Module d'enseignement "Nature et culture 3
Licence Sciences et Humanités
Aix-Marseille Université
Novembre 2020

Enquête bioinformatique sur les origines de SARS-CoV-2

CM2 – Inférer la phylogénie de SARS-CoV-2 à partir des séquences génomiques et protéiques

*Cours donné par **Jacques van Helden**, **Emese Meglécz** et **Gabriel Neve***

Sur base d'une enquête menée par Erwan Salard, José Haloy, Didier Casane, Etienne Decroly et Jacques van Helden

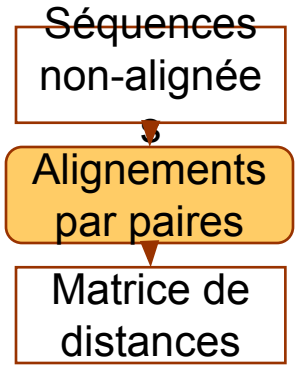
Alignement de séquences multiples

Alignement progressif

- L'approche la plus courante pour aligner des séquences multiples est de réaliser un ***alignement progressif***.
- L'algorithme procède en plusieurs étapes (détaillées dans les diapos suivantes):
  - Calculer une ***matrice de distances***, qui indique la distance entre chaque paire de séquences.
  - Construire un ***arbre guide*** qui regroupe en premier lieu les séquences les plus proches, et remonte en regroupant progressivement les séquences les plus éloignées.
  - Utiliser ce arbre pour aligner progressivement les séquences.
- Il s'agit d'une approche ***heuristique***
  - Cette approche est praticable pour un grand nombre de séquences, mais ne peut pas garantir de retourner l'alignement optimal.

Alignement progressif – 1ère étape: construction de la matrice de distance

- On effectue un alignement par paires entre chaque paire de protéines
  - Alignement par programmation dynamique ou par BLAST.
  - Nombre d'alignements = *n * (n – 1) / 2*
- A partir de chaque alignement par paire, calculer la distance entre les deux séquences.
  - $d_{i,j}=s_{i,j}/L_{j,j}$
    - $d_{j,j}$ distance entre les séquences *i* and *j*
    - $L_{j,j}$ longueur de l'alignement
    - $s_{j,j}$ nombre de substitutions
- Remarques
  - Les gaps ne sont pas pris en compte dans la métrique de distance
  - La matrice est symétrique: $d_{i,j} = d_{j,i}$
  - Les éléments diagonaux sont nuls: $d_{i,i} = 0$



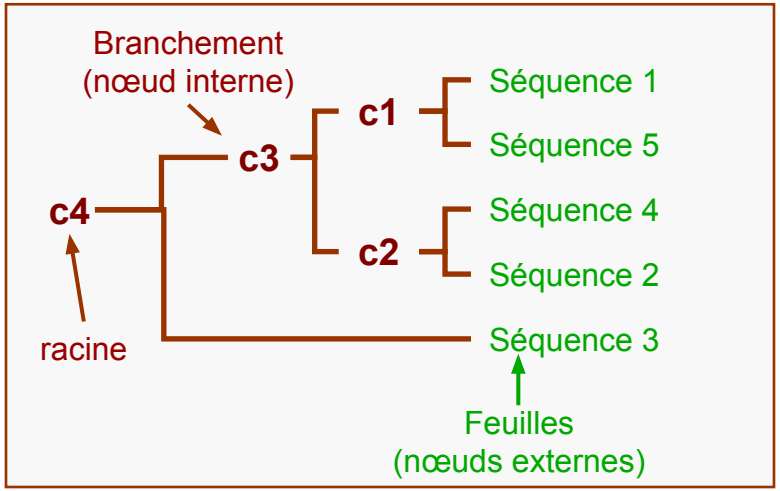
	seq 1	seq 2	...	seq n
seq 1	d1,1	d1,2	...	d1,n
seq 2	d2,1	d2,2	...	d2,n
...	...	...	...	...
seq n	dn,1	dn,2	...	dn,n

Principe de la construction de l'arbre-guide – Méthode UPGMA

Matrice de distance

	séquence 1	séquence 2	séquence 3	séquence 4	séquence 5
séquence 1	0.00	4.00	6.00	3.50	1.00
séquence 2	4.00	0.00	6.00	2.00	4.50
séquence 3	6.00	6.00	0.00	5.50	6.50
séquence 4	3.50	2.00	5.50	0.00	4.00
séquence 5	1.00	4.50	6.50	4.00	0.00

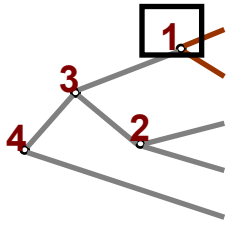
Arbre



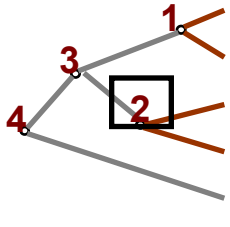
- Le clustering hiérarchique est une méthode de clustering agrégative.
  - Prend une matrice de distance en entrée
  - Regroupe progressivement les objets en allant des plus proches aux plus distants.
- Il existe plusieurs possibilités pour établir une règle d'agglomération, qui définit la distance entre deux groupes.
  - Liaison simple (***single linkage***): distance entre groupes A et B est la distance entre les plus proches de leurs éléments respectifs.
  - Liaison moyenne (***average linkage***): distance moyenne entre tous les objets des deux groupes (=**UPGMA**).
  - Liaison complète (***complete linkage***): distance entre les éléments les plus éloignés des groupes A et B.
- Algorithme
  - 1. Assigner chaque objet à un cluster séparé.
  - 2. Identifier la paire de clusters les plus proches, et les regrouper en un seul.
  - 3. Répéter la seconde étape jusqu'à ce qu'il ne reste qu'un seul cluster.
- Le résultat est un arbre, dont les nœuds intermédiaires correspondent aux clusters.
  - N objets → N-1 nœuds intermédiaires
- Les longueurs des branches représentent les distances entre clusters.

Alignement progressif – 3ème étape: alignement multiple

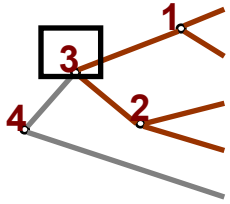
- On construit un alignement multiple en incorporant progressivement les séquences selon leur ordre de branchement dans l'arbre guide, en remontant des plus proches aux plus éloignées.



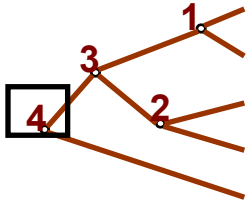
Seq5 **GATTGTAGTA**
Seq1 **GATGGTAGTA**
Seq4
Seq2
Seq3



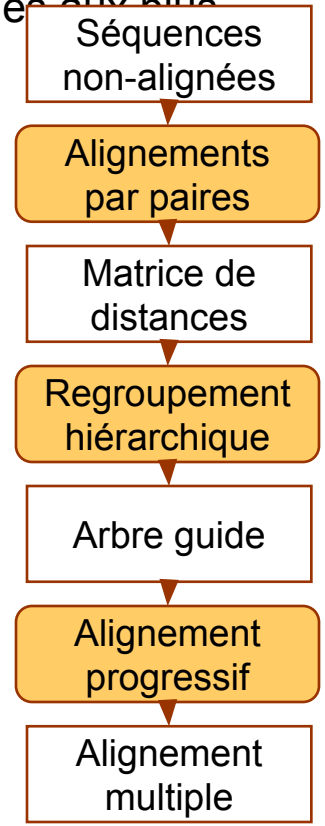
Seq5 GATTGTAGTA
Seq1 GATGGTAGTA
Seq4 **GATTGTTC--GTA**
Seq2 **GATTGTTCGGGTA**
Seq3



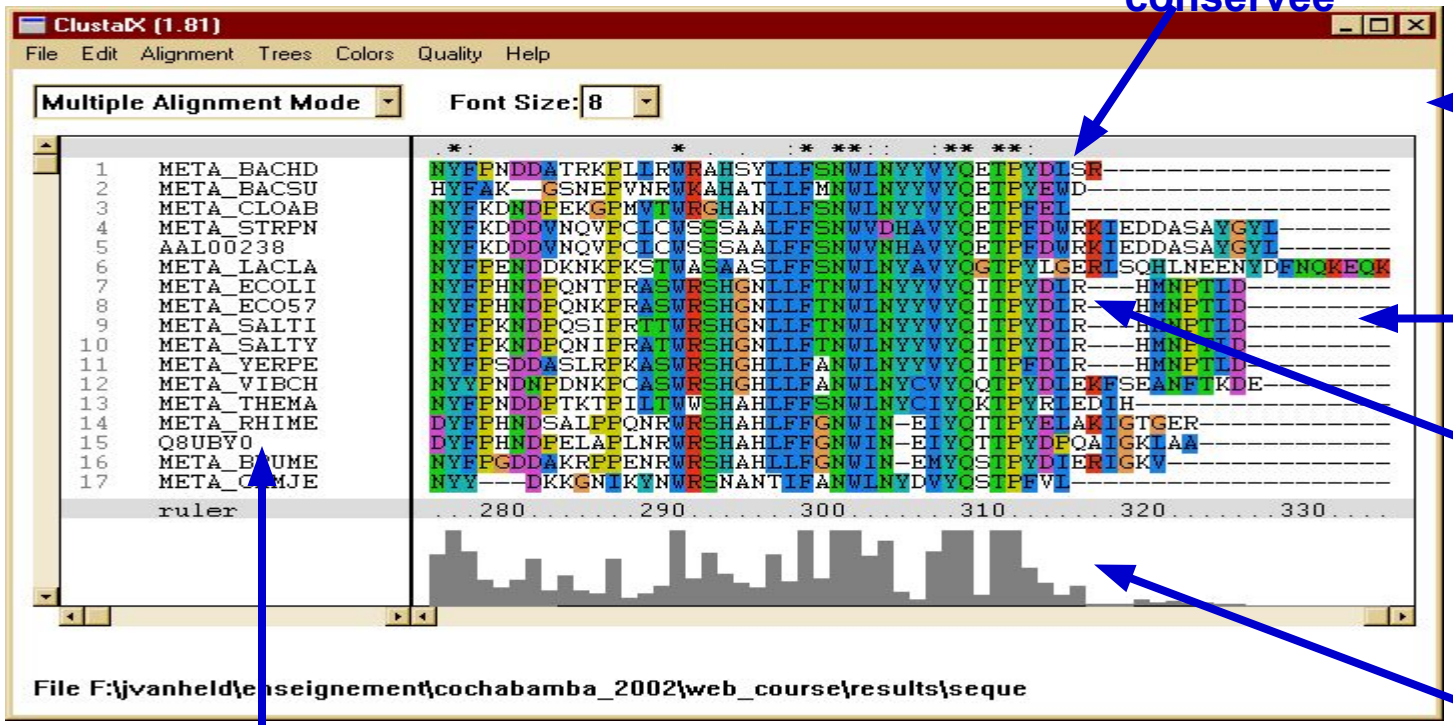
Seq5 **GATTGTA---GTA**
Seq1 **GATGGTA---GTA**
Seq4 **GATTGTTC--GTA**
Seq2 **GATTGTTCGGGTA**
Seq3



Seq5 GATTGTA-----GTA
Seq1 GATGGTA-----GTA
Seq4 GATTGTTC----GTA
Seq2 GATTGTTCGG--GTA
Seq3 **GATGGTAGGCGTGTA**



Alignement multiple global : Homoserine-O-dehydrogenase



Position conservée

Alignement multiple

Gap terminal

Gap interne

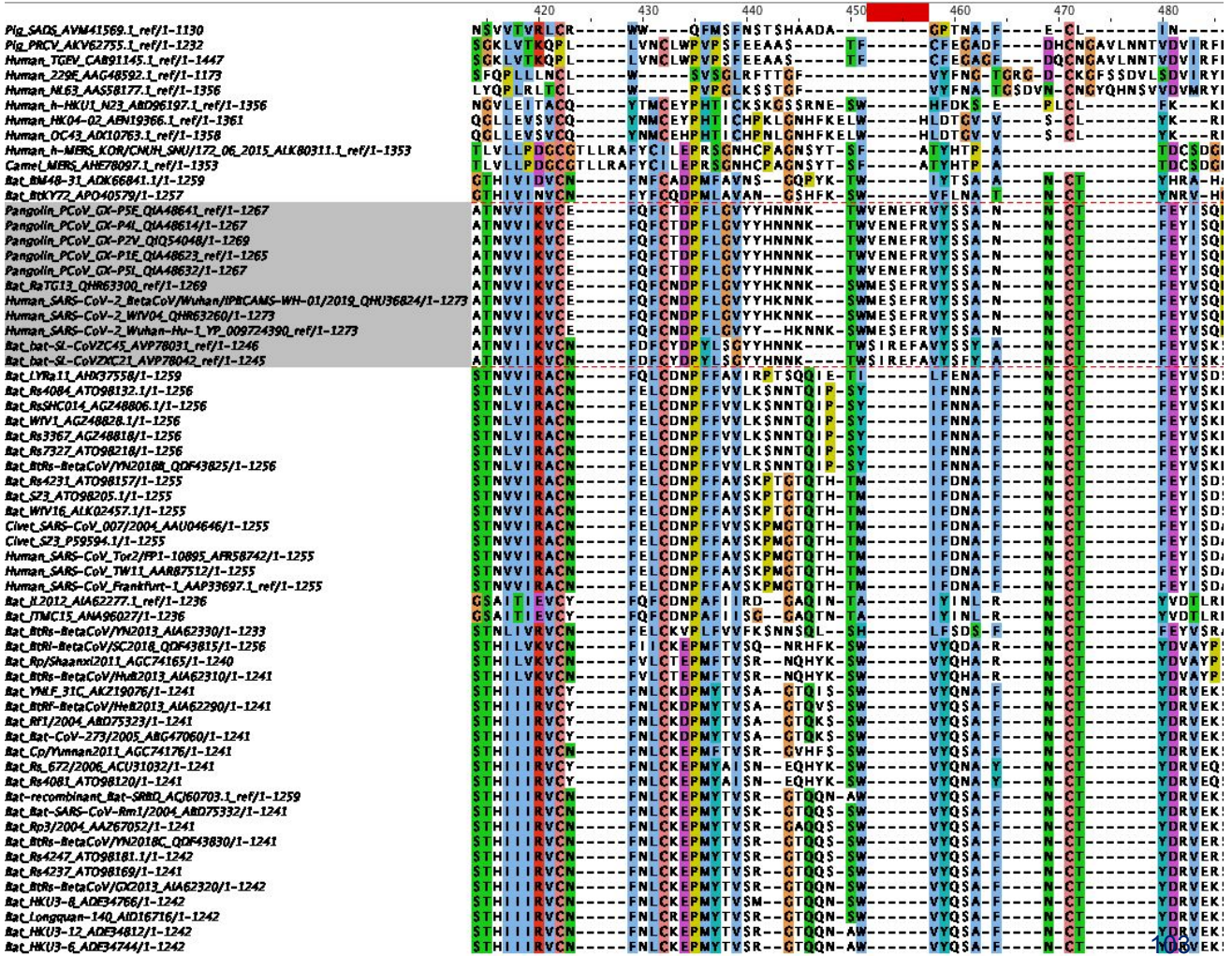
Colonnes de

Identifiants de séquences

Etude des insertions / délétions dans les alignements multiples de la protéine spicule (spike)

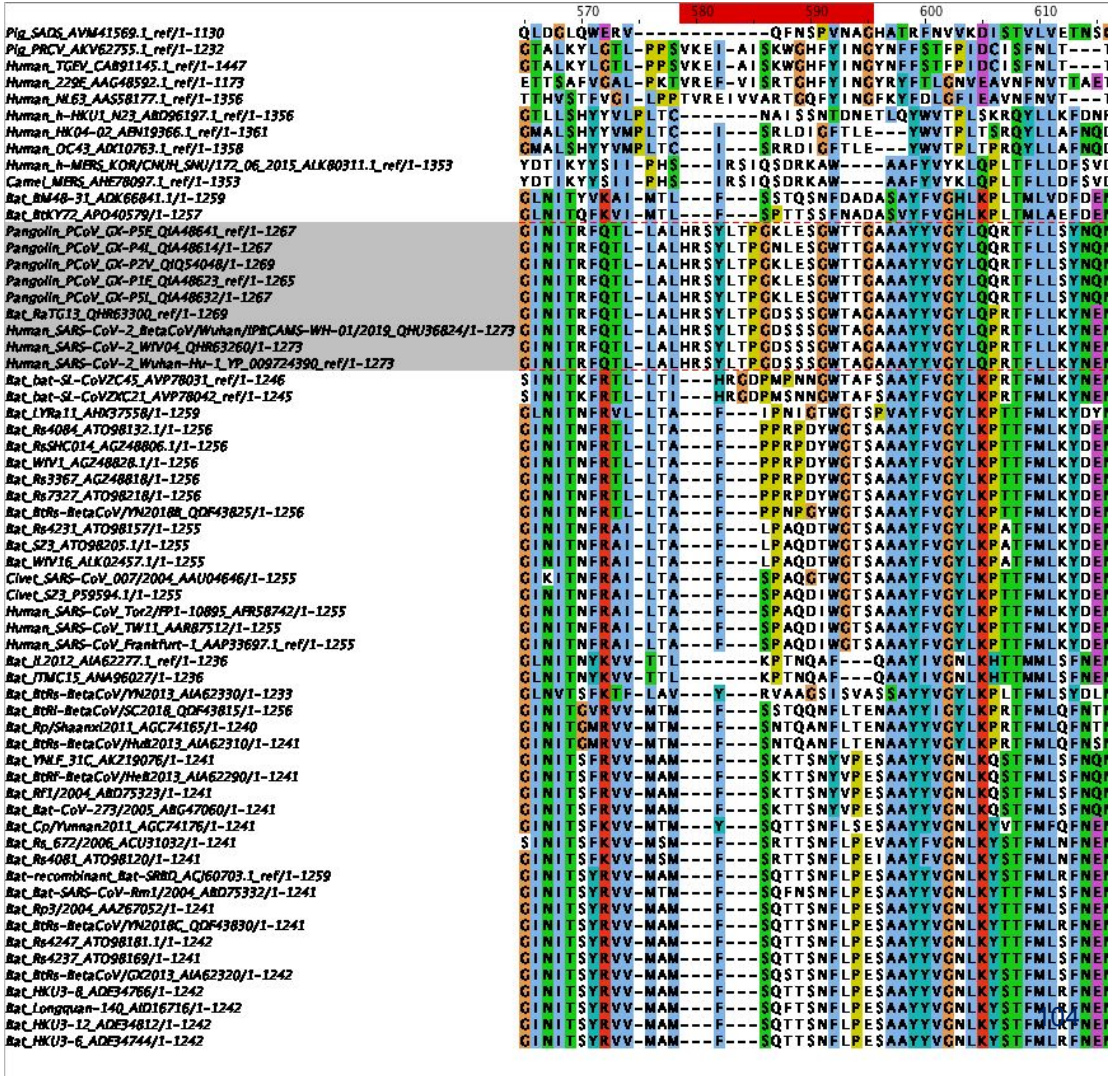
Insertion partagée entre tous les virus du groupe CoV-2

- Position: 153-158 de SARS-CoV-2
- Cette insertion se trouve chez les virus de pangolin + plusieurs chauve-souris
- Les résidus sont identiques entre SARS-CoV-2 et la souche RaTG13 de chauve-souris (la plus proche de SARS-CoV-2)
- Par contre elle présente 3 substitutions entre les souches de pangolin et SARS-CoV-2.



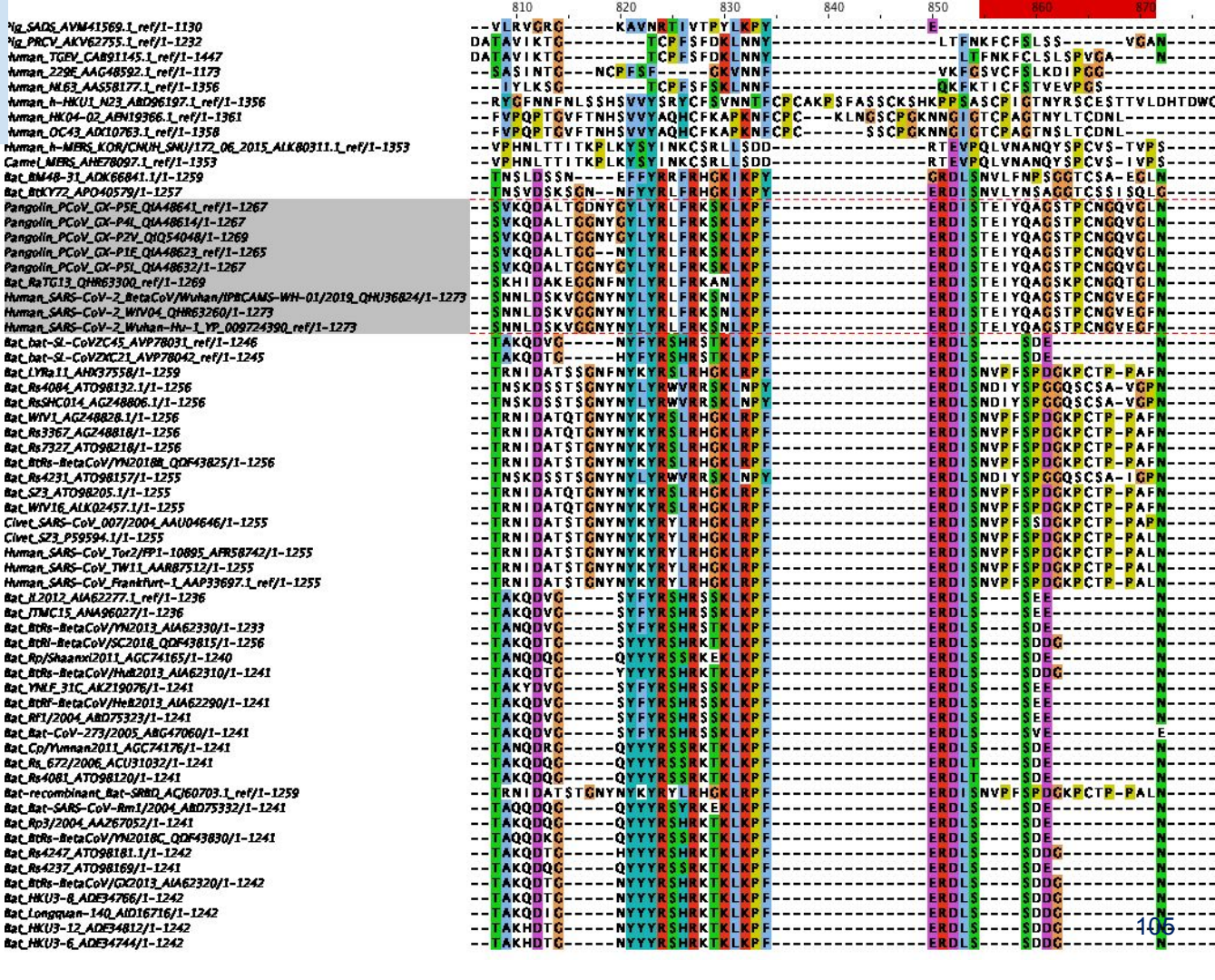
Insertion partagée par la majorité des virus du groupe CoV-2

- Position: 245-251 de SARS-CoV-2
- Cette insertion se trouve chez les virus de pangolin + la souche RaTG13 de chauve-souris
- Elle est cependant absente de 2 souches de chauves-souris appartenant au groupe CoV-2 : CoVZC45 et CoVZXC21
- Au-delà de l'insertion on trouve un bloc conservé (jusqu'à la position 595 de l'alignement).
- Au sein de ce bloc, une paire de résidus distingue les pangolins du groupe SARS2 + Bat RaTG13 .



