

Enquête bioinformatique sur les origines de SARS-CoV-2

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

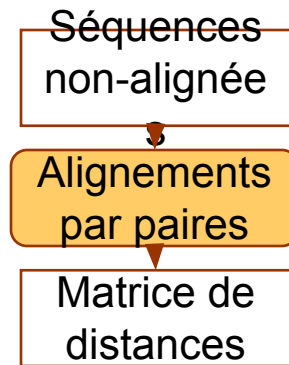
*Cours donné par **Jacques van Helden**, **Emese Meglécz** et **Gabriel Neve***

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

Alignement de séquences multiples

- 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.

- 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: $d_{i,i} = 0$



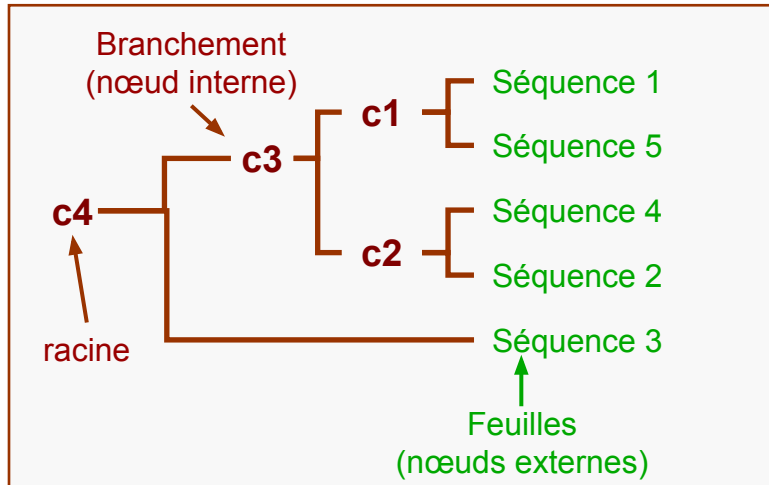
	seq 1	seq 2	...	seq n
seq 1	d1,1	d1,2	...	d1,n
seq 2	d2,1	d2,2	...	d2,n
...
seq n	dn,1	dn,2	...	dn,n

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

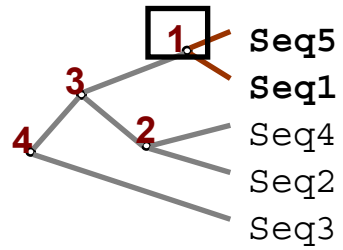
Arbre



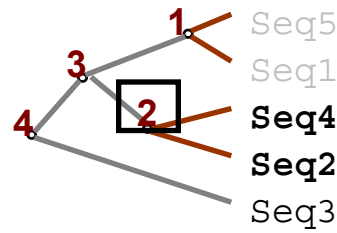
- Le clustering hiérarchique est une méthode de clustering agrégative.
 - Prend une matrice de distance en entrée
 - Regroupe progressivement les objets en allant des plus proches aux plus distants.
- Il existe plusieurs possibilités pour établir une règle d'agglomération, qui définit la distance entre 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.
- Algorithmes
 - 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 – 3^{ème} étape: alignement multiple

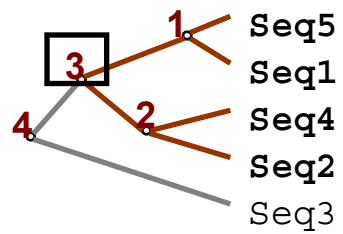
- On construit un alignement multiple en incorporant progressivement les séquences selon leur ordre de branchement dans l'arbre guide, en remontant des plus proches aux plus éloignées.



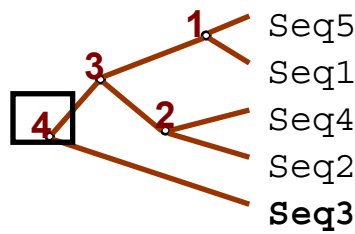
Seq5 GATTGTAGTA
Seq1 GATGTAGTA



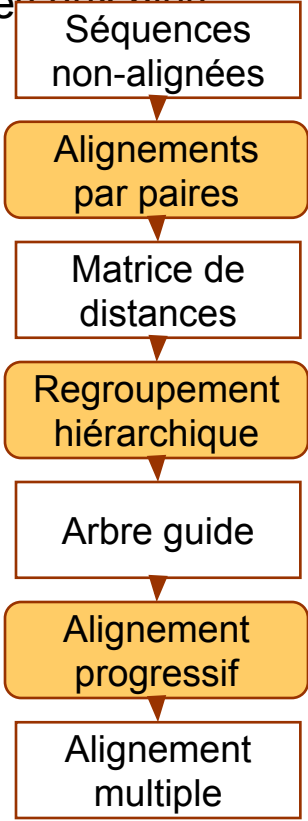
Seq5 GATTGTAGTA
Seq1 GATGTAGTA
Seq4 GATTGTTTC--GTA
Seq2 GATTGTTTCGGTA



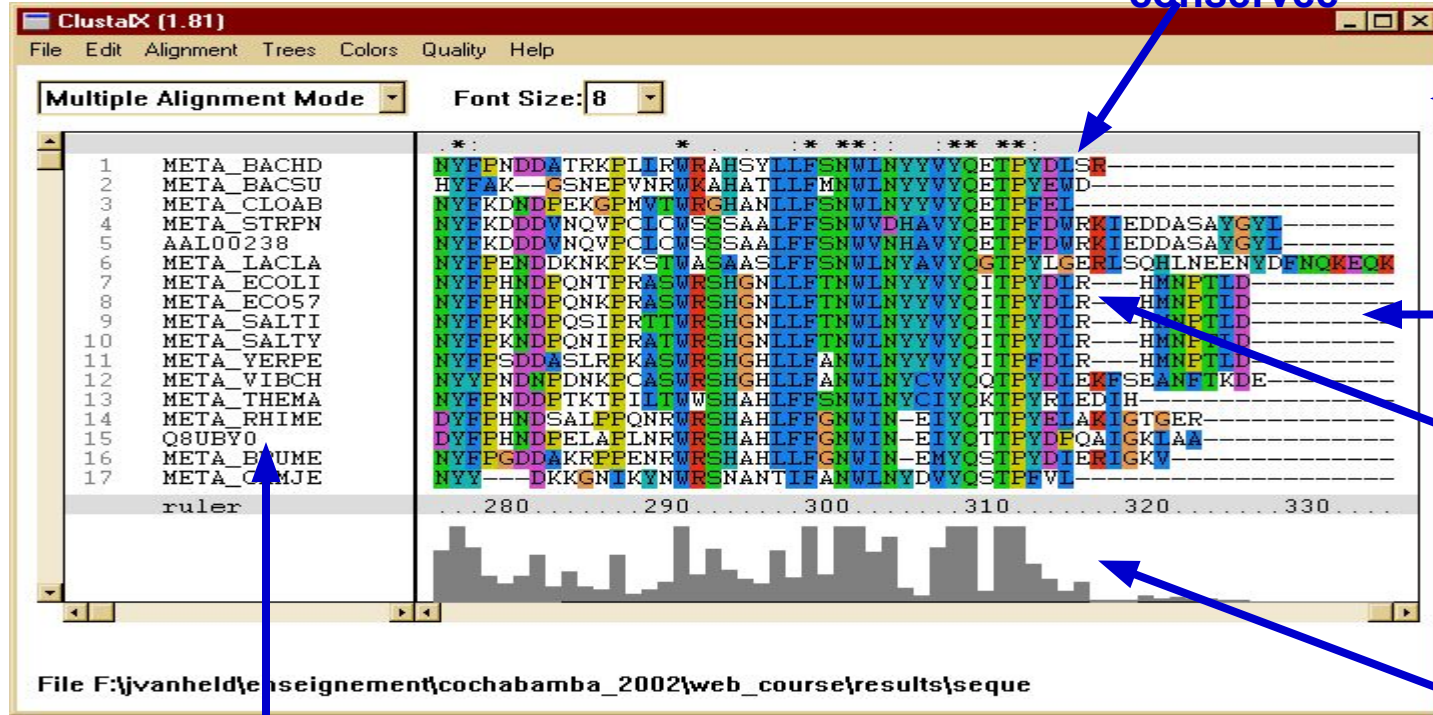
Seq5 GATTGTA---GTA
Seq1 GATGGTA---GTA
Seq4 GATTGTTTC--GTA
Seq2 GATTGTTTCGGTA
Seq3



Seq5 GATTGTA-----GTA
Seq1 GATGGTA-----GTA
Seq4 GATTGTTTC-----GTA
Seq2 GATTGTTTCGG--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.

