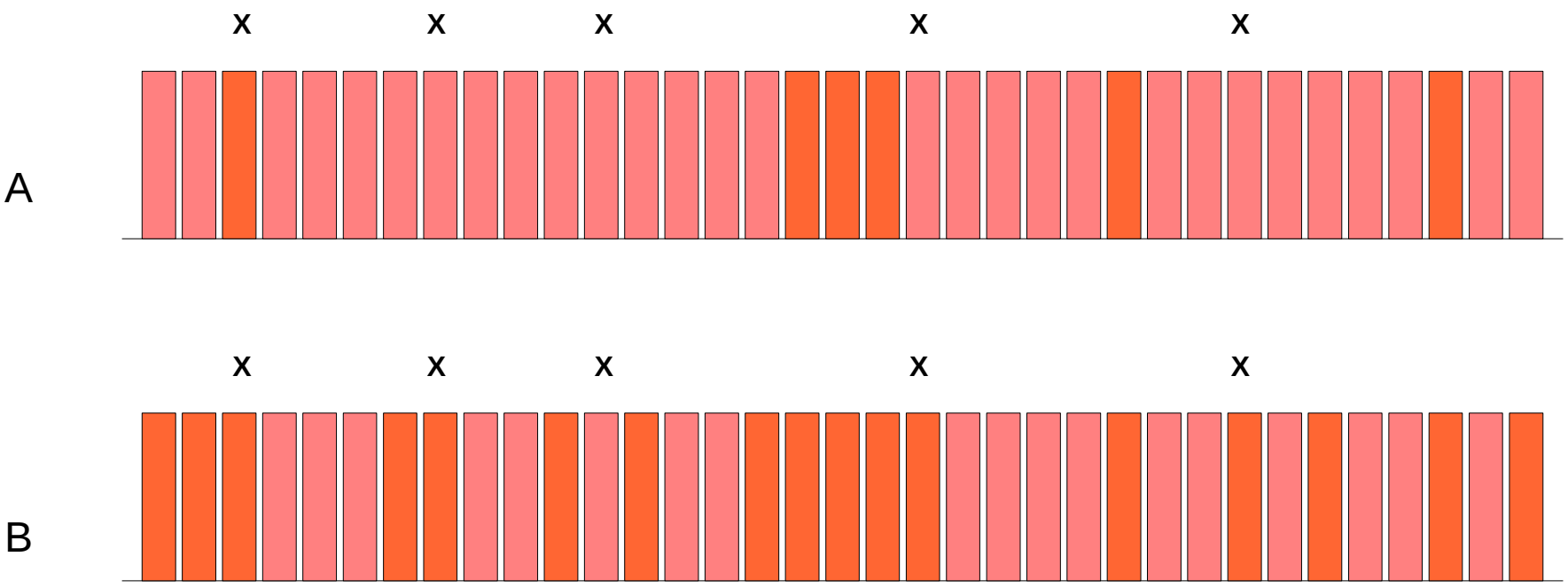

On regarde la fréquence des regroupements d'espèces dans tous les arbres trouvés.

C'est un estimateur du signal phylogénétique et de son homogénéité.




Arbre original

# Evaluation des arbres : bootstrap non paramétrique




C'est un estimateur du signal phylogénétique et de son homogénéité.

Etant donnée une liste de caractères associés à un ensemble d'entités, comment construire un arbre retraçant les liens évolutifs entre toutes ces entités ?

Comment proposer un scénario évolutif à partir de l'observation des différences et ressemblances ?

1. Les méthodes de parcimonie
2. Les méthodes phénétiques (de distance)
3. Les méthodes probabilistes (maximum de vraisemblance et Bayesiennes)

# Similitude et distance

Plus la ressemblance globale entre deux entité est importante, plus la parenté à de chances d'être proche.

Plus la similitude entre deux entités *i* et *j* est forte et plus la distance entre elles $d_{ij}$ est faible.

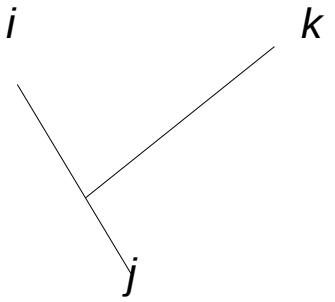
En outre les distances métriques doivent respecter les propriétés suivantes :

$d_{ij} > 0$ si $i <> j$

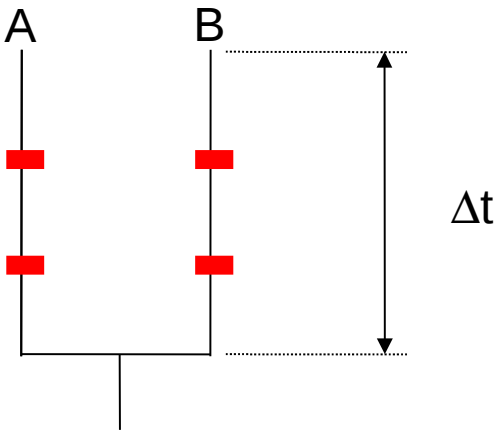
$d_{ij} = 0$ si $i = j$

$d_{ij} = d_{ji}$

$d_{ik} + d_{kj} >= d_{ij}$

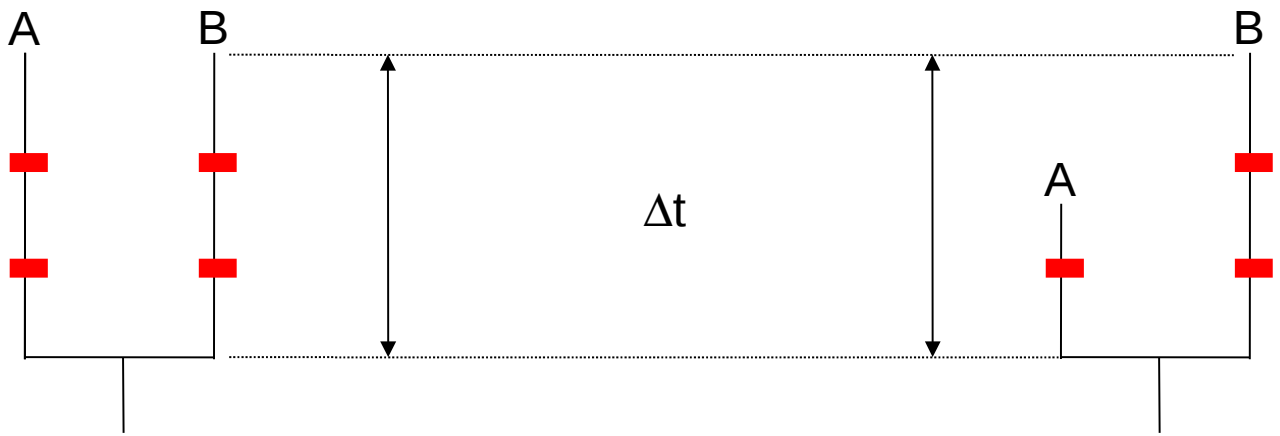


# Relation entre le temps et la distance




$$d = f(t)$$

# Relation entre le temps et la distance




$$d = r_i \cdot t$$

$r_i$ est le taux d'évolution pour l'UE *i*

# Calcul de la distance moléculaire entre deux séquences nucléotidiques

Seq 1 A T T G T A T G T C C T G T A T G C A A

Seq 2 A T T A T A T T T C G T G A A T G C A T

*p*-distance = $d_{(seq1,seq2)} = \frac{\text{\# de substitutions}}{\text{\# de résidus}} = \frac{5}{20} = 0{,}25$

Variations de la p-distance :

Nucléotides : 0 <= *p*-distance <= 0,75 ($^{3}/_{4}$)

Protéines : 0 <= *p*-distance <= 0,95 ($^{19}/_{20}$)

Seq 1 A T T G T A T G T C C T G T A T G C A A

Seq 2 A T T A T A T T T C G T G A A T G C A T

$$d_{(seq1,seq2)} = \frac{\text{\# de substitutions}}{\text{\# de résidus}} = \frac{5}{20} = 0{,}25$$

Seq 1 A T T G T A T G T C C T G T A T G C A A

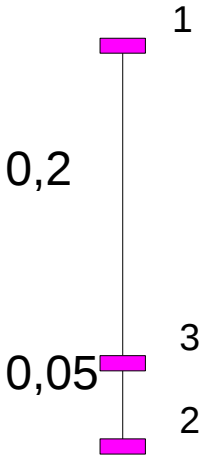
Seq 3 A T T A T A T T T C G T G T A T G C A T

$$d_{(seq1,seq3)} = \frac{\text{\# de substitutions}}{\text{\# de résidus}} = \frac{4}{20} = 0{,}20$$

Seq 2 A T T A T A T T T C G T G A A T G C A T

Seq 3 A T T A T A T T T C G T G T A T G C A T

$$d_{(seq1,seq3)} = \frac{\text{\# de substitutions}}{\text{\# de résidus}} = \frac{1}{20} = 0{,}05$$

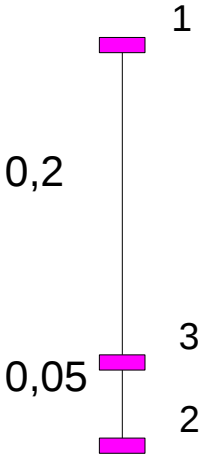


# Calcul de la distance moléculaire entre deux séquences nucléotidiques

Seq 1 A T T G T A T G T C C T G T A T G C A A

Seq 3 A T T A T A T T T C G T G T A T G C A T

$$d_{(seq1,seq3)} = \frac{\text{\# de substitutions}}{\text{\# de résidus}} = \frac{4}{20} = 0{,}20$$

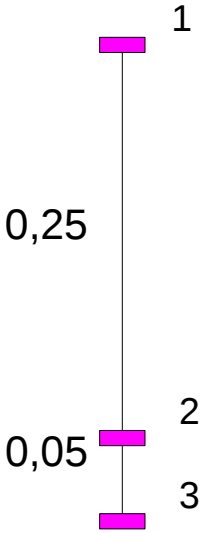



Seq 1 A T T G T A T G T C C T G T A T G C A A

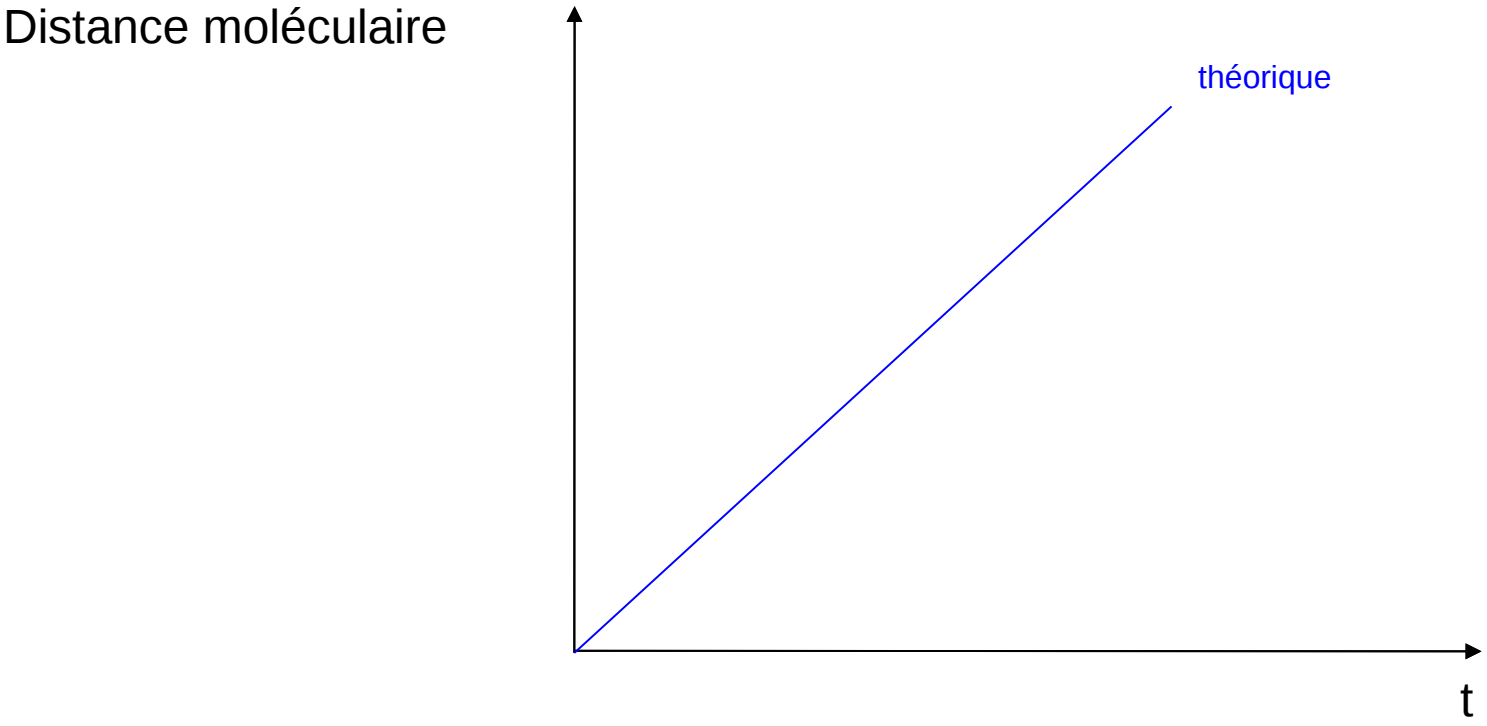
A T T A T A T T T C G T G A AT G C A T

Seq 3 A T T A T A T T T C G T G T A T G C A T

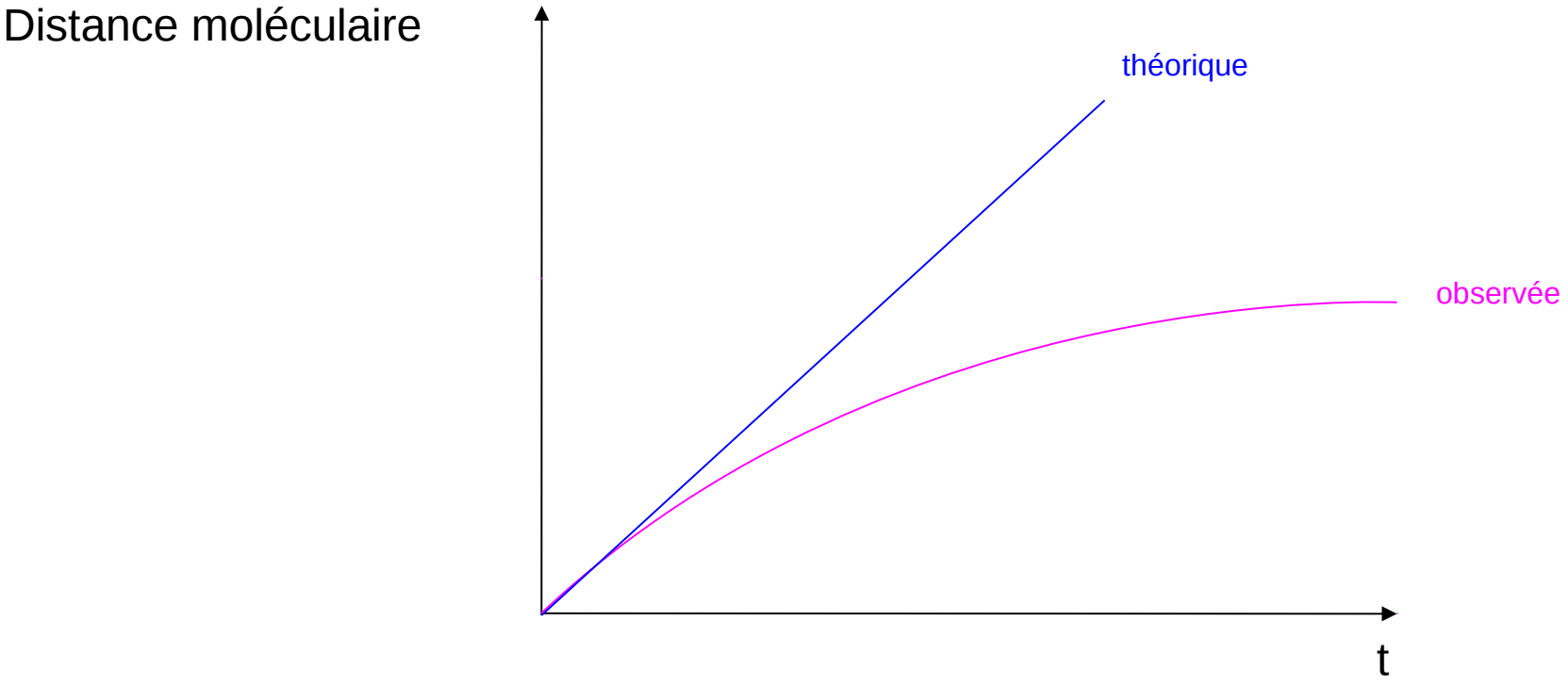
$$d_{(seq1,seq3)} = \frac{\text{\# de substitutions}}{\text{\# de résidus}} = \frac{6}{20} = 0{,}30$$