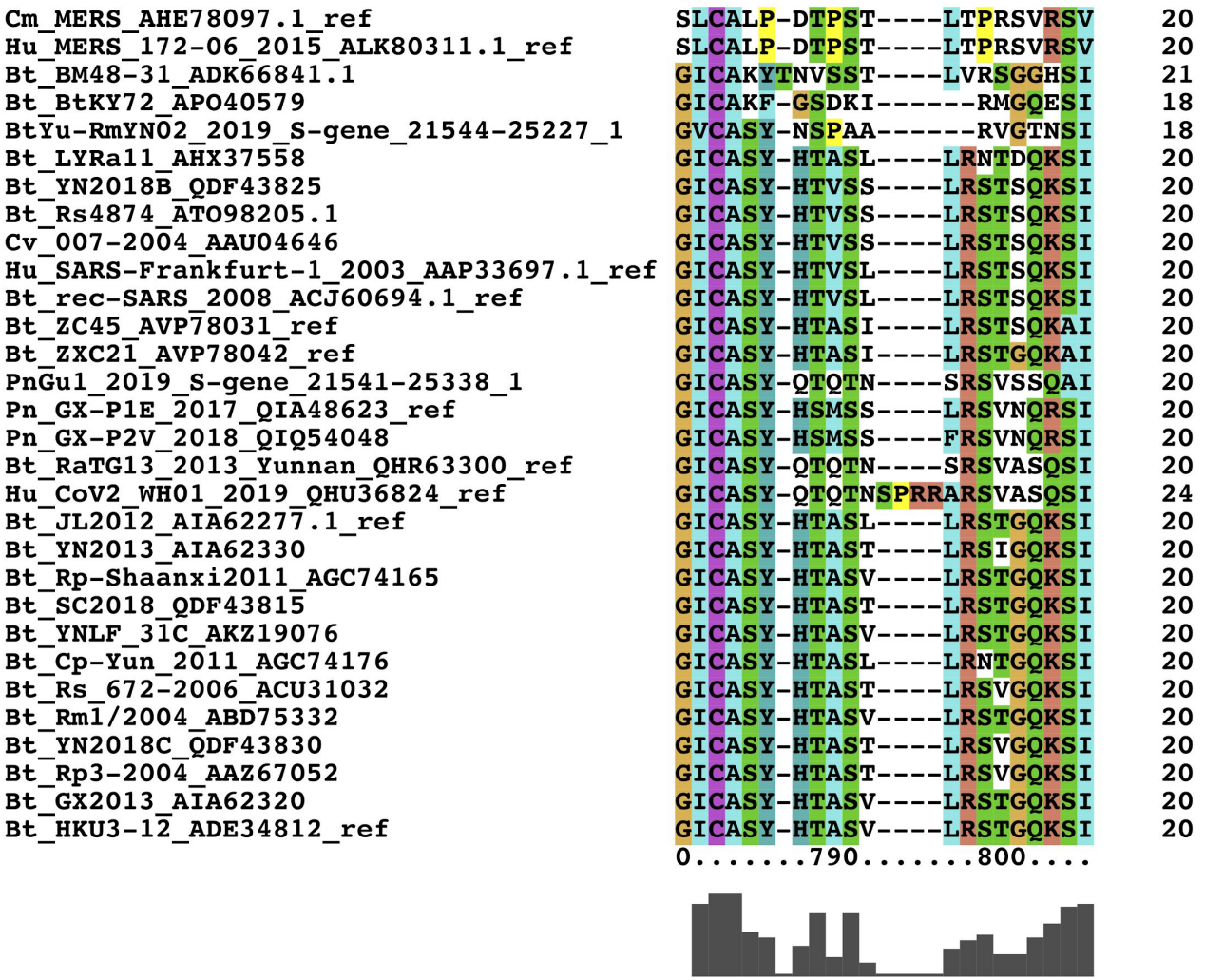
Insertion i3

- Position
  - 470-486 de SARS-CoV-2
  - 855-872 sur l'alignement
- Commune au groupe pangolin + Bat_RaTG13 + SARS-CoV-2
- 2 substitutions uniques à Bat_RaTG13



Insertion d'un site Furine (i4)

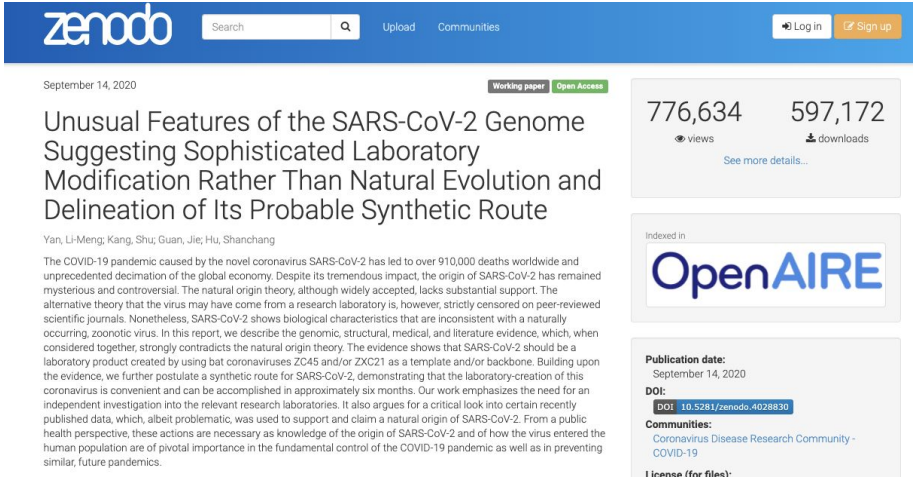
- Positions : 1181-1184 de l'alignement
- On trouve chez SARS-CoV-2 un site unique SPRRAR, qui résulte d'une insertion SPRR et d'une substitution L -> A
- La séquence PRRA correspond au motif reconnu par la furine (protéase).
- Cette insertion est à l'origine du site de clivage responsable du caractère particulièrement virulent de SARS-CoV-2

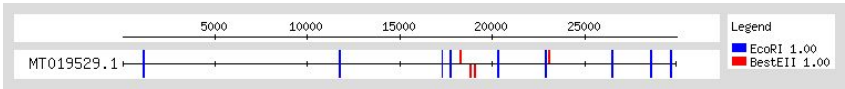


Un virus construit par ingénierie moléculaire ?

Un virus construit par ingénierie moléculaire ?

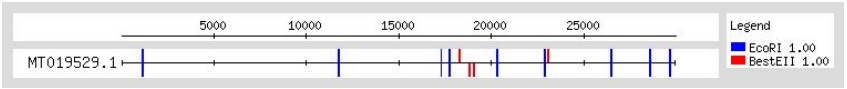
- Sept 2020: un preprint fait du bruit
- Li-Meng Yan
  - chercheuse chinoise
  - travaillait dans le laboratoire de référence de l'OMS pour la Chine
  - réfugiée aux Etats-Unis
- 600.000 téléchargements en 10 jours
- Arguments
  - Présence de sites de restriction dans le génome de SARS-Cov-2

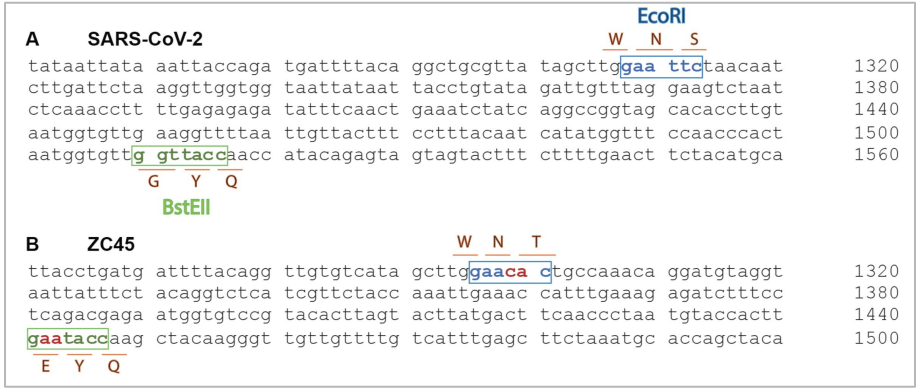




Des sites de restriction créés dans le génome de SARS-CoV-2 ?

- Li-Meng Yan détecte dans la séquence de SARS-CoV-2 des séquences correspondant aux sites de restriction de EcoRI et BstEII, souvent utilisés en biologie moléculaire pour créer de l'ADN recombinant au moyen d'enzymes de restriction ("ciseaux" moléculaires).
- Elle souligne que ces sites auraient pu être facilement créés à partir de sites très similaires présents dans le virus de chauve-souris Bat ZC45.
- Des questions se posent sur la signification de ce résultat
  - Probabilité de trouver ces sites aléatoirement ?
  - Les sites sont-ils uniques à SARS-CoV-2 dans le lignage des SARS-CoV ?
- Une première observation:
  - Le génome de SARS-CoV-2 contient 9 sites EcoRI et 4 sites BstEII
  - étant donné la courte taille des sites de restriction, et la taille du génome, on peut s'attendre à trouver de tels sites dans n'importe quelle séquence de 30 kilobases



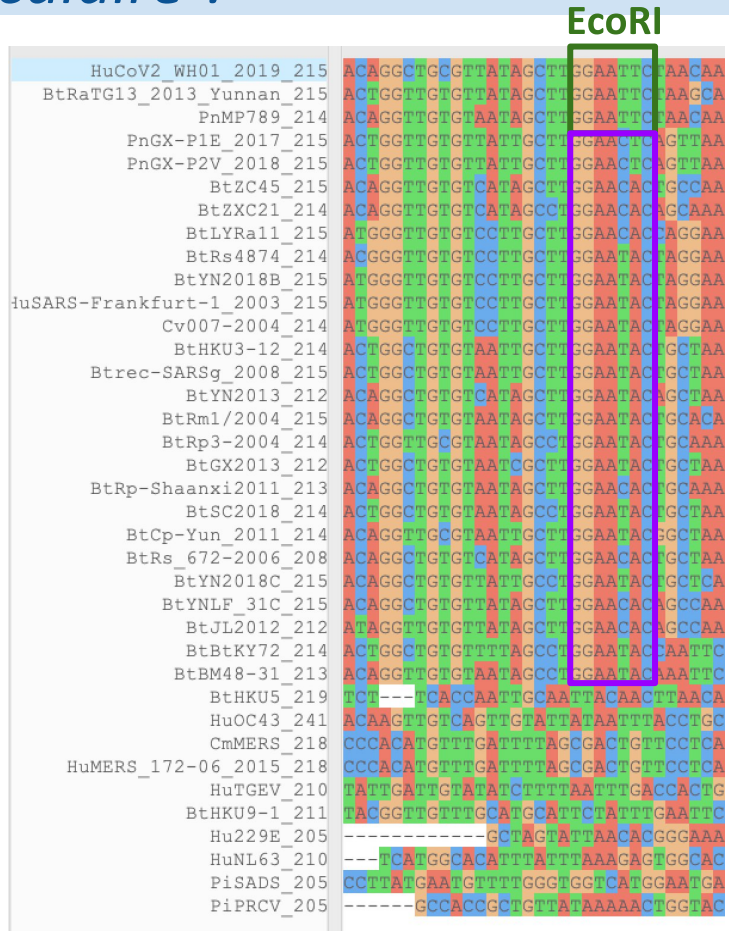


A. An EcoRI site is found at the 5'-end of the RBM and a BstEII site at the 3'-end.

B. Although these two restriction sites do not exist in the original spike gene of ZC45, they can be conveniently introduced given that the sequence discrepancy is small (2 nucleotides) in either case.

Un virus construit par ingénierie moléculaire ?

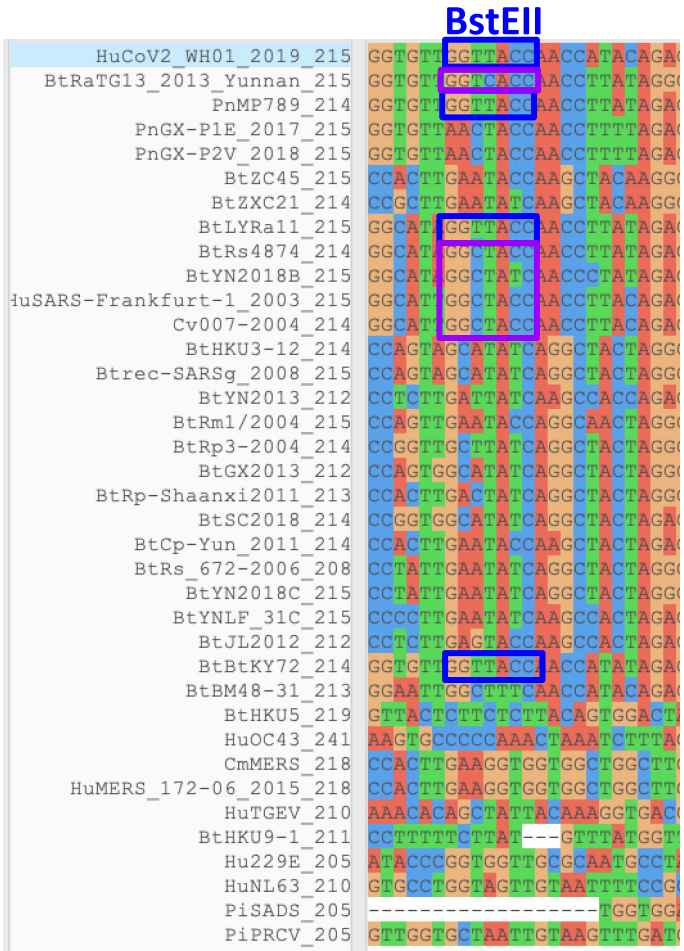
- Le site de restriction **EcoRI** (**séquence GAATTC**) se trouve dans le virus humain, mais également dans les virus les plus proches de chauve-souris (RatG13) et de pangolin (MP789).
- La majorité de ce site, ainsi que les séquences avoisinantes (**TTGGAAT*CT**) est conservée dans l'ensemble de la lignée SARS.
- Les séquences des autres SARS sont encore plus proche du site SARS-CoV-2 que celui de BtZ45 mis en avant par Li-Meng Yan.
- Il est donc plus vraisemblable que le site de SARS-CoV-2 soit apparu par une substitution d'un seul nucléotide à partir de n'importe quel SARS que par modification de 2 nucléotides du virus de chauve-souris BtZC45.



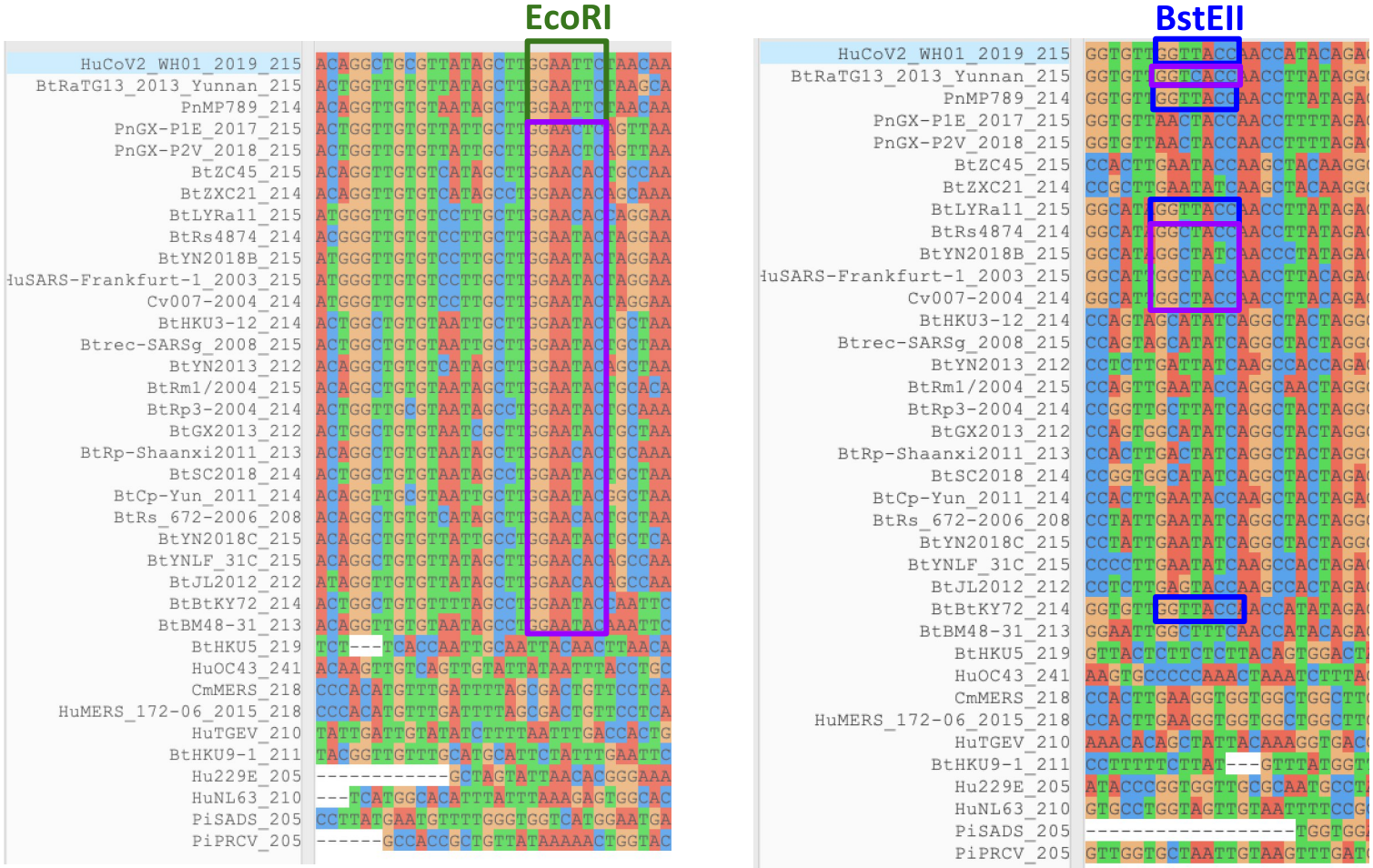
Quasi-identique à EcoRI

Un virus construit par ingénierie moléculaire ?

- Le site de restriction BstEII (**séquence GGTTACC**) se retrouve dans le virus humain, mais également dans les virus les plus proches de chauve-souris (RatG13) et de pangolin (MP789).
- Ce site se trouve également dans d'autres virus de la lignée, notamment ceux du SRAS humain (HuSARS-Frankfurt-1_2003), et de la civette (Cv007_2004).
- L'hypothèse de Li-Meng Yan (création à partir du génome de chauve-souris BtZC45) n'est donc pas convaincante.



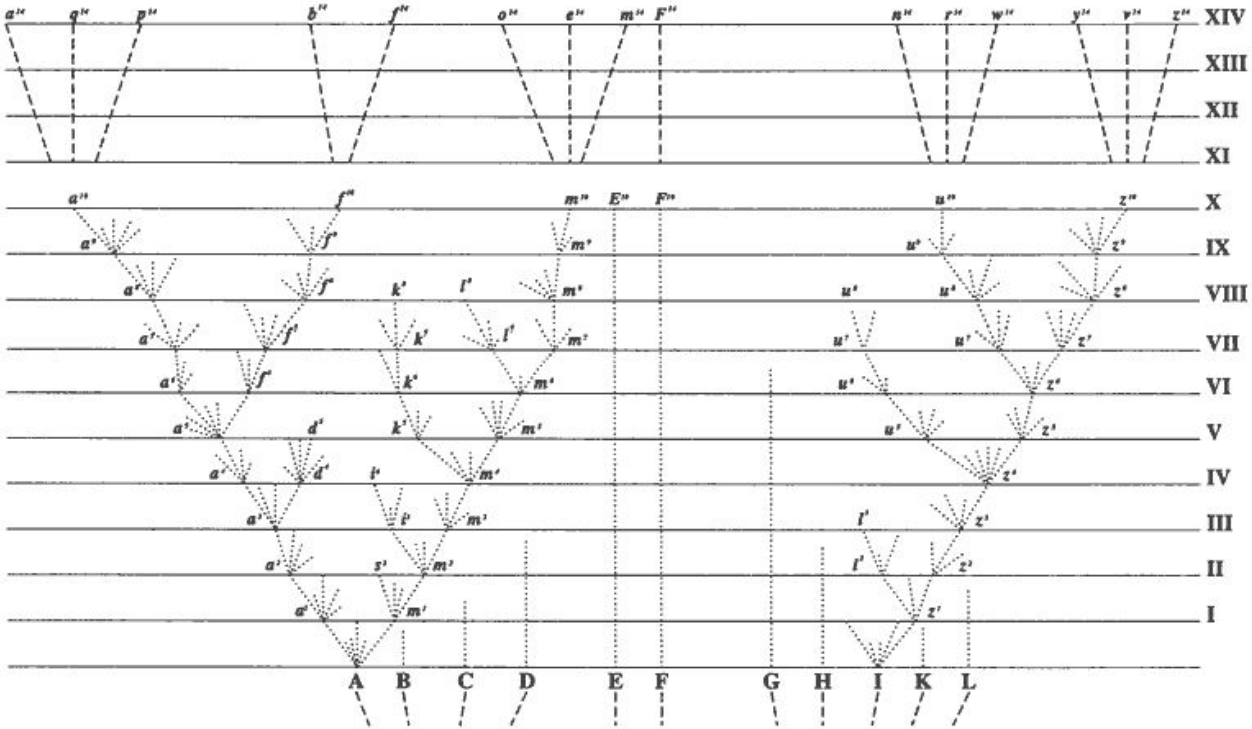
Co-occurrence des sites de restriction qui encadrent le RBD



Arbres de la vie

La divergence des caractères

La seule figure de l'Origine des Espèces (C.Darwin, 1859) est une représentation conceptuelle de l'arbre de la vie.

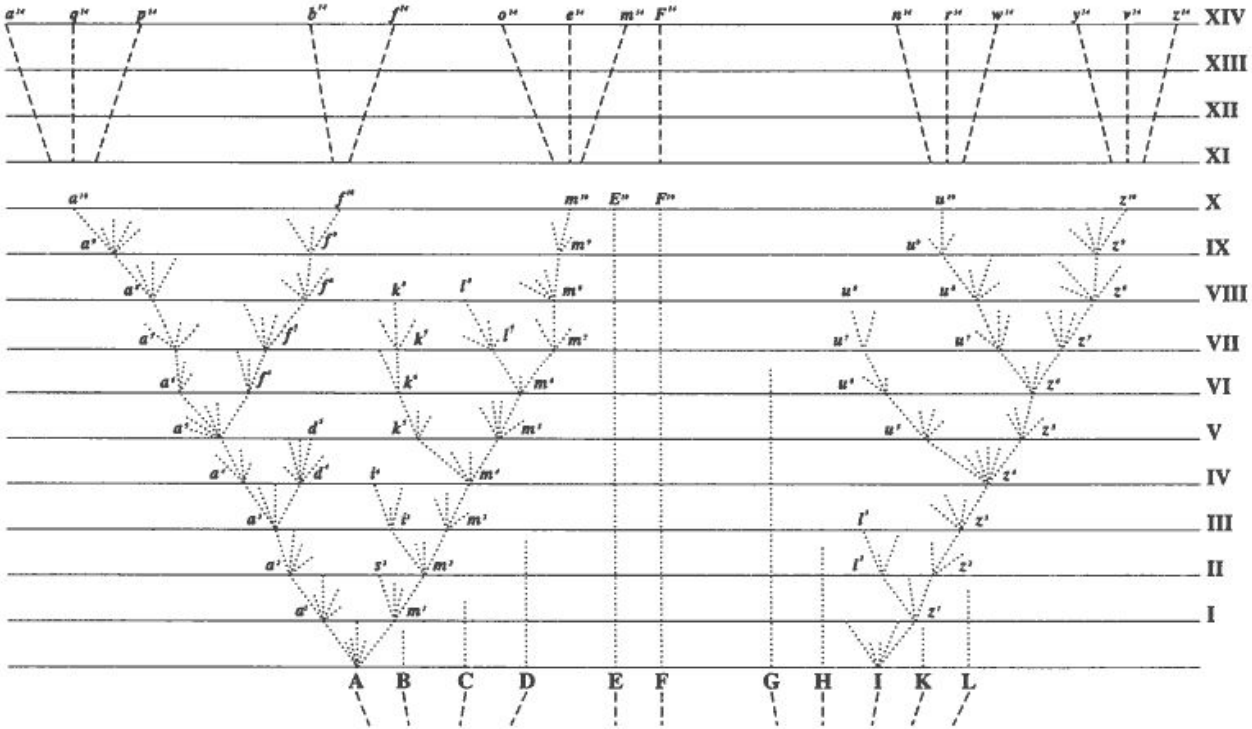


Source: Darwin, C. (1859). L'origine des espèces.