	420	430	440	450	460	470	480
Pig_SADS_AYV41569.1_ref/1-1130	N S V V T V R L C R	--- W W ---	Q F M S F N S T S H A A D A	--- G P T N A	F	--- E	C L
Pig_PRCV_AKV62755.1_ref/1-1232	S G K L V I K Q F	--- L V N C L W P V P S F E E A A S	--- T F	--- C F E G A D F	--- H C N C A V L N N T V D V I R F I		
Human_TGVEV_CA891145.1_ref/1-1447	S G K L V I K Q F	--- L V N C L W P V P S F E E A A S	--- T F	--- C F E G A D F	--- Q D C N C A V L N N T V D V I R F I		
Human_229E_AAG48592.1_ref/1-1173	S F Q P L L L N C L	--- W ---	S V S G L R F T T G F	--- Y F Y N G	T G R G	--- C K C F S S D V L S D V I R F I	
Human_MJ63_AA558177.1_ref/1-1356	L Y Q P L R L T C L	--- W ---	P V P G L K S S T G F	--- Y F Y N A	T G S D V Y	--- C N G Y Q H N S V D V I R F I	
Human_h-HKU1_N23_ABD96197.1_ref/1-1356	N G V L E I T A C Q	--- Y T M C E Y P H T I C K S K G S R N E	--- S W	--- H F D K S	E	--- P L C L	F K
Human_HK04-02_ABN19366.1_ref/1-1361	Q G L L E M S V C Q	--- N M M C E Y P H T I C H P K L G N H F K E L W	---	H D T G V	V	--- S	C L
Human_OC43_ADX10763.1_ref/1-1358	Q G L L E M S V C Q	--- N M M C E Y P H T I C H P N L G N H F K E L W	---	H D T G V	V	--- S	C L
Human_h-MERS_KOR/CNMH_SNU/172_06_2015_ALK80311.1_ref/1-1353	T L V L L P D G C G T L L R A	--- F Y C I L E P R S G N H C P A G N S Y T	---	S F	--- A	Y H T P	A
Camel_MERS_AHE78097.1_ref/1-1353	T L V L L P D G C G T L L R A	--- F Y C I L E P R S G N H C P A G N S Y T	---	S F	--- A	Y H T P	A
Rat_BM48-31_ADK66841.1/1-1259	G T H I V I D Y C N	--- F N F C A D P M A F A W N S	--- G Q F Y K	--- T W	---	I T S A	A
Rat_RoKY22_APO40579/1-1257	G T H I V I D Y C N	--- F Y F C Q D P M A F A W N S	--- G S H F K	--- S W	---	V F L N A	I
Pangolin_PCoV_GX-PSE_QIA48641.1_ref/1-1267	A T N V V I K V C E	--- F Q F C T D P F L G Y Y H N N N K	---	T W V E N E F R Y Y S S A	---	N	---
Pangolin_PCoV_GX-P4L_QIA48614/1-1267	A T N V V I K V C E	--- F Q F C T D P F L G Y Y H N N N K	---	T W V E N E F R Y Y S S A	---	N	---
Pangolin_PCoV_GX-P2V_QHQ54048/1-1269	A T N V V I K V C E	--- F Q F C T D P F L G Y Y H N N N K	---	T W V E N E F R Y Y S S A	---	N	---
Pangolin_PCoV_GX-P1E_QIA48623.1_ref/1-1265	A T N V V I K V C E	--- F Q F C T D P F L G Y Y H N N N K	---	T W V E N E F R Y Y S S A	---	N	---
Pangolin_PCoV_GX-P5L_QIA48632/1-1267	A T N V V I K V C E	--- F Q F C T D P F L G Y Y H N N N K	---	T W V E N E F R Y Y S S A	---	N	---
Rat_RaTG13_QHR63300.ref/1-1269	A T N V V I K V C E	--- F Q F C D P F L G Y Y H N N N K	---	S W M E S E F R Y Y S S A	---	N	---
Human_SARS-CoV-2_BetaCoV/Whan/IPBCAMS-WH-01/2019_QHU36824/1-1273	A T N V V I K V C E	--- F Q F C D P F L G Y Y H N N N K	---	S W M E S E F R Y Y S S A	---	N	---
Human_SARS-CoV-2_WIV04_QHR63260/1-1273	A T N V V I K V C E	--- F Q F C D P F L G Y Y H N N N K	---	S W M E S E F R Y Y S S A	---	N	---
Human_SARS-CoV-2_Whan-hu-1_YP_009724390.ref/1-1273	A T N V V I K V C E	--- F Q F C D P F L G Y Y H N N N K	---	S W M E S E F R Y Y S S A	---	N	---
Rat_bat-SI-CoVZC45_AVP7803.1_ref/1-1246	A T N V I I K V C N	--- F D F Y D P F L G Y Y H N N K	---	H K N N K	---	S W S I R E F A N Y S S A	---
Rat_bat-SI-CoVZMC12_AVP78042.ref/1-1245	A T N V I I K V C N	--- F D F Y D P F L G Y Y H N N K	---	H K N N K	---	S W S I R E F A N Y S S A	---
Rat_LYba11_AHK37558/1-1259	S T N V L V R V C N	--- F E L C D N P F F V W L K S N N T I P	---	S Y	---	F N N A	F
Rat_Ro4084_ATO98132.1/1-1256	S T N V L V R V C N	--- F E L C D N P F F V W L K S N N T I P	---	S Y	---	F N N A	F
Rat_RoSHC014_AG248806.1/1-1256	S T N V L V R V C N	--- F E L C D N P F F V W L K S N N T I P	---	S Y	---	F N N A	F
Rat_WIV1_AG248828.1/1-1256	S T N V L V R V C N	--- F E L C D N P F F V W L K S N N T I P	---	S Y	---	F N N A	F
Rat_Ro3367_AG248818/1-1256	S T N V L V R V C N	--- F E L C D N P F F V W L K S N N T I P	---	S Y	---	F N N A	F
Rat_Ro7927_ATO98218/1-1256	S T N V L V R V C N	--- F E L C D N P F F V W L K S N N T I P	---	S Y	---	F N N A	F