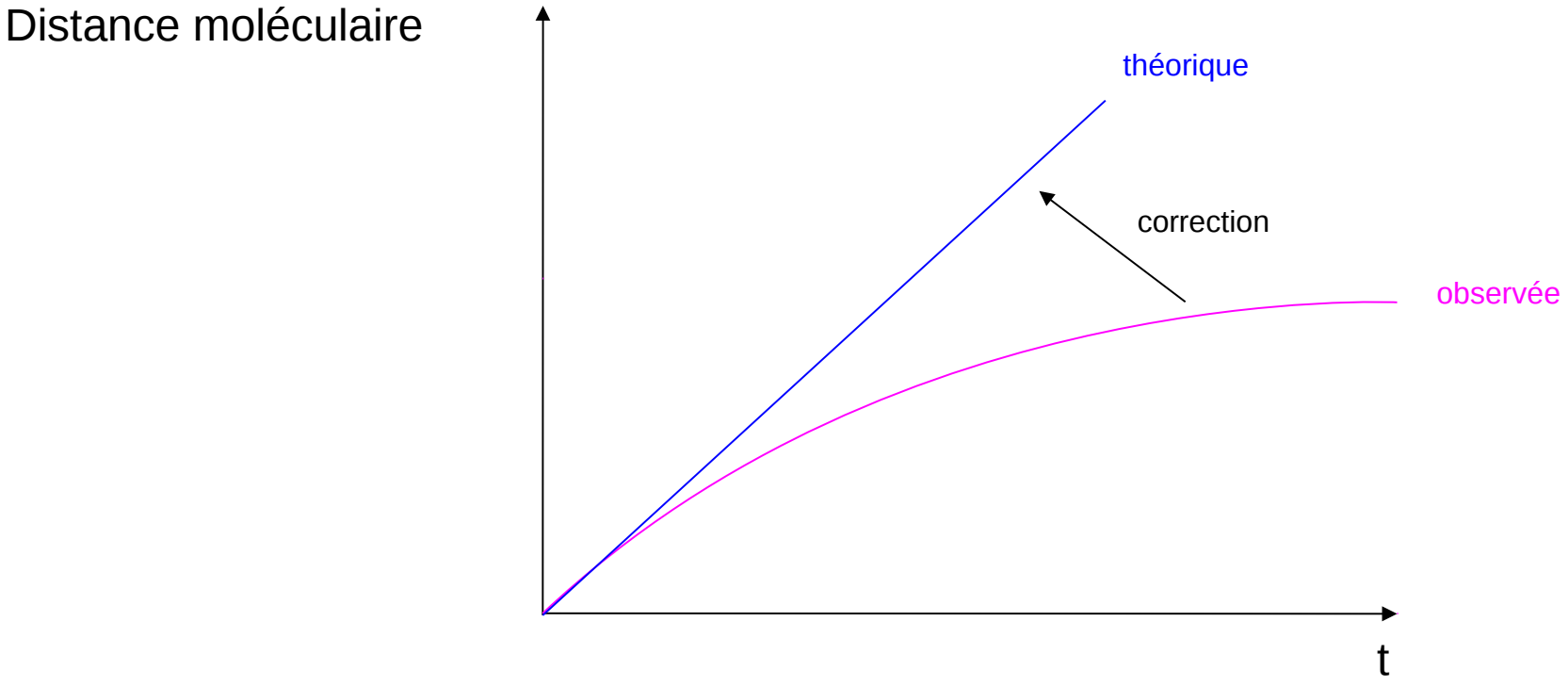
# Calcul de la distance entre deux séquences nucléotidiques



# Calcul de la distance moléculaire entre deux séquences nucléotidiques



# Calcul de la distance entre deux séquences nucléotidiques

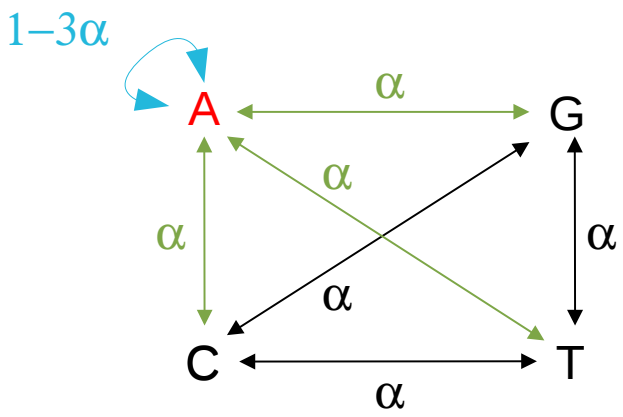


# Calcul de la distance entre deux séquences nucléotidiques

## Le modèle à un paramètre de Jukes-Cantor (1969) : JC69

Toutes les substitutions sont équiprobables

Les proportions de chacune des 4 bases sont identiques




A → A  C/G/T → A

$$p_A(t+1)=p_A(t)(1-3\alpha) + [1-p_A(t)]\alpha = -4\alpha p_A(t) + p_A(t) + \alpha$$

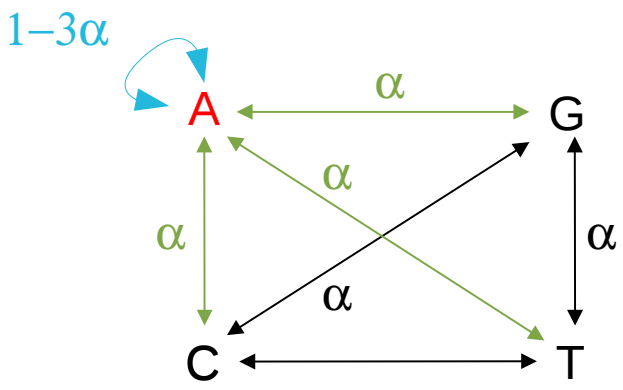
*(t discret)* $\Delta p_A = p_A(t+1) - p_A(t) = -4\alpha p_A(t) + \alpha = \dfrac{dp_A(t)}{dt}$ *(t continu)*

# Calcul de la distance entre deux séquences nucléotidiques

## Le modèle à un paramètre de Jukes-Cantor (1969) : JC69

Toutes les substitutions sont équiprobables

Les proportions de chacune des 4 bases sont identiques




*(t continu)* $$\frac{dp_A(t)}{dt} = -4\alpha p_A(t) + \alpha$$

La solution de cette équation différentielle est de la forme :

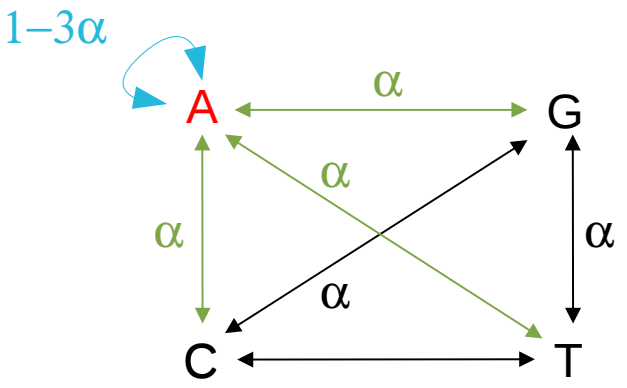
$$P_A(t) = \frac{1}{4} + \left[ P_A(0) - \frac{1}{4} \right] e^{-4\alpha t}$$

# Calcul de la distance entre deux séquences nucléotidiques

## Le modèle à un paramètre de Jukes-Cantor (1969) : JC69

Toutes les substitutions sont équiprobables

Les proportions de chacune des 4 bases sont identiques




$$P_A(t) = \frac{1}{4} + [\, P_A(0) - \frac{1}{4} \,]\, e^{-4\alpha t}$$

$$\text{Si } P_A(0) = 1 \text{ alors } P_A(t) = \frac{1}{4} + \frac{3}{4}\, e^{-4\alpha t} = p_{AA}(t) = p_{ii}(t)$$

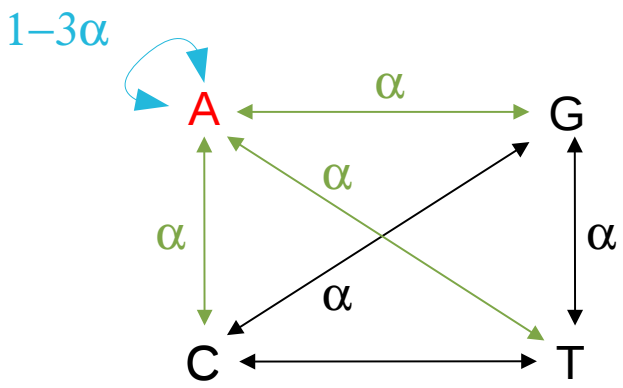
$$\text{Si } P_A(0) = 0 \text{ alors } P_A(t) = \frac{1}{4} - \frac{1}{4}\, e^{-4\alpha t} = p_{C/G/T \to A}(t) = p_{ij}(t)$$

# Calcul de la distance entre deux séquences nucléotidiques

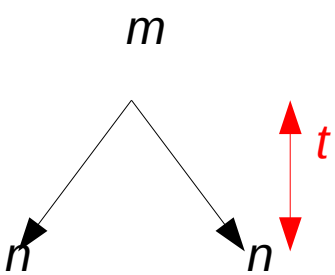
## Le modèle à un paramètre de Jukes-Cantor (1969) : JC69

Toutes les substitutions sont équiprobables

Les proportions de chacune des 4 bases sont identiques




Entre deux séquences la probabilité qu'il y ait le même nucléotide à la même position :




$$p_{id} = p_{mA}^{2} + p_{mT}^{2} + p_{mC}^{2} + p_{mG}^{2} = P_{ii}^{2} + 3.p_{ij}^{2} \qquad (m = A/C/G/T)$$

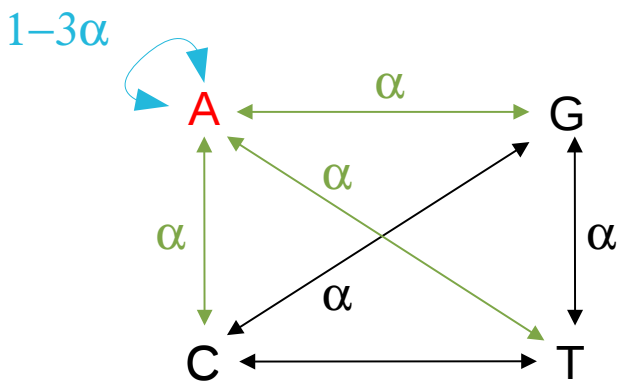
$$p_{id} = \left[\frac{1}{4} + \frac{3}{4} e^{-4\alpha t}\right]^2 + 3.\left[\frac{1}{4} - \frac{1}{4} e^{-4\alpha t}\right]^2 = \frac{1}{4} + \frac{3}{4} e^{-8\alpha t}$$

# Calcul de la distance entre deux séquences nucléotidiques

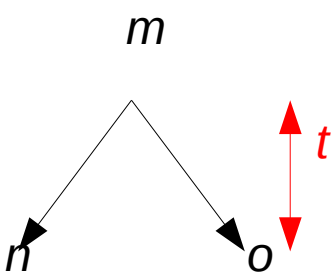
Le modèle à un paramètre de Jukes-Cantor (1969) : JC69

Toutes les substitutions sont équiprobables

Les proportions de chacune des 4 bases sont identiques




Entre deux séquences la probabilité qu'il y ait des nucléotides différents à la même position :




$$P_{nid} = 1 - p_{id} = 1 - \left[\frac{1}{4} + \frac{3}{4} e^{-8\alpha t}\right] = \frac{3}{4}\left(1 - e^{-8\alpha t}\right)$$

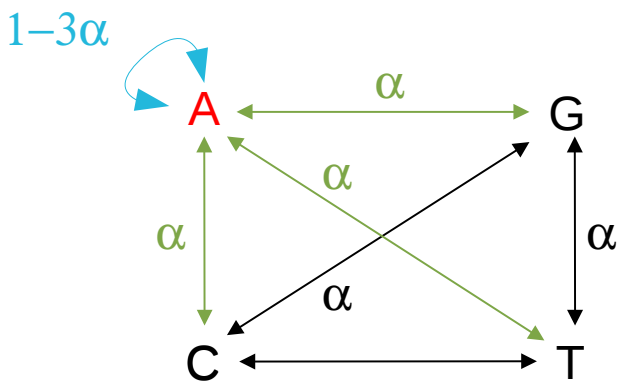
Ou encore, $8\alpha t = -\mathrm{Ln}\left(1 - \frac{4}{3} p_{nid}\right)$ où $\alpha$ et $t$ sont inconnus.

# Calcul de la distance entre deux séquences nucléotidiques

Le modèle à un paramètre de Jukes-Cantor (1969) : JC69

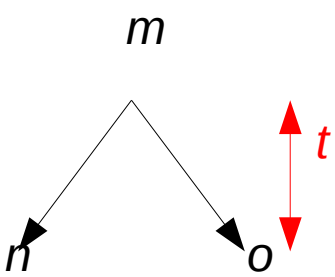
Toutes les substitutions sont équiprobables

Les proportions de chacune des 4 bases sont identiques




Entre deux séquences la probabilité qu'il y ait des nucléotides différents à la même position :

$$8\alpha t = -\,\mathrm{Ln}\left(1 - \frac{4}{3}\, p_{nid}\right)$$




Pour une séquence, le nombre de substitution par site est : $3\alpha$

*Pour deux séquences séparées depuis un temps t,*

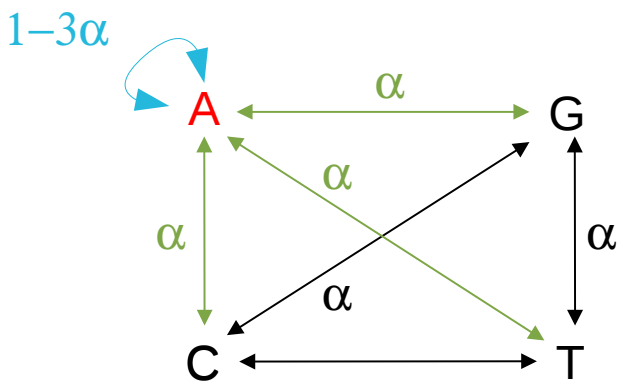
*ce nombre est égal à : $2.3\alpha t = 6\alpha t = K$*

Calcul de la distance entre deux séquences nucléotidiques

Le modèle à un paramètre de Jukes-Cantor (1969) : JC69

Toutes les substitutions sont équiprobables

Les proportions de chacune des 4 bases sont identiques




Entre deux séquences la probabilité qu'il y ait des nucléotides différents à la même position :

$$d'où,\ \frac{4}{3}K = -\,\mathrm{Ln}\left(1 - \frac{4}{3}\,p_{nid}\right)$$

$$et\ \ K = -\frac{3}{4}\,\mathrm{Ln}\left(1 - \frac{4}{3}\,p_{nid}\right)$$

Nombre de substitutions estimées

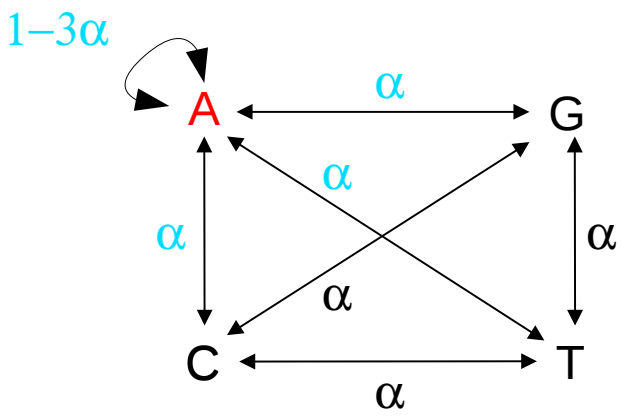
p-distance

# Calcul de la distance entre deux séquences nucléotidiques

Le modèle à un paramètre de Jukes-Cantor (1969) : JC69

Toutes les substitutions sont équiprobables

Les proportions de chacune des 4 bases sont identiques




Soit p la proportion de substitutions observées (*p*-distance) et k le nombre de substitutions estimées entre deux séquences, alors :

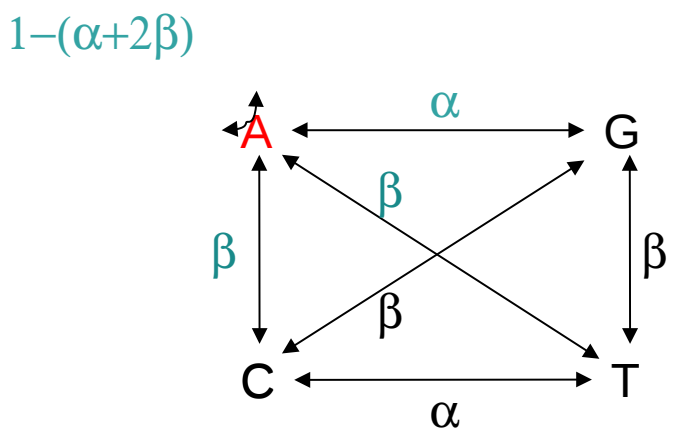
$$k = -\frac{3}{4} \ln\left(1 - \frac{4}{3} p\right)$$

# Calcul de la distance entre deux séquences nucléotidiques

## Le modèle à deux paramètres de Kimura (1980) : K2P

On distingue les probabilités de transition (AG et CT) et de transversion (AC,AT,GC et GT).

Les proportions de chacune des 4 bases sont identiques




Soit p la proportion de transitions observées, q la proportion de transversions observées

et k le nombre de substitutions estimées entre les deux séquences, alors :

$$k = -\frac{1}{2}\ln(1-2p-q) - \frac{1}{4}\ln(1-2q)$$

Calcul de la distance entre deux séquences nucléotidiques

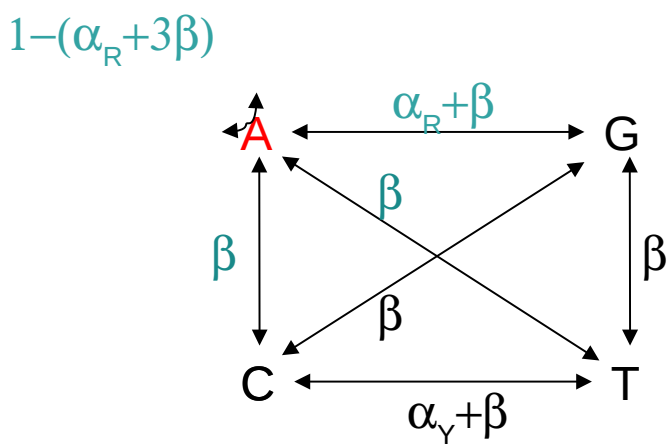
Le modèle à six paramètres de Tamura et Nei (1993) : TN93

On distingue : la probabilité de transition (AG): $\alpha_R$

la probabilité de transition (CT) : $\alpha_Y$

la probabilité de substitution : $\beta$

Les proportions de chacune des 4 bases ne sont pas forcément identiques ($\pi_A$, $\pi_C$, $\pi_G$ et $\pi_T$).