La divergence des caractères

Il s’agit d’un arbre synchrone : chaque niveau horizontal représente un moment donné.

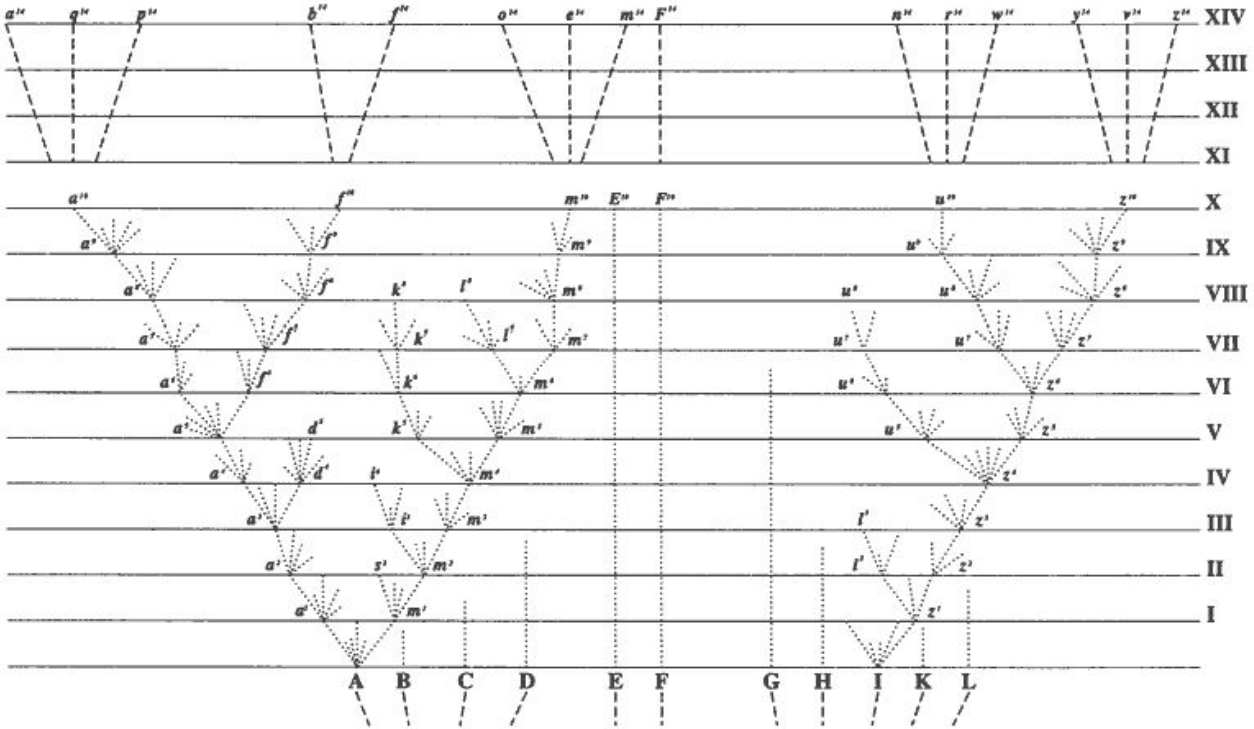
- La racine correspond aux époques les plus anciennes.
- Le niveau le plus élevé correspond au présent.
- A chaque époque on trouve des organismes de différents niveaux de complexité. La hauteur ne représente donc pas une complexité ou un “niveau d’ évolution”



Source: Darwin, C. (1859). L'origine des espèces.

La divergence des caractères

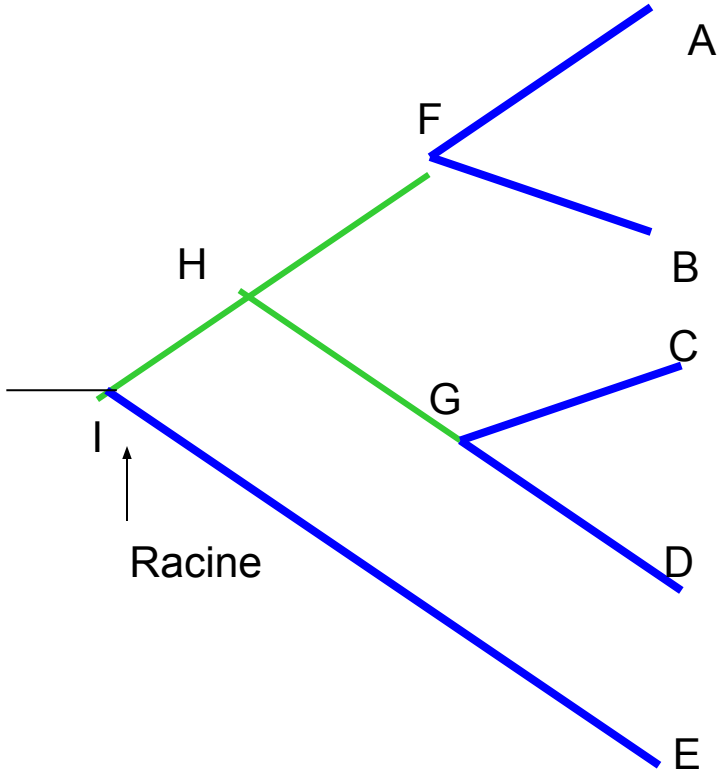
- La plupart des branches sont abortives
- **Evolution graduelle** par accumulation de variations (mutations) le long des branches.
- Juste après un branchement, on a de très petites différences entre les variétés.
- Les observations dont on dispose sont généralement fragmentaires.
- Elles ne sont pas forcément placées sur une trajectoire linéaire depuis un ancêtre donné jusqu'aux espèces actuelles.



Source: Darwin, C. (1859). L'origine des espèces.

Unités taxonomiques opérationnelles (OTU) et hypothétiques (HTU)

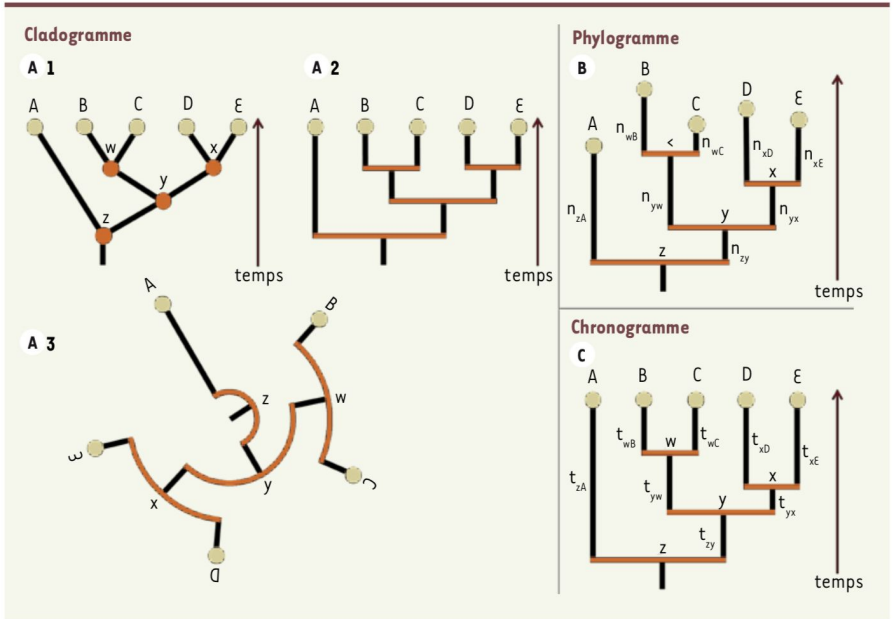
- Les relations évolutives entre les objets étudiés (espèces, organes, séquences) sont représentées par des arbres phylogénétiques
- Les arbres sont des graphes composés de *noeuds* et de *branches*
  - Noeuds = unités taxonomiques
    - Feuilles ou ***OTU*** = ***Unités Taxonomiques Opérationnelles*** (A, B, C, D, E), espèces existantes.
    - Noeuds internes ou HTU = Unités taxonomiques Hypothétiques (F, G, H, I), correspondent aux espèces ancestrales.
  - Branches = relations de parenté(ancêtre/descendants) entre unités taxonomiques
    - Branches internes
    - Branches externes
- On appelle ***topologie*** l'ensemble des branchements de l'arbre.



Source: Emese Meglézc

Représentations arborescentes des histoires évolutives

- On représente les histoires évolutives sous forme d'arbre
- Différents types de représentation peuvent être utilisés selon les cas.
  - Bifurcations triangulaires ou rectangulaires
  - Disposition radiale
- Selon les cas, les longueurs des branches représentent
  - Le nombre de divergences (cladogramme)
  - le nombre de différences entre deux espèces (phylogramme)
  - Le temps de divergence (chronogramme)

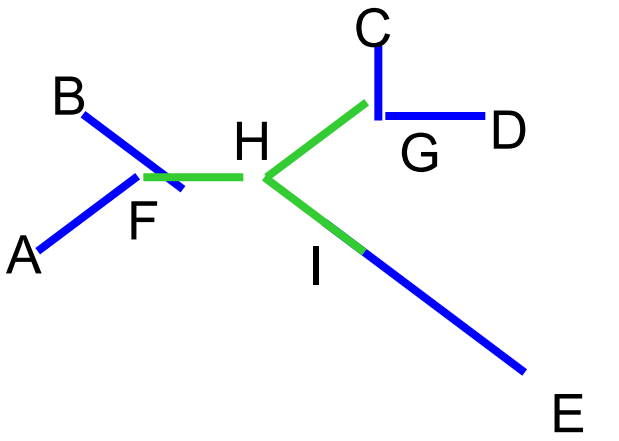


Casane & Laurenti (2011). Penser la biologie dans un cadre phylogénétique L'exemple de l'évolution des vertébrés. médecine/sciences

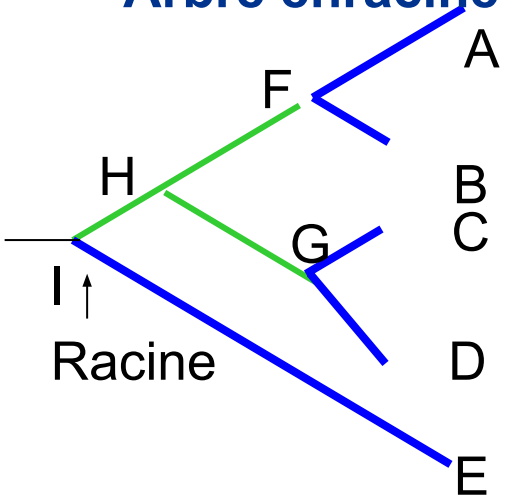
Arbres enracinés ou non enracinés

- Les arbres non-enracinés ne sont pas réellement des arbres phylogénétiques car ils n'ont pas de direction temporelle
  -> indiquent les distances, mais pas les relations de parenté entre les noeuds.
- La ***racine*** définit un une orientation de l'arbre, et donc un chemin évolutif unique vers chaque feuille.
- Elle symbolise le *dernier ancêtre commun* (*i.e*. le plus récent) de toutes les OTU.

Arbre non-enraciné



Arbre enraciné



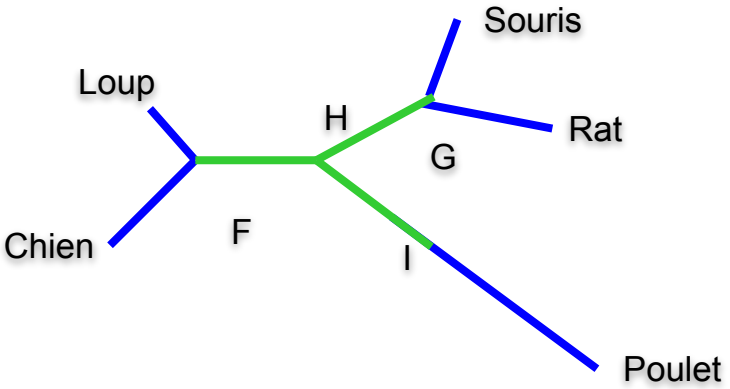
Combien d'arbres ?

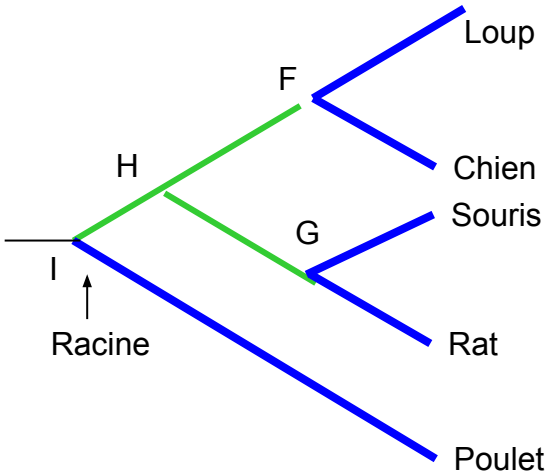
- Le nombre d'arbres possibles augmente de façon vertigineuse en fonction du nombre d'éléments terminaux (qu'ils représentent des molécules ou des espèces).
- Un seul de ces arbres correspond à l'histoire évolutive réelle.
- Puisqu'on ne dispose pas a priori de cet arbre, on doit l'***inférer*** à partir des éléments actuels (les unités taxonomiques opérationnelles, UTO).

n	Nb arbres enracinés	Nb arbres non-enracinés
2	1	1
3	3	1
4	15	3
5	105	15
6	945	105
7	10,395	945
8	135,135	10,395
9	2,027,025	135,135
10	3.45E+07	2,027,025
11	6.55E+08	3.45E+07
12	1.37E+10	6.55E+08
13	3.16E+11	1.37E+10
14	7.91E+12	3.16E+11
15	2.13E+14	7.91E+12
16	6.19E+15	2.13E+14
17	1.92E+17	6.19E+15
18	6.33E+18	1.92E+17
19	2.22E+20	6.33E+18
20	8.20E+21	2.22E+20
	$N_R = \frac{(2n-3)!}{2^{n-2}(n-2)!}$	$N_U = \frac{(2n-5)!}{2^{n-3}(n-3)!}$

Comment enraciner un arbre phylogénétique ?

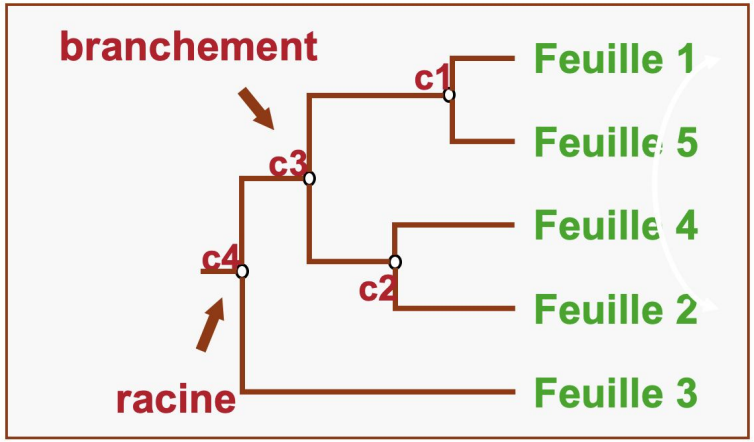
- Connaissance *a priori* de la feuille la plus externe parmi les OTU étudiées (« *outgroup* »)
  - Exemple: chien, loup, souris, rat et poulet
  - Sur base des connaissances biologiques, on décide que le **Groupe extérieur** est le poulet
- Sans connaissance *a priori* du OTU les plus externes parmi les OTU étudiées
  - Enracinement au poids moyen: on enracine l'arbre sur la branche qui minimise la moyenne des distances aux feuilles.
  - Ceci implique une hypothèse d'***horloge moléculaire***: on considère que le taux de mutation est constant au cours de l'évolution, et égal entre les branches.
  - Cette hypothèse n'est généralement pas très réaliste, il s'agit d'une approximation.

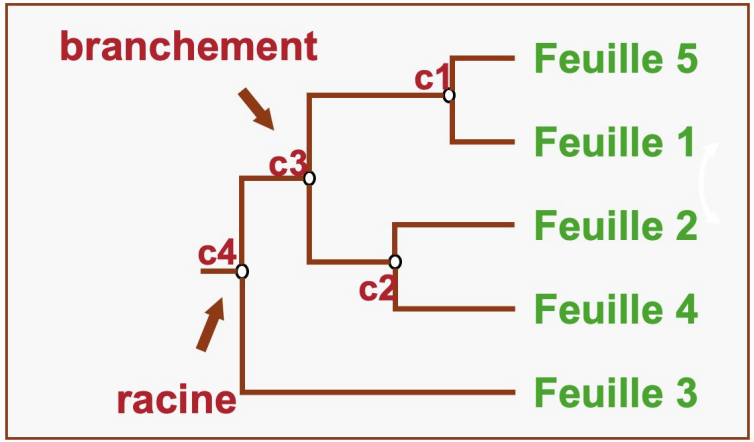




- Adapté d'après Emese Meglézc

Isomorphisme sur un arbre

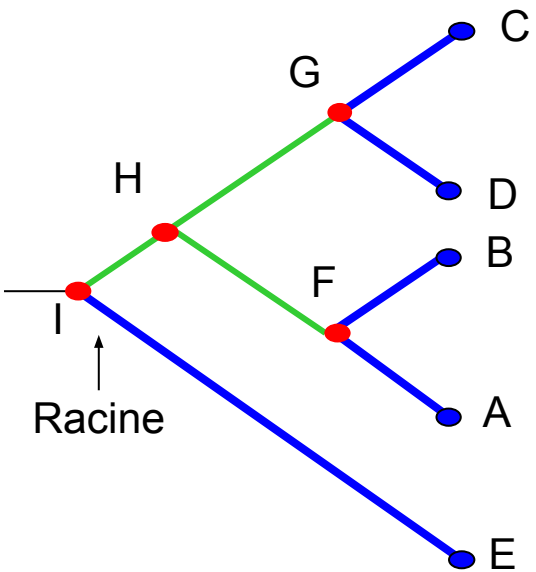


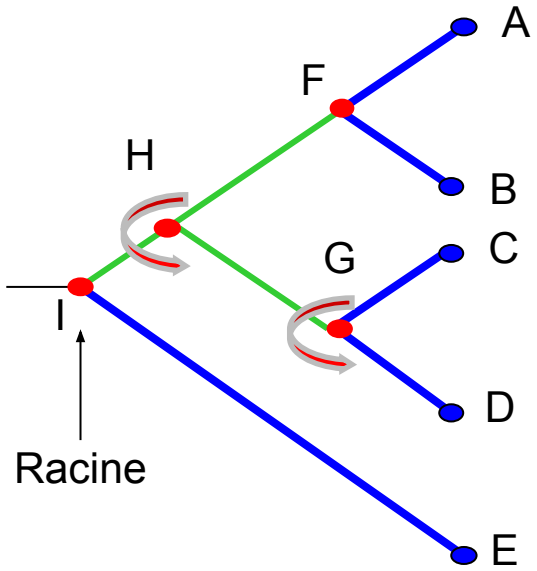


- Dans un arbre, les deux enfants de chaque branche peuvent être interchangés.
- Le résultat est un arbre ***isomorphique***, considéré équivalent à l'arbre initial.
- Les deux arbres de cauche sont équivalents.
- Cependant
  - Arbre du dessus: les feuilles 1 et 2 sont très éloignées.
  - Arbre du dessous: les feuilles 1 et 2 sont voisines.
- Les distances verticales entre deux nœuds ne reflètent pas leur distance réelle !
- La distance entre deux nœuds est la somme des longueurs des branches qui les séparent.

L'isomorphisme des arbres phylogénétiques

- Il faut éviter le piège d'évaluer les distance entre feuilles sur base de leur proximité verticale.
  - Les structures ci-dessous sont absolument identiques.
  - Pourtant les feuilles B et D semblent voisines sur le graphe de gauche, et éloignées sur celui de droite.
- Pour évaluer la distance entre deux nœuds d'un arbre , il faut prendre en compte la longueur totale du chemin le plus court pour les rejoindre (somme des longueurs de branches).

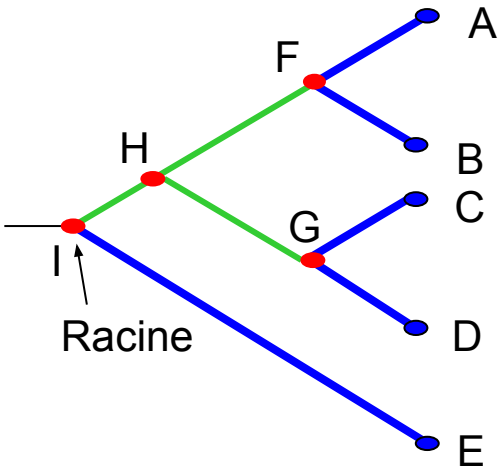




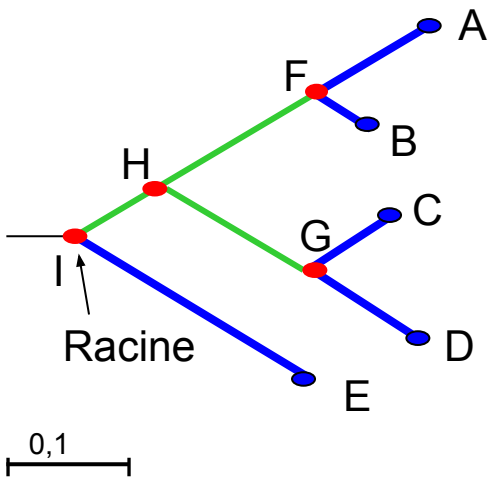
- Source: Emese Meglézc

Échelle d'un arbre phylogénétique

- **Cladogramme**
  - Représentation sans échelle
  - L'arbre indique uniquement l'ordre des branchements.
  - Les longueurs de branches ne sont pas proportionnelles au nombre de changements évolutifs.



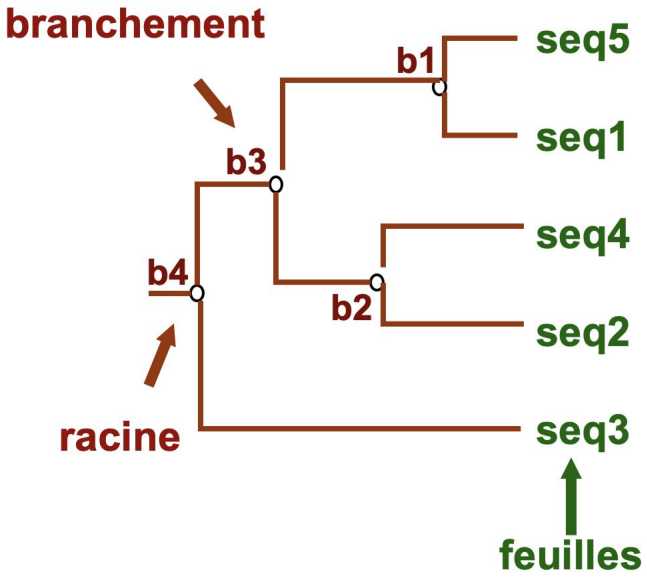
- **Phylogramme**
  - Représentation avec échelle
  - L'arbre indique les distances évolutives entre nœuds.
  - Les longueurs de branches sont proportionnelles au nombre d'événements évolutifs (substitutions ou substitution/sites).