Rat_RoRo-BetaCoV/YN2010R_QDF43825/1-1256	S T N V V I R V C N	--- F E L C D N P F F A M S K P T G T Q T H	---	T M	---	F D N A	F
Rat_Ro4231_ATO98157/1-1255	S T N V V I R V C N	--- F E L C D N P F F A M S K P T G T Q T H	---	T M	---	F D N A	F
Rat_S23_ATO98205.1/1-1255	S T N V V I R V C N	--- F E L C D N P F F A M S K P T G T Q T H	---	T M	---	F D N A	F
Rat_WIV16_AIK02457.1/1-1255	S T N V V I R V C N	--- F E L C D N P F F A M S K P T G T Q T H	---	T M	---	F D N A	F
Civet_SARS-CoV_007/2004_AAU04646/1-1255	S T N V V I R V C N	--- F E L C D N P F F V W S K P M G T Q T H	---	T M	---	F D N A	F
Civet_S23_P59594.1/1-1255	S T N V V I R V C N	--- F E L C D N P F F A M S K P M G T Q T H	---	T M	---	F D N A	F
Human_SARS-CoV_Tor2/FP1-10895_APR58742/1-1255	S T N V V I R V C N	--- F E L C D N P F F A M S K P M G T Q T H	---	T M	---	F D N A	F
Human_SARS-CoV_TW11_AA87512/1-1255	S T N V V I R V C N	--- F E L C D N P F F A M S K P M G T Q T H	---	T M	---	F D N A	F
Human_SARS-CoV_Frankfurt-LAAP33697.1_ref/1-1255	S T N V V I R V C N	--- F E L C D N P F F A M S K P M G T Q T H	---	T M	---	F D N A	F
Ra1.2012_AIA62277.1_ref/1-1236	G S A I T I E V C Y	--- F Q F C D N P A F I I R D	---	G A Q I N	---	T A	I
Rat_TMC15_ANA96027/1-1236	G S A I T I E V C Y	--- F Q F C D N P A F I I S G	---	G A Q I N	---	T A	I
Rat_RoRo-BetaCoV/YN2013_AIA62330/1-1233	S T N L I V R V C N	--- F E L C K V P L F V W F K S N N S Q L	---	S H	---	L F S D S	F
Rat_RoRo-BetaCoV/SC2018_QDF43815/1-1256	S T H I L V K V C N	--- F I I C K E P M F T V S Q	---	N R H F K	---	S W	Y Y Q D A
Rat_Ro/Shaanx2011_AGC74163/1-1240	S T H I L V K V C N	--- F V L C T E P M F T V S R	---	N Q H Y K	---	S W	V Y Q H A
Rat_RoRo-BetaCoV/HuB2013_AIA62310/1-1241	S T H I L V K V C N	--- F V L C T E P M F T V S R	---	N Q H Y K	---	S W	V Y Q H A
Rat_YNF1_31C_AK219076/1-1241	S T H I I I R V C Y	--- F N L C K D P M Y T V S A	---	G T Q I S	---	S W	V Y Q N A
Rat_RoRo-BetaCoV/HeB2013_AIA62290/1-1241	S T H I I I R V C Y	--- F N L C K D P M Y T V S A	---	G T Q V S	---	S W	V Y Q S A
Rat_RY1/2004_ABD75323/1-1241	S T H I I I R V C Y	--- F N L C K D P M Y T V S A	---	G T Q K S	---	S W	V Y Q S A
Rat_Bat-CoV-273/2005_ABG47060/1-1241	S T H I I I R V C Y	--- F N L C K D P M Y T V S A	---	G T Q K S	---	S W	V Y Q S A
Rat_Co/Yunnan2011_AGC74176/1-1241	S T H I I I R V C N	--- F N L C K E P M Y T V S R	---	G V H F S	---	S W	V Y Q S A
Rat_Ro_672/2006_ACU31032/1-1241	S T H I I I R V C Y	--- F N L C K E P M Y A I S N	---	E Q H Y K	---	S W	V Y Q N A
Rat_Ro4081_ATO98120/1-1241	S T H I I I R V C Y	--- F N L C K E P M Y A I S N	---	E Q H Y K	---	S W	V Y Q N A
Rat-recombinant_Bat-SRBD_ACJ60703.1_ref/1-1259	S T H I I I R V C N	--- F N L C K E P M Y T V S R	---	G T Q Q N	---	A W	V Y Q S A
Rat_Bat-SARS-CoV-Rm1/2004_ABD75332/1-1241	S T H I I I R V C N	--- F N L C K E P M Y T V S K	---	G T Q Q S	---	S W	V Y Q S A
Rat_Ro3/2004_AA267052/1-1241	S T H I I I R V C N	--- F N L C K E P M Y T V S R	---	G A Q Q S	---	S W	V Y Q S A
Rat_RoRo-BetaCoV/YN2018C_QDF43830/1-1241	S T H I I I R V C N	--- F N L C K E P M Y T V S R	---	G T Q Q S	---	S W	V Y Q S A
Rat_Ro4247_ATO98181.1/1-1242	S T H I I I R V C N	--- F N L C K E P M Y T V S R	---	G T Q Q S	---	S W	V Y Q S A
Rat_Ro4237_ATO98169/1-1241	S T H I I I R V C N	--- F N L C K E P M Y T V S R	---	G T Q Q S	---	S W	V Y Q S A
Rat_RoRo-BetaCoV/GQ2013_AIA62320/1-1242	S T H I I I R V C N	--- F N L C K E P M Y T V S R	---	G T Q Q N	---	S W	V Y Q S A
Rat_HKU3-8_ADE34766/1-1242	S T H I I I R V C N	--- F N L C K E P M Y T V S M	---	G T Q Q N	---	S W	V Y Q S A
Rat_Longquan-14Q_AID16716/1-1242	S T H I I I R V C N	--- F N L C R E P M Y T V S R	---	G T Q Q N	---	S W	V Y Q S A
Rat_HKU3-12_ADE34812/1-1242	S T H I I I R V C N	--- F N L C K E P M Y T V S R	---	G T Q Q N	---	A W	V Y Q S A
Rat_HKU3-6_ADE34744/1-1242	S T H I I I R V C N	--- F N L C K E P M Y T V S R	---	G T Q Q N	---	A W	V Y Q S A