Si $\alpha_R=\alpha_Y$ alors nous somme dans le modèle de Felsenstein (1984) : F84

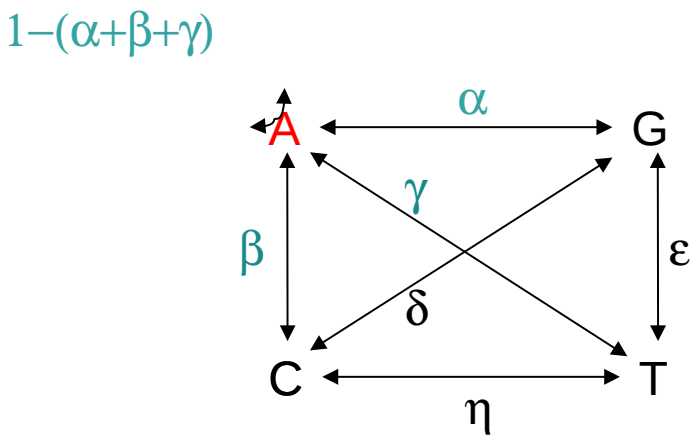
Si $\alpha_r/\alpha_Y = \pi_r/\pi_y$ alors nous sommes dans le modèle Hasegawa, Kishino and Yano (1985) : HKY

# Calcul de la distance entre deux séquences nucléotidiques

Le modèle General Time Reversible : GTR

On distingue 6 probabilités de substitution.

Les proportions de chacune des 4 bases ne sont pas forcément identiques ($\pi_A$, $\pi_C$, $\pi_G$ et $\pi_T$).

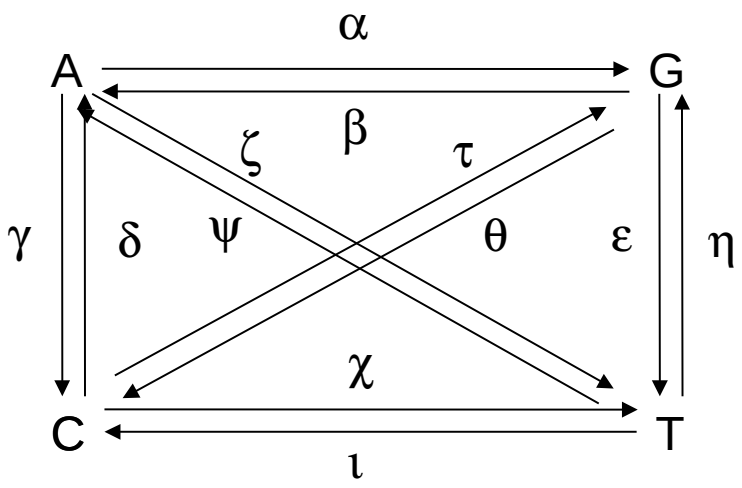


Calcul de la distance entre deux séquences nucléotidiques

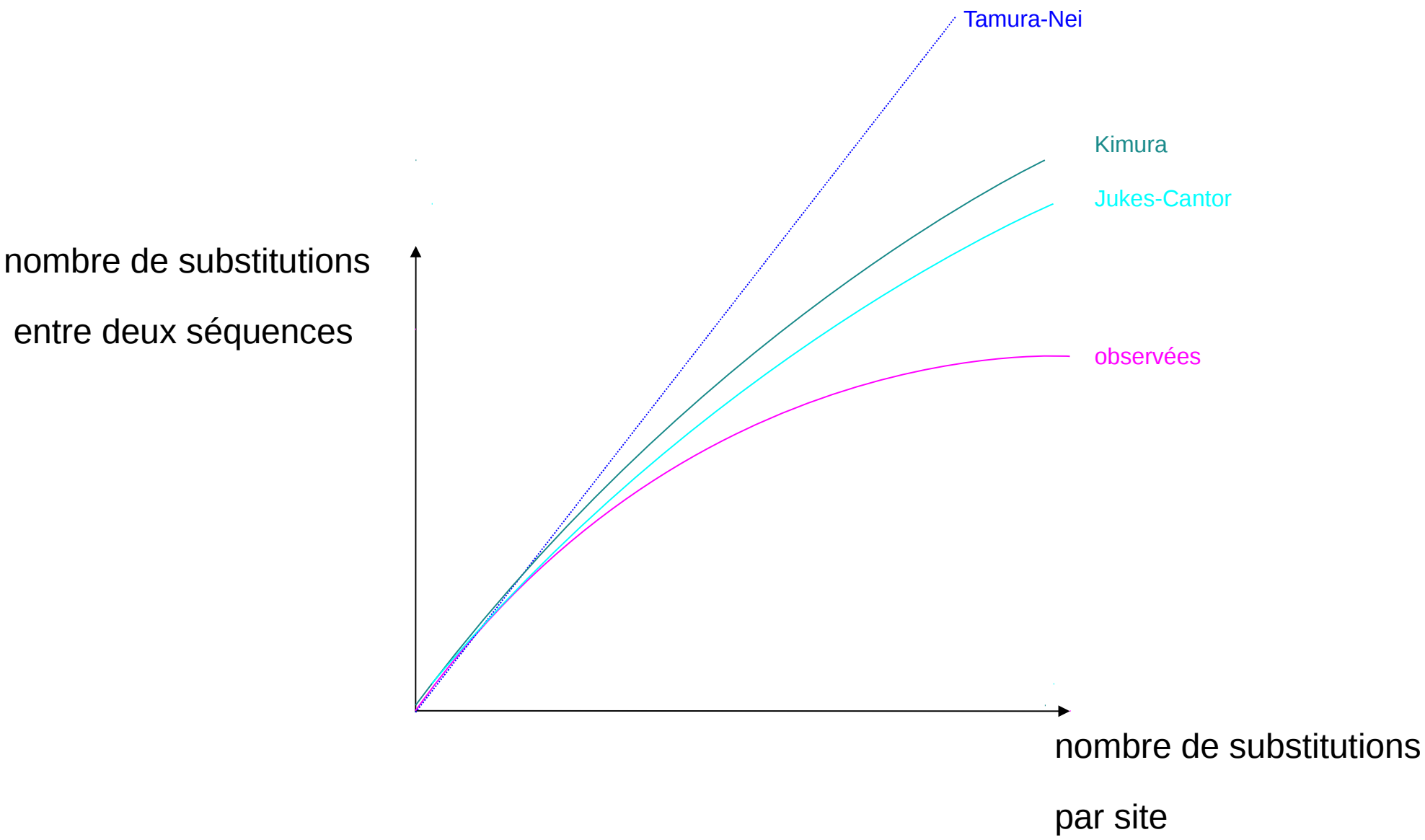
Le modèle General à 12 paramètres :

On distingue 12 probabilités de substitution.

Les proportions de chacune des 4 bases ne sont pas forcément identiques ($\pi_A$, $\pi_C$, $\pi_G$ et $\pi_T$).



# Calcul de la distance entre deux séquences nucléotidiques



# Calcul de la distance entre deux séquences nucléotidiques

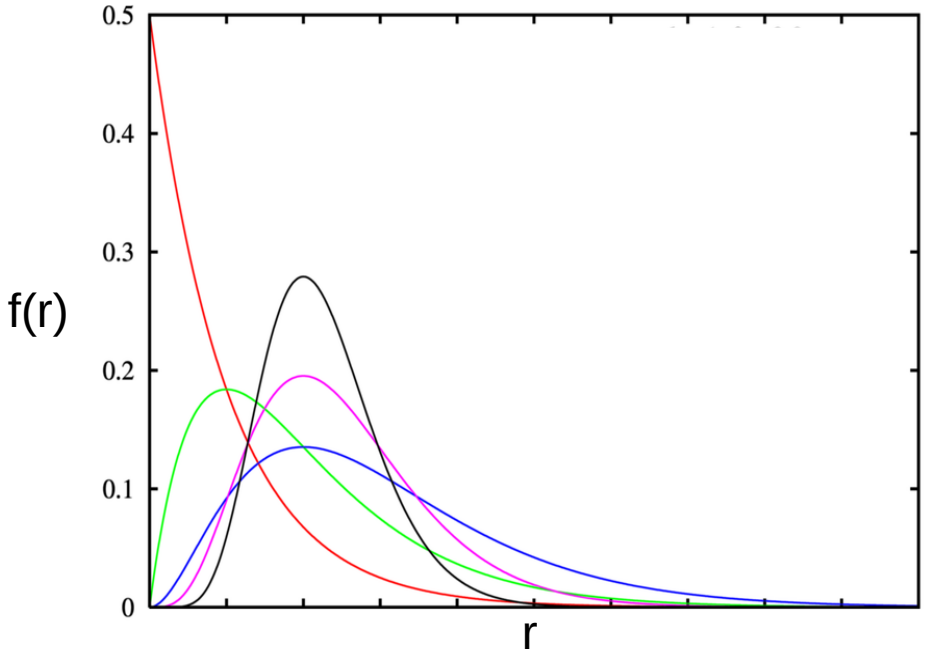
Tous les sites n'évoluent pas avec la même vitesse :

$\bar{r}$ = taux de substitution pour un site donné

$\alpha = \bar{r}^2/V(r)$ (paramètre gamma)

$b = \bar{r}/V(r)$ (facteur d'échelle)

$$f(r) = \frac{b^\alpha}{\Gamma(a)} e^{-br} r^{\alpha-1} \quad \text{avec} \quad \Gamma(a) = \int_0^\infty e^{-t} t^{\alpha-1} dt$$



# Calcul de la distance entre deux séquences protéiques

On utilise des matrices de substitution pour évaluer la distance entre les

paires de séquences protéiques :

Dayhoff : Dayhoff, Eck and Park, 1972

Blosum62 : Henikoff and Henikoff, 1992

JTT : Jones, Taylor and Thornton, 1992

WAG : Whelan and Goldman, 2001

mtREV : Adachi and Hasegawa, 1996

Des corrections existent également :

poisson

kimura

gamma

# Construire un arbre en utilisant les méthodes de distances

## Les méthodes agglomératives

1 . UPGMA

2 . Neighbor-Joining

# **U**nweighted **P**air **G**roup **M**ethod with **A**rithmetic **M**ean (UPGMA)

## Alignement multiple

| | |
|---|---|
| **A** | GCTTGTCCGTTACGAT |
| **B** | ACTTGTCTGTTACGAT |
| **C** | ACTTGTCCGAAACGAT |
| **D** | ACTTGACCGTTTCCTT |
| **E** | AGATGACCGTTTCGAT |
| **F** | ACTACACCCTTATGAG |

## Matrice de distance

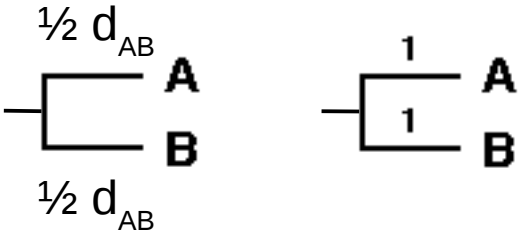
| | A | B | C | D | E |
|---|---|---|---|---|---|
| B | 2 | | | | |
| C | 4 | 4 | | | |
| D | 6 | 6 | 6 | | |
| E | 6 | 6 | 6 | 4 | |
| F | 8 | 8 | 8 | 8 | 8 |

# **U**nweighted **P**air **G**roup **M**ethod with **A**rithmetic **M**ean (UPGMA)

$$d_{AX} = d_{BX} = \frac{d_{AB}}{2}$$

Matrice de distance

| | A | B | C | D | E |
|---|---|---|---|---|---|
| B | 2 | | | | |
| C | 4 | 4 | | | |
| D | 6 | 6 | 6 | | |
| E | 6 | 6 | 6 | 4 | |
| F | 8 | 8 | 8 | 8 | 8 |




On regroupe les OTUs les plus proches et on calcule une nouvelle matrice.

# **U**nweighted **P**air **G**roup **M**ethod with **A**rithmetic **M**ean (UPGMA)

Matrice de distance $M_i$

| | A | B | C | D | E |
|---|---|---|---|---|---|
| B | 2 | | | | |
| C | 4 | 4 | | | |
| D | 6 | 6 | 6 | | |
| E | 6 | 6 | 6 | 4 | |
| F | 8 | 8 | 8 | 8 | 8 |

$$d_{AB,X} = \frac{d_{A,X} + d_{B,X}}{2}$$

Matrice de distance $M_{i+1}$

| | A B | C | D | E |
|---|---|---|---|---|
| C | | | | |
| D | | 6 | | |
| E | | 6 | 4 | |
| F | | 8 | 8 | 8 |

# **U**nweighted **P**air **G**roup **M**ethod with **A**rithmetic **M**ean (UPGMA)

Matrice de distance $M_i$

| | A | B | C | D | E |
|---|---|---|---|---|---|
| B | 2 | | | | |
| C | 4 | 4 | | | |
| D | 6 | 6 | 6 | | |
| E | 6 | 6 | 6 | 4 | |
| F | 8 | 8 | 8 | 8 | 8 |

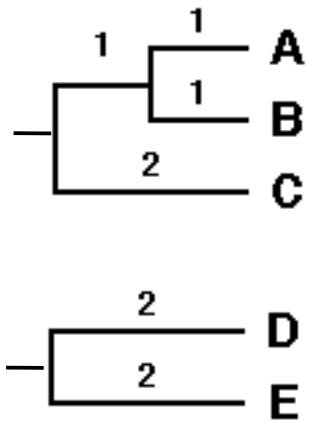
Matrice de distance $M_{i+1}$

| | A,B | C | D | E |
|---|---|---|---|---|
| C | 4 | | | |
| D | 6 | 6 | | |
| E | 6 | 6 | 4 | |
| F | 8 | 8 | 8 | 8 |

# **U**nweighted **P**air **G**roup **M**ethod with **A**rithmetic **M**ean (UPGMA)

Matrice de distance $M_{i+1}$

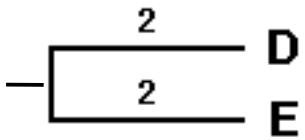
| | A,B | C | D | E |
|---|---|---|---|---|
| C | 4 | | | |
| D | 6 | 6 | | |
| E | 6 | 6 | 4 | |
| F | 8 | 8 | 8 | 8 |



# **U**nweighted **P**air **G**roup **M**ethod with **A**rithmetic **M**ean (UPGMA)

Matrice de distance $M_{i+1}$

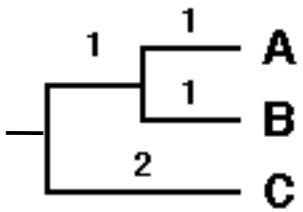
| | A,B | C | D | E |
|---|---|---|---|---|
| C | 4 | | | |
| D | 6 | 6 | | |
| E | 6 | 6 | 4 | |
| F | 8 | 8 | 8 | 8 |



# **U**nweighted **P**air **G**roup **M**ethod with **A**rithmetic **M**ean (UPGMA)

Matrice de distance $M_{i+2}$

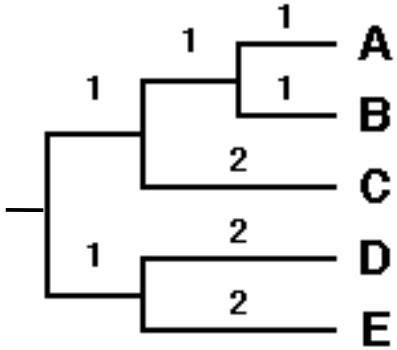
| | A,B | C | D,E |
|---|---|---|---|
| C | 4 | | |
| D,E | 6 | 6 | |
| F | 8 | 8 | 8 |



# **U**nweighted **P**air **G**roup **M**ethod with **A**rithmetic **M**ean (UPGMA)

Matrice de distance $M_{i+3}$

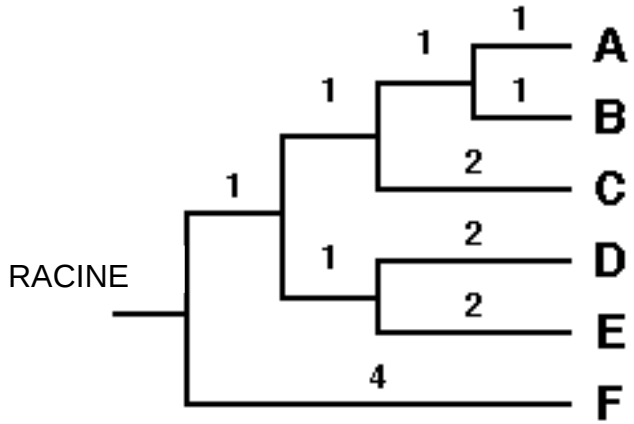
| | (A,B),C | D,E |
|---|---|---|
| D,E | 6 | |
| F | 8 | 8 |



# **U**nweighted **P**air **G**roup **M**ethod with **A**rithmetic **M**ean (UPGMA)

Matrice de distance $M_{i+4}$

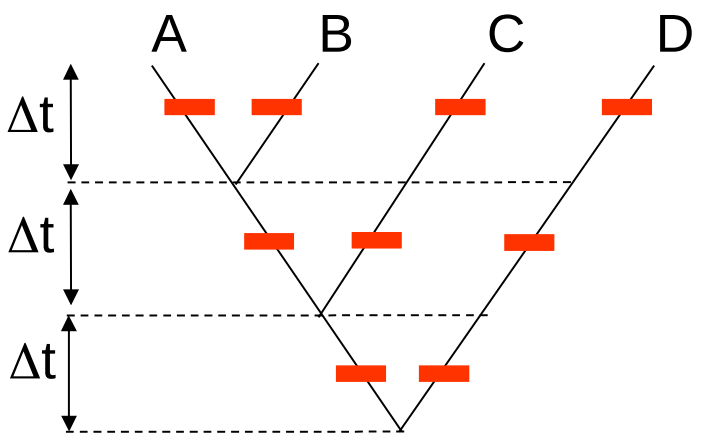
| | ((A,B),C),(D,E) |
|---|---|
| F | 8 |