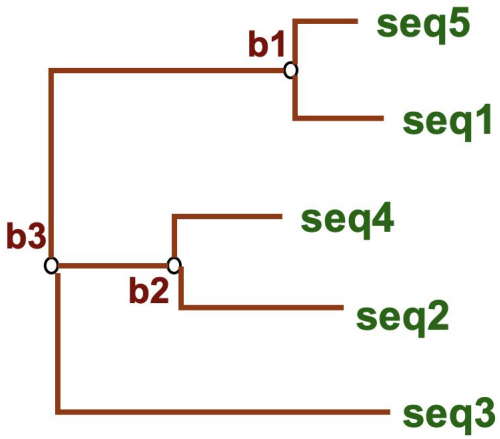
Source: Emese Meglézc

Calcul de la distance sur un arbre

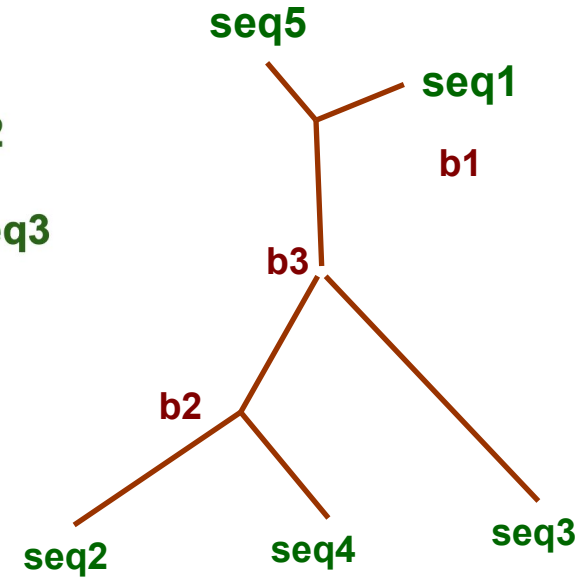
Arbre enraciné



Arbre non-enraciné



Arbre non-enraciné



- La distance entre deux nœuds est la somme des longueurs des branches qui les séparent.

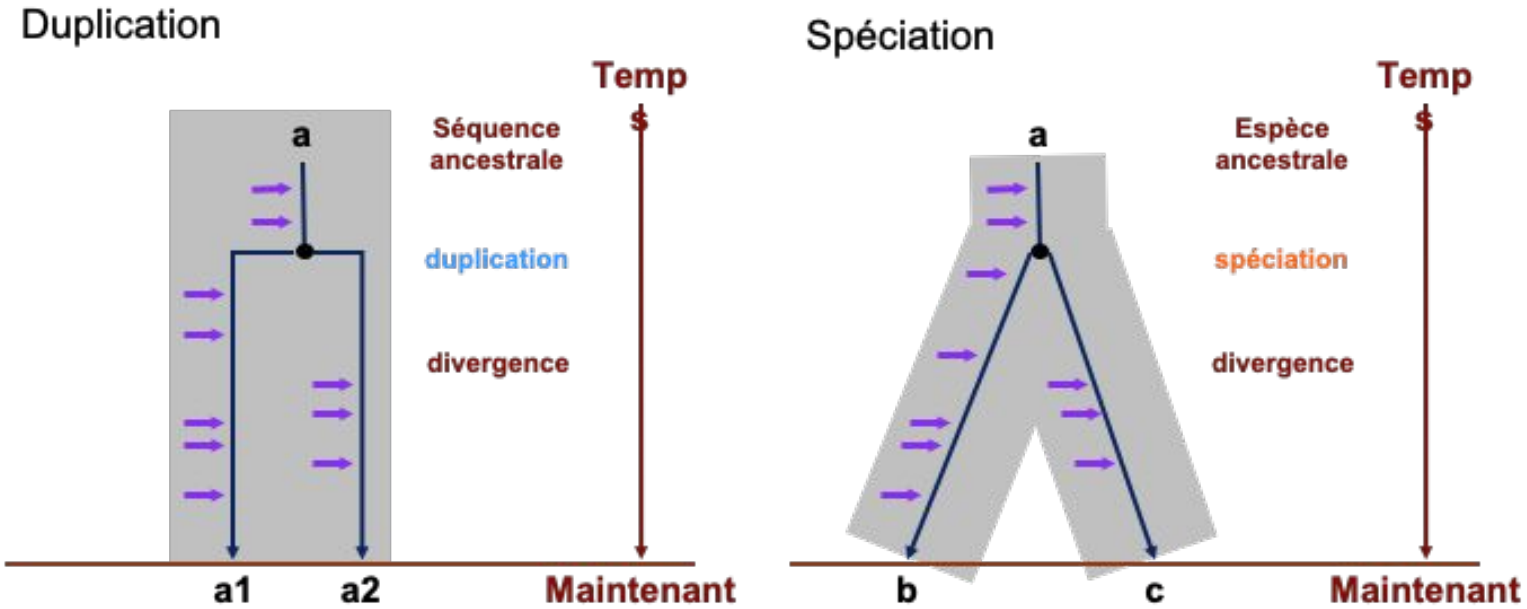
Concepts: homologie, paralogie, orthologie

Événements évolutifs

- Pour l'analyse de la phylogénie moléculaire, nous porterons un intérêt tout particulier à deux événements évolutifs particuliers: duplication et spéciation.
- ***Duplication***
  - Une duplication est une mutation qui génère un dédoublement d'une partie de l'ADN génomique. La duplication peut recouvrir l'ensemble du génome (formation de polyploïdes), un chromosome entier, ou un fragment de chromosome de taille plus ou moins grande.
  - Les duplications peuvent éventuellement entraîner l'apparition de copies multiples d'un ou plusieurs gènes, provoquant ainsi une certaine redondance de l'information génétique.
  - Dans certains cas, l'une des copies dupliquées du gène acquiert, par accumulation de mutations, de nouvelles caractéristiques qui lui permettent d'assumer une nouvelle fonction. Ce mécanisme, appelé duplication divergence, est à l'origine de la diversification des fonctions biologiques.
- ***Spéciation***
  - Processus évolutif qui résulte en la formation d'espèces distinctes à partir d'une espèce unique.
- Les événements de duplication et spéciation suscitent l'apparition de copies multiples à partir d'une seule séquence, soit au sein d'une même espèce (duplication), soit au sein des espèces distinctes dérivées de la spéciation. Ces séquences, dont la similarité résulte d'une séquence ancestrale commune, sont dites ***homologues***

Scénarios évolutifs

- Nous disposons de deux séquences, et nous supposons qu'elles divergent d'un ancêtre commun.
- La divergence peut résulter
  - d'une ***duplication*** (création de deux copies du gène dans le même génome)
  - ou d'une ***spéciation*** (formation d'espèces séparées à partir d'une espèce unique).
- Les flèches violettes indiquent les mutations (substitutions, délétions, insertions) qui s'accumulent au sein d'une séquence particulière au cours de son histoire évolutive. Ces mutations sont à l'origine de la diversification des séquences, des structures et des fonctions.



Homologie et analogie

- La similarité entre deux traits (organes, séquences) peut s'interpréter par deux hypothèses alternatives: homologie et analogie.
- **Homologie**
  - La similarité s'explique par le fait que les deux séquences divergent d'un ancêtre commun.
  - Les différences entre les deux caractères homologues résultent de l'accumulation de mutations à partir de l'ancêtre commun. Il s'agit donc d'une évolution par ***divergence évolutive***.
- **Analogie**
  - Ressemblance entre deux traits (organes, séquence) qui ne résulte pas d'une origine ancestrale commune (par opposition à l'homologie).
  - Les traits similaires sont apparus de façon ***indépendante***. Leur ressemblance peut éventuellement manifester l'effet d'une pression évolutive qui a sélectionné les mêmes propriétés.
  - Dans ce cas, on parle de ***convergence évolutive***.

Inférence d'homologie

- Inférence
  - Avant d'affirmer que deux séquences sont homologues, nous devrions pouvoir retracer leur histoire jusqu'à leur ancêtre commun.
  - Nous ne pouvons malheureusement pas disposer des séquences de toutes les espèces disparues. Il est donc impossible de démontrer formellement l'homologie.
  - Cependant, nous pouvons appuyer l'hypothèse d'homologie sur une analyse de la vraisemblance d'un scénario évolutif (taux de mutations, niveaux de similarités).
  - L'inférence d'homologie est toujours attachée à un certain ***risque de faux positifs.*** Les modèles évolutifs nous permettent d'estimer ce risque.

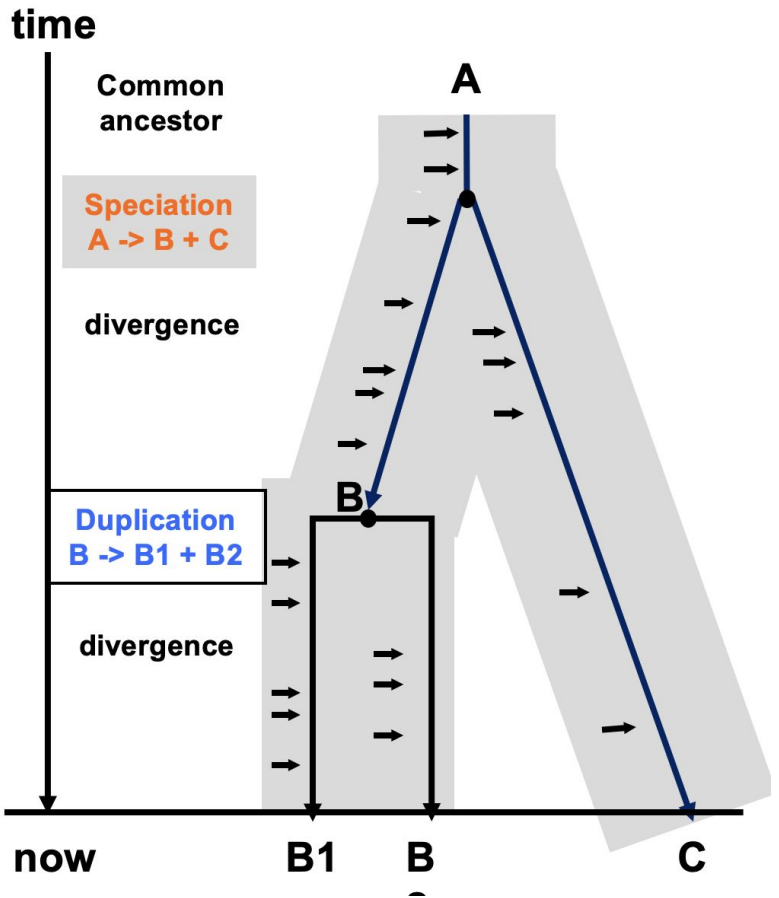
Inférence d'homologie

L'homologie est une relation logique (soit vraie, soit fausse).

- Deux séquences sont homologues (possèdent des caractères communs parce qu'elles dérivent d'un ancêtre commun) ou elles ne le sont pas.
- Il est donc complètement inapproprié de parler de « niveau d'homologie » ou « pourcentage d'homologie ».
- La formulation correcte
  - On observe un certain niveau de similarité entre deux séquences (pourcentages de résidus identiques, pourcentages de résidus « similaires »).
  - Sur cette base, on évalue deux scénarios évolutifs: cette similarité peut provenir d'une évolution convergente (analogie) ou divergente à partir d'un ancêtre commun (homologie).
  - Si la deuxième hypothèse est la plus vraisemblable, on *infère* que les séquences sont homologues.

Orthologie versus paralogie

- Zvelebil & Baum (2000) fournissent une définition claire et opérationnelle des concepts d'orthologie et paralogie.
  - **Orthologues:** séquences dont le dernier ancêtre commun précède immédiatement un événement de spéciation.
  - **Paralogues** séquences dont le dernier ancêtre commun précède immédiatement un événement de duplication
- Exemples:
  - B et C sont ***orthologues***, car leur dernier ancêtre commun (A) précède un événement de ***spéciation*** (A → B + C).
  - B1 et B2 sont ***paralogues*** car le premier événement évolutif qui succède à leur dernier ancêtre commun (B) est une ***duplication*** (B → B1 + B2).

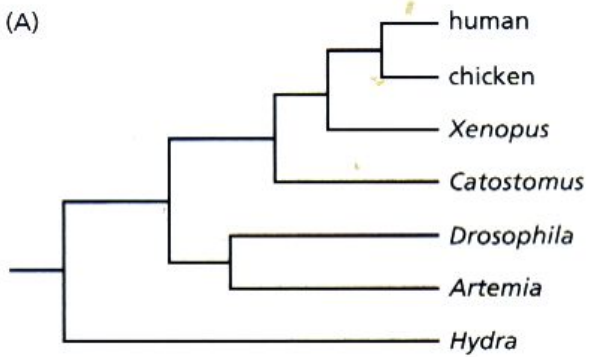


Approches d'inférence phylogénétique

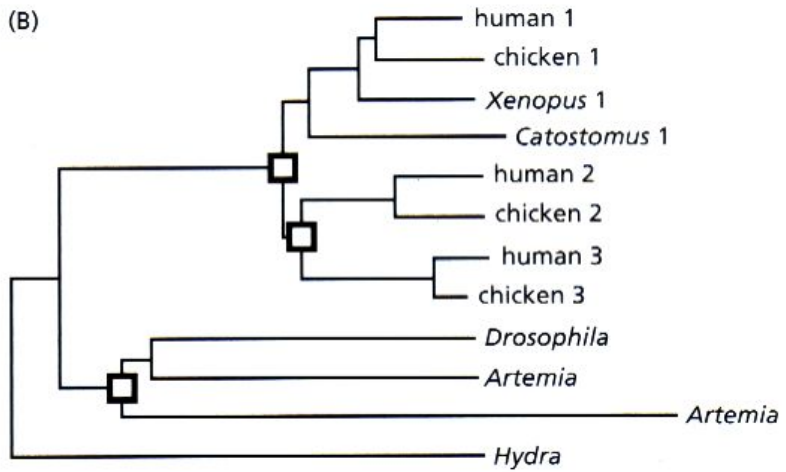
Inférence phylogénique à partir de séquences moléculaires

- En partant d'une famille de séquences macromoléculaires (ADN, ARN, protéines), on peut construire des arbres phylogéniques.
- En comparant l'arbre des molécules et l'arbre des espèces, on peut inférer l'histoire évolutive de cette famille de séquences.
- Nous reviendrons plus tard sur cet exemple, en expliquant les méthodes bioinformatiques permettant d'inférer des arbres moléculaires à partir de séquences, et les façons d'interpréter ces arbres en tenant compte de la filiation des espèces.

L'arbre des espèces



Arbre des molécules



Source: Zvelebil, M.J. and Baum, J.O. (2008) Understanding Bioinformatics. Garland Science, New York and London.

Phylogénomique

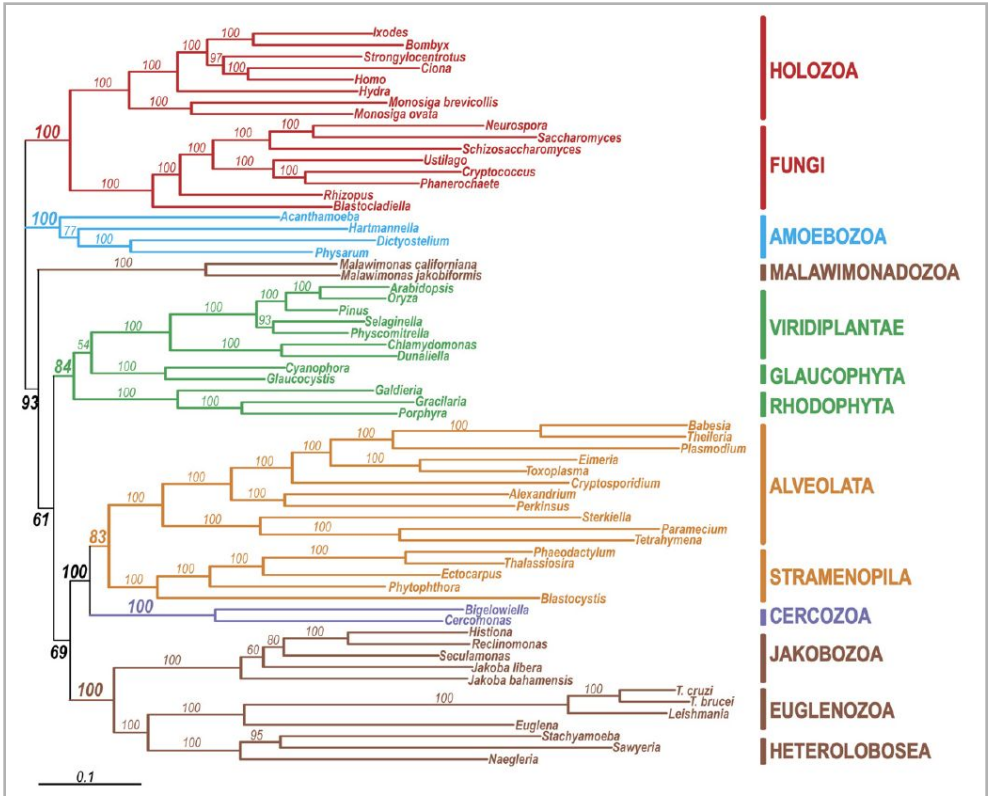


Figure 1. Maximum-Likelihood Tree of Eukaryotes

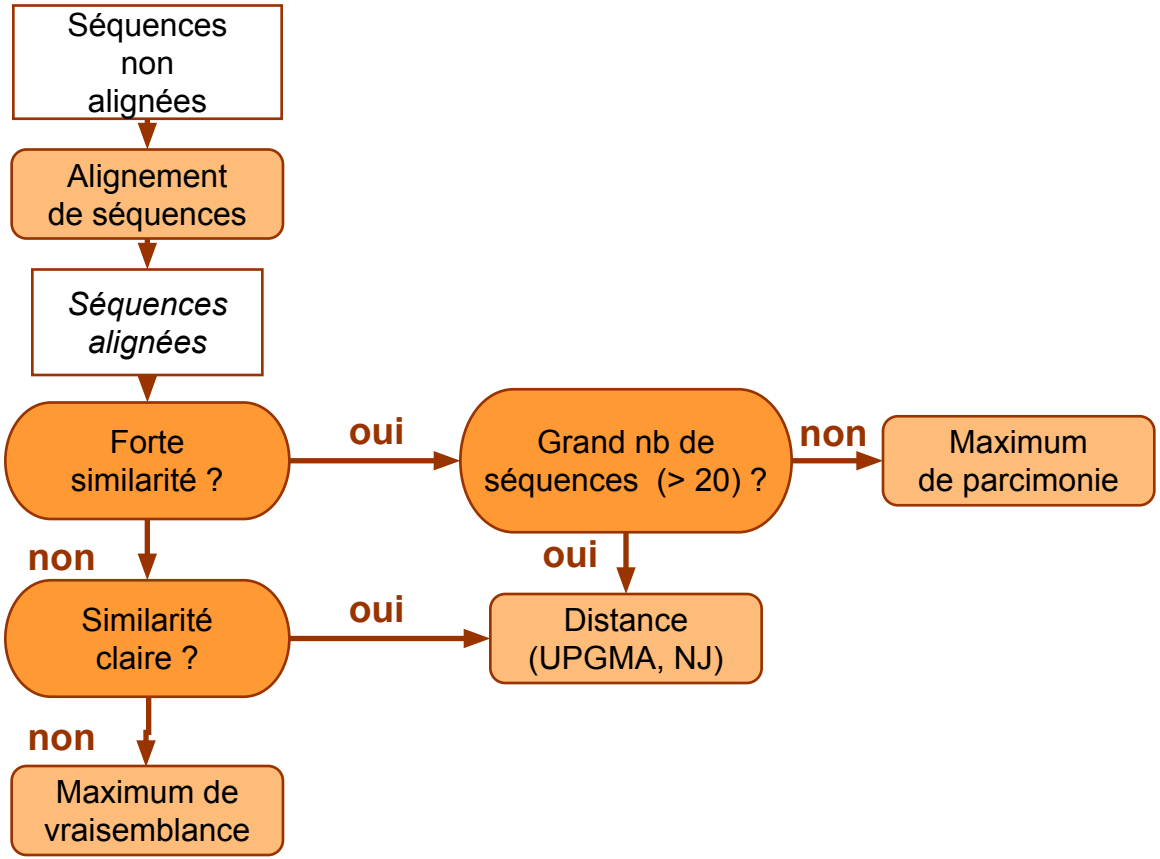
The tree includes 64 species and is based on 143 concatenated nucleus-encoded proteins (31,604 amino acid positions). Numbers indicate support values of RaxML analysis (100 replicates) with the WAG + F + Γ model. Posterior probabilities obtained in the Bayesian Inference with MrBayes are 1.0 for all branches. The scale bar denotes the estimated number of amino acid substitutions per site. The tree was rooted according to a gene fusion [13, 16].

- En phylogénie moléculaire, une approche classique consiste à se concentrer sur un gène considéré comme représentatif, et à construire un arbre sur base de la divergence de séquence de ce gène.
- Ces approches peuvent maintenant être généralisées en comparant les séquences de plusieurs centaines de gènes (ci-contre, arbre basé sur 143 familles de protéines).
- Elles permettent d'inférer des phylogénies entre organismes très éloignés (règnes différents), et d' établir ainsi des scénarios concernant les premières étapes de la diversification des êtres vivants.

- Source: Rodríguez-Ezpeleta et al. Curr Biol (2007) vol. 17 (16) pp. 1420-5
- Toward resolving the eukaryotic tree: the phylogenetic positions of jakobids and cercozoans

Inférence phylogénétique par comparaison de séquences

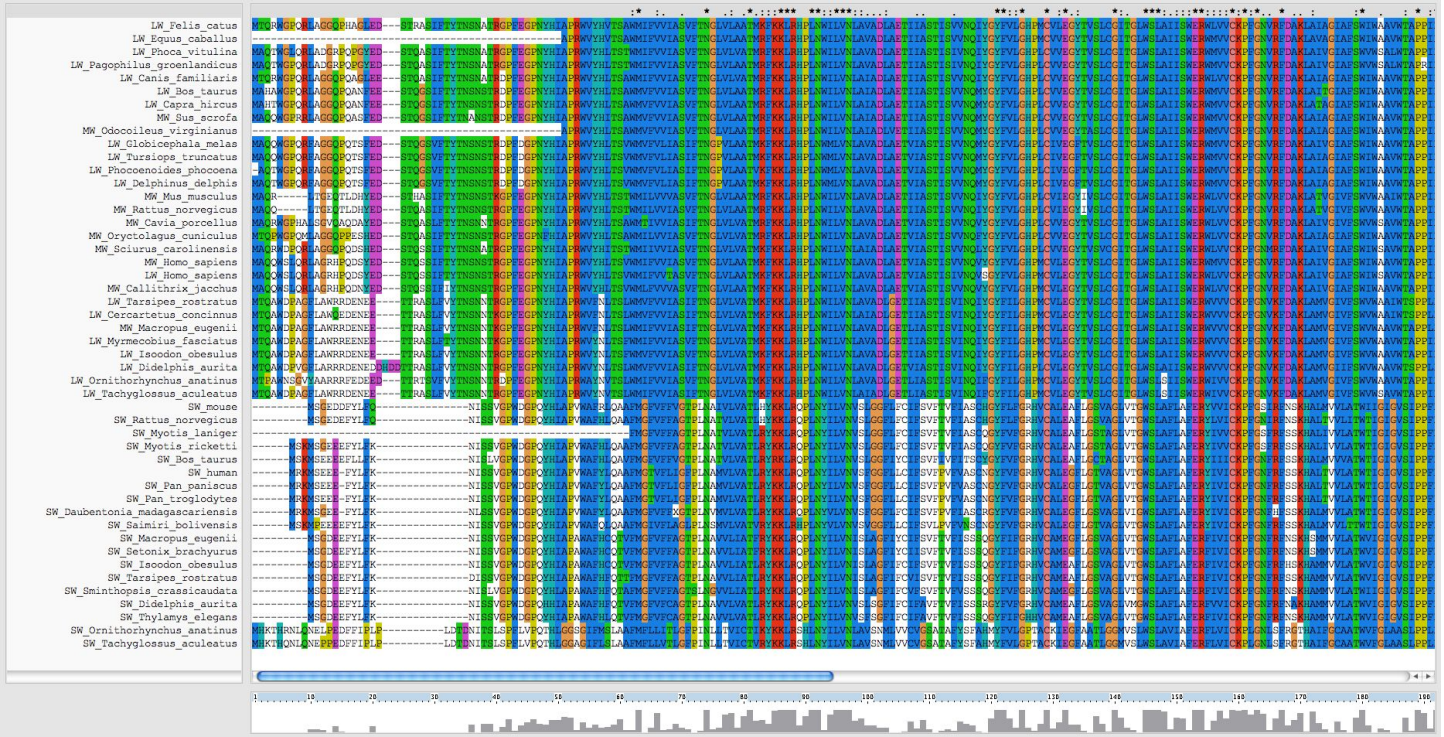
- Il existe plusieurs méthodes pour inférer un arbre évolutif à partir de séquences.
  - Maximum de parcimonie
  - Distance
  - Maximum de vraisemblance
- On part toujours d'un jeu de séquences alignées (alignement multiple).
- Le choix de la méthode dépend du nombre de séquences, et de leur degré de similarité.