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 .

Pig_SADS_AVM41569.1_ref/1-1130	Q L D G L Q W R V Y T T	---Q F N S P V N A G H A T R F N V M Y K I T S V L V E T N S I
Pig_PRCV_AKV62755.1_ref/1-1232	G T A L K Y V G T L P P S V K E E A I S K W G H F Y I N G Y N F F T F P D C I S F N L T	---G T A L K Y V G T L P P S V K E E A I S K W G H F Y I N G Y N F F T F P D C I S F N L T
Human_TGEV_CA891145.1_ref/1-1447	E T T S A F V G A L P K I V R E E V I S R T G H F Y I N G Y R Y F T L G N V E A V N F N V T A E	---E T T S A F V G A L P K I V R E E V I S R T G H F Y I N G Y R Y F T L G N V E A V N F N V T A E
Human_NL63_AA58177.1_ref/1-1356	T H V S T F V G I L P P T V R E I V V A R T G Q F Y I N G F K Y F D L G F I E A V N F N V T	---T H V S T F V G I L P P T V R E I V V A R T G Q F Y I N G F K Y F D L G F I E A V N F N V T
Human_h-HKUY_N23_ABD96197.1_ref/1-1356	G T L L S H Y V L P L T C --- I --- S R L D I G F T L E --- Y W V T P L T S R Q Y L L A F N Q I	---G T L L S H Y V L P L T C --- I --- S R L D I G F T L E --- Y W V T P L T S R Q Y L L A F N Q I
Human_HK04-02_ABN19366.1_ref/1-1361	G M A L S H Y V M P L T C --- I --- S R R D I G F T L E --- Y W V T P L T P R O Y L L A F N Q I	---G M A L S H Y V M P L T C --- I --- S R R D I G F T L E --- Y W V T P L T P R O Y L L A F N Q I
Human_OC43_ADC10763.1_ref/1-1358	Y D T I K Y Y S I I P H S --- I R S I Q D S R K A W --- A A F Y Y V K L Q P L T F L L D F S V I	---Y D T I K Y Y S I I P H S --- I R S I Q D S R K A W --- A A F Y Y V K L Q P L T F L L D F S V I
Human_h-MERS_KOR/CNUH_SNU/172_06_2015_ALK80311.1_ref/1-1353	G L N I T Y K K A I M T L F --- F --- S T S Q S N F D A D A S A Y F V G H K L P L M L V D F D E	---G L N I T Y K K A I M T L F --- F --- S T S Q S N F D A D A S A Y F V G H K L P L M L V D F D E
Camel_MERS_AHE78097.1_ref/1-1353	G L N I T Y K K V I M T L F --- F --- S P T T S S F N A D A S Y F V G H K L P L M L A E F D E	---G L N I T Y K K V I M T L F --- F --- S P T T S S F N A D A S Y F V G H K L P L M L A E F D E
Bat_BN48-31_ADK66841.1/1-1259	G I N I T R F Q T L L A L H R S Y L T P G K L E S G W T T G A A A Y Y V G Y L Q P R T F L L S Y N Q I	---G I N I T R F Q T L L A L H R S Y L T P G K L E S G W T T G A A A Y Y V G Y L Q P R T F L L S Y N Q I
Bat_BtkY72_APO40579/1-1257	G I N I T R F Q T L L A L H R S Y L T P G N L E S G W T T G A A A Y Y V G Y L Q P R T F L L S Y N Q I	---G I N I T R F Q T L L A L H R S Y L T P G N L E S G W T T G A A A Y Y V G Y L Q P R T F L L S Y N Q I
Pangolin_PCoV_GX-P5E_QJA48641_ref/1-1267	G I N I T R F Q T L L A L H R S Y L T P G K L E S G W T T G A A A Y Y V G Y L Q P R T F L L S Y N Q I	---G I N I T R F Q T L L A L H R S Y L T P G K L E S G W T T G A A A Y Y V G Y L Q P R T F L L S Y N Q I
Pangolin_PCoV_GX-P4L_QJA4861/1-1267	G I N I T R F Q T L L A L H R S Y L T P G N L E S G W T T G A A A Y Y V G Y L Q P R T F L L S Y N Q I	---G I N I T R F Q T L L A L H R S Y L T P G N L E S G W T T G A A A Y Y V G Y L Q P R T F L L S Y N Q I
Pangolin_PCoV_GX-P2V_QIQ5404/1-1269	G I N I T R F Q T L L A L H R S Y L T P G K L E S G W T T G A A A Y Y V G Y L Q P R T F L L S Y N Q I	---G I N I T R F Q T L L A L H R S Y L T P G K L E S G W T T G A A A Y Y V G Y L Q P R T F L L S Y N Q I
Pangolin_PCoV_GX-P1E_QJA48623_ref/1-1265	G I N I T R F Q T L L A L H R S Y L T P G K L E S G W T T G A A A Y Y V G Y L Q P R T F L L S Y N Q I	---G I N I T R F Q T L L A L H R S Y L T P G K L E S G W T T G A A A Y Y V G Y L Q P R T F L L S Y N Q I
Pangolin_PCoV_GX-P5L_QJA48632/1-1267	G I N I T R F Q T L L A L H R S Y L T P G K L E S G W T T G A A A Y Y V G Y L Q P R T F L L S Y N Q I	---G I N I T R F Q T L L A L H R S Y L T P G K L E S G W T T G A A A Y Y V G Y L Q P R T F L L S Y N Q I
Bat_RaTG13_QHR63300_ref/1-1269	G I N I T R F Q T L L A L H R S Y L T P G D S S S G W T T A G A A A Y Y V G Y L Q P R T F L L K Y N E I	---G I N I T R F Q T L L A L H R S Y L T P G D S S S G W T T A G A A A Y Y V G Y L Q P R T F L L K Y N E I
Human_SARS-CoV-2_BetaCoV/Wuhan/HPRCAMS-WH-01/2019_QHU36824/1-1273	G I N I T R F Q T L L A L H R S Y L T P G D S S S G W T T A G A A A Y Y V G Y L Q P R T F L L K Y N E I	---G I N I T R F Q T L L A L H R S Y L T P G D S S S G W T T A G A A A Y Y V G Y L Q P R T F L L K Y N E I
Human_SARS-CoV-2_WIV04_QHR63260/1-1273	G I N I T R F Q T L L A L H R S Y L T P G D S S S G W T T A G A A A Y Y V G Y L Q P R T F L L K Y N E I	---G I N I T R F Q T L L A L H R S Y L T P G D S S S G W T T A G A A A Y Y V G Y L Q P R T F L L K Y N E I
Human_SARS-CoV-2_Wuhan-Hu-1_YP_009724390_ref/1-1273	S I N I T K F R T L L T I --- H R G D P M S N N G W T A F S A A Y F V G Y L K P R T F M L K Y N E I	---S I N I T K F R T L L T I --- H R G D P M S N N G W T A F S A A Y F V G Y L K P R T F M L K Y N E I
Bat_dat-SL-CoVZC45_AVP78031_ref/1-1246	G L N I T N F R V L L T A --- F --- I P N I G T W G T S P A A Y F V G Y L K P T F M L K Y D Y I	---G L N I T N F R V L L T A --- F --- I P N I G T W G T S P A A Y F V G Y L K P T F M L K Y D Y I
Bat_dat-SL-CoVZXC21_AVP78042_ref/1-1245	G I N I T N F R T L L T A --- F --- P R P D Y W G T S A A A Y F V G Y L K P T F M L K Y D E I	---G I N I T N F R T L L T A --- F --- P R P D Y W G T S A A A Y F V G Y L K P T F M L K Y D E I
Bat_CyRa11_AH93755/1-1259	G I N I T N F R T L L T A --- F --- P R P D Y W G T S A A A Y F V G Y L K P T F M L K Y D E I	---G I N I T N F R T L L T A --- F --- P R P D Y W G T S A A A Y F V G Y L K P T F M L K Y D E I
Bat_Rs4084_ATO98132.1/1-1256	G I N I T N F R T L L T A --- F --- P R P D Y W G T S A A A Y F V G Y L K P T F M L K Y D E I	---G I N I T N F R T L L T A --- F --- P R P D Y W G T S A A A Y F V G Y L K P T F M L K Y D E I
Bat_RsSHC014_AG248806.1/1-1256	G I N I T N F R T L L T A --- F --- P R P D Y W G T S A A A Y F V G Y L K P T F M L K Y D E I	---G I N I T N F R T L L T A --- F --- P R P D Y W G T S A A A Y F V G Y L K P T F M L K Y D E I
Bat_WIV1_AG248828.1/1-1256	G I N I T N F R T L L T A --- F --- P R P D Y W G T S A A A Y F V G Y L K P T F M L K Y D E I	---G I N I T N F R T L L T A --- F --- P R P D Y W G T S A A A Y F V G Y L K P T F M L K Y D E I
Bat_Rs3367_AG248818/1-1256	G I N I T N F R T L L T A --- F --- P R P D Y W G T S A A A Y F V G Y L K P T F M L K Y D E I	---G I N I T N F R T L L T A --- F --- P R P D Y W G T S A A A Y F V G Y L K P T F M L K Y D E I
Bat_Rs7327_ATO98218/1-1256	G I N I T N F R T L L T A --- F --- P R P D Y W G T S A A A Y F V G Y L K P T F M L K Y D E I	---G I N I T N F R T L L T A --- F --- P R P D Y W G T S A A A Y F V G Y L K P T F M L K Y D E I
Bat_BtRs-BetaCoV/YN2018R_QDF43825/1-1256	G I N I T N F R A I L T A --- F --- L P A Q D T W G T S A A A Y F V G Y L K P A T F M L K Y D E I	---G I N I T N F R A I L T A --- F --- L P A Q D T W G T S A A A Y F V G Y L K P A T F M L K Y D E I
Bat_Rs4231_ATO98157/1-1255	G I N I T N F R A I L T A --- F --- L P A Q D T W G T S A A A Y F V G Y L K P A T F M L K Y D E I	---G I N I T N F R A I L T A --- F --- L P A Q D T W G T S A A A Y F V G Y L K P A T F M L K Y D E I