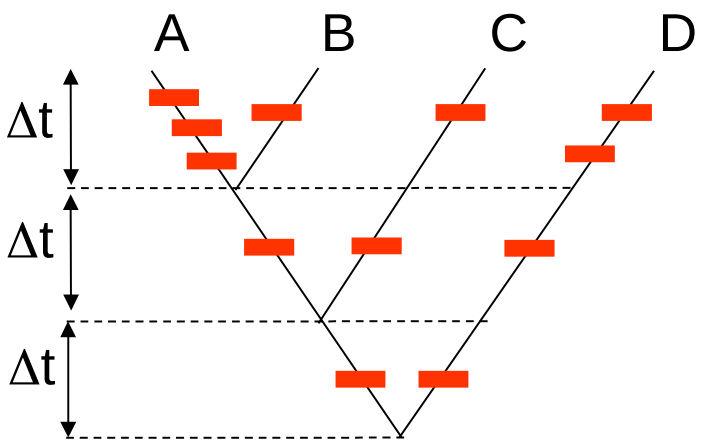


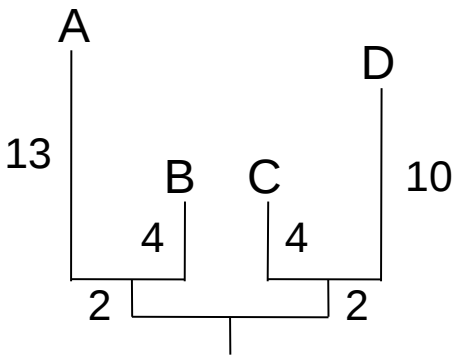
La méthode UPGMA :

- est rapide et simple

- mais sensible à l'horloge moléculaire (MCH, Zuckerkandl and Pauling, 1962)



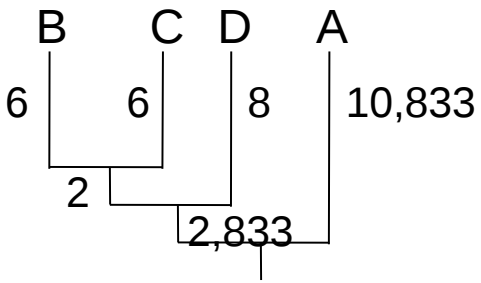



Arbre vrai

| | A | B | C |
|---|---|---|---|
| B | 17 | | |
| C | 21 | 12 | |
| D | 27 | 18 | 14 |

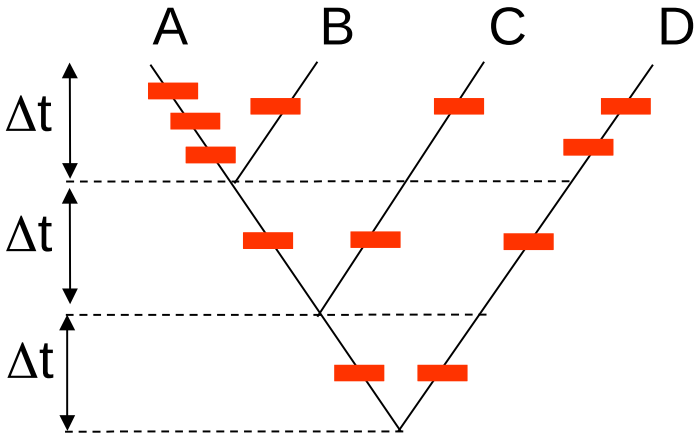
Matrice de distance

(arborée)



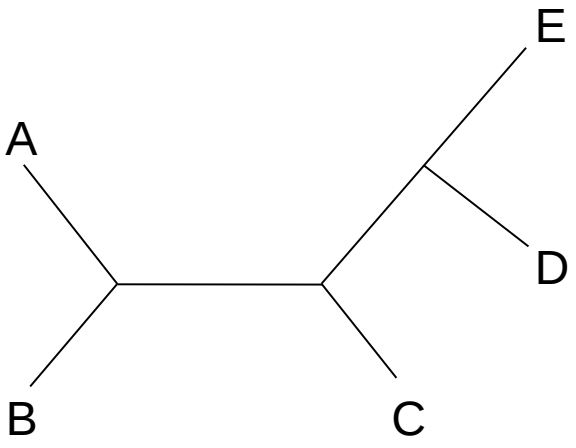

Arbre UPGMA

(LBA)



# **N**eighbor **J**oining (NJ)

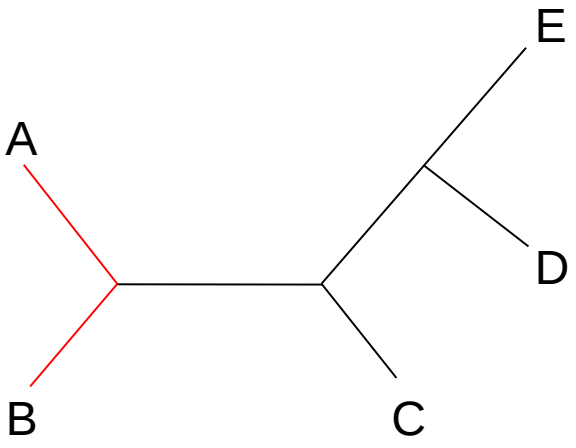
(Saitou and Nei, *Mol. Biol. Evol.* 1987)




Dans un arbre non raciné, deux taxons *i* et *j* sont considérés comme voisin s'il ne sont séparés que par un seul noeud interne.

# **N**eighbor **J**oining (NJ)

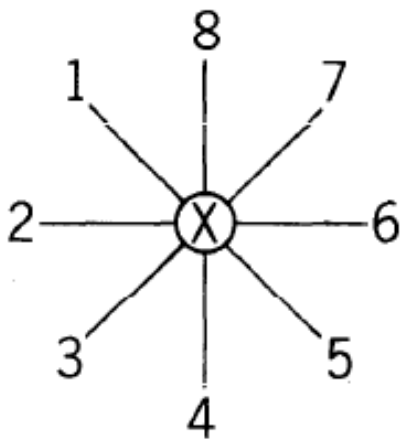
(Saitou and Nei, *Mol. Biol. Evol.* 1987)




Dans un arbre non raciné, deux taxons *i* et *j* sont considérés comme voisin s'il ne sont séparés que par un seul noeud interne.

# **N**eighbor **J**oining (NJ)

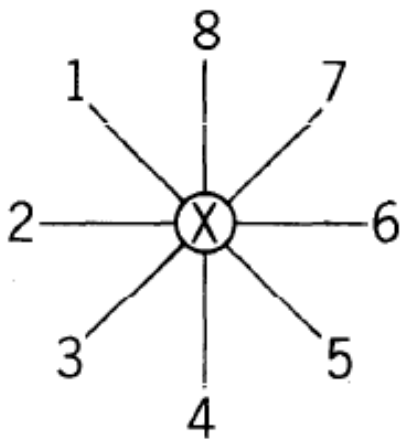
(Saitou and Nei, *Mol. Biol. Evol.* 1987)




$$S_0 = \sum_{i=1}^{n} L_{ix}$$

# **N**eighbor **J**oining (NJ)

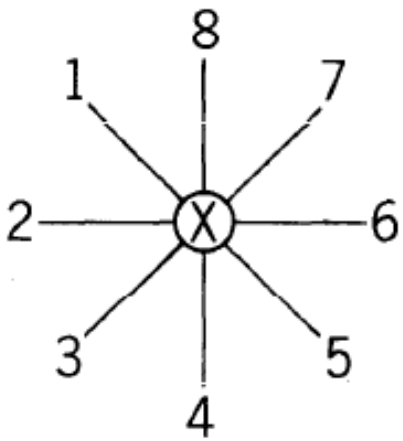
(Saitou and Nei, *Mol. Biol. Evol.* 1987)




$$S_0 = \sum_{i=1}^{n} L_{ix}$$

# **N**eighbor **J**oining (NJ)

(Saitou and Nei, *Mol. Biol. Evol.* 1987)

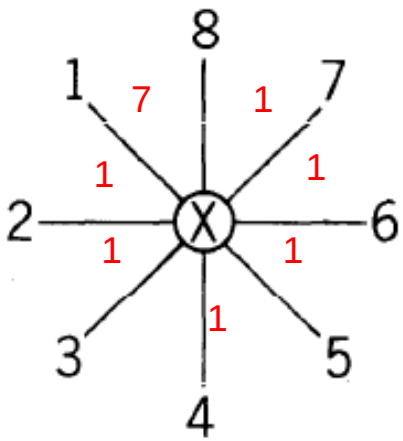



$$S_0 = \sum_{i=1}^{n} L_{ix}$$

$$\sum_{i<j}^{n} d_{ij} = \sum_{i<j}^{n} (d_{ix}+d_{xj})$$

# **N**eighbor **J**oining (NJ)

(Saitou and Nei, *Mol. Biol. Evol.* 1987)

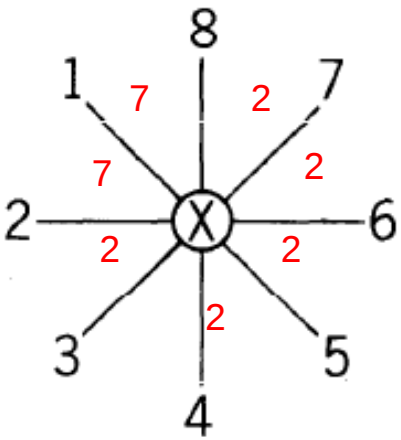



$$S_0 = \sum_{i=1}^{n} L_{ix}$$

$$\sum_{i<j}^{n} d_{ij} = \sum_{i<j}^{n} (d_{ix}+d_{xj})$$

# **N**eighbor **J**oining (NJ)

(Saitou and Nei, *Mol. Biol. Evol.* 1987)

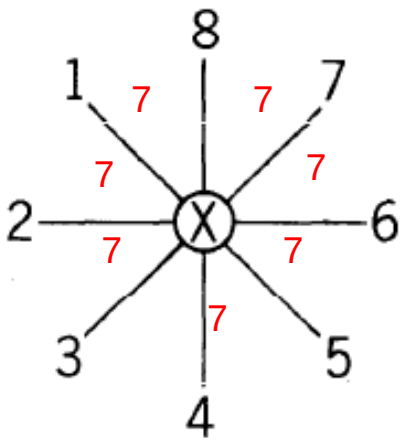



$$S_0 = \sum_{i=1}^{n} L_{ix}$$

$$\sum_{i<j}^{n} d_{ij} = \sum_{i<j}^{n} (d_{ix}+d_{xj})$$

# **N**eighbor **J**oining (NJ)

(Saitou and Nei, *Mol. Biol. Evol.* 1987)




$$S_0 = \sum_{i=1}^{n} L_{ix}$$

$$\sum_{i<j}^{n} d_{ij} = \sum_{i<j}^{n} (d_{ix}+d_{xj})$$

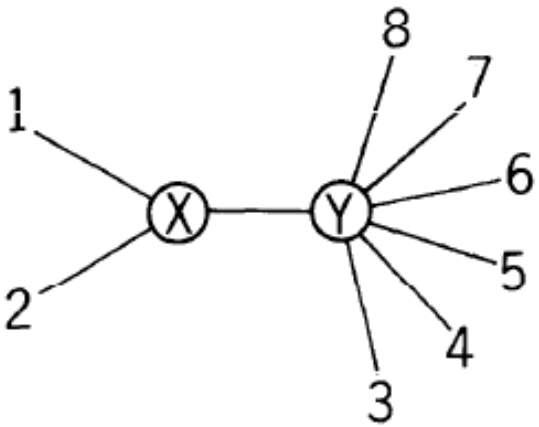
$$= (n-1) \sum_{i=1}^{n} d_{ix}$$

$$= (n-1)\, S_0$$

$$S_0 = \frac{1}{(n-1)} \sum_{i<j}^{n} d_{ij}$$

# **N**eighbor **J**oining (NJ)

(Saitou and Nei, *Mol. Biol. Evol.* 1987)




$$S_{12} = L_{1x} + L_{2x} + L_{xy} + \sum_{k=3}^{n} L_{ky}$$

$$d_{12} = L_{1x} + L_{2x}$$

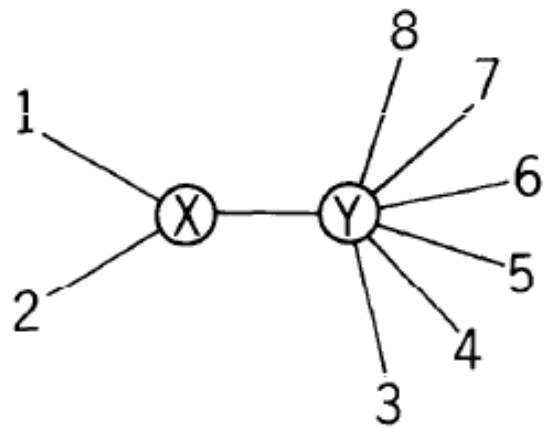
$$\sum_{k=3}^{n} L_{ky} = \frac{1}{n-3} \sum_{3 \le i < j}^{n} d_{ij}$$

$$S_{12} = d_{12} + L_{xy} + \frac{1}{n-3} \sum_{3 \le i < j}^{n} d_{ij}$$

$$L_{xy} = \frac{1}{2\,(n-2)} \left( \sum_{i=3}^{n} ( d_{1i} + d_{2i} ) - ( n-2 )( L_{1x} + L_{2x} ) - 2 \sum_{i=3}^{n} d_{iy} \right)$$

# **N**eighbor **J**oining (NJ)

(Saitou and Nei, *Mol. Biol. Evol.* 1987)



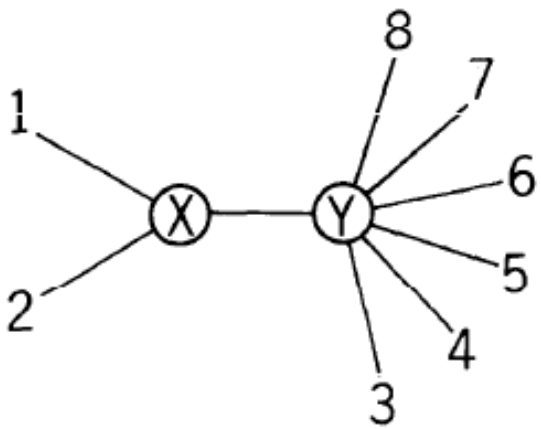

$$S_{ij} = L_{ix} + L_{jx} + L_{xy} + \sum_{k}^{n}{}_{i,j} L_{ky}$$

$$S_{ij} = d_{ij} + L_{xy} + \sum_{k}^{n}{}_{i,j} L_{ky}$$

Il reste donc a trouver le couple *ij* qui minimise $S_{ij}$

# **N**eighbor **J**oining (NJ)

(Saitou and Nei, *Mol. Biol. Evol.* 1987)




$$d_{xi} = \frac{1}{2\,(n-2)} \left( (n-2)\, d_{ij} + \sum_{k=1}^{n} d_{ik} - \sum_{k=1}^{n} d_{jk} \right)$$

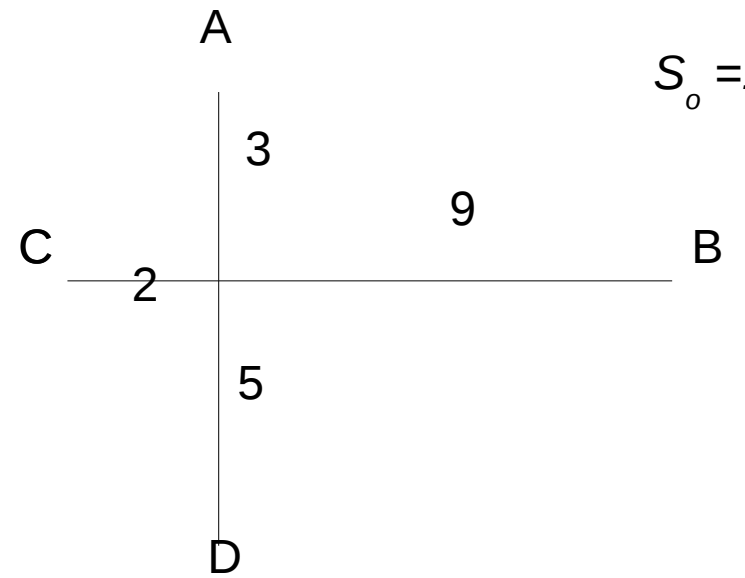
$$d_{xj} = \frac{1}{2\,(n-2)} \left( (n-2)\, d_{ij} - \sum_{k=1}^{n} d_{ik} + \sum_{k=1}^{n} d_{jk} \right)$$

$$d_{xk} = \frac{d_{ix} + d_{jx} - d_{ij}}{2} \qquad \text{pour } k \quad i,j$$

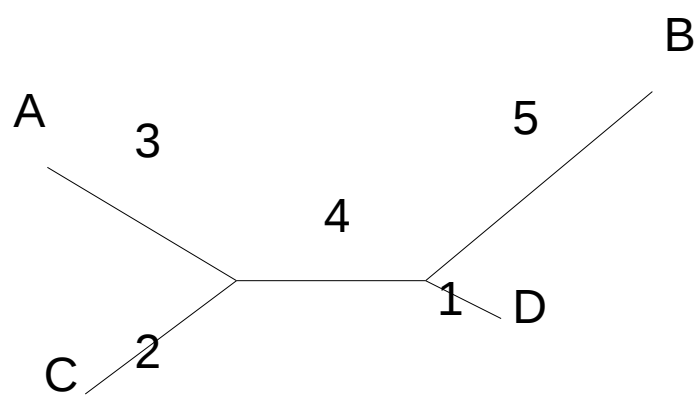
# **N**eighbor **J**oining (NJ)