- Source: Mount (2000)

Exemple : la famille des opsines

- Pour inférer un arbre phylogénétique à partir d'une famille de séquences, on part toujours d'un alignement multiple.
- La figure ci-dessous montre la première partie d'un alignement multiple entre 50 opsines de mammifère.
- A l'œil nu, on distingue déjà 2 groupes évidents.
  - Dessus: opsines sensibles aux ondes moyennes (vert) ou longues (rouge)



Matrice de pourcentages d'identité (opsines mammaliennes, export de clustalx)

Mammalian opsins - percent identify matrix generated by clustalx

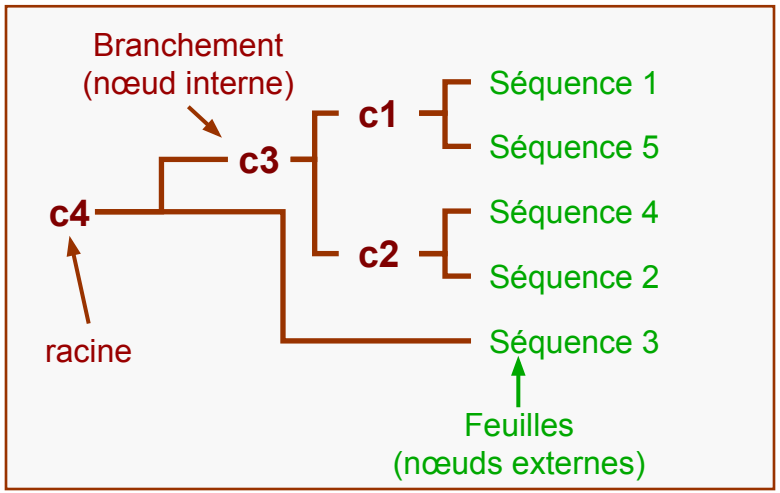
	1 LW_Felis_catus	2 LW_Equus_caballus	3 LW_Phoca_vitulina	4 LW_Pagophilus_groenlandicus	5 LW_Canis_familiaris	6 LW_Bos_taurus	7 LW_Capra_hircus	8 MW_Sus_scrofa	9 MW_Odocoileus_virginianus	10 LW_Globicephala_melas	11 LW_Tursiops_truncatus	12 LW_Phocoenoides_phocoena	13 LW_Delphinus_delphis	14 MW_Mus_musculus	15 MW_Rattus_norvegicus	16 MW_Cavia_porcellus	17 MW_Oryctolagus_cuniculus	18 MW_Sciurus_carolinensis	19 MW_Homo_sapiens	20 LW_Homo_sapiens	21 MW_Callithrix_jacchus	22 LW_Tarsipes_rostratus	23 LW_Cercartetus_concinnus	24 MW_Macropus_eugenii	25 MW_Myrmecobius_fasciatus	26 LW_Isoodon_obesulus	27 LW_Didelphis_aurita	28 LW_Ornithorhynchus_anatinus	29 LW_Tachyglossus_aculeatus	30 SW_mouse	31 SW_Rattus_norvegicus	32 SW_Myotis_laniger	33 SW_Myotis_ricketti	34 SW_Bos_taurus	35 SW_human	36 SW_Pan_paniscus	37 SW_Pan_troglodytes	38 SW_Daubentonia_madagascariensis	39 SW_Saimiri_bolivensis	40 SW_Macropus_eugenii	41 SW_Setonix_brachyurus	42 SW_Isoodon_obesulus	43 SW_Tarsipes_rostratus	44 SW_Sminthopsis_crassicaudata	45 SW_Didelphis_aurita	46 SW_Thylamys_elegans	47 SW_Ornithorhynchus_anatinus	48 SW_Tachyglossus_aculeatus
1 LW_Felis_catus	100	97	92	92	95	93	93	93	95	90	90	89	90	88	89	87	91	90	90	91	91	83	84	84	86	85	84	86	85	41	42	42	42	42	42	42	42	41	42	42	43	43	42	44	41	42	41	40
2 LW_Equus_caballus	97	100	95	95	96	96	96	96	96	92	93	92	92	91	92	90	93	91	93	91	92	85	86	87	89	88	86	89	88	41	42	43	44	42	43	43	42	43	43	44	43	44	45	42	43	42	42	42
3 LW_Phoca_vitulina	92	95	100	99	94	92	92	92	91	92	88	88	88	89	90	88	90	91	90	91	90	82	82	82	84	83	83	83	83	41	42	43	42	42	42	42	42	43	43	43	43	43	42	43	42	43	42	41
4 LW_Pagophilus_groenlandicus	92	95	99	100	93	92	92	91	92	88	88	88	88	89	90	88	89	91	90	91	89	81	81	82	83	83	82	82	82	40	41	42	42	42	41	41	41	41	42	42	43	43	43	41	42	41	41	41
5 LW_Canis_familiaris	95	96	94	93	100	95	95	94	95	90	90	89	90	89	90	90	93	92	90	89	90	83	83	83	85	84	83	84	84	41	43	43	43	42	42	42	42	42	42	43	43	43	44	43	41	43	41	41
6 LW_Bos_taurus	93	96	92	92	95	100	99	97	97	93	94	93	93	88	89	87	90	90	89	89	91	81	82	81	83	83	82	83	82	41	43	42	43	42	42	42	42	42	42	43	43	43	44	41	43	41	40	
7 LW_Capra_hircus	93	96	92	92	95	99	100	97	97	93	94	93	93	88	89	87	90	90	89	89	90	81	82	81	83	83	82	82	82	41	42	42	43	42	42	42	42	42	43	43	43	43	44	41	43	41	40	
8 MW_Sus_scrofa	93	96	91	91	94	97	97	100	97	94	94	93	93	88	89	87	90	90	89	90	90	80	81	80	82	82	81	82	82	41	42	41	42	42	41	41	41	42	42	42	43	43	43	44	41	42	40	40
9 MW_Odocoileus_virginianus	95	96	92	92	95	97	97	97	100	93	93	93	93	89	89	89	91	89	91	89	92	84	85	85	87	86	85	86	85	41	42	42	44	42	42	42	42	42	42	43	42	43	45	42	42	42	42	42
10 LW_Globicephala_melas	90	92	88	88	90	93	93	94	93	100	99	98	99	86	87	86	88	88	88	88	89	80	80	80	81	81	81	81	80	41	41	40	42	42	41	41	41	41	42	42	43	42	43	44	41	42	40	39
11 LW_Tursiops_truncatus	90	93	88	88	90	94	94	94	93	99	100	99	99	86	87	86	88	88	88	88	89	79	80	79	81	80	81	81	80	41	41	40	42	42	41	41	41	41	42	42	43	43	43	44	41	42	40	39
12 LW_Phocoenoides_phocoena	89	92	88	88	89	93	93	93	93	98	99	100	98	85	86	85	88	87	87	87	88	78	79	78	80	80	80	80	79	40	41	40	41	41	41	41	41	41	42	42	42	42	43	41	42	40	39	
13 LW_Delphinus_delphis	90	92	88	88	90	93	93	93	93	99	99	98	100	85	86	85	88	87	87	87	88	79	79	79	80	80	80	80	80	40	41	39	41	41	41	41	41	41	42	42	42	43	40	41	40	39		
14 MW_Mus_musculus	88	91	89	89	89	88	88	88	89	86	86	85	85	100	98	91	90	90	88	88	89	82	83	82	84	83	83	83	82	41	43	42	43	43	43	43	43	44	43	44	44	42	43	43	42			
15 MW_Rattus_norvegicus	89	92	90	90	90	89	89	89	89	87	87	86	86	98	100	92	91	91	90	89	90	83	83	83	85	84	84	84	83	41	43	42	43	43	43	43	43	44	43	44	44	42	43	43	42			
16 MW_Cavia_porcellus	87	90	88	88	90	87	87	87	89	86	86	85	85	91	92	100	91	91	89	87	88	82	83	83	84	83	83	83	83	41	43	43	43	43	42	42	42	42	43	43	43	43	44	45	42	43	42	41
17 MW_Oryctolagus_cuniculus	91	93	90	89	93	90	90	90	91	88	88	88	88	90	91	91	100	92	88	88	87	80	80	80	82	81	80	81	81	41	42	42	42	42	42	42	41	42	42	42	43	43	44	41	42	41	41	
18 MW_Sciurus_carolinensis	90	91	91	91	92	90	90	90	89	88	88	87	87	90	91	91	92	100	90	91	90	80	81	80	82	82	82	82	82	41	43	42	42	42	42	42	42	42	43	43	43	44	45	42	43	41	40	
19 MW_Homo_sapiens	90	93	90	90	90	89	89	89	91	88	88	87	87	88	90	89	88	90	100	96	96	82	82	83	85	84	84	85	83	42	43	43	44	44	43	43	43	44	44	44	44	44	45	43	44	43	42	
20 LW_Homo_sapiens	91	91	91	91	89	89	89	90	89	88	88	87	87	88	89	87	88	91	96	100	96	82	82	82	84	83	83	85	83	41	42	41	42	42	42	42	41	42	42	43	43	43	44	42	43	42	42	
21 MW_Callithrix_jacchus	91	92	90	89	90	91	90	90	92	89	89	88	88	89	90	88	87	90	96	96	100	83	83	83	85	85	85	86	84	41	43	42	43	43	42	42	42	43	43	43	43	43	45	42	43	42	41	
22 LW_Tarsipes_rostratus	83	85	82	81	83	81	81	80	84	80	79	78	79	82	83	82	80	80	82	82	83	100	98	96	95	96	96	90	92	42	43	43	42	42	42	42	42	43	43	44	43	43	43	42	44	42	41	
23 LW_Cercartetus_concinnus	84	86	82	81	83	82	82	81	85	80	80	79	79	83	83	83	80	81	82	82	83	98	100	96	95	95	95	89	91	42	43	43	43	43	42	42	42	42	43	43	44	43	44	43	43	44	42	41
24 MW_Macropus_eugenii	84	87	82	82	83	81	81	80	85	80	79	78	79	82	83	83	80	80	83	82	83	96	96	100	97	95	94	90	92	42	42	43	42	42	42	42	42	43	43	44	43	44	43	42	44	42	42	
25 LW_Myrmecobius_fasciatus	86	89	84	83	85	83	83	82	87	81	81	80	80	84	85	84	82	82	85	84	85	95	95	97	100	97	94	91	92	42	42	43	42	42	42	42	42	43	43	44	43	43	43	42	44	41	40	
26 LW_Isoodon_obesulus	85	88	83	83	84	83	83	82	86	81	80	80	80	83	84	83	81	82	84	83	85	96	95	95	97	100	95	91	92	42	42	42	42	42	41	41	41	41	42	42	43	43	43	42	41	43	40	40
27 LW_Didelphis_aurita	84	86	83	82	83	82	82	81	85	81	81	80	80	83	84	83	80	82	84	83	85	96	95	94	94	95	100	90	91	42	42	42	42	42	41	41	41	42	42	43	43	43	44	43	42	43	41	40
28 LW_Ornithorhynchus_anatinus	86	89	83	82	84	83	82	82	86	81	81	80	80	83	84	83	81	82	85	85	86	90	89	90	91	91	90	100	93	40	41	42	42	42	41	41	41	41	42	43	44	43	44	43	42	43	41	40
29 LW_Tachyglossus_aculeatus	85	88	83	82	84	82	82	82	85	80	80	79	80	82	83	83	81	82	83	83	84	92	91	92	92	92	91	93	100	39	40	42	41	41	41	41	41	41	42	43	43	43	42	41	42	42	40	39
30 SW_mouse	41	41	41	40	41	41	41	41	41	41	41	40	40	41	41	41	41	41	42	41	42	42	42	42	42	42	40	39	100	96	92	90	87	86	86	86	88	85	86	85	87	85	86	85	87	51	51	
31 SW_Rattus_norvegicus	42	42	42	41	43	43	42	42	42	41	41	41	41	43	43	43	42	43	43	43	42	42	42	42	42	41	40	96	100	91	91	87	87	87	87	88	86	86	85	88	86	85	85	87	51	50		
32 SW_Myotis_laniger	42	43	43	42	43	42	42	42	41	42	40	40	40	39	42	42	43	42	42	43	41	42	43	43	43	43	42	42	42	42	92	91	100	99	92	90	90	90	93	87	89	88	91	91	90	88	91	51	51
33 SW_Myotis_ricketti	42	44	42	42	43	43	43	42	44	42	42	41	41	43	43	43	42	42	44	42	43	42	43	42	42	42	42	41	90	91	99	100	91	89	89	89	91	88	88	88	89	89	88	86	90	49	49	
34 SW_Bos_taurus	42	42	42	42	42	42	42	42	42	42	42	41	41	43	43	43	42	42	44	42	43	42	43	42	42	42	42	41	87	87	92	91	100	86	86	86	87	86	86	86	86	86	83	86	49	48		
35 SW_human	42	43	42	41	42	42	42	41	42	41	41	41	41	43	43	43	42	42	43	42	42	42	42	42	42	41	41	41	86	87	90	89	86	100	100	100	91	92	83	83	85	83	85	82	86	52	51	
36 SW_Pan_paniscus	42	43	42	41	42	42	42	41	42	41	41	41	41	43	43	42	42	42	43	42	42	42	42	42	42	41	41	41	86	87	90	89	86	100	100	100	91	92	83	83	85	83	85	82	86	52	51	
37 SW_Pan_troglodytes	42	43	42	41	42	42	42	41	42	41	41	41	41	43	43	42	42	42	43	42	42	42	42	42	42	41	41	41	86	87	90	89	86	100	100	100	91	92	83	83	85	83	85	82	86	52	51	
38 SW_Daubentonia_madagascariensis	41	42	42	41	42	42	42	42	42	42	41	41	41	43	43	42	41	42	43	41	42	42	42	42	42	41	42	41	88	88	93	91	87	91	91	91	100	88	85	85	86	86	86	84	87	50	49	
39 SW_Saimiri_bolivensis	42	43	43	42	42	42	42	42	42	42	42	42	41	44	44	43	42	42	44	42	43	43	43	43	42	42	42	41	85	86	87	88	86	92	92	92	88	100	84	83	84	84	84	82	84	52	51	
40 SW_Macropus_eugenii	42	43	43	42	42	42	43	43	42	42	42	42	42	41	43	43	43	42	43	44	42	43	43	43	43	43	42	42	86	86	89	88	86	83	83	83	85	84	100	99	95	94	93	91	94	51	51	
41 SW_Setonix_brachyurus	43	44	43	43	43	43	43	43	43	43	43	42	42	44	44	43	42	43	44	43	43	44	44	44	43	43	44	43	85	85	88	88	86	83	83	83	85	83	99	100	95	95	93	90	94	50	50	
42 SW_Isoodon_obesulus	43	43	43	43	43	43	43	43	42	42	42	42	42	43	43	43	43	43	44	43	43	43	43	43	42	43	43	43	87	88	91	89	86	85	85	85	86	84	95	95	100	95	94	91	95	50	51	
43 SW_Tarsipes_rostratus	42	44	43	43	43	43	43	43	43	43	43	43	42	42	44	44	44	43	44	44	43	43	43	44	44	43	43	44	85	86	91	89	86	83	83	83	86	84	94	95	95	100	94	90	93	51	51	
44 SW_Sminthopsis_crassicaudata	44	45	43	43	44	44	44	44	45	44	44	43	43	44	44	45	44	45	45	44	45	43	43	43	43	42	43	43	86	85	90	88	86	85	85	85	86	84	93	93	94	94	100	90	92	51	51	
45 SW_Didelphis_aurita	41	42	42	41	42	42	41	41	41	42	41	41	40	42	42	42	41	42	43	42	42	42	43	42	42	41	42	42	41	85	85	88	86	83	82	82	82	84	82	91	90	91	90	90	100	94	49	49
46 SW_Thylamys_elegans	42	43	43	42	43	43	43	42	42	42	42	42	42	41	43	43	43	42	43	44	43	43	44	44	44	43	43	43	42	87	87	91	90	86	86	86	86	87	84	94	94	95	93	92	94	100	50	50
47 SW_Ornithorhynchus_anatinus	41	42	42	41	41	41	41	40	42	40	40	40	40	43	43	42	41	41	43	42	42	42	42	41	40	41	41	40	51	51	51	49	49	52	52	52	50	52	51	50	50	51	51	49	50	100	96	
48 SW_Tachyglossus_aculeatus	40	42	41	41	41	40	40	40	42	39	39	39	39	42	42	41	41	40	42	42	41	41	41	42	40	40	40	39	51	50	51	49	48	51	51	51	49	51	51	50	51	51	51	49	50	96	100	

Principe de la construction de l'arbre

Matrice de distance

	séquence 1	séquence 2	séquence 3	séquence 4	séquence 5
séquence 1	0.00	4.00	6.00	3.50	1.00
séquence 2	4.00	0.00	6.00	2.00	4.50
séquence 3	6.00	6.00	0.00	5.50	6.50
séquence 4	3.50	2.00	5.50	0.00	4.00
séquence 5	1.00	4.50	6.50	4.00	0.00

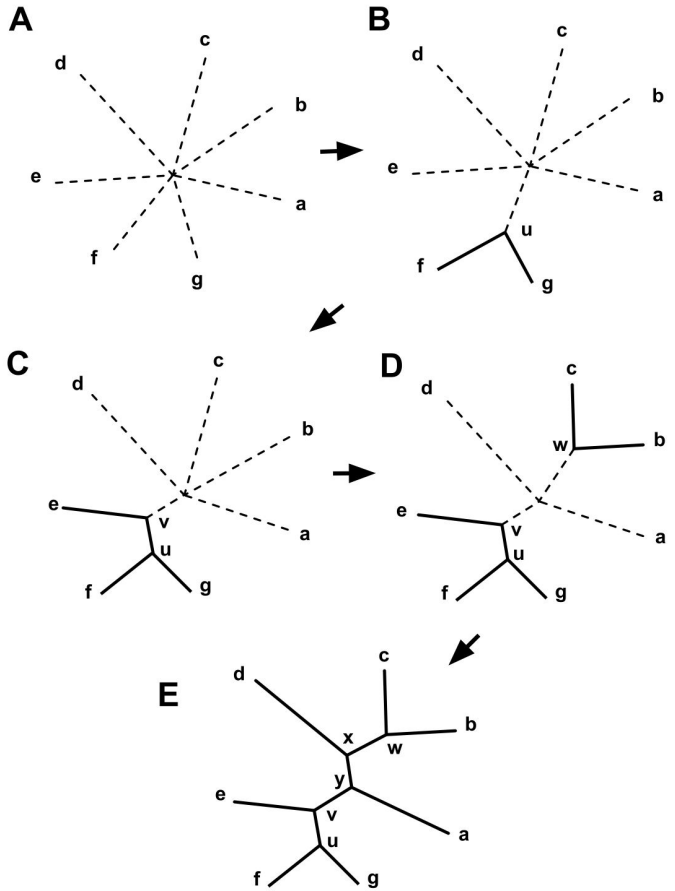
Arbre



- Le clustering hiérarchique est une méthode de clustering agrégative.
  - Prend une matrice de distance en entrée
  - Regroupe progressivement les objets en allant des plus proches aux plus distants.
- Il existe plusieurs possibilités pour établir une règle d'agglomération, qui définit la distance entre deux groupes.
  - Liaison simple (***single linkage***): distance entre groupes A et B est la distance entre les plus proches de leurs éléments respectifs.
  - Liaison moyenne (***average linkage***): distance moyenne entre tous les objets des deux groupes (=**UPGMA**).
  - Liaison complète (***complete linkage***): distance entre les éléments les plus éloignés des groupes A et B.
- Algorithme
  - 1. Assigner chaque objet à un cluster séparé.
  - 2. Identifier la paire de clusters les plus proches, et les regrouper en un seul.
  - 3. Répéter la seconde étape jusqu'à ce qu'il ne

Neighbour joining (NJ) - Méthode

- Développé par Saitou et Nei (1987) est une approximation de l'algorithme pour trouver l'arbre le plus court (minimum évolution)
- Principe:
  - A chaque étape, rechercher le couple d'UTO qui minimise la longueur totale de l'arbre

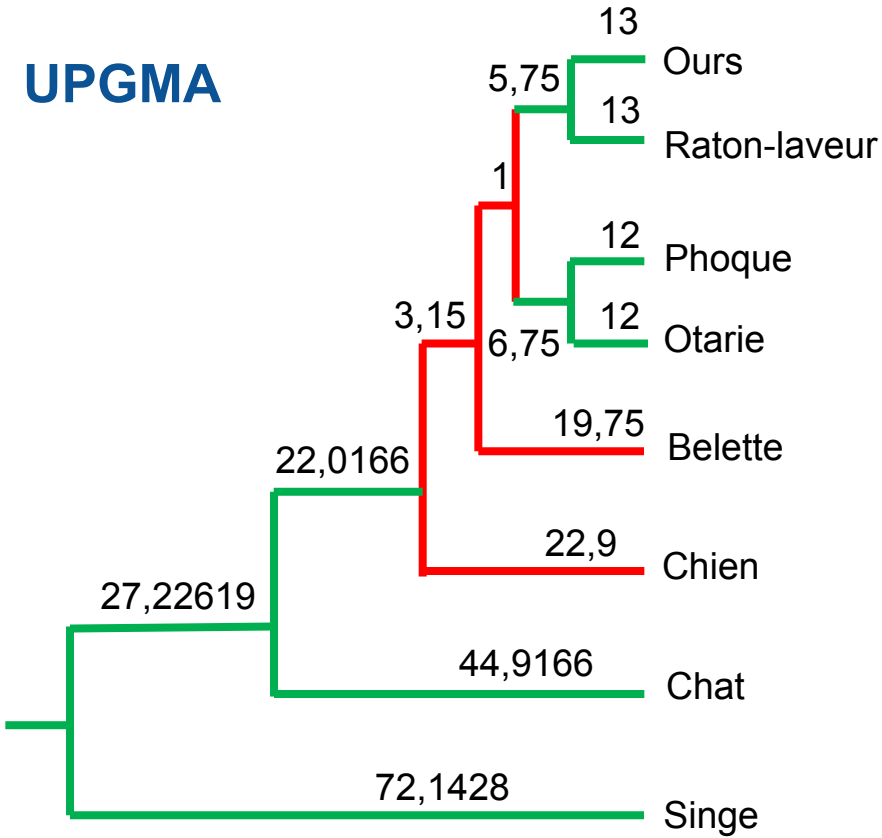


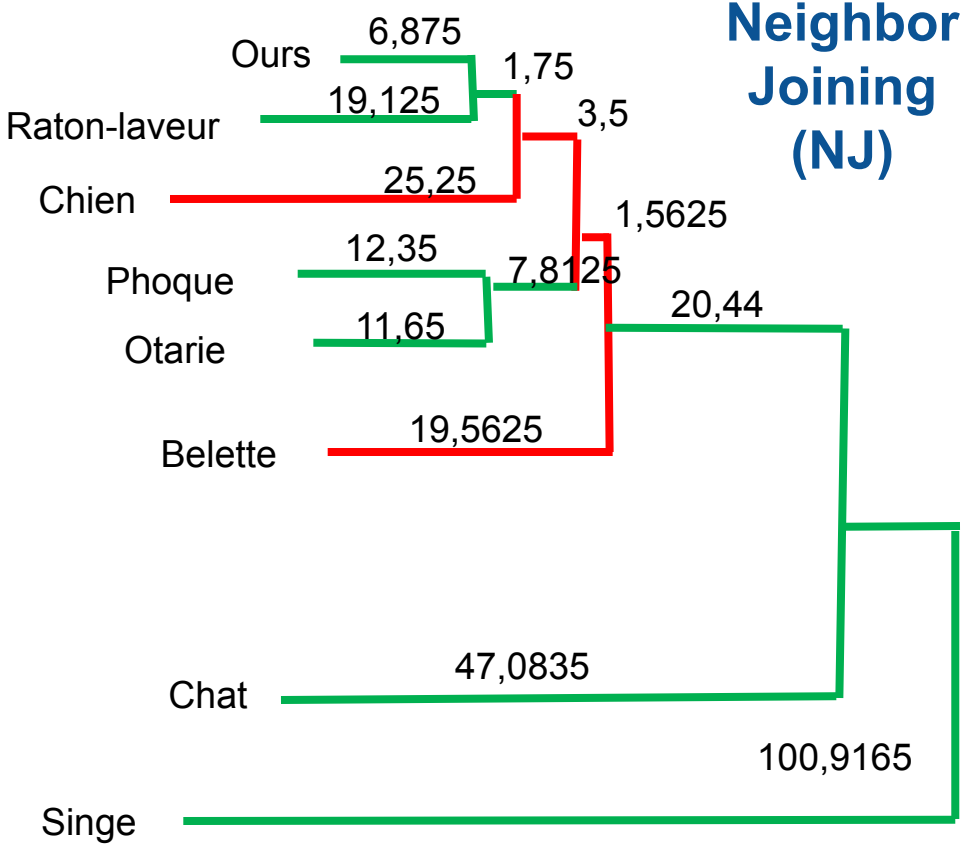
- Diapo: Emese Meglecz

Propriétés de la méthode NJ

- Méthode rapide et simple qui permet de travailler avec un très grand nombre de taxons.
- Les arbres ne sont pas enracinés.
- Les longueurs des branches sont informatives (phylogramme).
- Bonne approximation de la méthode du minimum d'évolution (l'arbre le plus court).
- Retrouve l'arbre vrai si la matrice de distances est un reflet exact des distances évolutives (malheureusement ce n'est pas souvent le cas).
- Ne dépend pas d'hypothèse de l'horloge moléculaire, donc la méthode est applicable dans les cas où le taux d'évolution varie entre les lignées.