Bat_S23_ATO98205.1/1-1255	G I N I T N F R A I L T A --- F --- L P A Q D T W G T S A A A Y F V G Y L K P A T F M L K Y D E I	---G I N I T N F R A I L T A --- F --- L P A Q D T W G T S A A A Y F V G Y L K P A T F M L K Y D E I
Bat_WIV16_ALK02457.1/1-1255	G I K I T N F R A I L T A --- F --- S P A Q D I W G T S A A A Y F V G Y L K P T F M L K Y D E I	---G I K I T N F R A I L T A --- F --- S P A Q D I W G T S A A A Y F V G Y L K P T F M L K Y D E I
Civet_SARS-CoV_007/2004_AAU04646/1-1255	G I N I T N F R A I L T A --- F --- S P A Q D I W G T S A A A Y F V G Y L K P T F M L K Y D E I	---G I N I T N F R A I L T A --- F --- S P A Q D I W G T S A A A Y F V G Y L K P T F M L K Y D E I
Civet_S23_P59594.1/1-1255	G I N I T N F R A I L T A --- F --- S P A Q D I W G T S A A A Y F V G Y L K P T F M L K Y D E I	---G I N I T N F R A I L T A --- F --- S P A Q D I W G T S A A A Y F V G Y L K P T F M L K Y D E I
Human_SARS-CoV_Tor2/FP1-10895_AFR58742/1-1255	G I N I T N F R A I L T A --- F --- S P A Q D I W G T S A A A Y F V G Y L K P T F M L K Y D E I	---G I N I T N F R A I L T A --- F --- S P A Q D I W G T S A A A Y F V G Y L K P T F M L K Y D E I
Human_SARS-CoV_TW11_AA887512/1-1255	G I N I T N F R A I L T A --- F --- S P A Q D I W G T S A A A Y F V G Y L K P T F M L K Y D E I	---G I N I T N F R A I L T A --- F --- S P A Q D I W G T S A A A Y F V G Y L K P T F M L K Y D E I
Human_SARS-CoV_Frankfurt-1_AAP33697.1_ref/1-1255	G L N I T N Y K V Y T T L --- K P T N Q A F --- Q A A Y I V G N L K H T M N M L S F N E I	---G L N I T N Y K V Y T T L --- K P T N Q A F --- Q A A Y I V G N L K H T M N M L S F N E I
Bat_JTMC15_AMA96027/1-1236	G L N I T N Y K V Y T T L --- K P T N Q A F --- Q A A Y I V G N L K H T M N M L S F N E I	---G L N I T N Y K V Y T T L --- K P T N Q A F --- Q A A Y I V G N L K H T M N M L S F N E I
Bat_BtRs-BetaCoV/YN2013_AIA62390/1-1233	G I N I T S Y R V Y M T M --- F --- S T Q Q N F L T E N A A Y Y I G Y L K P R T F M L Q F N T I	---G I N I T S Y R V Y M T M --- F --- S T Q Q N F L T E N A A Y Y I G Y L K P R T F M L Q F N T I
Bat_BtRs-BetaCoV/SC2018_QDF43815/1-1256	G I N I T G M R V Y M T M --- F --- S N T Q A N F L T E N A A Y Y V G Y L K P R T F M L Q F N T I	---G I N I T G M R V Y M T M --- F --- S N T Q A N F L T E N A A Y Y V G Y L K P R T F M L Q F N T I
Bat_Rp/Shaanxi2011_AGC74163/1-1240	G I N I T S F R V Y M A M --- F --- S K T T S N V Y P E S A A Y Y V G N L K Y T F M L S F N E I	---G I N I T S F R V Y M A M --- F --- S K T T S N V Y P E S A A Y Y V G N L K Y T F M L S F N E I
Bat_BtRs-BetaCoV/HuB2013_AIA62310/1-1241	G I N I T S F R V Y M A M --- F --- S K T T S N V Y P E S A A Y Y V G N L K Y T F M L S F N E I	---G I N I T S F R V Y M A M --- F --- S K T T S N V Y P E S A A Y Y V G N L K Y T F M L S F N E I
Bat_YNF5_31C_AK219076/1-1241	G I N I T S F R V Y M A M --- F --- S K T T S N V Y P E S A A Y Y V G N L K Y T F M L S F N E I	---G I N I T S F R V Y M A M --- F --- S K T T S N V Y P E S A A Y Y V G N L K Y T F M L S F N E I
Bat_BtRs-BetaCoV/HeB2013_AIA62290/1-1241	G I N I T S F R V Y M A M --- F --- S K T T S N V Y P E S A A Y Y V G N L K Y T F M L S F N E I	---G I N I T S F R V Y M A M --- F --- S K T T S N V Y P E S A A Y Y V G N L K Y T F M L S F N E I
Bat_RF1/2004_ABD75323/1-1241	G I N I T S F K V Y M T M --- Y --- S Q T T S N F L S E S A A Y Y V G N L K Y T F M F Q F N E I	---G I N I T S F K V Y M T M --- Y --- S Q T T S N F L S E S A A Y Y V G N L K Y T F M F Q F N E I
Bat_Bat-CoV-273/2005_ABG47060/1-1241	S I N I T S F K V Y M S M --- F --- S R T T S N F L P E I A A Y F V G N L K Y S T F M L N F N E I	---S I N I T S F K V Y M S M --- F --- S R T T S N F L P E I A A Y F V G N L K Y S T F M L N F N E I
Bat_Cp/Yunnan2011_AGC74176/1-1241	G I N I T S Y R V Y M A M --- F --- S Q T T S N F L P E S A A Y Y V G N L K Y T F M L S F N E I	---G I N I T S Y R V Y M A M --- F --- S Q T T S N F L P E S A A Y Y V G N L K Y T F M L S F N E I
Bat_Rs_672/2006_ACU31032/1-1241	G I N I T S Y R V Y M A M --- F --- S Q T T S N F L P E S A A Y Y V G N L K Y T F M L S F N E I	---G I N I T S Y R V Y M A M --- F --- S Q T T S N F L P E S A A Y Y V G N L K Y T F M L S F N E I
Bat_Rs4081_ATO98120/1-1241	G I N I T S Y R V Y M A M --- F --- S Q T T S N F L P E S A A Y Y V G N L K Y T F M L S F N E I	---G I N I T S Y R V Y M A M --- F --- S Q T T S N F L P E S A A Y Y V G N L K Y T F M L S F N E I
Bat-recombinant_Bat-SRBD_AG160703.1_ref/1-1259	G I N I T S Y R V Y M A M --- F --- S Q T T S N F L P E S A A Y Y V G N L K Y T F M L S F N E I	---G I N I T S Y R V Y M A M --- F --- S Q T T S N F L P E S A A Y Y V G N L K Y T F M L S F N E I
Bat_Bat-SARS-CoV-Rm1/2004_ABD75332/1-1241	G I N I T S Y R V Y M A M --- F --- S Q T T S N F L P E S A A Y Y V G N L K Y T F M L S F N E I	---G I N I T S Y R V Y M A M --- F --- S Q T T S N F L P E S A A Y Y V G N L K Y T F M L S F N E I
Bat_Rp3/2004_AA267052/1-1241	G I N I T S Y R V Y M A M --- F --- S Q T T S N F L P E S A A Y Y V G N L K Y T F M L S F N E I	---G I N I T S Y R V Y M A M --- F --- S Q T T S N F L P E S A A Y Y V G N L K Y T F M L S F N E I
Bat_BtRs-BetaCoV/YN2018C_QDF43830/1-1241	G I N I T S Y R V Y M A M --- F --- S Q T T S N F L P E S A A Y Y V G N L K Y T F M L S F N E I	---G I N I T S Y R V Y M A M --- F --- S Q T T S N F L P E S A A Y Y V G N L K Y T F M L S F N E I
Bat_Rs4247_ATO98181.1/1-1242	G I N I T S Y R V Y M A M --- F --- S Q T T S N F L P E S A A Y Y V G N L K Y T F M L S F N E I	---G I N I T S Y R V Y M A M --- F --- S Q T T S N F L P E S A A Y Y V G N L K Y T F M L S F N E I
Bat_Rs4237_ATO98169/1-1241	G I N I T S Y R V Y M A M --- F --- S Q T T S N F L P E S A A Y Y V G N L K Y T F M L S F N E I	---G I N I T S Y R V Y M A M --- F --- S Q T T S N F L P E S A A Y Y V G N L K Y T F M L S F N E I
Bat_BtRs-BetaCoV/GX2013_AIA62320/1-1242	G I N I T S Y R V Y M A M --- F --- S Q T T S N F L P E S A A Y Y V G N L K Y T F M L S F N E I	---G I N I T S Y R V Y M A M --- F --- S Q T T S N F L P E S A A Y Y V G N L K Y T F M L S F N E I
Bat_HKU3-8_ADE34766/1-1242	G I N I T S Y R V Y M A M --- F --- S Q T T S N F L P E S A A Y Y V G N L K Y S T F M L R F N E I	---G I N I T S Y R V Y M A M --- F --- S Q T T S N F L P E S A A Y Y V G N L K Y S T F M L R F N E I
Bat_Longquan-140_AID16716/1-1242	G I N I T S Y R V Y M A M --- F --- S Q T T S N F L P E S A A Y Y V G N L K Y S T F M L R F N E I	---G I N I T S Y R V Y M A M --- F --- S Q T T S N F L P E S A A Y Y V G N L K Y S T F M L R F N E I
Bat_HKU3-12_ADE34812/1-1242	G I N I T S Y R V Y M A M --- F --- S Q T T S N F L P E S A A Y Y V G N L K Y S T F M L R F N E I	---G I N I T S Y R V Y M A M --- F --- S Q T T S N F L P E S A A Y Y V G N L K Y S T F M L R F N E I
Bat_HKU3-6_ADE34744/1-1242	G I N I T S Y R V Y M A M --- F --- S Q T T S N F L P E S A A Y Y V G N L K Y S T F M L R F N E I	---G I N I T S Y R V Y M A M --- F --- S Q T T S N F L P E S A A Y Y V G N L K Y S T F M L R F N E I