(Saitou and Nei, *Mol. Biol. Evol.* 1987)

| $d_{ij}$ | A | B | C |
|---|---|---|---|
| B | 12 | | |
| C | 5 | 11 | |
| D | 8 | 14 | 7 |




$S_o = 19$




$S_{AC} = 15$

# **N**eighbor **J**oining (NJ)

(Saitou and Nei, *Mol. Biol. Evol.* 1987)

Si l'on pose :

$$T = \sum_{i<j}^{n} d_{ij}$$

$$R_i = \sum_{k=1}^{n} d_{ik}$$

$$R_j = \sum_{k=1}^{n} d_{jk}$$

alors :

$$S_o = \frac{T}{(n-1)}$$

$$S_{ij} = \frac{2T - R_i - R_j}{2\,(n-2)} + \frac{d_{ij}}{2}$$

$$d_{xi} = \frac{1}{2\,(n-2)} \left( (n-2)\, d_{ij} + R_i - R_j \right)$$

$$d_{xj} = \frac{1}{2\,(n-2)} \left( (n-2)\, d_{ij} - R_i + R_j \right)$$

$$d_{xk} = \frac{d_{ix} + d_{jx} - d_{ij}}{2}$$ pour $k$ $i,j$

# **N**eighbor **J**oining (NJ)

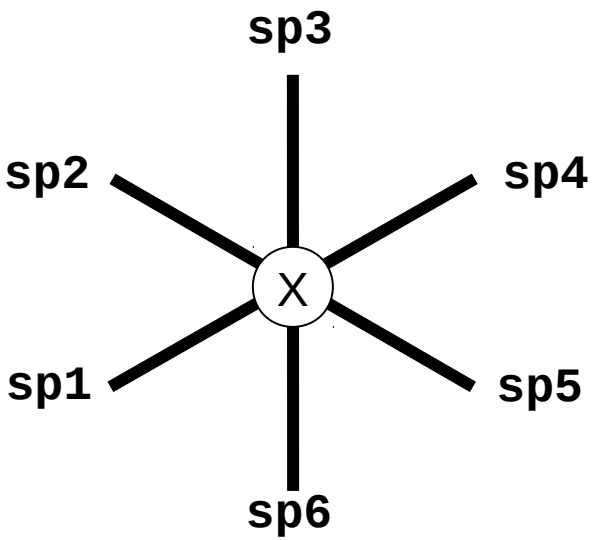
(Saitou and Nei, *Mol. Biol. Evol.* 1987)

Matrice de distance ($D_0$)

| | **sp1** | **sp2** | **sp3** | **sp4** | **sp5** |
|---|---|---|---|---|---|
| **sp2** | 9 | | | | |
| **sp3** | 12 | 7 | | | |
| **sp4** | 15 | 10 | 5 | | |
| **sp5** | 20 | 15 | 10 | 11 | |
| **sp6** | 16 | 11 | 6 | 7 | 8 |

T = 162

n = 6



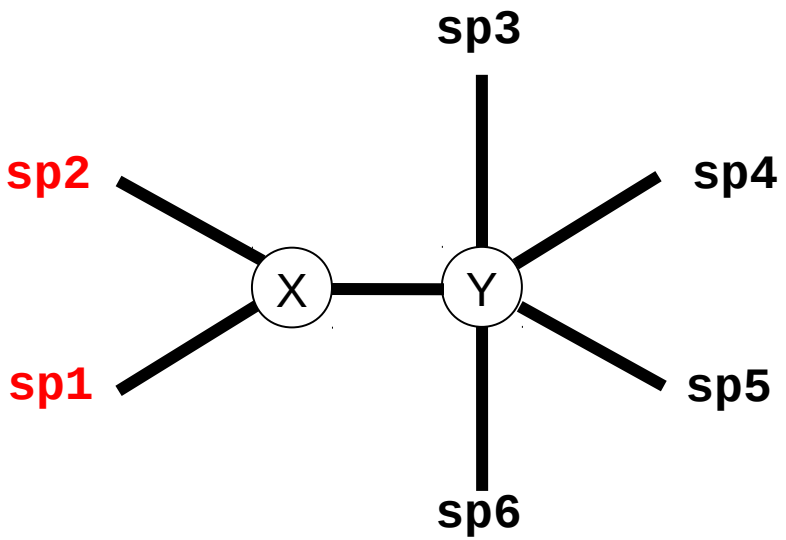

$$S_0 = \frac{162}{(6-1)} = 32{,}4$$

# **N**eighbor **J**oining (NJ)

## (Saitou and Nei, *Mol. Biol. Evol.* 1987)

Matrice de distance ($D_0$)

| | **sp1** | **sp2** | **sp3** | **sp4** | **sp5** |
|---|---|---|---|---|---|
| **sp2** | 9 | | | | |
| **sp3** | 12 | 7 | | | |
| **sp4** | 15 | 10 | 5 | | |
| **sp5** | 20 | 15 | 10 | 11 | |
| **sp6** | 16 | 11 | 6 | 7 | 8 |




$T = 162$

$n = 6$

$R_1 = 72$

$R_2 = 52$

$d_{12} = 9$

$$S_{12} = \frac{(2x162)-72-52}{2\,(6-2)} + \frac{9}{2} = 29{,}5$$

# **N**eighbor **J**oining (NJ)

## (Saitou and Nei, *Mol. Biol. Evol.* 1987)

Matrice de distance ($D_0$)

| | **sp1** | **sp2** | **sp3** | **sp4** | **sp5** |
|---|---|---|---|---|---|
| **sp2** | 9 | | | | |
| **sp3** | 12 | 7 | | | |
| **sp4** | 15 | 10 | 5 | | |
| **sp5** | 20 | 15 | 10 | 11 | |
| **sp6** | 16 | 11 | 6 | 7 | 8 |

T = 162

n = 6

Matrice $S_{ij}$

| | **sp1** | **sp2** | **sp3** | **sp4** | **sp5** |
|---|---|---|---|---|---|
| **sp2** | 29,5 | | | | |
| **sp3** | 32,5 | 32,5 | | | |
| **sp4** | 33,0 | 33,0 | 32,0 | | |
| **sp5** | 33,5 | 33,5 | 32,5 | 32,0 | |
| **sp6** | 33,5 | 33,5 | 32,5 | 32,0 | 30,5 |

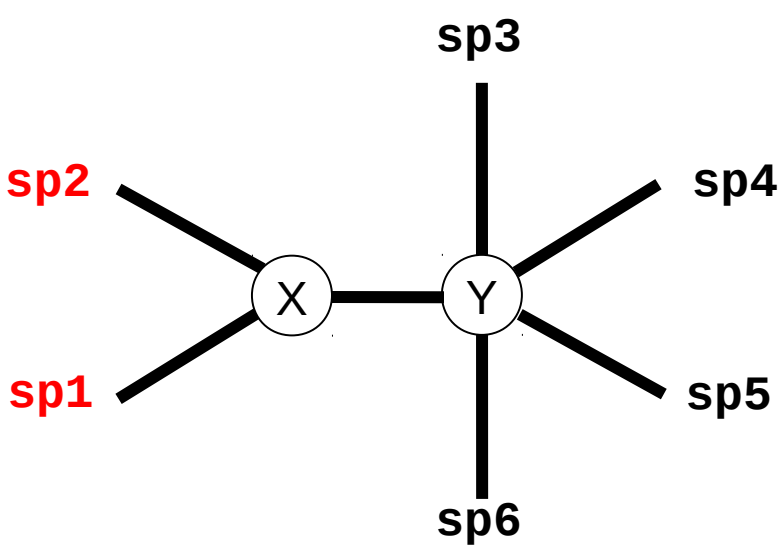
# **N**eighbor **J**oining (NJ)

## (Saitou and Nei, *Mol. Biol. Evol.* 1987)

Matrice de distance ($D_0$)

| | **sp1** | **sp2** | **sp3** | **sp4** | **sp5** |
|---|---|---|---|---|---|
| **sp2** | 9 | | | | |
| **sp3** | 12 | 7 | | | |
| **sp4** | 15 | 10 | 5 | | |
| **sp5** | 20 | 15 | 10 | 11 | |
| **sp6** | 16 | 11 | 6 | 7 | 8 |

$S_{12}$ = 29,5




T = 162

n = 6

$R_1 = 72$

$R_2 = 52$

$d_{12} = 9$

$$d_{1X} = \frac{1}{2\,(6-2)}\Big((6-2)\,9 + 72 - 52\Big) = 7$$

$$d_{2X} = \frac{1}{2\,(6-2)}\Big((6-2)\,9 - 72 + 52\Big) = 2$$

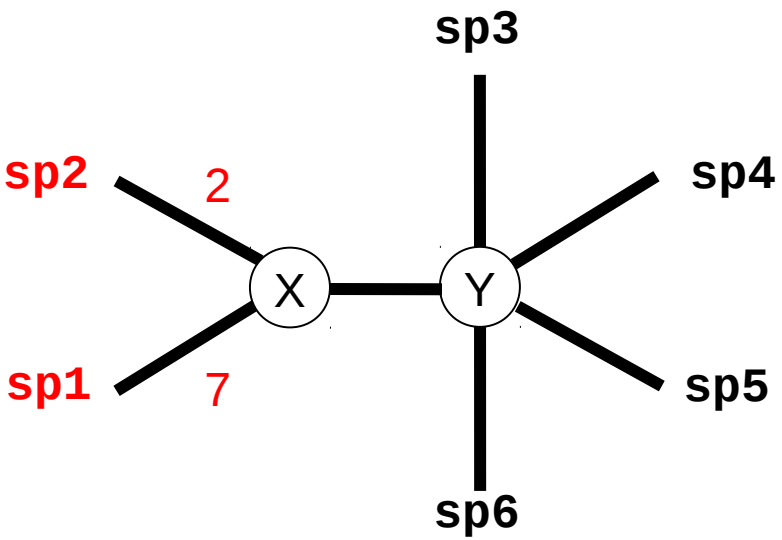
# **N**eighbor **J**oining (NJ)

## (Saitou and Nei, *Mol. Biol. Evol.* 1987)

Matrice de distance ($D_0$)

| | **sp1** | **sp2** | **sp3** | **sp4** | **sp5** |
|---|---|---|---|---|---|
| **sp2** | 9 | | | | |
| **sp3** | 12 | 7 | | | |
| **sp4** | 15 | 10 | 5 | | |
| **sp5** | 20 | 15 | 10 | 11 | |
| **sp6** | 16 | 11 | 6 | 7 | 8 |

$S_{12}$ = 29,5




T = 162

n = 6

$R_1 = 72$

$R_2 = 52$

$d_{12} = 9$

$$d_{xk} = \frac{d_{1x} + d_{2x} - 9}{2}$$

$$d_{x3} = \frac{12 + 7 - 9}{2} = 5$$

# **N**eighbor **J**oining (NJ)

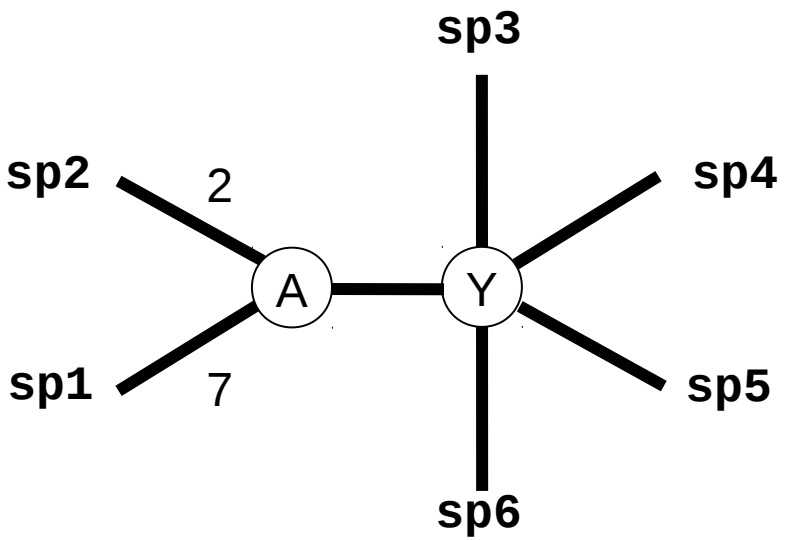
(Saitou and Nei, *Mol. Biol. Evol.* 1987)

Matrice de distance ($D_1$)

| | **A** | **sp3** | **sp4** | **sp5** |
|---|---|---|---|---|
| **sp3** | 5 | | | |
| **sp4** | 8 | 5 | | |
| **sp5** | 13 | 10 | 11 | |
| **sp6** | 9 | 6 | 7 | 8 |

T = 82

n = 6



# **N**eighbor **J**oining (NJ)

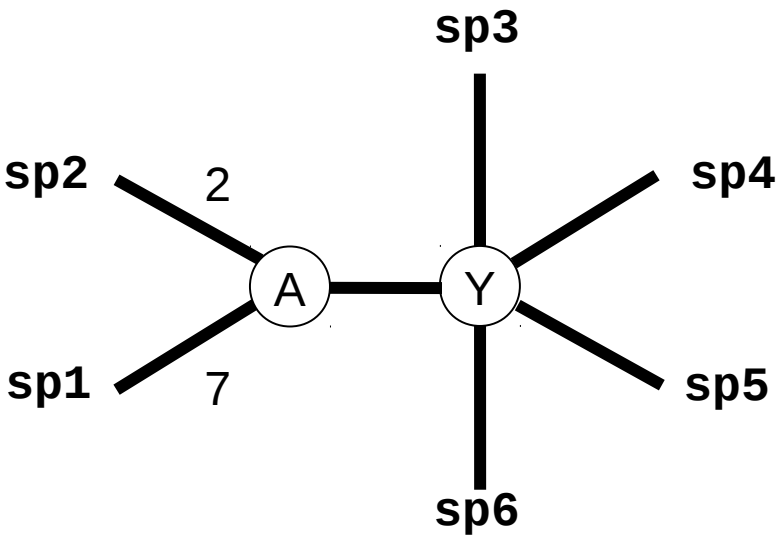
(Saitou and Nei, *Mol. Biol. Evol.* 1987)

Matrice de distance ($D_1$)

| | **A** | **sp3** | **sp4** | **sp5** |
|---|---|---|---|---|
| **sp3** | 5 | | | |
| **sp4** | 8 | 5 | | |
| **sp5** | 13 | 10 | 11 | |
| **sp6** | 9 | 6 | 7 | 8 |

T = 82

n = 5

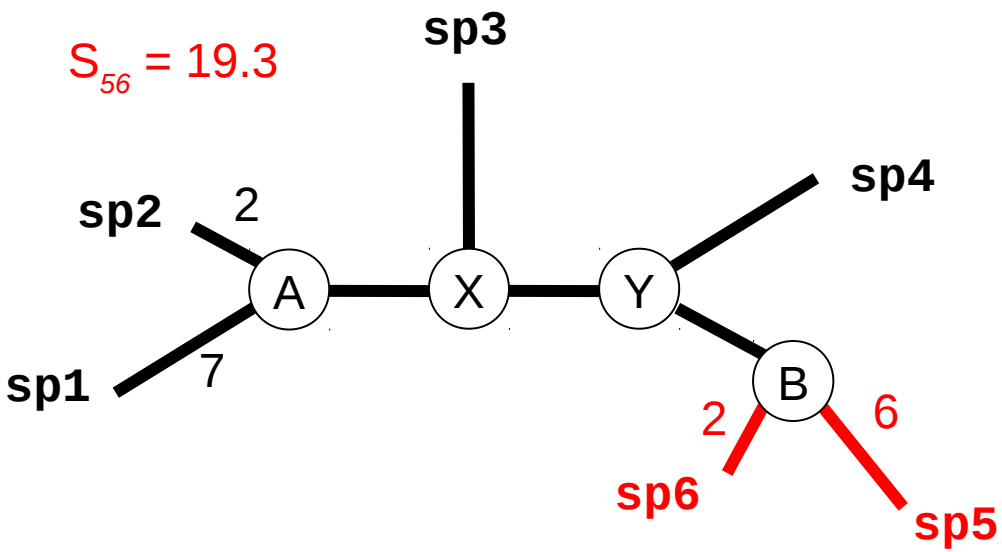



Matrice $S_{ij}$ ($D_1$)

| | **A** | **sp3** | **sp4** | **sp5** |
|---|---|---|---|---|
| **sp3** | 19,7 | | | |
| **sp4** | 20,3 | 20,3 | | |
| **sp5** | 21,0 | 21,0 | 20,7 | |
| **sp6** | 21,0 | 21,0 | 20,7 | 19,3 |

# **N**eighbor **J**oining (NJ)

(Saitou and Nei, *Mol. Biol. Evol.* 1987)




Matrice de distance ($D_2$)