Comparaison UPGMA - Neighbour Joining

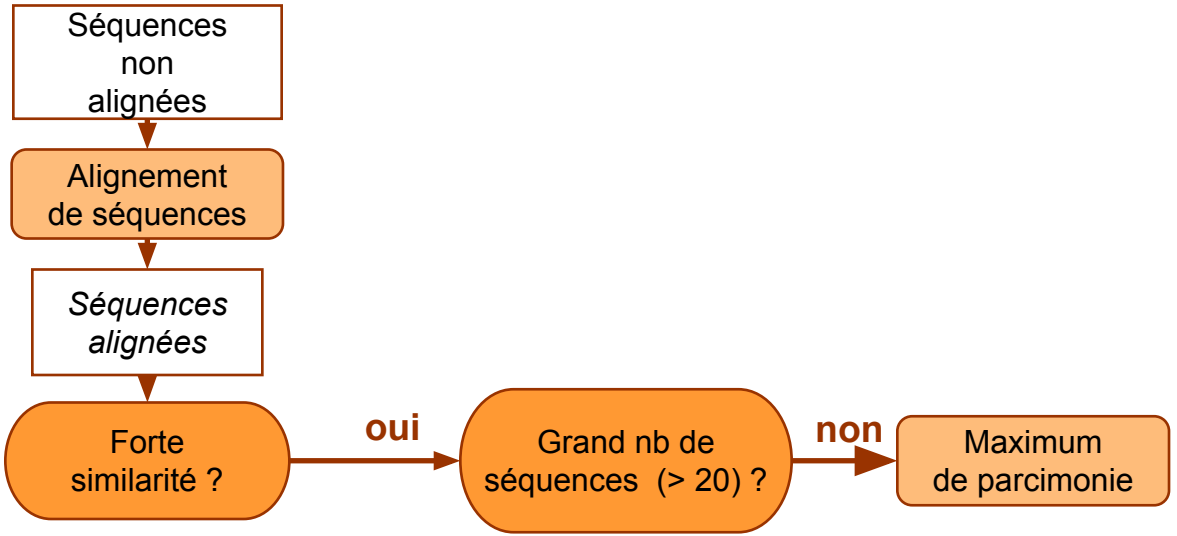




Inférence phylogénétique par la méthode du maximum de parcimonie

Inférence phylogénétique par comparaison de séquences

- Approches alternatives
  - Maximum de parcimonie
  - Distance
  - Maximum de vraisemblance



- Source: Mount (2000)

Maximum de parcimonie - Méthode

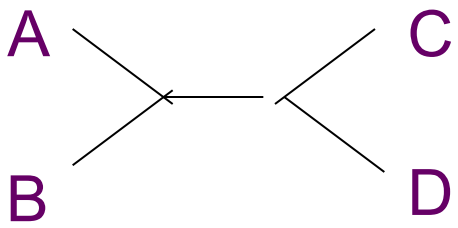
- Principe:
  - Identifier la topologie T qui implique le plus petit nombre de changements évolutifs suffisant à rendre compte des différences observées entre les OTU étudiées.
  - Utilise des états de caractères discrets => L'arbre le plus parcimonieux => plus court chemin conduisant aux états de caractères observés

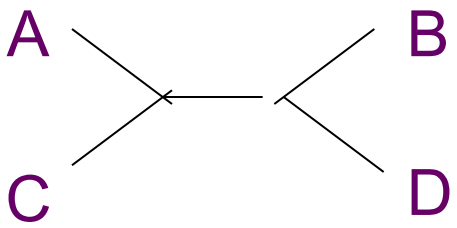
- Algorithme
  - Construction de tous les arbre possibles.
  - Pour chaque site (position de l'alignement), on compte le nombre de substitutions nécessaires pour expliquer chaque arbre.
  - On retient l'arbre qui nécessite le plus petit nombre de substitutions au total (en tenant compte de tous les sites).

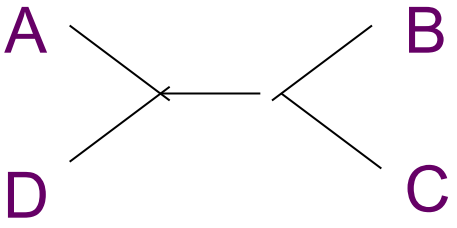
- Caractéristique des arbres obtenus
  - Solutions multiples => plusieurs arbres avec le même nombre minimum de changements peuvent être obtenus.
  - La longueur des branches ne reflète par la distance évolutive (arbre sans échelle).
  - Arbres non enracinés.

Déterminer toutes les topologies possibles
4 UTO => 3 arbres non racinés

Espèces	Séquences								
	1	2	3	4	5	6	7	8	9
A	A	A	G	A	G	T	T	C	A
B	A	G	C	C	G	T	T	C	T
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	T



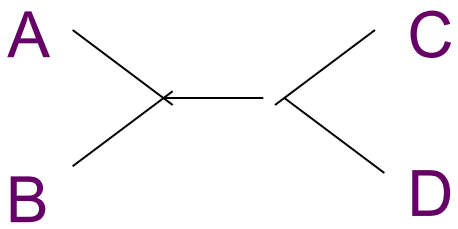


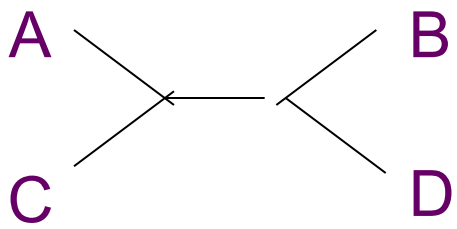


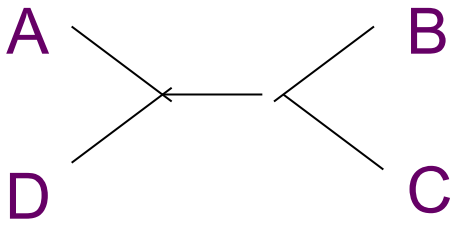
Maximum de parcimonie – classification des sites

- Caractère ***invariant***: toutes les OTU possèdent le même état de caractères pour un site donné.
- Caractère ***variable***
  - ***Non informatif*** si les états de caractères à ce site ne favorisent aucune topologie parmi l'ensemble des topologies possibles
  - ***Informatif*** si les états de caractères à ce site favorise une (ou plusieurs) topologie(s) parmi l'ensemble des topologies possibles

Espèces	Séquences								
	1	2	3	4	5	6	7	8	9
A	A	A	G	A	G	T	T	C	A
B	A	G	C	C	G	T	T	C	T
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	T

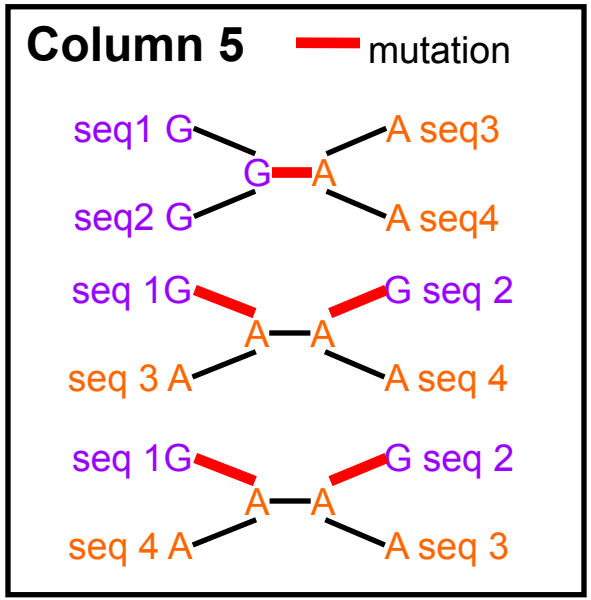






Maximum de parcimonie

position	1	2	3	4	5	6	7	8	9
seq1	A	A	G	A	G	T	G	C	A
seq2	A	G	C	C	G	T	G	C	G
seq3	A	G	A	T	A	T	C	C	A
seq4	A	G	A	G	A	T	C	C	G

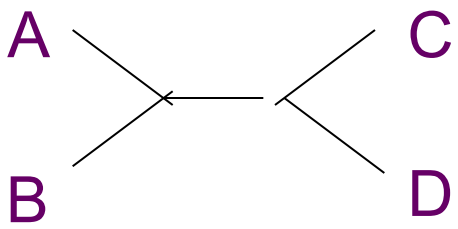


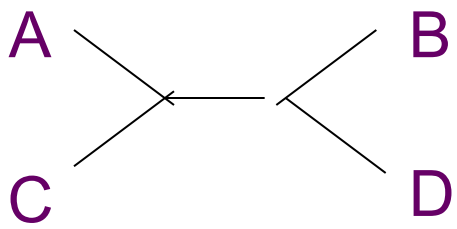
Adapted from Mount (2000)

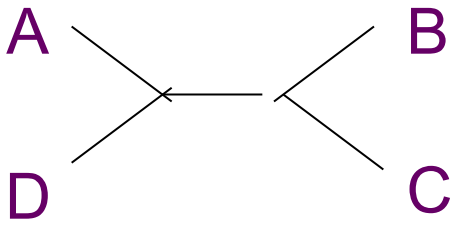
- Pour chacune des colonnes de l'alignement, tous les arbres possibles sont évalués.
- Pour chaque colonne informative, l'arbre qui présente le plus petit nombre de mutations est retenu.
- On retient ensuite l'arbre qui correspond au nombre le plus élevé de colonnes (consensus)
- Note: cette approche peut éventuellement retourner plusieurs arbres *ex aequos*.

Déterminer toutes les topologies possibles
4 UTO => 3 arbres non racinés

Espèces	Séquences								
	1	2	3	4	5	6	7	8	9
A	A	A	G	A	G	T	T	C	A
B	A	G	C	C	G	T	T	C	T
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	T







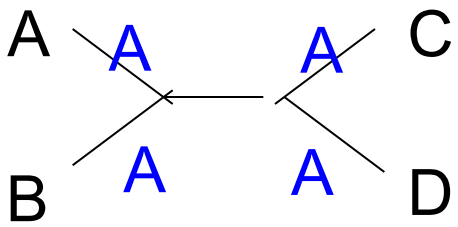
Maximum de parcimonie - Méthode

Pour un caractère donné, on compte le nombre de changements évolutifs (CE) pour chaque topologie possible.

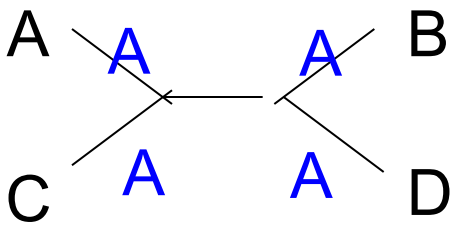
Étude du caractère n°1

	1	2	3	4	5	6	7	8	9
A	A	A	G	A	G	T	T	C	A
B	A	G	C	C	G	T	T	C	T
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	T

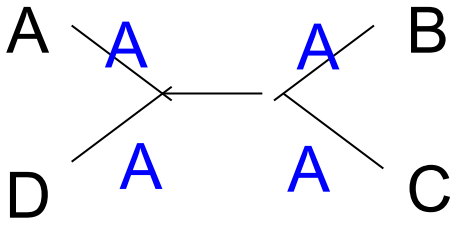
Caractère constant (même état de caractère à tous les sites).
Caractère **non informatif** : ne favorise aucune topologie par rapport à une autre.



Nb CE= 0



Nb CE= 0

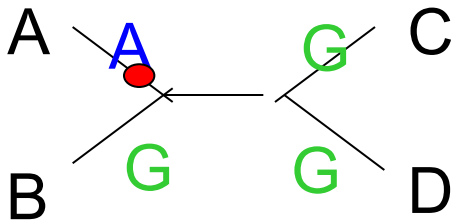


Nb CE= 0

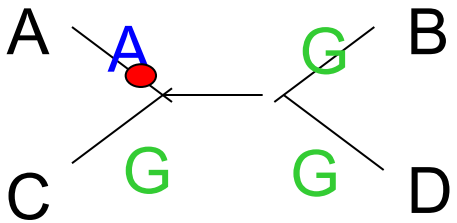
Étude du caractère n°2

	1	2	3	4	5	6	7	8	9
A	A	A	G	A	G	T	T	C	A
B	A	G	C	C	G	T	T	C	T
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	T

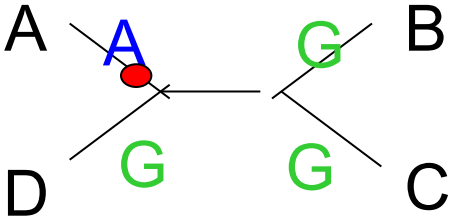
Caractère **variable** mais **non informatif**.
Caractère ne favorisant aucune topologie par rapport à une autre.



Nb CE= 1



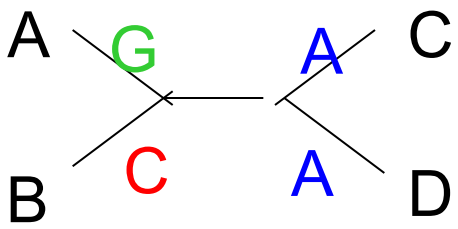
Nb CE= 1

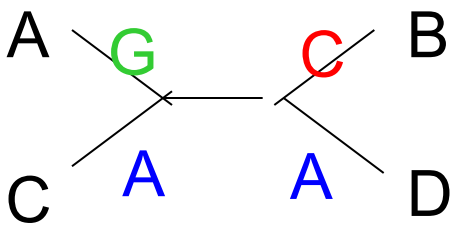


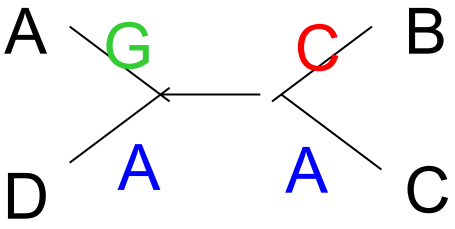
Nb CE= 1

Étude du caractère n°3

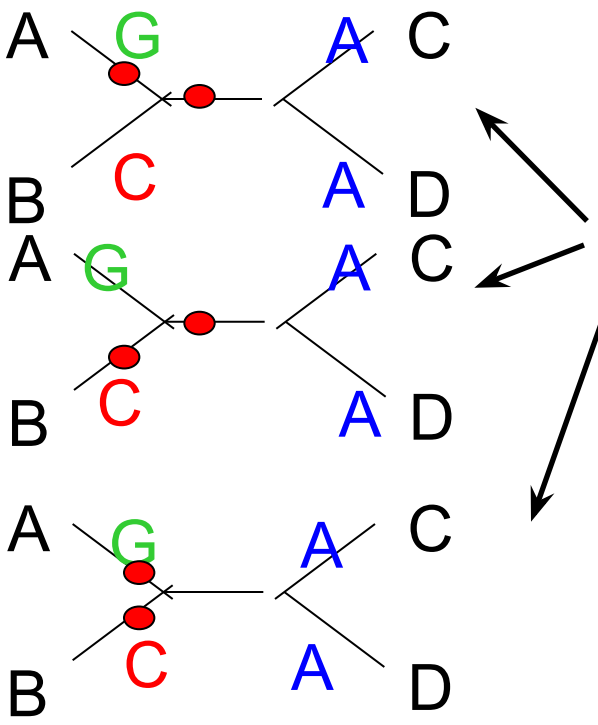
	1	2	3	4	5	6	7	8	9
A	A	A	G	A	G	T	T	C	A
B	A	G	C	C	G	T	T	C	T
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	T





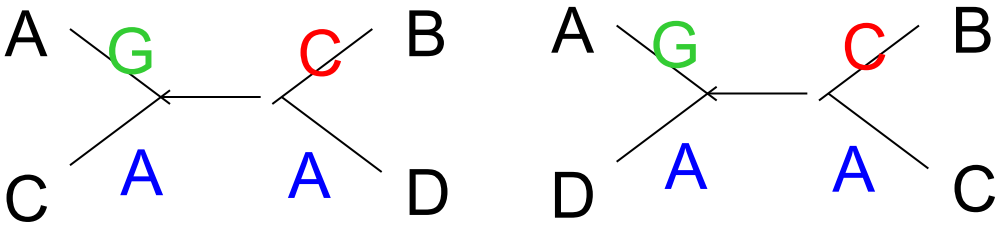


Étude du caractère n°3



Arbre 1

	1	2	3	4	5	6	7	8	9
A	A	A	G	A	G	T	T	C	A
B	A	G	C	C	G	T	T	C	T
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	T

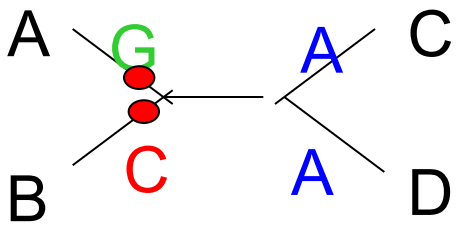


Nb CE= 2

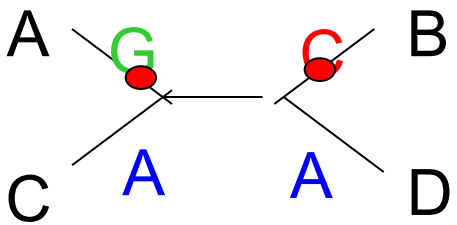
Étude du caractère n°3

	1	2	3	4	5	6	7	8	9
A	A	A	G	A	G	T	T	C	A
B	A	G	C	C	G	T	T	C	T
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	T

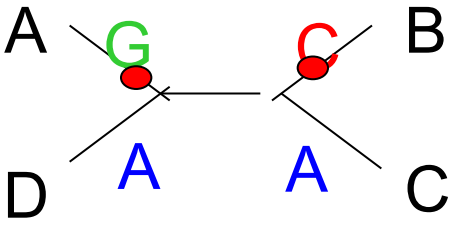
Caractère **variable** mais **non informatif**: tous les scénarios « coûtent » 2 CE.
Caractère ne favorisant aucune topologie par rapport à une autre.



Nb CE= 2



Nb CE= 2

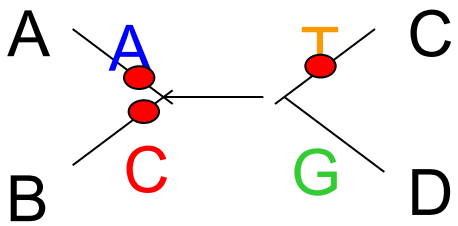


Nb CE= 2

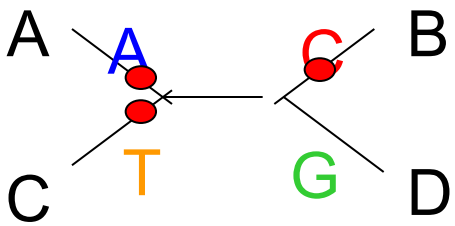
Étude du caractère n°4

	1	2	3	4	5	6	7	8	9
A	A	A	G	A	G	T	T	C	A
B	A	G	C	C	G	T	T	C	T
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	T

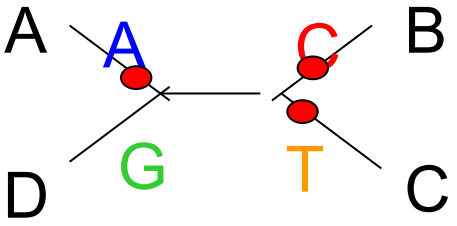
Caractère **variable** mais **non informatif**.
Caractère ne favorisant aucune topologie par rapport à une autre.



Nb CE= 3



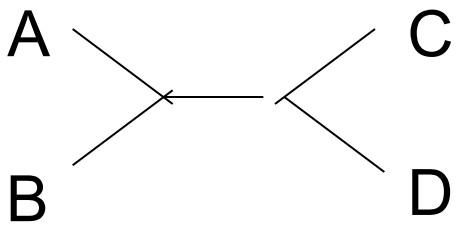
Nb CE= 3



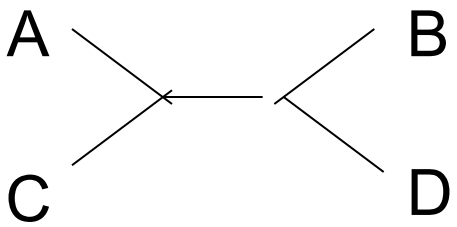
Nb CE= 3

Étude du caractère n°5

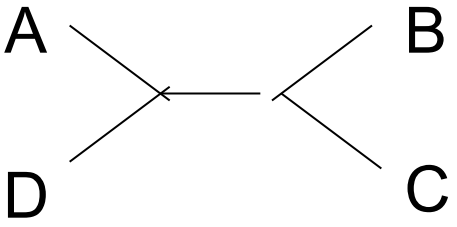
	1	2	3	4	5	6	7	8	9
A	A	A	G	A	G	T	T	C	A
B	A	G	C	C	G	T	T	C	T
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	T



Nb CE= ?



Nb CE= ?