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

```
%ig_SARS_AVM41569.1_ref/1-1130
%ig_PRCV_AKV62755.1_ref/1-1232
%man_TGEV_CAB91145.1_ref/1-1447
%man_229E_AA648592.1_ref/1-1173
%man_NL63_AA558177.1_ref/1-1356
%man_h-hKU1_N23_ABD96197.1_ref/1-1356
%man_HK04-02_ABN19366.1_ref/1-1361
%man_OC43_ADI0763.1_ref/1-1358
%man_h-MERS_KOR/CNH/SNU/172_06_2015_AIK80311.1_ref/1-1353
%man_MERS_AHE78097.1_ref/1-1353
%Bat_BM48-31_ADK66041.1/1-1259
%Bat_BTKY122_APO40579/1-1257
%Pangolin_PCoV_GX-PSE_QIA48641_ref/1-1267
%Pangolin_PCoV_GX-P41_QIA48614/1-1267
%Pangolin_PCoV_GX-P2V_QIQ54048/1-1269
%Pangolin_PCoV_GX-P1E_QIA48623_ref/1-1265
%Pangolin_PCoV_GX-PSL_QIA48632/1-1267
%Bat_RaTG13_QHR63300_ref/1-1269
%Human_SARS-CoV-2_BetaCoV/Wuhan/HPBCAMS-WH-01/2019_QHU36824/1-1273
%Human_SARS-CoV-2_WIV04_QHR63260/1-1273
%Human_SARS-CoV-2_Wuhan-Hu-L_YP_009724390_ref/1-1273
%Bat_bat-SL-CoVZC45_AVP78031_ref/1-1246
%Bat_bat-SL-CoVZXC21_AVP78042_ref/1-1245
%Bat_LYRa11_AHX97558/1-1259
%Bat_Rs4084_ATO98132.1/1-1256
%Bat_RsSHC014_AG248806.1/1-1256
%Bat_WIV1_AG24882B.1/1-1256
%Bat_Rs3367_AG24881B/1-1256
%Bat_Rs7227_ATO98218/1-1256
%Bat_BtRs-BetaCoV/YN2018L_QDF43825/1-1256
%Bat_Rs4231_ATO98157/1-1255
%Bat_S23_ATO98205.1/1-1255
%Bat_WIV16_ALK02457.1/1-1255
%ClveC_SARS-CoV_007/2004_AAU04646/1-1255
%ClveC_S23_P59594.1/1-1255
%Human_SARS-CoV_Tor2/HP1-10895_AFR58742/1-1255
%Human_SARS-CoV_TW11_AAR87512/1-1255
%Human_SARS-CoV_Frankfurt_1_AAP33697.1_ref/1-1255
%Bat_H2012_AIA62277.1_ref/1-1236
%Bat_JTMC15_AIA96027/1-1236
%Bat_BtRs-BetaCoV/YN2013_AIA62330/1-1233
%Bat_BtRi-BetaCoV/SC2018_QDF43815/1-1256
%Bat_Rp/Shaanxi2011_AGC74165/1-1240
%Bat_BtRs-BetaCoV/HuB2013_AIA62310/1-1241
%Bat_YNF_E_31C_AK219076/1-1241
%Bat_BtRi-BetaCoV/HuB2013_AIA62290/1-1241
%Bat_RF1/2004_ABD75323/1-1241
%Bat_Bat-CoV-273/2005_ABG47060/1-1241
%Bat_Cp/Wuhan2011_AGC74176/1-1241
%Bat_Rs_672/2006_ACU31032/1-1241
%Bat_Rs4081_ATO98120/1-1241
%Bat-recombinant_Bat-SRBD_ACJ60703.1_ref/1-1259
%Bat_Bat-SARS-CoV-Rm1/2004_ABD75332/1-1241
%Bat_Rp3/2004_AAZ67052/1-1241
%Bat_BtRs-BetaCoV/YN2018C_QDF43830/1-1241
%Bat_Rs4247_ATO98181.1/1-1242
%Bat_Rs4237_ATO98169/1-1241
%Bat_BtRs-BetaCoV/GX2013_AIA62320/1-1242
%Bat_HKU3-B_ADE34766/1-1242
%Bat_Longman-140_AID16716/1-1242
%Bat_HKU3-12_ADE34812/1-1242
%Bat_HKU3-6_ADE34744/1-1242
```

```
810 820 830 840 850 860 870
--VLRVGRG---KAVNRITVTRFLKPY--LTFNKFCFLSS--VGAN
DAVAVIKTG---CFPSFDLNNY--LTFNFKCLSLSPVGA--N
DAVAVIKTG---CFPSFDLNNY--VKFGSVCFLKDIPLGG--
SASINTG---NCFPSF--GKVNFF--YKFKTICFTVPEVPS--
IYLLKSG---CFPSFKLNNF--
RFGNFNPLSSSHSVVSRVCFSVNNTGCPKAPKSFASSCKSHKPPSASCPIGTNYRSCSTTVLDHTDWC
FVFPQRTGVFTNHSSVVAQHCFKFAPKNCPC--KLNKSGCPGKNNIGCTCPAGTNYLTCNDL--
FVFPQRTGVFTNHSSVVAQHCFKFAPKNCPC--SSCPGKNNIGCTCPAGTNYLTCNDL--
VPHNLTITIKPLKYSVINCKSRLLSDD--RTFEVPLQVLMANQYSPCVS--VTPS
VPHNLTITIKPLKYSVINCKSRLLSDD--RTFEVPLQVLMANQYSPCVS--VTPS
--TNSLSDSN---EFFYRFRFHGKIKPY--GKDLNVLNFPNGGTCSA--EGLN
--TNSVDSKSGN---NFYRFLRHGKIKPY--ERDLSNVLNYSAGGTCSSISGLG
--SVKQDALTGDNYGYLYRFLRKSXLKF--ERDLSSTEIYQAGSTPFCNGQVGLN
--SVKQDALTGDNYGYLYRFLRKSXLKF--ERDLSSTEIYQAGSTPFCNGQVGLN
--SVKQDALTGDNYGYLYRFLRKSXLKF--ERDLSSTEIYQAGSTPFCNGQVGLN
--SVKQDALTGG---NYLYRFLRKSXLKF--ERDLSSTEIYQAGSTPFCNGQVGLN
--SVKQDALTGDNYGYLYRFLRKSXLKF--ERDLSSTEIYQAGSTPFCNGQVGLN
--SKHIDAKEGGFNLYRFLRKANLKF--ERDLSSTEIYQAGSKPQNGVGLN
--SNLNSKVGQGNMYLYRFLRKSXLKF--ERDLSSTEIYQAGSTPFCNGVEGFN
--SNLNSKVGQGNMYLYRFLRKSXLKF--ERDLSSTEIYQAGSTPFCNGVEGFN
--SNLNSKVGQGNMYLYRFLRKSXLKF--ERDLSSTEIYQAGSTPFCNGVEGFN
--TAKQDVG---SYFYRSHRSXLKPF--ERDLS--DE--N
--TAKQDTG---HYFYRSHRSXLKPF--ERDLS--DE--N
--TRNIDATSSGNFNKYRSLRHGKLRP--ERDLSNVVFPDQKPCFP--PAFN
--TNSKDSSTSGNMYLYRWYRSLKLPY--ERDLSNDIYSPGGSCSA--VGFN
--TNSKDSSTSGNMYLYRWYRSLKLPY--ERDLSNDIYSPGGSCSA--VGFN
--TRNIDATQTNMYNKYRSLRHGKLRP--ERDLSNVVFPDQKPCFP--PAFN
--TRNIDATQTNMYNKYRSLRHGKLRP--ERDLSNVVFPDQKPCFP--PAFN
--TRNIDATQTNMYNKYRSLRHGKLRP--ERDLSNVVFPDQKPCFP--PAFN
--TRNIDATQTNMYNKYRSLRHGKLRP--ERDLSNVVFPDQKPCFP--PAFN
--TRNIDATQTNMYNKYRSLRHGKLRP--ERDLSNVVFPDQKPCFP--PAFN
--TRNIDATQTNMYNKYRSLRHGKLRP--ERDLSNVVFPDQKPCFP--PAFN
--TAKQDVG---SYFYRSHRSXLKPF--ERDLS--EE--N
--TAKQDVG---SYFYRSHRSXLKPF--ERDLS--EE--N
--TAKQDVG---SYFYRSHRSXLKPF--ERDLS--DE--N
--TAKQDTG---SYFYRSHRSXLKPF--ERDLS--DDG--N
--TAKQDQG---QYYRSHRSXLKPF--ERDLS--DE--N
--TAKQDTG---YFYRSHRSXLKPF--ERDLS--DDG--N
--TAKYDVG---SYFYRSHRSXLKPF--ERDLS--EE--N
--TAKYDVG---SYFYRSHRSXLKPF--ERDLS--EE--N
--TAKQDVG---SYFYRSHRSXLKPF--ERDLS--EE--N
--TAKQDVG---SYFYRSHRSXLKPF--ERDLS--EE--N
--TAKQDVG---SYFYRSHRSXLKPF--ERDLS--VE--E
--TANQDRG---QYYRSHRSXLKPF--ERDLS--DE--N
--TAKQDQG---QYYRSHRSXLKPF--ERDLS--DE--N
--TAKQDQG---QYYRSHRSXLKPF--ERDLS--DE--N
--TRNIDATSTGNMYNKYRSLRHGKLRP--ERDLSNVVFPDQKPCFP--PAFN
--TAKQDQG---QYYRSHRSXLKPF--ERDLS--DE--N
--TAKQDQG---QYYRSHRSXLKPF--ERDLS--DE--N
--TAKQDKG---QYYRSHRSXLKPF--ERDLS--DE--N
--TAKQDTG---HYFYRSHRSXLKPF--ERDLS--DDG--N
--TAKQDQG---QYYRSHRSXLKPF--ERDLS--DE--N
--TAKQDTG---NYFYRSHRSXLKPF--ERDLS--DDG--N
--TAKQDTG---NYFYRSHRSXLKPF--ERDLS--DDG--N
--TAKHDIG---NYFYRSHRSXLKPF--ERDLS--DDG--N
--TAKHDTG---NYFYRSHRSXLKPF--ERDLS--DDG--N
```