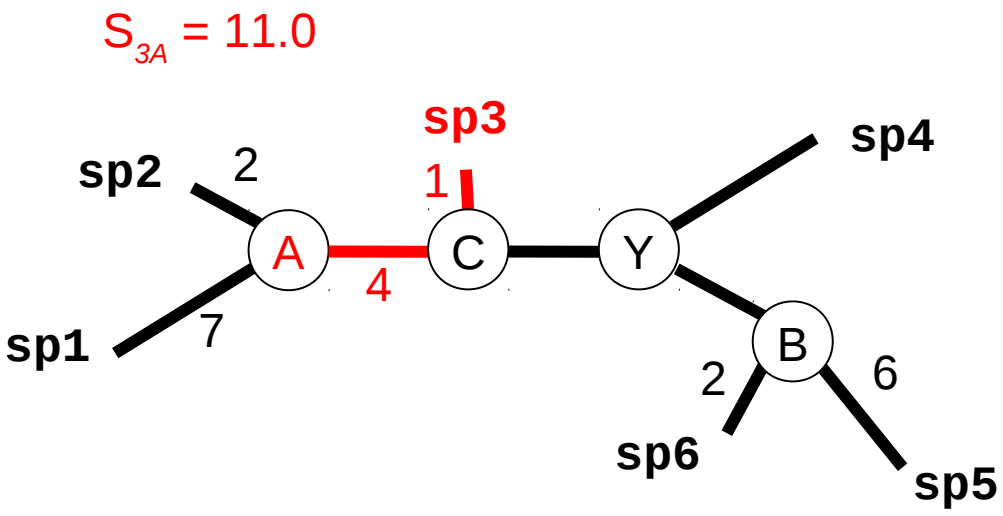
| | **A** | **sp3** | **sp4** |
|---|---|---|---|
| **sp3** | 5 | | |
| **sp4** | 8 | 5 | |
| **B** | 7 | 4 | 5 |

Matrice de $S_{ij}$ ($D_2$)

| | **A** | **sp3** | **sp4** |
|---|---|---|---|
| **sp3** | 11,0 | | |
| **sp4** | 11,5 | 11,5 | |
| **B** | 11,5 | 11,5 | 11,0 |

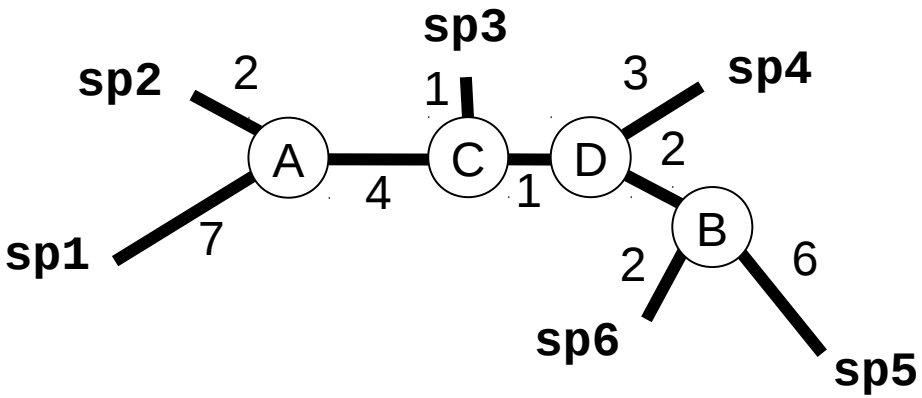
# **N**eighbor **J**oining (NJ)

(Saitou and Nei, *Mol. Biol. Evol.* 1987)



# **N**eighbor **J**oining (NJ)

(Saitou and Nei, *Mol. Biol. Evol.* 1987)

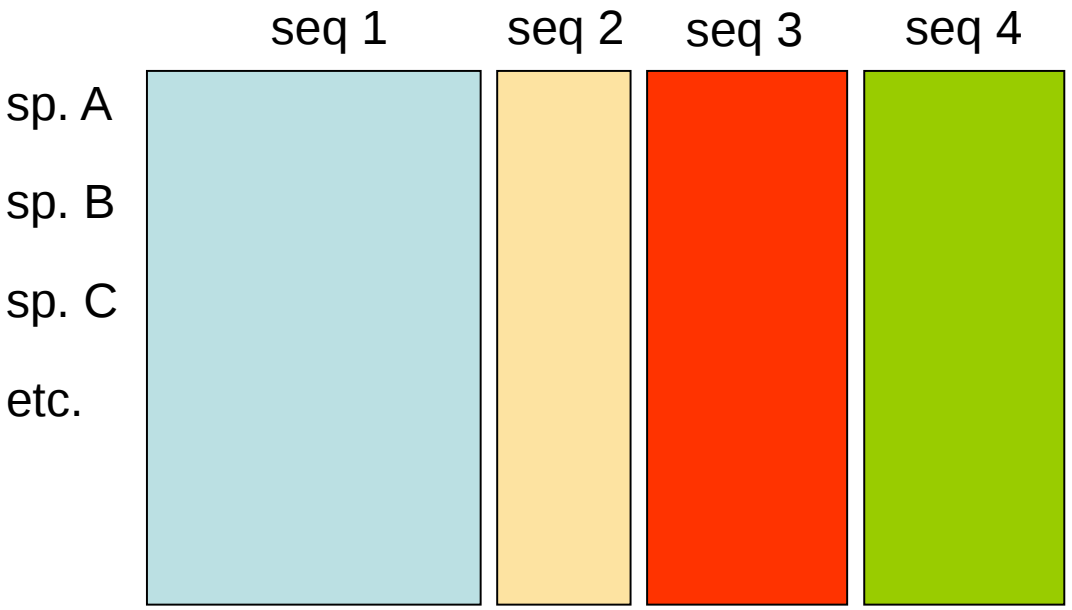


# Méthodes de distance

Très rapide

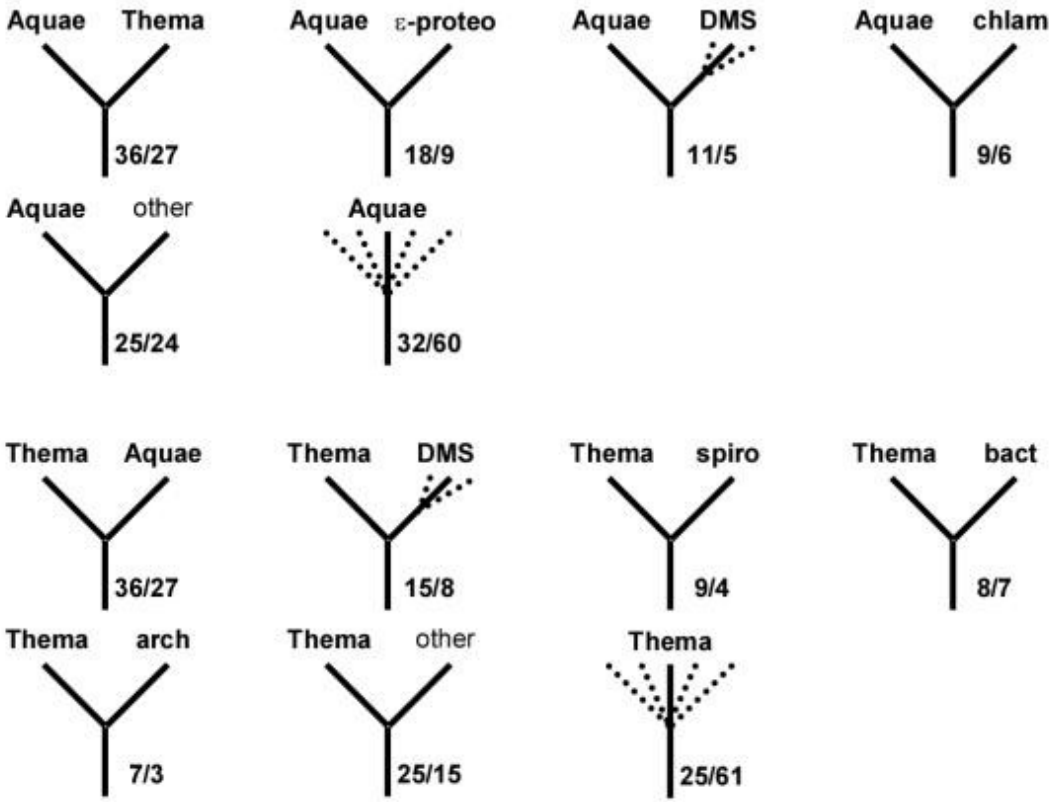
Correction possible des distances selon les modèles d'évolution

# Méthode dite de «Total Evidence»



# Approaches based on multiple trees (Super tree)

congruence



# Approaches based on multiple trees (Super tree)

## Matrix Representation Parsimony (MRP)

T1 : ((A,B),(C,D));

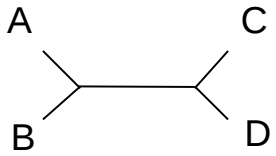
T2 : (A,B,C);

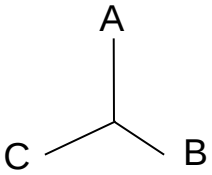
| | | |
|---|---|---|
| A | 1 | 0 |
| B | 1 | 0 |
| C | 0 | 1 |
| D | 0 | 1 |

T1

| |
|---|
| 1 |
| 1 |
| 1 |
| ? |

T2



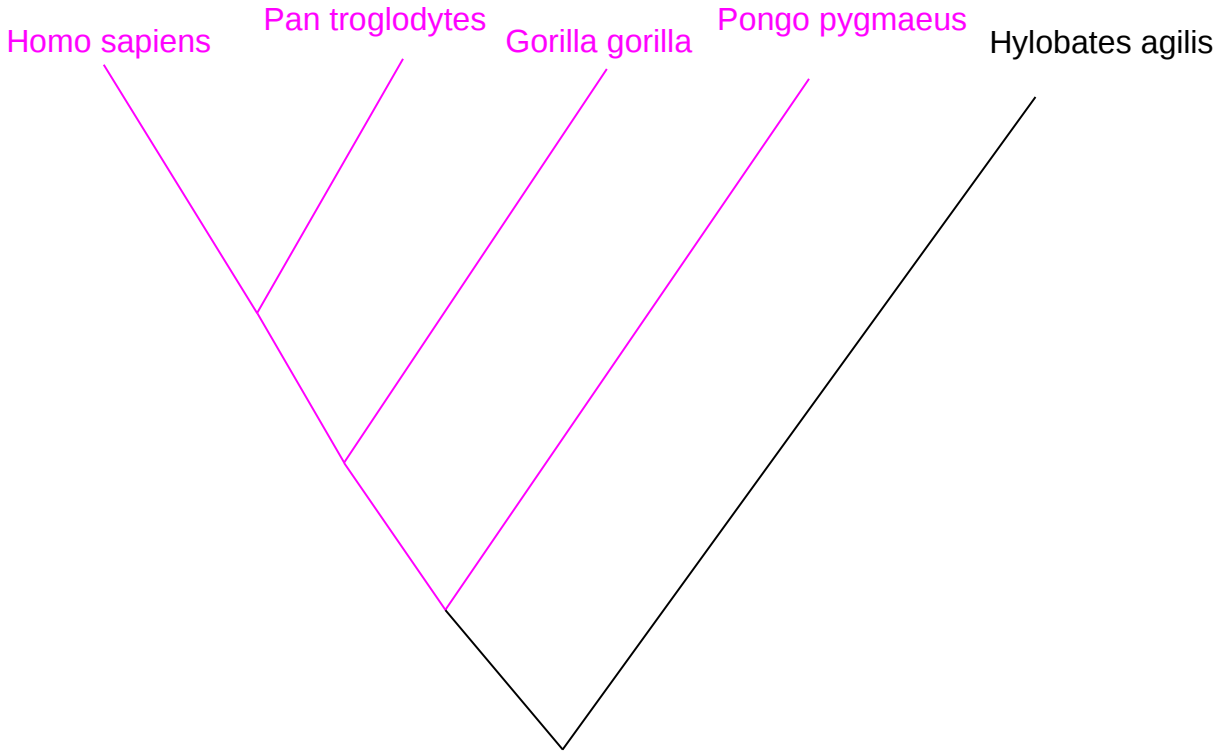




T1 : ((A,B),(C,D));

T2 : (A,B,C);

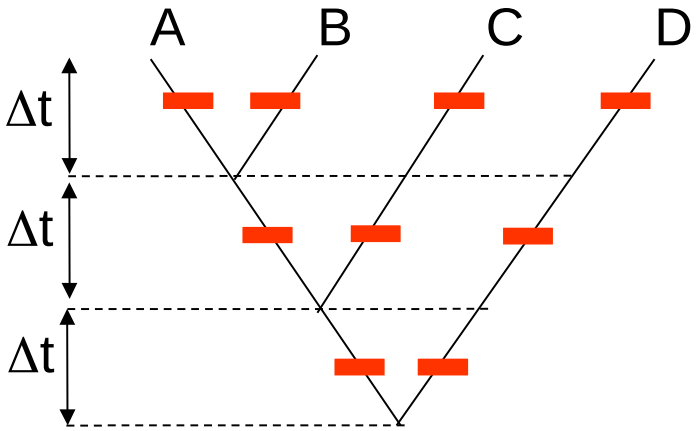
# Comment raciner un arbre phylogénétique ?

Utiliser un groupe externe :



# Horloge moléculaire

Utiliser l'horloge moléculaire :



Etant donnée une liste de caractères associés à un ensemble d'entités, comment construire un arbre retraçant les liens évolutifs entre toutes ces entités ?

Comment proposer un scénario évolutif à partir de l'observation des différences et ressemblances ?

1. Les méthodes de parcimonie
2. Les méthodes phénétiques (de distance)
3. Les méthodes probabilistes (maximum de vraisemblance et Bayesiennes)

Méthode de maximum de vraisemblance

$$P(D \text{ et } H) = P(D|H)\,P(H) = P(H|D)\,P(D)$$

Probabilité conditionnelle

# Méthode de maximum de vraisemblance

$$P(D \text{ et } H) = P(D|H)\,P(H) = P(H|D)\,P(D)$$

Probabilité conditionnelle

fonction de vraisemblance de H

probabilité marginale (a priori) de H

$$P(H|D) = \frac{P(D|H)\,P(H)}{P(D)}$$

Théorème de Bayes

probabilité a posteriori de H sachant D

probabilité marginale (a priori) de D

# Méthode de maximum de vraisemblance

$$P(D \text{ et } H) = P(D|H)\,P(H) = P(H|D)\,P(D)$$

Probabilité conditionnelle

fonction de vraisemblance de H

probabilité marginale (a priori) de H

$$P(H|D) = \frac{P(D|H)\,P(H)}{P(D)}$$

Théorème de Bayes

probabilité a posteriori de H sachant D

probabilité marginale (a priori) de D

# Méthode de maximum de vraisemblance

La vraisemblance est donc la probabilité d'observer un jeu de données (D) sachant une hypothèse H :

$$L = P(D|H)$$

On considère que l'hypothèse pour laquelle cette probabilité est maximale est celle qui explique le mieux les données.

# Méthode de maximum de vraisemblance

<u>Exemple</u> : Soit p la probabilité d'obtenir pile, quelle est la vraisemblance d'observer le résultat suivant si on réalise 6 lancés ?

D : Pile Face Face Pile Pile Pile

$L = P(D|H) = ?$

# Méthode de maximum de vraisemblance

<u>Exemple</u> : Soit p la probabilité d'obtenir pile, quelle est la vraisemblance d'observer le résultat suivant si on réalise 6 lancés ?

D : Pile Face Face Pile Pile Pile

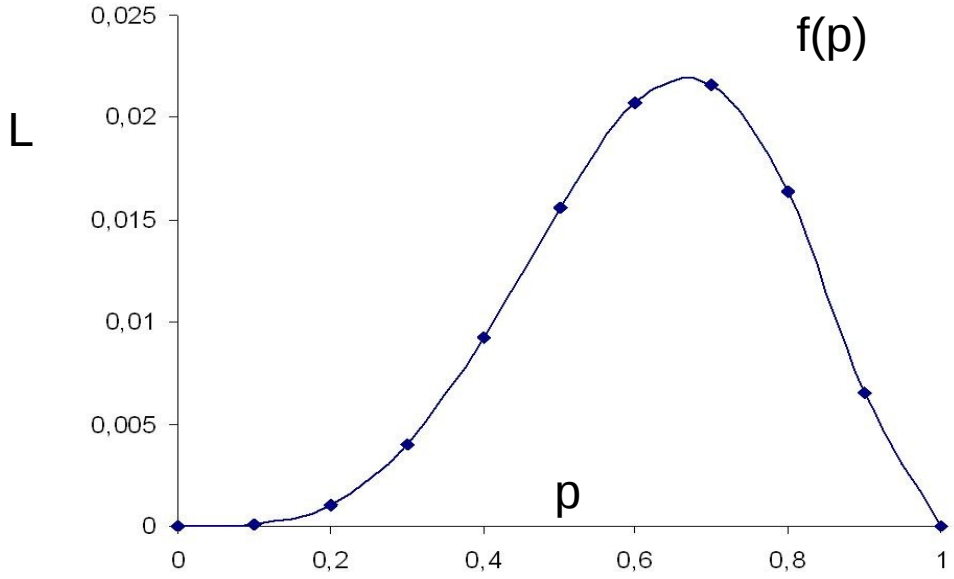
$L = P(D|H) = P(D|p) = p\,(1-p)\,(1-p)\,p\,p\,p = p^4(1-p)^2$

# Méthode de maximum de vraisemblance

<u>Exemple</u> : Soit p la probabilité d'obtenir pile, quelle est la vraisemblance d'observer le résultat suivant si on réalise 6 lancés ?