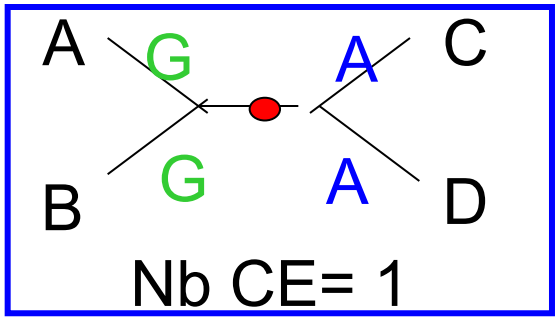


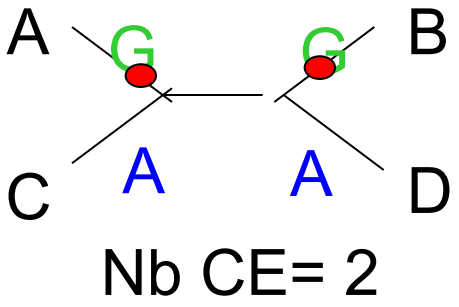
Nb CE= ?

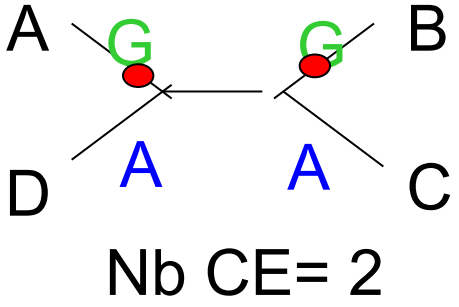
Étude du caractère n°5

	1	2	3	4	5	6	7	8	9
A	A	A	G	A	G	T	T	C	A
B	A	G	C	C	G	T	T	C	T
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	T

Caractère variable et **informatif** (au moins 2 états de caractère sont partagés par au moins 2 OTU).
Caractère favorisant la première topologie par rapport aux deux autres.







Maximum de parcimonie – classification des sites

- Caractère ***invariant***: toutes les OTU possèdent le même état de caractères pour un site donné

- Caractère ***variable***
  - ***Non informatif*** si les états de caractères à ce site ne favorisent aucune topologie parmi l'ensemble des topologies possibles
  - ***Informatif*** si les états de caractères à ce site favorise une (ou plusieurs) topologie(s) parmi l'ensemble des topologies possibles

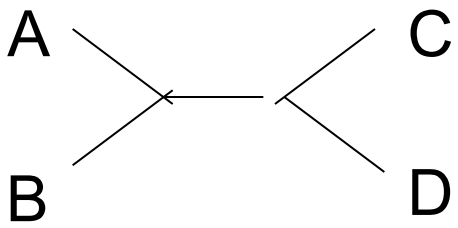
Maximum de parcimonie - Méthode

On compte ensuite le nombre total de mutations nécessaires pour chaque topologie.

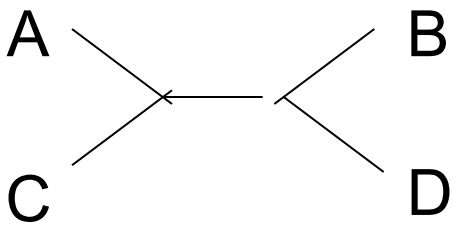
Bilan:
T1 = 0+1+2+3+1+0+1+0+2=10
T2 = 0+1+2+3+2+0+2+0+1=11
T3 = 0+1+2+3+2+0+2+0+2=12

	1	2	3	4	5	6	7	8	9
A	A	A	G	A	G	T	T	C	A
B	A	G	C	C	G	T	T	C	T
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	T

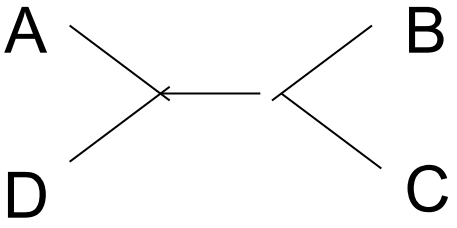
L'arbre le plus parcimonieux = arbre 1



Nb CE= 10



Nb CE= 11



Nb CE= 12

Maximum de parcimonie - désavantages

- Le nombre d'arbres possibles augmente rapidement avec le nombre d'UTOs (séquences).
  - Dans les exemples qui précèdent nous avons analysé 4 séquences.
  - Pour analyser ne fût-ce que 20 séquences, on se trouve confronté à un nombre astronomique de possibilités.
- La parcimonie repose intrinsèquement sur une hypothèse de l'horloge moléculaire => suppose que toutes les branches ont évolué avec la même vitesse.
- Cette méthode fonctionne seulement avec les séquences très conservées.

n	Nb arbres enracinés	Nb arbres non-enracinés
2	1	1
3	3	1
4	15	3
5	105	15
6	945	105
7	10,395	945
8	135,135	10,395
9	2,027,025	135,135
10	3.45E+07	2,027,025
11	6.55E+08	3.45E+07
12	1.37E+10	6.55E+08
13	3.16E+11	1.37E+10
14	7.91E+12	3.16E+11
15	2.13E+14	7.91E+12
16	6.19E+15	2.13E+14
17	1.92E+17	6.19E+15
18	6.33E+18	1.92E+17
19	2.22E+20	6.33E+18
20	8.20E+21	2.22E+20
	$N_R = \frac{(2n-3)!}{2^{n-2}(n-2)!}$	$N_U = \frac{(2n-5)!}{2^{n-3}(n-3)!}$

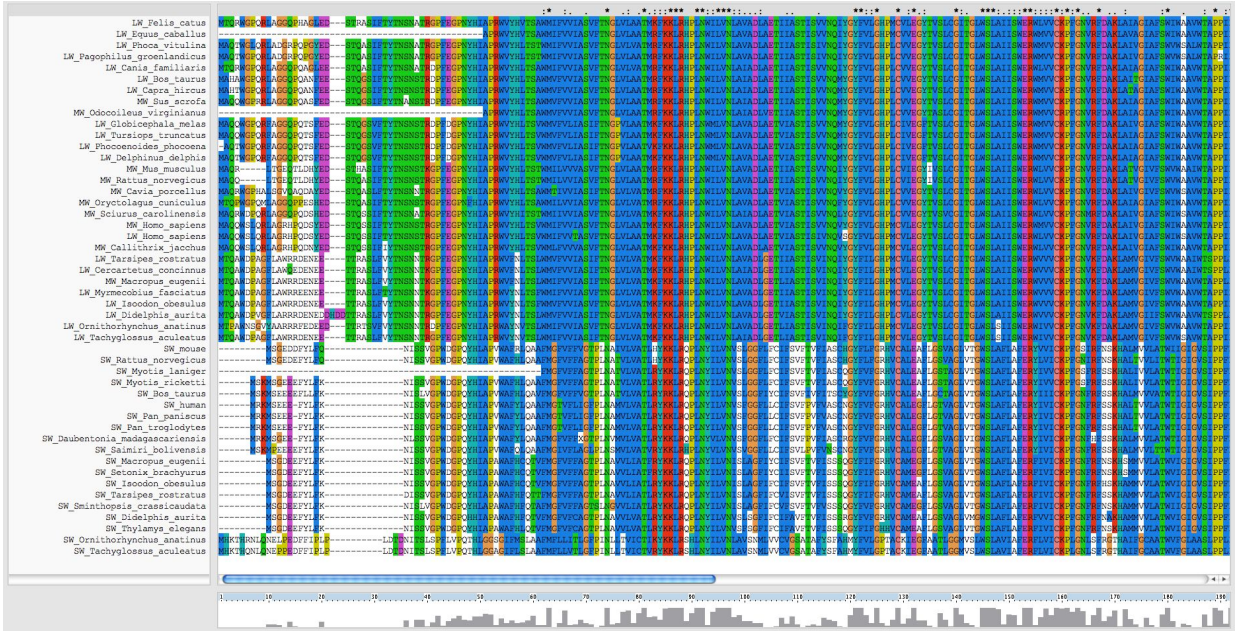
- Résumé des méthodes d'inférence d'arbre implémentées dans PHYLIP.
- Note: le temps de calcul augmente drastiquement quand on passe de méthodes de voisinage (NJ, UPGMA: temps quadratique) aux méthodes de kitsch ou fitch (puissance 4 de la longueur des séquences).

Phylip program	method	rooted tree	time	accuracy	remarks
fitch	Fitch-Margoliah	no	O(n^4)	higher	loss of accuracy when the tree contains long branches
kitsch	Fitch-Margoliah	yes	O(n^4)	higher	
neighbor	neighbour-joining	no	O(n^2)	lower	suitable when rate of evolution varies among branches
neighbor	UPGMA	yes	O(n^2)	lower	assumes constant rate of evolution along the banches

Evaluation de la robustesse de l'inférence: le "bootstrap"

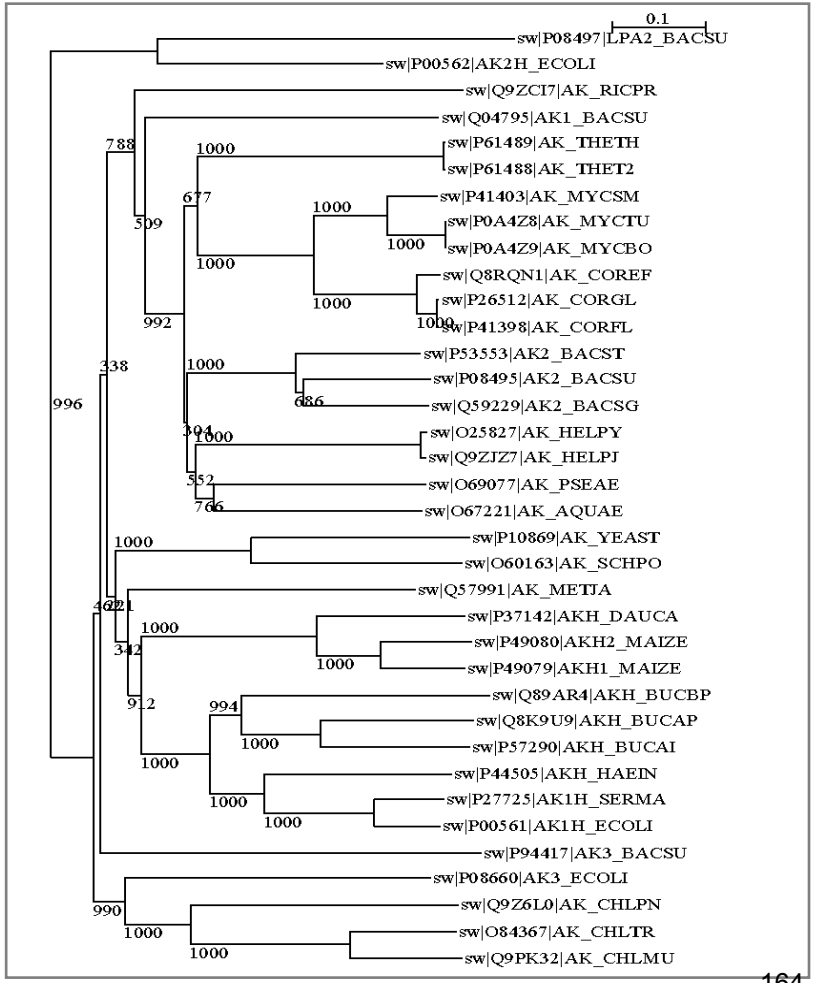
Quelle est la fiabilité d'un arbre inféré ?

- On se base sur les colonnes d'un alignement multiple pour inférer un arbre phylogénétique cohérent avec les différences entre groupes de séquences.
- Cependant, selon les colonnes choisies on peut observer des variations de séquence qui touchent des sous-groupes différents.
- Comment évaluer la robustesse de l'inférence par rapport aux particularité des échantillons disponibles ?



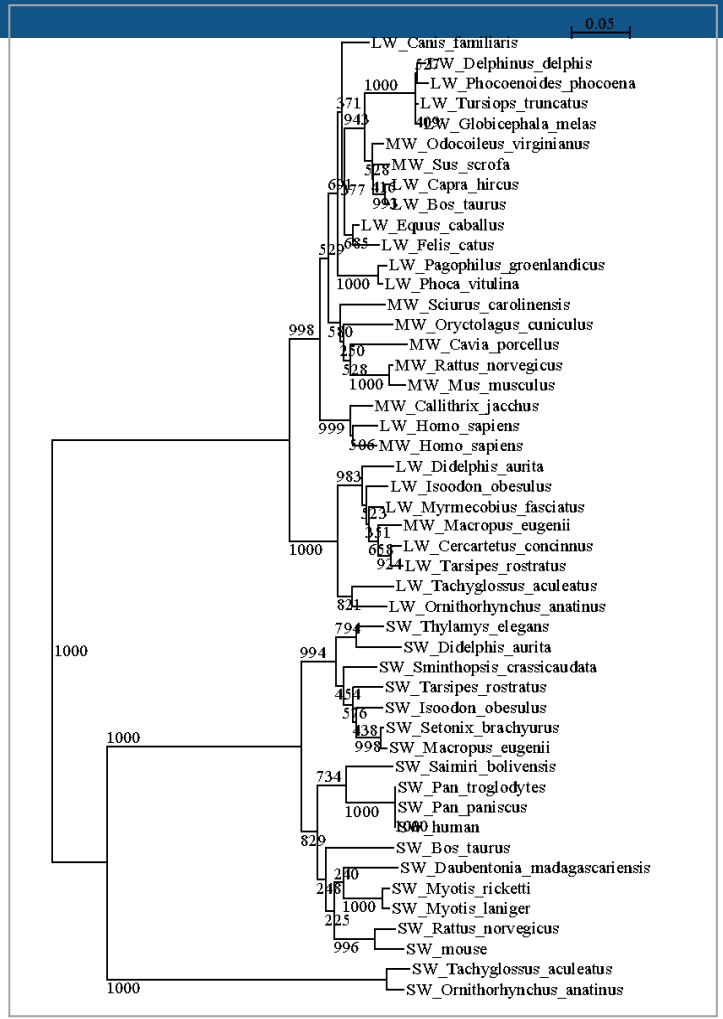
Bootstrapping

- Dans certains cas, les données ne permettent pas d'inférer la phylogénie de façon fiable.
- Pour évaluer la fiabilité de l'inférence, on peut appliquer la méthode du ***bootstrapping***.
  - Etant donné un alignement de N séquences et M colonnes, on effectue une sélection aléatoire de M colonnes ***avec remise***. Certaines colonnes sont donc tirées plusieurs fois, et d'autres aucune fois.
  - On calcule un arbre avec les colonnes échantillonnées.
  - On répète l'opération un bon nombre de fois (1000), et on compte le nombre de fois où chaque branchement de l'arbre original se reproduit.

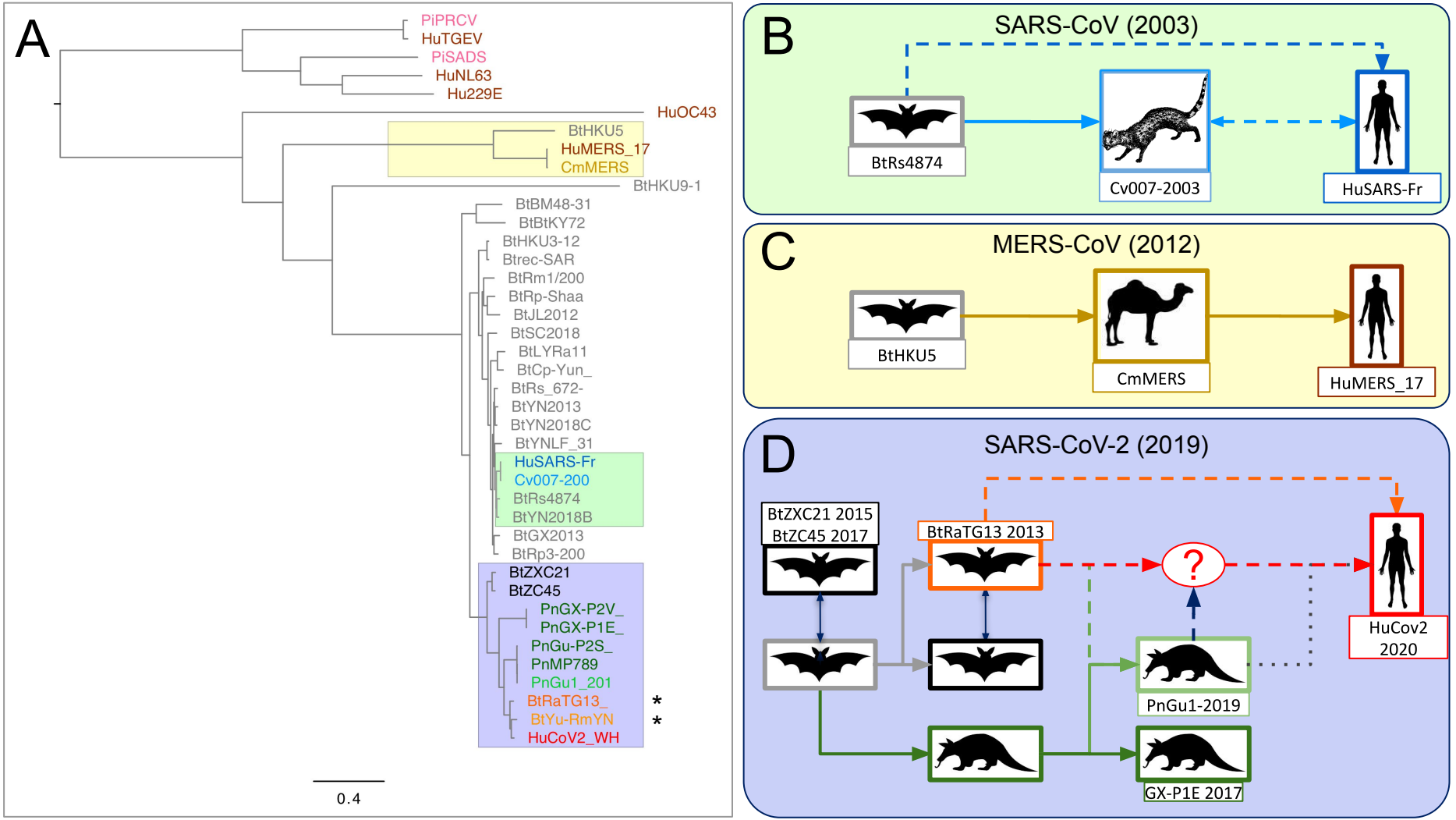


Bootstrapping

- Le phylogramme permet d'identifier les relations entre longueurs des branches et valeurs de bootstrap.
- Les valeurs de bootstrap sont cependant moins faciles à lire que sur un cladogramme (où toutes les branches ont la même longueur).

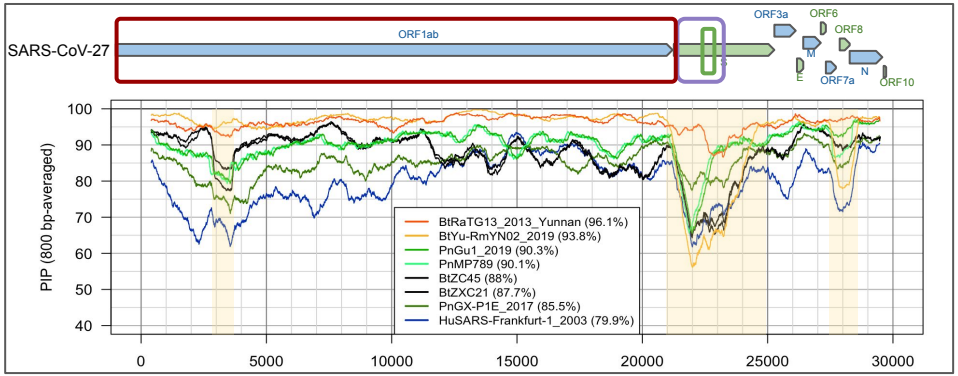


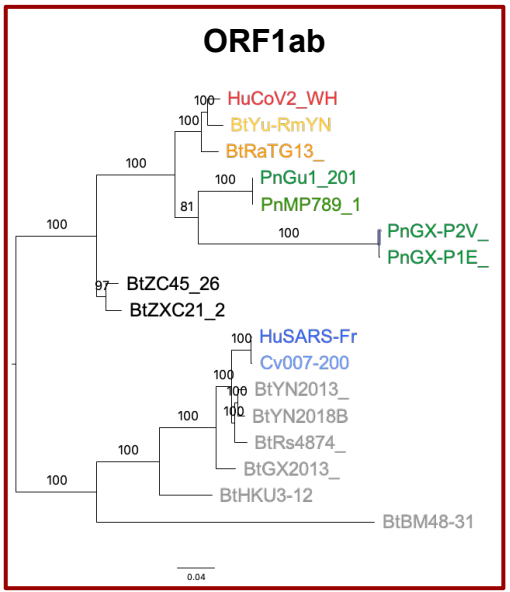
Phylogénie des coronavirus

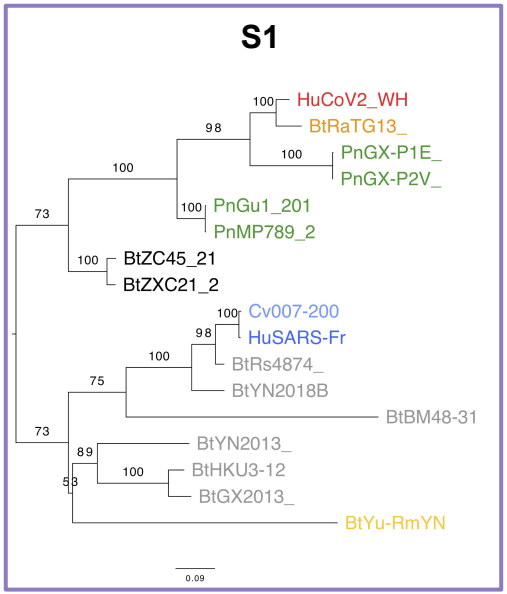

A
PiPRCV
HuTGEV
PiSADS
HuNL63
Hu229E
HuOC43
BtHKU5
HuMERS_17
CmMERS
BtHKU9-1
BtBM48-31
BtBtKY72
BtHKU3-12
Btrec-SAR
BtRm1/200
BtRp-Shaa
BtJL2012
BtSC2018
BtLYRa11
BtCp-Yun_
BtRs_672-
BtYN2013
BtYN2018C
BtYNLF_31
HuSARS-Fr
Cv007-200
BtRs4874
BtYN2018B
BtGX2013
BtRp3-200
BtZXC21
BtZC45
PnGX-P2V_
PnGX-P1E_
PnGu-P2S_
PnMP789
PnGu1_201
BtRaTG13_
BtYu-RmYN
HuCoV2_WH
*
*
0.4
B
SARS-CoV (2003)
BtRs4874
Cv007-2003
HuSARS-Fr
C
MERS-CoV (2012)
BtHKU5
CmMERS
HuMERS_17
D
SARS-CoV-2 (2019)
BtZXC21 2015
BtZC45 2017
BtRaTG13 2013
?
HuCov2
2020
PnGu1-2019
GX-P1E 2017

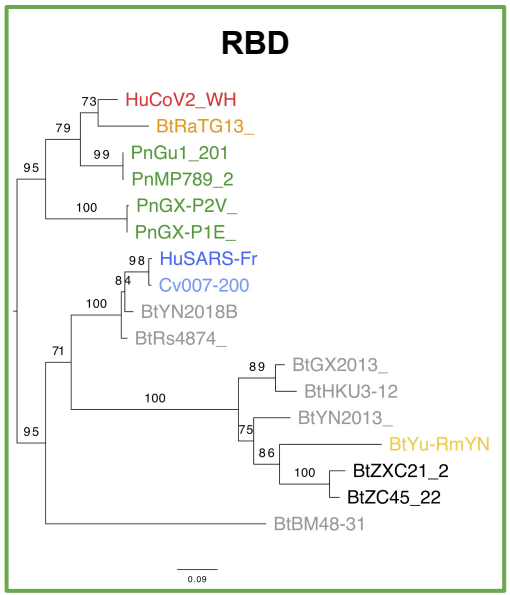
Comparaison entre coronavirus - gène S

Profils de pourcentages de positions identiques (PPI) entre différents génomes de coronavirus (d'humain, de chauve-souris et de pangolin) et SARS-CoV-2 (la référence qui correspond à 100% sur toute la largeur).