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

Cm_MERS_AHE78097.1_ref	SLCALP-DTPST----LTPRSVRSV	20
Hu_MERS_172-06_2015_ALK80311.1_ref	SLCALP-DTPST----LTPRSVRSV	20
Bt_BM48-31_ADK66841.1	GICAKYTNVSSST----LVRSGGHSI	21
Bt_BtKY72_APO40579	GICAKF-GSDKI-----RMGOESI	18
BtYu-RmYN02_2019_S-gene_21544-25227_1	GVCASY-NSPAA-----RVGTNSI	18
Bt_LYRa11_AHX37558	GICASY-HTASL----LRNTDQKSI	20
Bt_YN2018B_QDF43825	GICASY-HTVSS----LRSTSQKSI	20
Bt_Rs4874_ATO98205.1	GICASY-HTVSS----LRSTSQKSI	20
Cv_007-2004_AAU04646	GICASY-HTVSS----LRSTSQKSI	20
Hu_SARS-Frankfurt-1_2003_AAP33697.1_ref	GICASY-HTVSL----LRSTSQKSI	20
Bt_rec-SARS_2008_ACJ60694.1_ref	GICASY-HTVSL----LRSTSQKSI	20
Bt_ZC45_AVP78031_ref	GICASY-HTASI----LRSTSQKAI	20
Bt_ZXC21_AVP78042_ref	GICASY-HTASI----LRSTGQKAI	20
PnGu1_2019_S-gene_21541-25338_1	GICASY-QTQTN----SRSVSSQAI	20
Pn_GX-P1E_2017_QIA48623_ref	GICASY-HSMSS----LRSVNORSI	20
Pn_GX-P2V_2018_QIQ54048	GICASY-HSMSS----FRSVNORSI	20
Bt_RaTG13_2013_Yunnan_QHR63300_ref	GICASY-QTQTN----SRSVASQSI	20
Hu_CoV2_WH01_2019_QHU36824_ref	GICASY-QTQTN SPRR ARSVASQSI	24
Bt_JL2012_AIA62277.1_ref	GICASY-HTASL----LRSTGQKSI	20
Bt_YN2013_AIA62330	GICASY-HTAST----LRSIGQKSI	20
Bt_Rp-Shaanxi2011_AGC74165	GICASY-HTASV----LRSTGQKSI	20
Bt_SC2018_QDF43815	GICASY-HTAST----LRSTGQKSI	20
Bt_YNLF_31C_AKZ19076	GICASY-HTASV----LRSTGQKSI	20
Bt_Cp-Yun_2011_AGC74176	GICASY-HTASL----LRNTGQKSI	20
Bt_Rs_672-2006_ACU31032	GICASY-HTAST----LRSVGQKSI	20
Bt_Rm1/2004_ABD75332	GICASY-HTASV----LRSTGQKSI	20
Bt_YN2018C_QDF43830	GICASY-HTAST----LRSVGQKSI	20
Bt_Rp3-2004_AAZ67052	GICASY-HTAST----LRSVGQKSI	20
Bt_GX2013_AIA62320	GICASY-HTASV----LRSTGQKSI	20
Bt_HKU3-12_ADE34812_ref	GICASY-HTASV----LRSTGQKSI	20

0.....790.....800....



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

The screenshot shows the Zenodo preprint page for the paper. The header includes the Zenodo logo, a search bar, and navigation links for 'Upload' and 'Communities'. The paper title is prominently displayed, along with statistics for 776,634 views and 597,172 downloads. The authors listed are Yan, Li-Meng; Kang, Shu; Guan, Jie; Hu, Shanchang. The abstract discusses the COVID-19 pandemic and the authors' hypothesis of a laboratory origin for SARS-CoV-2. A timeline visualization at the bottom shows restriction sites for EcoRI and BestE1I on the MT019529.1 sequence.

zenodo Search Upload Communities Log in Sign up

September 14, 2020 Working paper Open Access

Unusual Features of the SARS-CoV-2 Genome Suggesting Sophisticated Laboratory Modification Rather Than Natural Evolution and Delineation of Its Probable Synthetic Route

Yan, Li-Meng; Kang, Shu; Guan, Jie; Hu, Shanchang

The COVID-19 pandemic caused by the novel coronavirus SARS-CoV-2 has led to over 910,000 deaths worldwide and unprecedented decimation of the global economy. Despite its tremendous impact, the origin of SARS-CoV-2 has remained mysterious and controversial. The natural origin theory, although widely accepted, lacks substantial support. The alternative theory that the virus may have come from a research laboratory is, however, strictly censored on peer-reviewed scientific journals. Nonetheless, SARS-CoV-2 shows biological characteristics that are inconsistent with a naturally occurring, zoonotic virus. In this report, we describe the genomic, structural, medical, and literature evidence, which, when considered together, strongly contradicts the natural origin theory. The evidence shows that SARS-CoV-2 should be a laboratory product created by using bat coronaviruses ZC45 and/or ZXC21 as a template and/or backbone. Building upon the evidence, we further postulate a synthetic route for SARS-CoV-2, demonstrating that the laboratory-creation of this coronavirus is convenient and can be accomplished in approximately six months. Our work emphasizes the need for an independent investigation into the relevant research laboratories. It also argues for a critical look into certain recently published data, which, albeit problematic, was used to support and claim a natural origin of SARS-CoV-2. From a public health perspective, these actions are necessary as knowledge of the origin of SARS-CoV-2 and of how the virus entered the human population are of pivotal importance in the fundamental control of the COVID-19 pandemic as well as in preventing similar, future pandemics.