D : Pile Face Face Pile Pile Pile

$L = P(D|H) = P(D|p) = p\,(1-p)\,(1-p)\,p\,p\,p = p^4(1-p)^2$

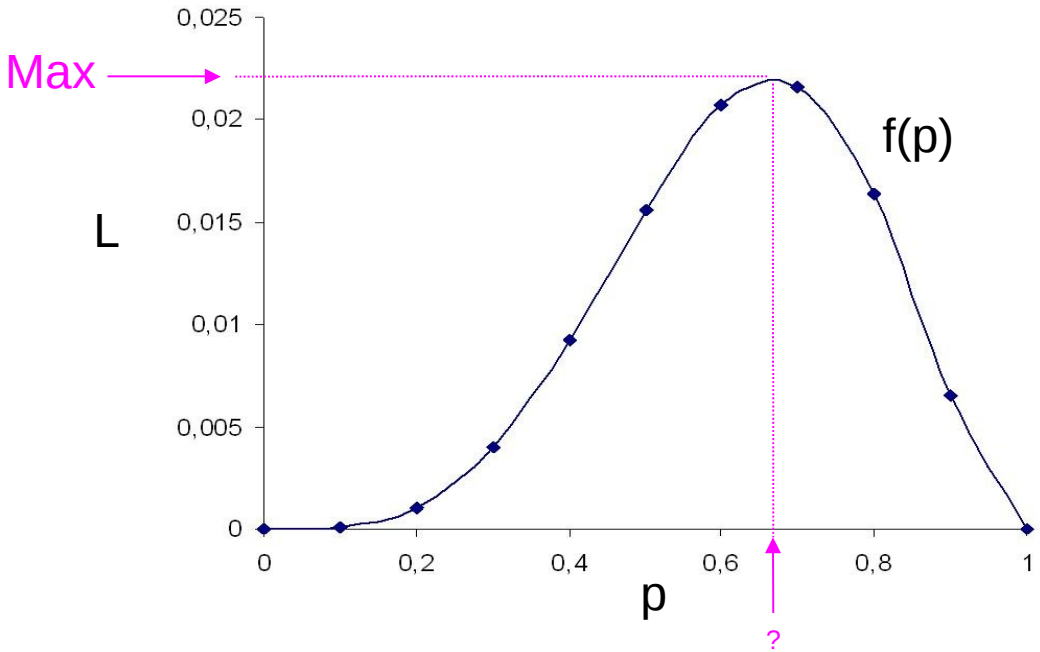


# Méthode de maximum de vraisemblance

<u>Exemple</u> : Soit p la probabilité d'obtenir pile, quelle est la vraisemblance d'observer le résultat suivant si on réalise 6 lancés ?

D : Pile Face Face Pile Pile Pile

$L = P(D|H) = P(D|p) = p\,(1-p)\,(1-p)\,p\,p\,p = p^4(1-p)^2$

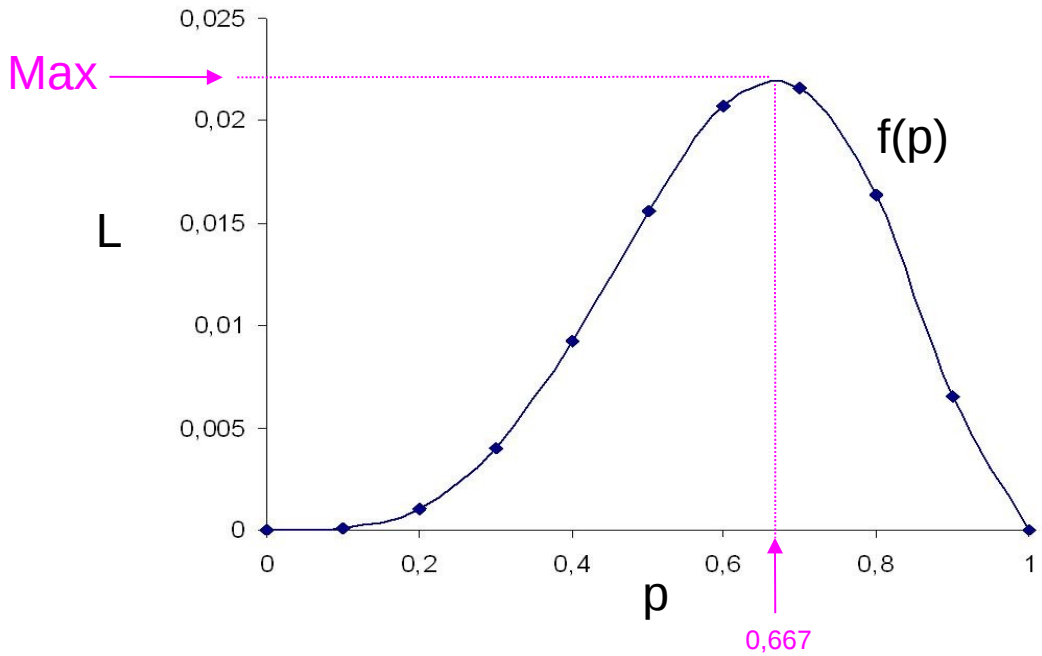


# Méthode de maximum de vraisemblance

<u>Exemple</u> : Soit p la probabilité d'obtenir pile, quelle est la vraisemblance d'observer le résultat suivant si on réalise 6 lancés ?

D : Pile Face Face Pile Pile Pile

$L = P(D|H) = P(D|p) = p\,(1-p)\,(1-p)\,p\,p\,p = p^4(1-p)^2$




$\ln L = \ln [\, p^4 (1-p)^2 \,] = 4 \ln p + 2 \ln (1-p)$

$$\frac{d\,(\ln L)}{d\,p} = \frac{4}{p} - \frac{2}{(1-p)} = 0$$

$$p = \frac{4}{6} = \frac{2}{3} = 0{,}666666667$$

La vraisemblance est maximale quand p, le paramètre du modèle est égal à 2/3.

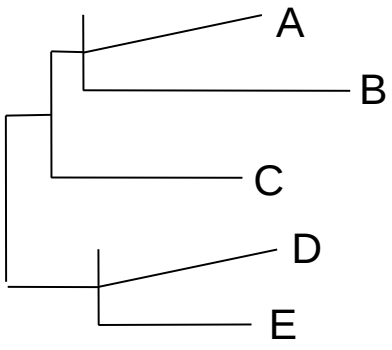
# Méthode de maximum de vraisemblance

En phylogénie moléculaire, les données seront composées d'un ensemble de caractères (des séquences) et l'hypothèse sera constituée par une topologie et un modèle d'évolution.

| | |
|---|---|
| Seq A | AAGCGTATGCGCGAATGC |
| Seq B | AAGCGTATGCGCGAATGC |
| Seq C | ATGCGTATGCGCGAATGC |
| Seq D | ATGCGTATGAGTGAATGC |
| Seq E | ATGCGTATGAGTGAATGC |

$\underbrace{\hspace{10em}}_{m}$

Modèle d'évolution des caractères →




$\tau$

$$L = P(D|\tau, M)$$

# Méthode de maximum de vraisemblance

En phylogénie moléculaire, les données seront composées d'un ensemble de caractères (des séquences) et l'hypothèse sera constituée par une topologie et un modèle d'évolution.

Seq A AAGCGTATGCGCGAATGC

Seq B AAGCGTATGCGCGAATGC

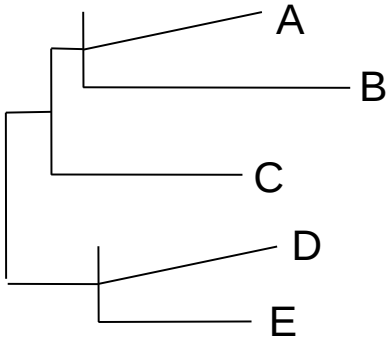
Seq C ATGCGTATGCGCGAATGC

Seq D ATGCGTATGAGTGAATGC

Seq E ATGCGTATGAGTGAATGC

$m$

Modèle d'évolution des caractères




$\tau$

Les *m* caractères étant considérés comme indépendants alors

$$L = P(D|\tau,M) = \prod_{i=1}^{m} P(D^{(i)}|\tau,M) = \prod_{i=1}^{m} L^{(i)}$$

$$\ln L = \ln\left(\prod_{i=1}^{m} L^{(i)}\right) = \sum_{i=1}^{m} \ln L^{(i)}$$

# Méthode de maximum de vraisemblance

En phylogénie moléculaire, les données seront composées d'un ensemble de caractères (des séquences) et l'hypothèse sera constituée par une topologie et un modèle d'évolution.

Seq A AAGCGTATGCGCGAATGC

Seq B AAGCGTATGCGCGAATGC

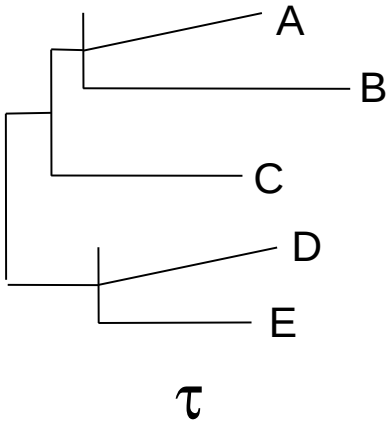
Seq C ATGCGTATGCGCGAATGC

Seq D ATGCGTATGAGTGAATGC

Seq E ATGCGTATGAGTGAATGC

$m$

Modèle d'évolution des caractères




Les *m* caractères étant considérés comme indépendants alors

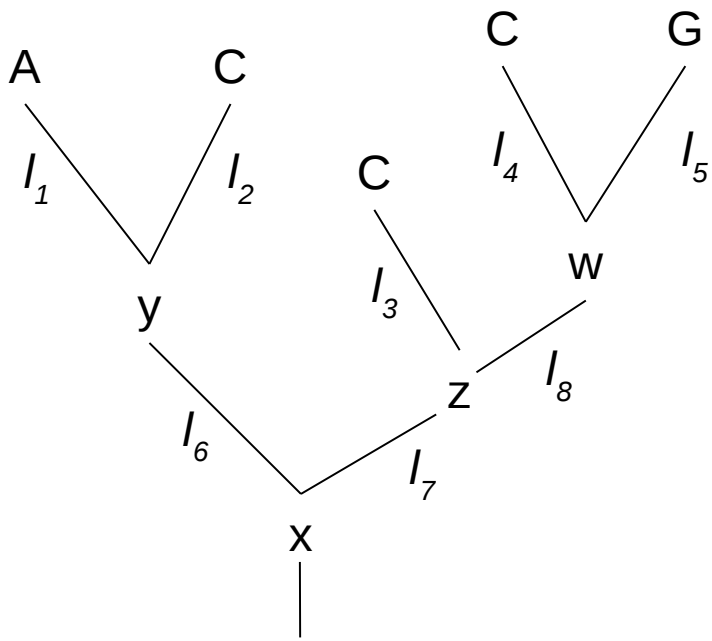
$$L = P(D|\tau,M) = \prod_{i=1}^{m} P(D^{(i)}|\tau,M) = \prod_{i=1}^{m} L^{(i)}$$

$$\ln L = \ln\left(\prod_{i=1}^{m} L^{(i)}\right) = \sum_{i=1}^{m} \ln L^{(i)}$$

# Méthode de maximum de vraisemblance

Indépendance des différents sites *i*

Indépendance de l'évolution des différentes feuilles



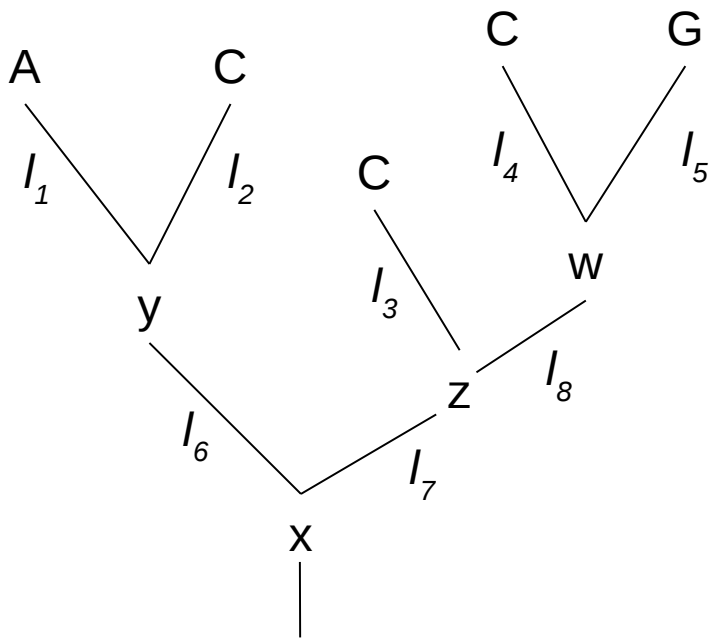

$$P(D^{(i)}|\tau,M) = \sum_x \sum_y \sum_z \sum_w P(A,C,C,C,G,x,y,z,w|\tau,M)$$

$$P(A,C,C,C,G,x,y,w,z|\tau,M) = P(x)$$

$$P(y|x,l_6)\, P(A|y,l_1)\, P(C|y,l_2)$$

$$P(z|x,l_7)\, P(C|z,l_3)\, P(w|z,l_8)\, P(C|w,l_4)\, P(G|w,l_5)$$

# Méthode de maximum de vraisemblance



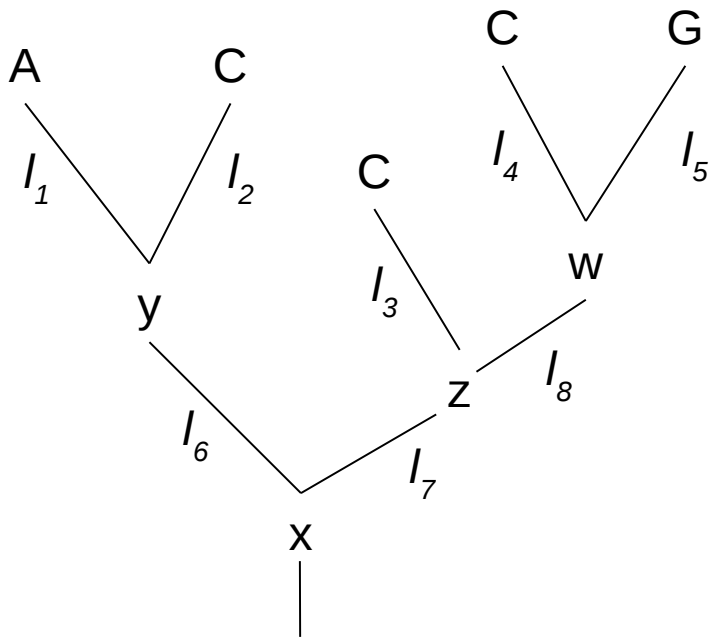

$$P(D^{(i)}|\tau,M) = \sum_x \sum_y \sum_z \sum_w P(A,C,C,C,G,x,y,z,w|\tau,M)$$

$$P(D^{(i)}|\tau,M) = \sum_x \sum_y \sum_z \sum_w \quad P(x)$$

$$P(y|x,l_6)\ P(A|y,l_1)\ P(C|y,l_2)$$

$$P(z|x,l_7)\ P(C|z,l_3)\ P(w|z,l_8)\ P(C|w,l_4)\ P(G|w,l_5)$$

# Méthode de maximum de vraisemblance



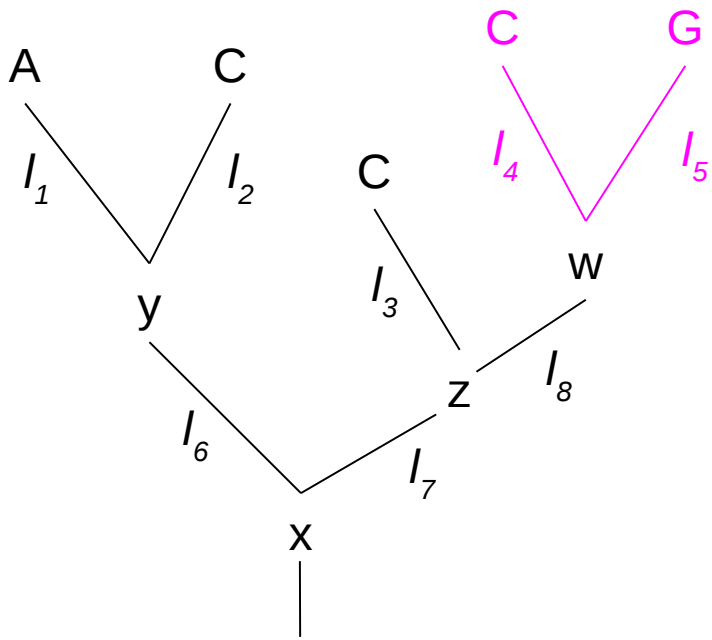

$$P(D^{(i)}|\tau,M) = \sum_x \sum_y \sum_z \sum_w P(A,C,C,C,G,x,y,z,w|\tau,M)$$

$$P(D^{(i)}|\tau,M) = \sum_x P(x)$$

$$\left(\sum_y P(y|x,l_6)\,(A|y,l_1)\,P(C|y,l_2)\right.$$

$$\left.\left(\sum_z P(z|x,l_7)\,P(C|z,l_3)\,\left(\sum_w P(w|z,l_8)\,P(C|w,l_4)\,P(G|w,l_5)\right)\right)\right)$$