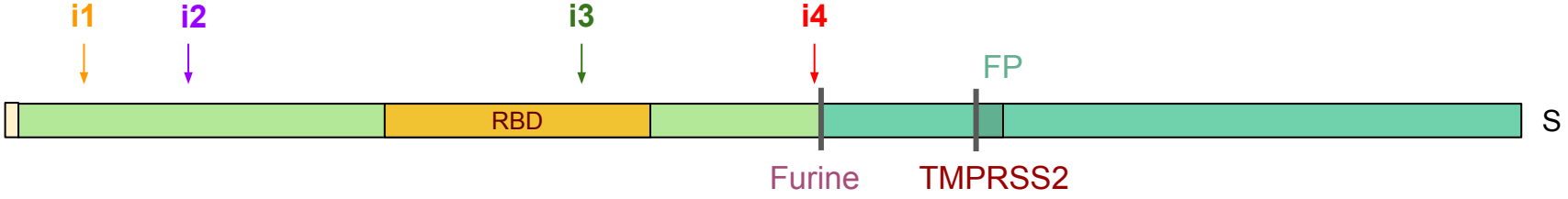


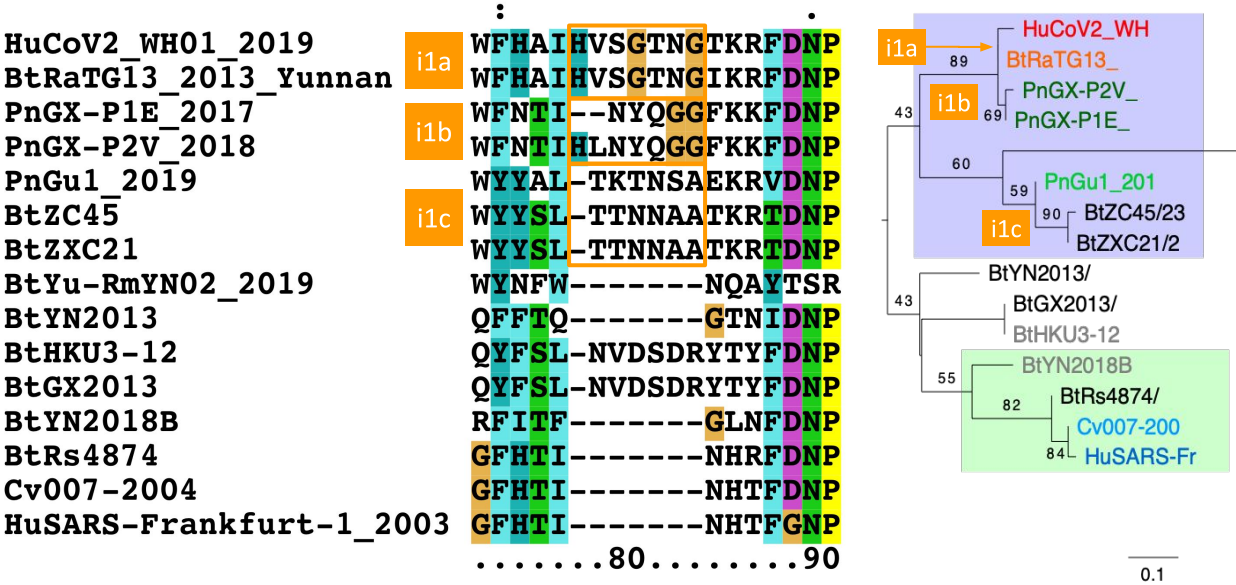




Sallard, E., Halloy, J., Casane, D., van Helden, J., Decroly, E. Retrouver les origines du SARS-COV-2 dans les phylogénies de coronavirus. Médecine/Sciences 36, Sept-Août 2020.

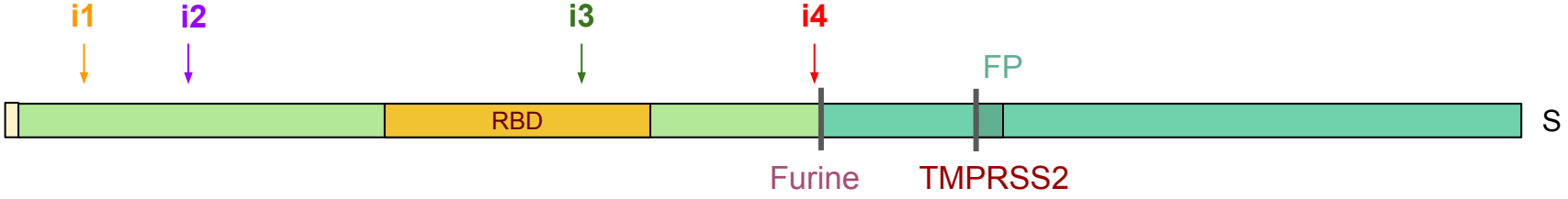
Insertion 1

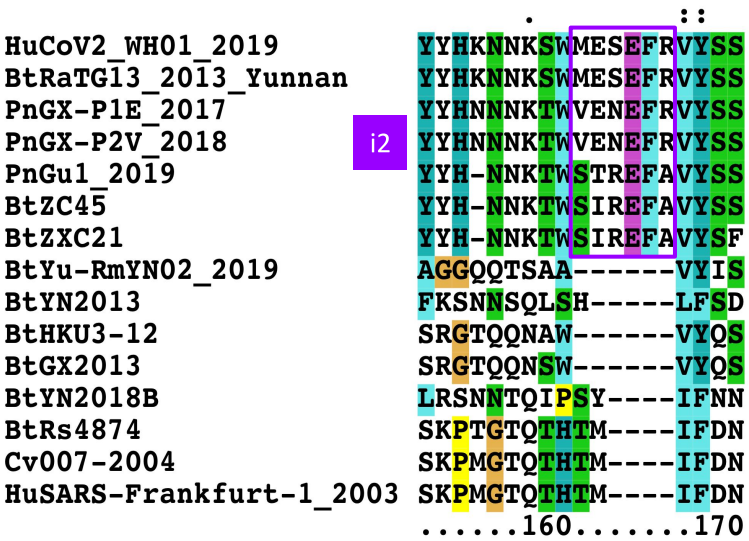


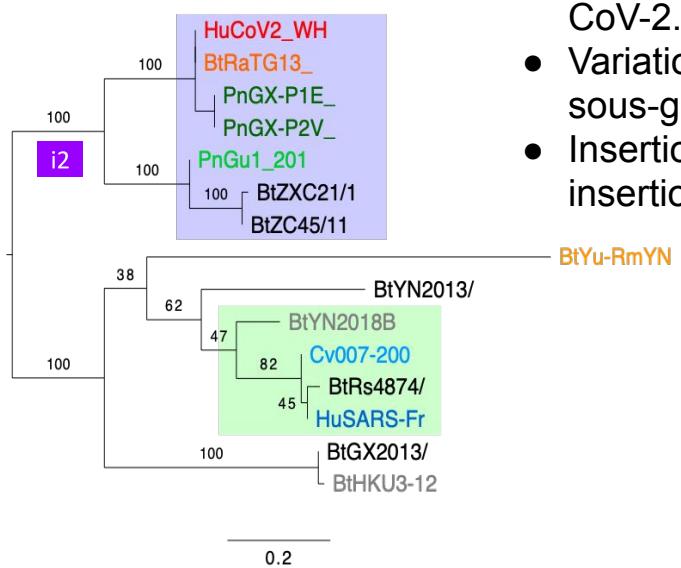


- Insertion retrouvée dans tous les génomes de la branche CoV-2.
- Variations selon les sous-groupes.
- Insertions indépendantes ou insertion suivie de mutations ?

Insertion 2

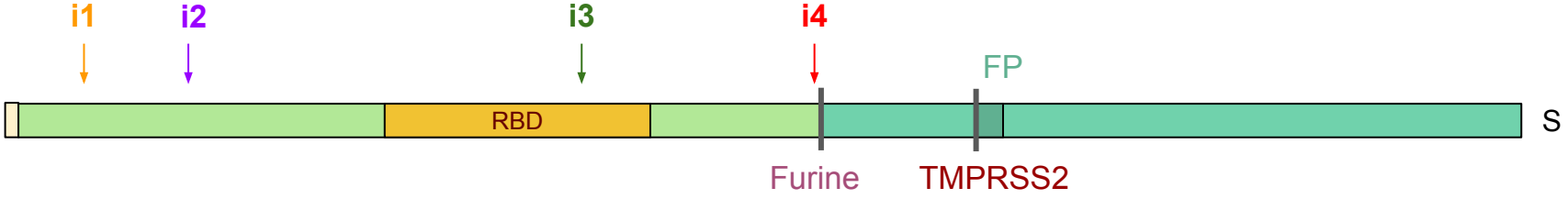


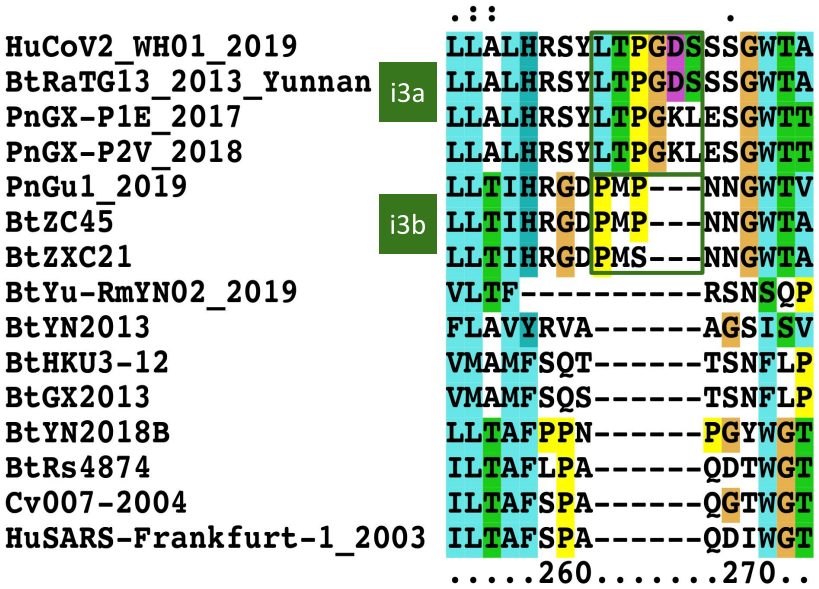


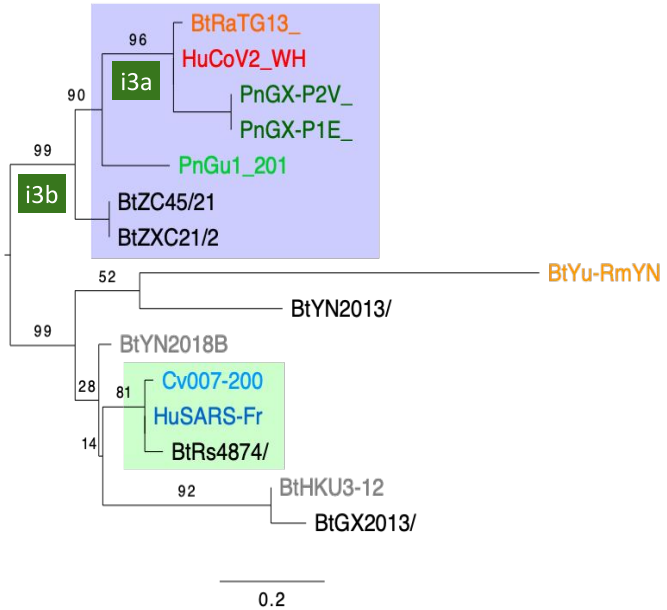


- Insertion retrouvée dans tous les génomes de la branche CoV-2.
- Variations selon les sous-groupes.
- Insertions indépendantes ou insertion suivie de mutations ?

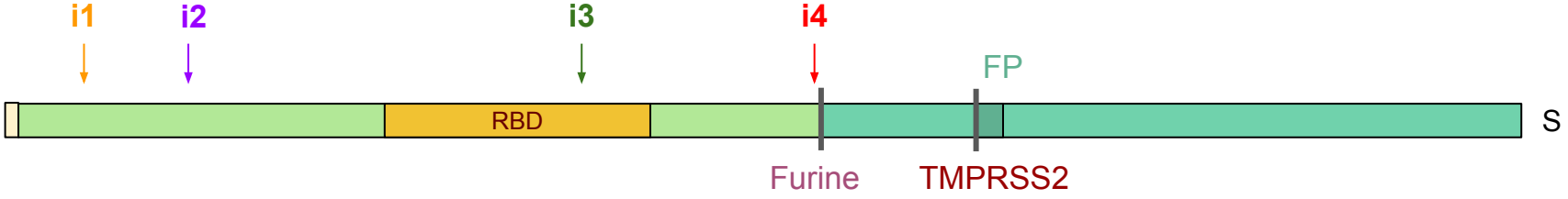
Insertion 3


i1
i2
i3
i4
FP
RBD
S
Furine
TMPRSS2

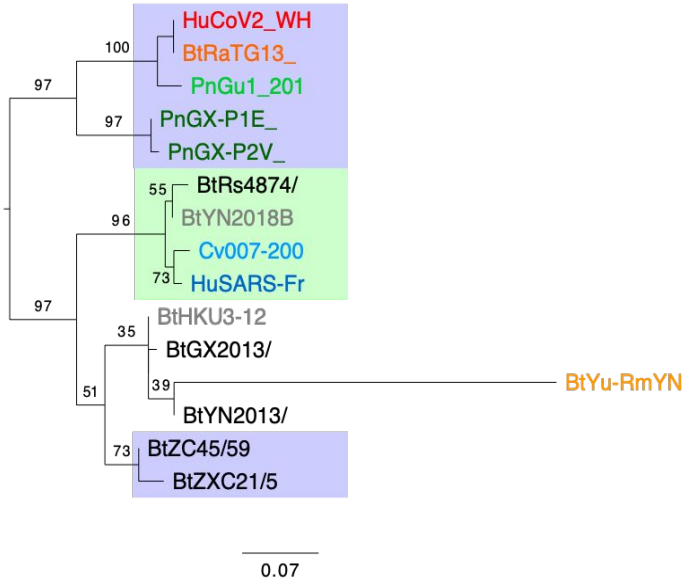

.:: .
HuCoV2_WH01_2019 LLALHRSYLTPGDSSSGWTA
BtRaTG13_2013_Yunnan LLALHRSYLTPGDSSSGWTA
PnGX-P1E_2017 LLALHRSYLTPGKLESGWTT
PnGX-P2V_2018 LLALHRSYLTPGKLESGWTT
PnGu1_2019 LLTIHRGDPMP---NNGWTV
BtZC45 LLTIHRGDPMP---NNGWTA
BtZXC21 LLTIHRGDPMS---NNGWTA
BtYu-RmYN02_2019 VLTF----------RSNSQP
BtYN2013 FLAVYRVA------AGSISV
BtHKU3-12 VMAMFSQT------TSNFLP
BtGX2013 VMAMFSQS------TSNFLP
BtYN2018B LLTAFPPN------PGYWGT
BtRs4874 ILTAFLPA------QDTWGT
Cv007-2004 ILTAFSPA------QGTWGT
HuSARS-Frankfurt-1_2003 ILTAFSPA------QDIWGT
.....260.....270..
i3a
i3b

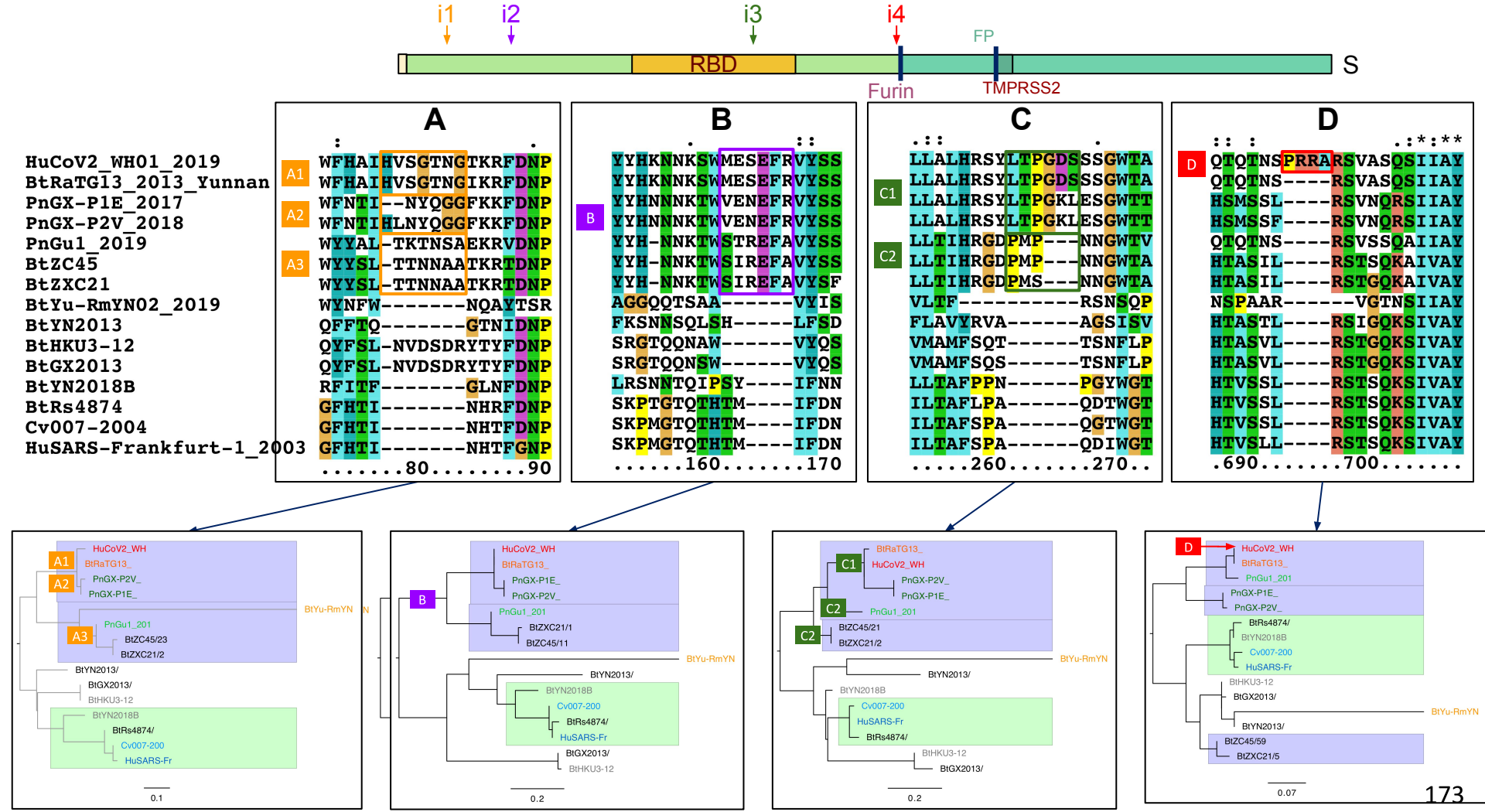

BtRaTG13_
HuCoV2_WH
PnGX-P2V_
PnGX-P1E_
PnGu1_201
BtZC45/21
BtZXC21/2
i3a
i3b
BtYu-RmYN
BtYN2013/
BtYN2018B
Cv007-200
HuSARS-Fr
BtRs4874/
BtHKU3-12
BtGX2013/
96
90
99
52
99
28
81
14
92
0.2

Insertion 4

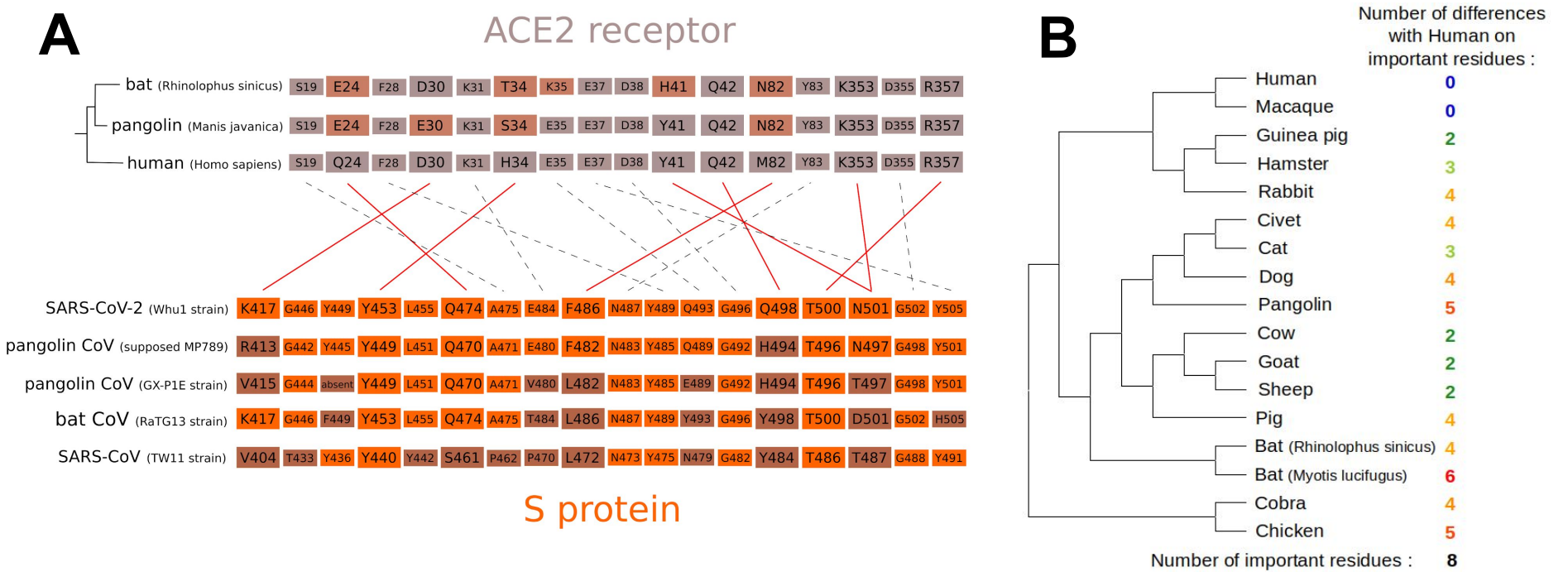



```
                                    :: :                 .:*:**
HuCoV2_WH01_2019            i4      QTQTNSPRRARSVASQSIIAY
BtRaTG13_2013_Yunnan                QTQTNS----RSVASQSIIAY
PnGX-P1E_2017                       HSMSSL----RSVNQRSIIAY
PnGX-P2V_2018                       HSMSSF----RSVNQRSIIAY
PnGu1_2019                          QTQTNS----RSVSSQAIIAY
BtZC45                              HTASIL----RSTSQKAIVAY
BtZXC21                             HTASIL----RSTGQKAIVAY
BtYu-RmYN02_2019                    NSPAAR------VGTNSIIAY
BtYN2013                            HTASTL----RSIGQKSIVAY
BtHKU3-12                           HTASVL----RSTGQKSIVAY
BtGX2013                            HTASVL----RSTGQKSIVAY
BtYN2018B                           HTVSSL----RSTSQKSIVAY
BtRs4874                            HTVSSL----RSTSQKSIVAY
Cv007-2004                          HTVSSL----RSTSQKSIVAY
HuSARS-Frankfurt-1_2003             HTVSLL----RSTSQKSIVAY
                                    .690.......700.......
```





Sallard, E., Halloy, J., Casane, D., van Helden, J., Decroly, E. Retrouver les origines du SARS-COV-2 dans les phylogénies de coronavirus. Médecine/Sciences 36, Sept-Août 2020.



Sallard, E., Halloy, J., Casane, D., van Helden, J., Decroly, E. Retrouver les origines du SARS-COV-2 dans les phylogénies de coronavirus. Médecine/Sciences 36, Sept-Août 2020.

Que peut-on conclure à ce stade ? Discussion en classe (virtuelle)