# Méthode de maximum de vraisemblance




$$P(D^{(i)}|\tau,M) = \sum_x \sum_y \sum_z \sum_w P(A,C,C,C,G,x,y,z,w|\tau,M)$$

$$P(D^{(i)}|\tau,M) = \sum_x P(x)$$

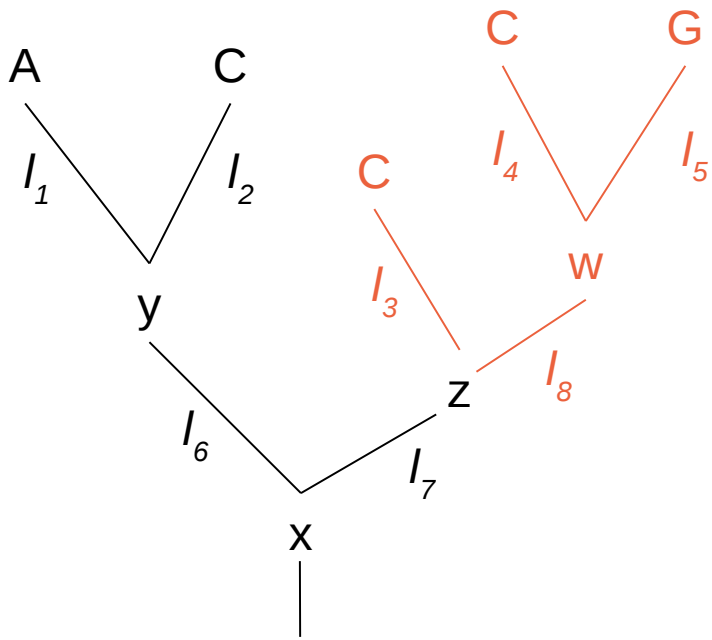
$$\Big(\sum_z P(y|x,l_6)\,(A|y,l_1)\,P(C|y,l_2)$$

$$\Big(\sum_y P(z|x,l_7)\,P(C|z,l_3)\,\Big(\sum_w P(w|z,l_8)\,P(C|w,l_4)\,P(G|w,l_5)\Big)\Big)\Big)$$

vraisemblance conditionnelle

$L_8^{(i)}(w)$

# Méthode de maximum de vraisemblance




$$P(D^{(i)}|\tau,M) = \sum_x \sum_y \sum_z \sum_w P(A,C,C,C,G,x,y,z,w|\tau,M)$$

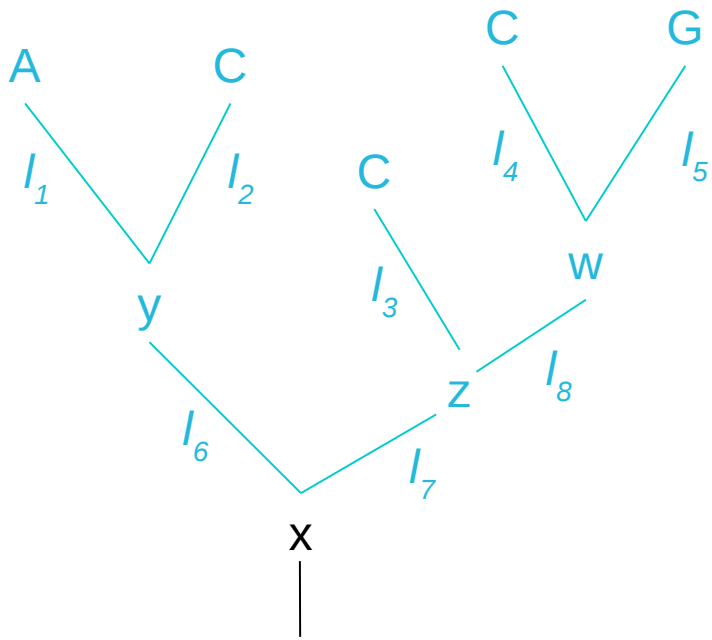
$$P(D^{(i)}|\tau,M) = \sum_x P(x)$$

$$\left(\sum_y P(y|x,l_6)\,(A|y,l_1)\,P(C|y,l_2)\right.$$

$$\left(\sum_z P(z|x,l_7)\,P(C|z,l_3)\,\left(\sum_w P(w|z,l_8)\,P(C|w,l_4)\,P(G|w,l_5)\right)\right))$$

$$L_7^{(i)}(z) = P(C|z,l_3) \sum_w P(w|z,l_8)\,L_8^{(i)}(w)$$

# Méthode de maximum de vraisemblance




$$P(D^{(i)}|\tau,M) = \sum_x \sum_y \sum_z \sum_w P(A,C,C,C,G,x,y,z,w|\tau,M)$$

$$P(D^{(i)}|\tau,M) = \sum_x P(x)$$

$$\left(\sum_y P(y|x,l_6)\ (A|y,l_1)\ P(C|y,l_2)\right.$$

$$\left(\sum_z P(z|x,l_7)\ P(C|z,l_3)\ \left(\sum_w P(w|z,l_8)\ P(C|w,l_4)\ P(G|w,l_5)\right)\right)$$

$$L_{root}^{(i)}(x) = \sum_y P(y|x,l_6)\ L_6^{(i)}(y)\ \sum_z P(z|x,l_7)\ L_7^{(i)}(z)$$

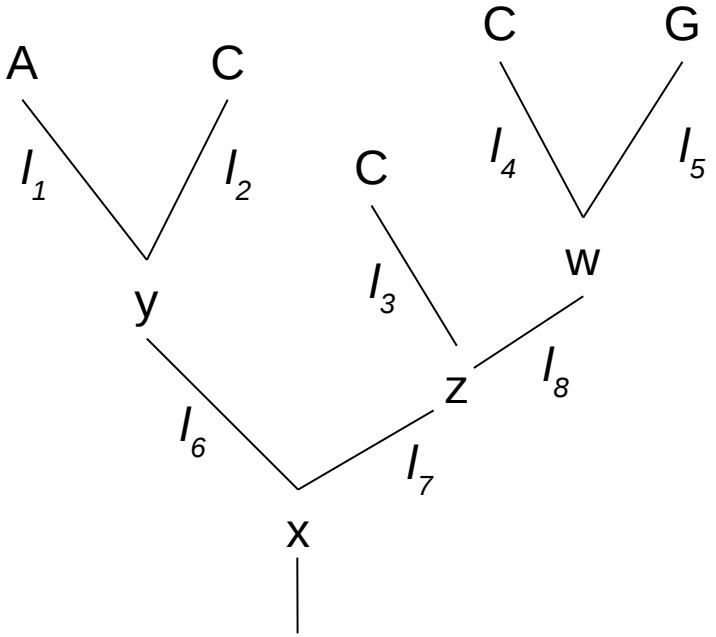
# Méthode de maximum de vraisemblance

$$P(D^{(i)}|\tau,M) = L^{(i)} = \sum_x \pi_x \, L_{root}^{(i)}(x)$$

probabilité à priori

d'après le modèle

d'évolution choisi




$$P(D^{(i)}|\tau,M) = \sum_x \sum_y \sum_z P(x)\, P(y|x,l_6)\, L_6^{(i)}(y)\, P(z|x,l_7)\, L_7^{(i)}(z)$$

or $\quad P(x)\, P(y|x,l_6) = P(y)\, P(x|y,l_6)$

$$P(D^{(i)}|\tau,M) = \sum_x \sum_y \sum_z P(y)\, P(x|y,l_6)\, L_6^{(i)}(y)\, P(z|x,l_7)\, L_7^{(i)}(z)$$

# Méthode de maximum de vraisemblance

En résumé :

1. On explore l'univers des topologies

2. Pour chaque topologie on cherche les longueurs de branches
   et les paramètre de modèle pour lesquels la vraisemblance est maximale

3. On retient l'arbre pour lequel la topologie, les longueurs de branches et les paramètres de modèles présentent la vraisemblance maximale. Après un temps variable on doit théoriquement converger vers cette valeur maximale

# Méthodes Bayesiennes

$$P(D \text{ et } H) = P(D|H)\,P(H) = P(H|D)\,P(D)$$

Probabilité conditionnelle

fonction de vraisemblance

de H

probabilité marginale

(a priori) de H

$$P(H|D) = \frac{P(D|H)\,P(H)}{P(D)}$$

Théorème de Bayes

probabilité a posteriori

de H sachant D

probabilité marginale

(a priori) de D

# Méthodes Bayesiennes

$$P(D \text{ et } H) = P(D|H)\,P(H) = P(H|D)\,P(D)$$

Probabilité conditionnelle

fonction de vraisemblance de H

probabilité marginale (a priori) de H

$$P(H|D) = \frac{P(D|H)\,P(H)}{P(D)}$$

Théorème de Bayes

probabilité a posteriori de H sachant D

probabilité marginale (a priori) de D

D'après la loi des probabilités totales (alternatives) $P(D) = \sum_{H} P(D \text{ et } H)$

# Méthodes Bayesiennes

$$P(D \text{ et } H) = P(D|H)\,P(H) = P(H|D)\,P(D)$$

Probabilité conditionnelle

fonction de vraisemblance de H

probabilité marginale (a priori) de H

$$P(H|D) = \frac{P(D|H)\,P(H)}{\sum_H P(H)\,P(D|H)}$$

Théorème de Bayes

probabilité a posteriori de H sachant D

probabilité marginale (a priori) de D

# Méthodes Bayesiennes

$$P(D \text{ et } H) = P(D|H)\,P(H) = P(H|D)\,P(D)$$

Probabilité conditionnelle

fonction de vraisemblance

de H

probabilité marginale

(a priori) de H

$$P(H|D) = \frac{P(D|H)\,P(H)}{\sum_H P(H)\,P(D|H)}$$

Théorème de Bayes

probabilité a posteriori

de H sachant D

probabilité marginale

(a priori) de D

# Méthodes Bayesiennes

$$P(D \text{ et } H) = P(D|H)\,P(H) = P(H|D)\,P(D)$$

Probabilité conditionnelle

fonction de vraisemblance

de H

probabilité marginale

(a priori) de H

$$P(H|D) = \frac{P(D|H)\,P(H)}{\sum_H P(H)\,P(D|H)}$$

Théorème de Bayes

probabilité a posteriori

de H sachant D

Echantillonnage

probabilité marginale

(a priori) de D

Méthodes Bayesiennes

Utilisation du rapport des vraisemblances

$$P(H_i|D) = \frac{P(D|H_i)\,P(H_i)}{\sum_H P(H)\,P(D|H)} \qquad P(H_j|D) = \frac{P(D|H_j)\,P(H_j)}{\sum_H P(H)\,P(D|H)}$$

$$R = \frac{P(D|H_i)\,P(H_i)}{P(D|H_j)\,P(H_j)} = \frac{P(D|H_i)}{P(D|H_j)}\,\frac{P(H_i)}{P(H_i)}$$

Si les hypothèses $H_i$ et $H_j$ sont équiprobables alors :

$$R = \frac{P(D|H_i)}{P(D|H_j)} = \frac{P(D|\tau_i)}{P(D|\tau_j)}$$

# Méthodes Bayesiennes

## Utilisation du rapport des vraisemblances

$$P(H_i|D) = \frac{P(D|H_i)\,P(H_i)}{\sum_H P(H)\,P(D|H)} \qquad P(H_j|D) = \frac{P(D|H_j)\,P(H_j)}{\sum_H P(H)\,P(D|H)}$$

$$R = \frac{P(D|H_i)\,P(H_i)}{P(D|H_j)\,P(H_j)} = \frac{P(D|H_i)}{P(D|H_j)}\,\frac{P(H_i)}{P(H_i)}$$

Si les hypothèses $H_i$ et $H_j$ sont équiprobables alors :

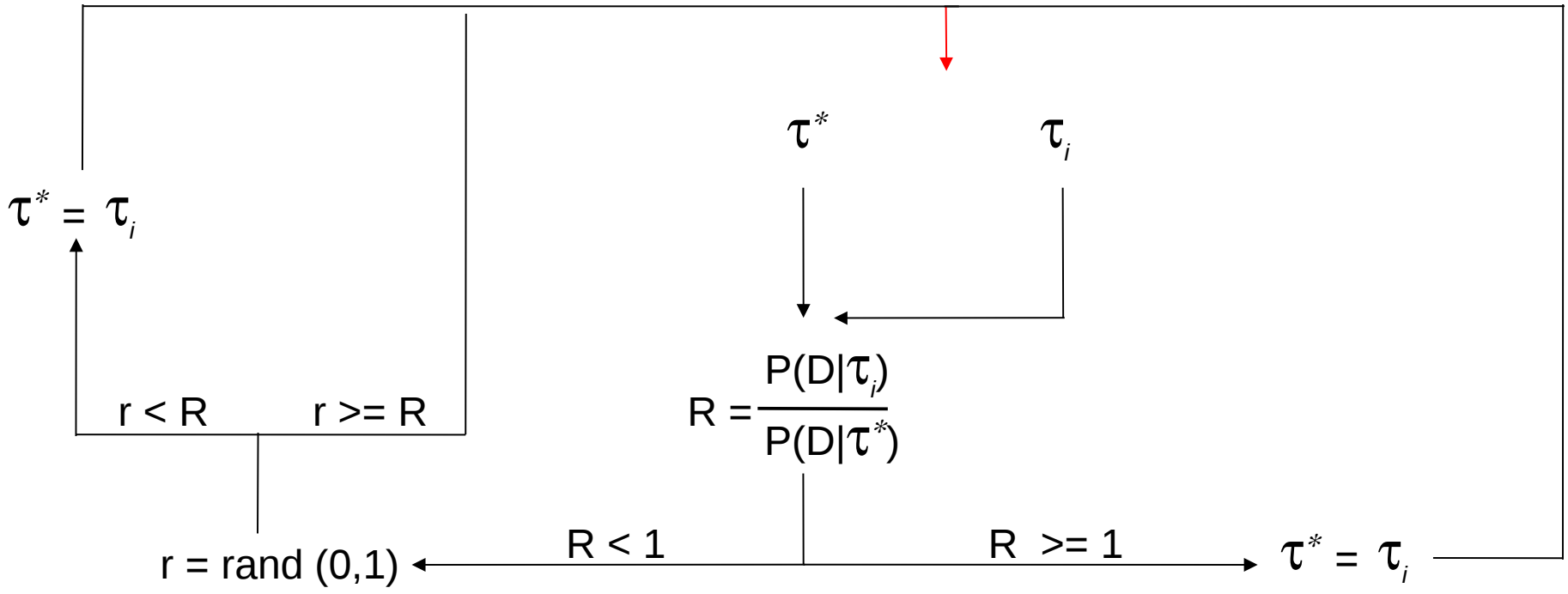
$$R = \frac{P(D|H_i)}{P(D|H_j)} = \frac{P(D|\tau_i)}{P(D|\tau_j)}$$

$L(\tau_i)$

$L(\tau_j)$

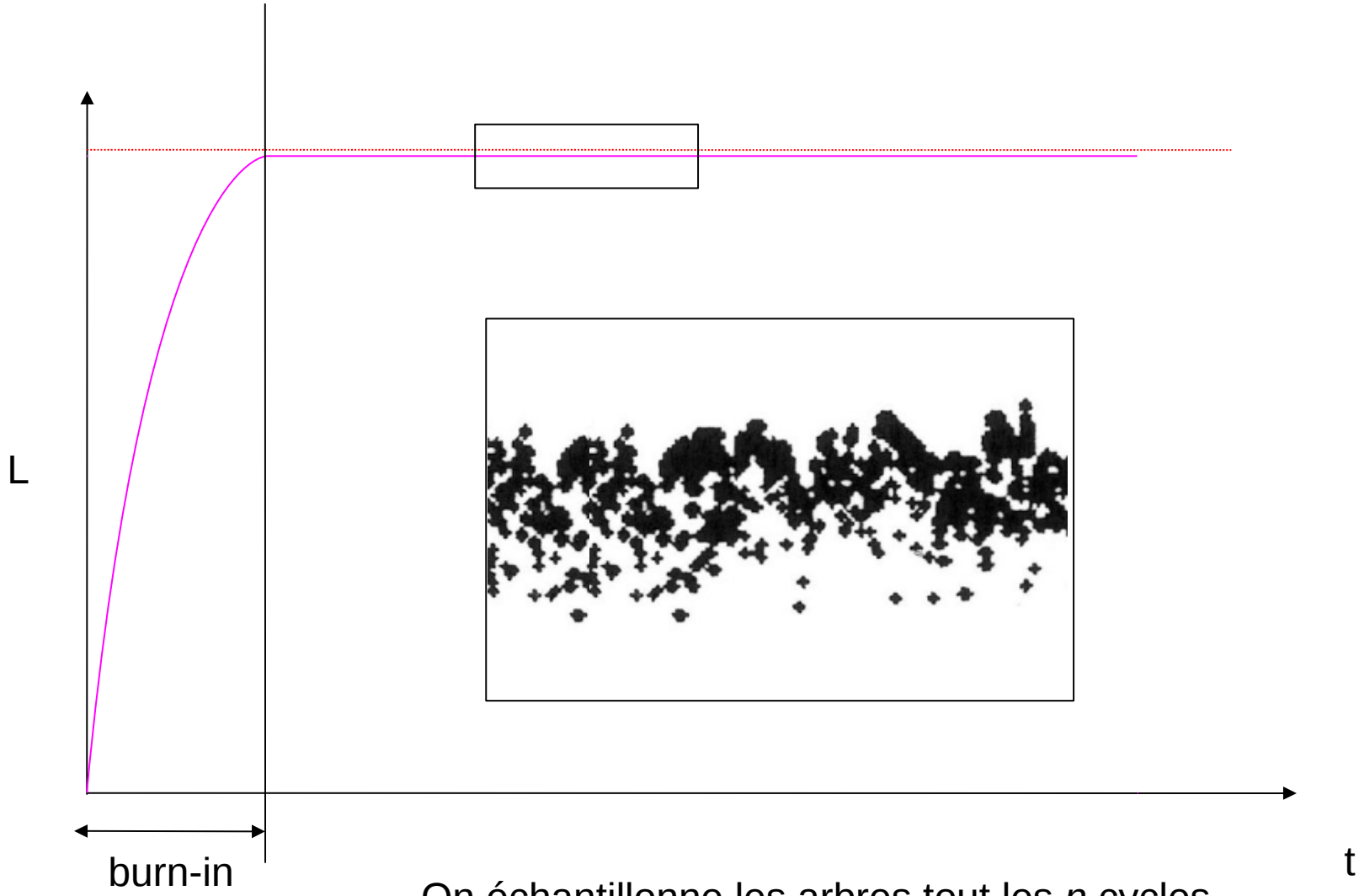
Méthodes Bayesiennes

Algorithme de Metropolis-Hastings (Markov Chain Monte Carlo, MCMC)



Méthodes Bayesiennes

Algorithme de Metropolis-Hastings (Markov Chain Monte Carlo, MCMC)




On échantillonne les arbres tout les *n* cycles

$P(\tau_i|D)=f(L(\tau_i))$

# Méthodes Bayesiennes

Résumé :

1. Construction de CM reliant les arbres entre eux

2. obtention d'un échantillonnage reflétant la distribution des probabilités postérieures

3. Sommation des données :

   - Recherche de l'arbre de plus grande probabilité postérieure

   - Construction d'un consensus à partir des arbres de plus grande probabilités postérieures

   - Probabilité d'un clade : $\Sigma$ des probabilités postérieures des arbres qui possèdent ce clade