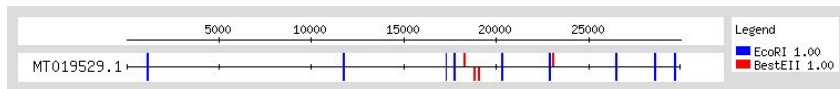
Indexed in OpenAIRE

Publication date: September 14, 2020
DOI: 10.5281/zenodo.4028830
Communities: Coronavirus Disease Research Community - COVID-19
License (for files):

Legend
■ EcoRI 1.00
■ BestE1I 1.00

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 SARS-CoV-2		EcoRI							
		W	N	S					
tataattata	aattaccaga	tgattttaca	ggctgcgta	tagcttg	gaa	ttgt	aaacaat	1320	
cttgattcta	aggttggtgg	taattataat	tacctgtata	gattgtttag	gaagtcta	aat		1380	
ctcaaacctt	ttagagagaga	tatttcaact	gaaatctatc	aggccggtag	cacacctt	gt		1440	
aatggtgttg	aaggttttaa	ttgttacttt	cctttacaat	catatggttt	ccaacc	caact		1500	
aatggtgttg	gttacc	aacc	atacagagta	gtagtacttt	cttttgaact	tctacat	gca	1560	
		G	Y	Q	BstEII				
B ZC45		EcoRI							
		W	N	T					
ttacctgatg	attttacagg	ttgtgtcata	gcttg	gaaca	ct	tgccaaaca	ggatgtaggt	1320	
aattatttct	acaggtctca	tcgttctacc	aaattgaaac	catttgaaag	agatccttcc			1380	
tcagacgaga	atggtgtccg	tacacttagt	acttatgact	tcaaccctaa	tgtaccactt			1440	
gaa	tacc	aaag	ctacaagggt	tgttgttttg	tcatttgagc	ttcta	aatg	accagctaca	1500
		E	Y	Q					

- 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.

	EcoRI
HuCoV2_WH01_2019_215	ACAGGCTGCGTTATAGCTTGAATTCGAACAA
BtRaTG13_2013_Yunnan_215	ACTGGTTGTGTTATAGCTTGAATTCGAACCA
PnMP789_214	ACAGGTTGTGTAATAGCTTGAATTCGAACAA
PnGX-P1E_2017_215	ACTGGTTGTGTTATTGCTTGAATTCAGTTAA
PnGX-P2V_2018_215	ACTGGTTGTGTTATTGCTTGAATTCAGTTAA
BtZC45_215	ACAGGTTGTGCATAGCTTGAACACAGCCAA
BtZXC21_214	ACAGGTTGTGCATAGCTTGAACACAGCCAA
BtLYRa11_215	ATGGGTTGTGCTTGGCTTGAACACAGGAA
BtRs4874_214	ACGGGTTGTGCTTGGCTTGAATACAGGAA
BtYN2018B_215	ATGGGTTGTGCTTGGCTTGAATACTAGGAA
HuSARS-Frankfurt-1_2003_215	ATGGGTTGTGCTTGGCTTGAATACAGGAA
Cv007-2004_214	ATGGGTTGTGCTTGGCTTGAATACTAGGAA
BtHKU3-12_214	ACTGGCTGTGTAATTGCTTGAATACAGCTAA
Btrec-SARSg_2008_215	ACTGGCTGTGTAATTGCTTGAATACAGCTAA
BtYN2013_212	ACAGGCTGTGCATAGCTTGAATACAGCTAA
BtRm1/2004_215	ACAGGCTGTGTAATAGCTTGAATACAGCA
BtRp3-2004_214	ACTGGTTGGTAATAGCTTGAATACAGCAAA
BtGX2013_212	ACTGGCTGTGTAATCGCTTGAATACAGCTAA
BtRp-Shaanxi2011_213	ACAGGCTGTGTAATAGCTTGAACACAGCAAA
BtSC2018_214	ACTGGCTGTGTAATAGCTTGAATACAGCTAA
BtCp-Yun_2011_214	ACAGGTTGGTAATTGCTTGAATACTAGCTAA
BtRs_672-2006_208	ACAGGCTGTGCATAGCTTGAACACAGCTAA
BtYN2018C_215	ACAGGCTGTGTTATTGCTTGAATACTAGCTAA
BtYNLF_31C_215	ACAGGCTGTGTTATAGCTTGAACACAGCCAA
BtJL2012_212	ATAGGTTGTGTTATAGCTTGAACACAGCCAA
BtBtKY72_214	ACTGGCTGTGTTTATAGCTTGAATACAGCTAA
BtBM48-31_213	ACAGGTTGTGTAATAGCTTGAATACAGCTAA
BtHKU5_219	TCT---TCACCAATTGCAATTACAACCTAACCA
HuOC43_241	ACAAGTTGTCAGTTGTAATTAATTTACCTGTC
CmMERS_218	CCCACATGTTTGTATTTAGCGACTGTTCCCTCA
HuMERS_172-06_2015_218	CCCACATGTTTGTATTTAGCGACTGTTCCCTCA
HuTGEV_210	TATTGATTGATATATCTTTTAAATTTGACCAGT
BtHKU9-1_211	TACGGTTGTTTGCATGCAATCTATTTGAAATTC
Hu229E_205	-----GCTAGTATTAAACACGGGAAA
HuNL63_210	---TCATGGCACATTTATTTAAAGAGTGGCAC
PiSADS_205	CCTTATGAATGTTTGGGTTGGTCAATGGAATGA
PiPRCV_205	-----GCCACCGCTGTTATAAAAACCTGGTAC

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.

BstEII

HuCoV2_WH01_2019_215	GGTGTGGTTACCAACCATACAGAA
BtRaTG13_2013_Yunnan_215	GGTGTGGTACCAACCTTATAGGA
PnMP789_214	GGTGTGGTTACCAACCTTATAGAA
PnGX-P1E_2017_215	GGTGTAACTACCAACCTTTTAGAA
PnGX-P2V_2018_215	GGTGTAACTACCAACCTTTTAGAA
BtZC45_215	CCACTTGAATACCAAGCTACAAGG
BtZXC21_214	CCGCTTGAATATCAAGCTACAAGG
BtLYRa11_215	GGCATGGTTACCAACCTTATAGAA
BtRs4874_214	GGCATGGCTACCAACCTTATAGAA
BtYN2018B_215	GGCATGGCTATCAACCTTATAGAA
HuSARS-Frankfurt-1_2003_215	GGCATGGCTACCAACCTTACAGAA
Cv007-2004_214	GGCATGGCTACCAACCTTACAGAA
BtHKU3-12_214	CCAGTAGCATATCAGGCTACTAGGA
Btrec-SARSg_2008_215	CCAGTAGCATATCAGGCTACTAGGA
BtYN2013_212	CCCTCTGATTATCAAGCCACCAGAA
BtRm1/2004_215	CCAGTGAATACCAAGGCACTAGGA
BtRp3-2004_214	CCGGTGGCTTATCAGGCTACTAGGA
BtGX2013_212	CCAGTGGCATATCAGGCTACTAGGA
BtRp-Shaanxi2011_213	CCACTTGAATATCAGGCTACTAGGA
BtSC2018_214	CCGGTGGCATATCAGGCTACTAGAA
BtCp-Yun_2011_214	CCACTTGAATACCAAGCTACTAGAA
BtRs_672-2006_208	CCATATTGAATATCAGGCTACTAGGA
BtYN2018C_215	CCATATTGAATATCAGGCTACTAGGA
BtYNLF_31C_215	CCCCTTGAATATCAAGCCACTAGAA
BtJL2012_212	CCCTCTGAGTACCAAGCCACTAGAA
BtBtKY72_214	GGTGTGGTTACCAACCATATAGAA
BtBM48-31_213	GGAAATTGGCTTTCAACCATACAGAA
BtHKU5_219	GTTACTCTTCTCTTACAGTGGACT
HuOC43_241	AAGTGCCCCCAAATAAATCTTTAA
CmMERS_218	CCACTTGAAGGTGGTGGCTGGCTT
HuMERS_172-06_2015_218	CCACTTGAAGGTGGTGGCTGGCTT
HuTGEV_210	AAACACAGCTATTACAAAGGTGAC
BtHKU9-1_211	CCTTTTCTTAT---GTTTATGGT
Hu229E_205	ATACCCGGTGGTTGCGCAATGCC
HuNL63_210	GTGCCTGGTAGTTGTAATTTTCCG
PiSADS_205	-----GGTGG
PiPRCV_205	GTTGGTGGCTAATTTGTAAGTTTGA

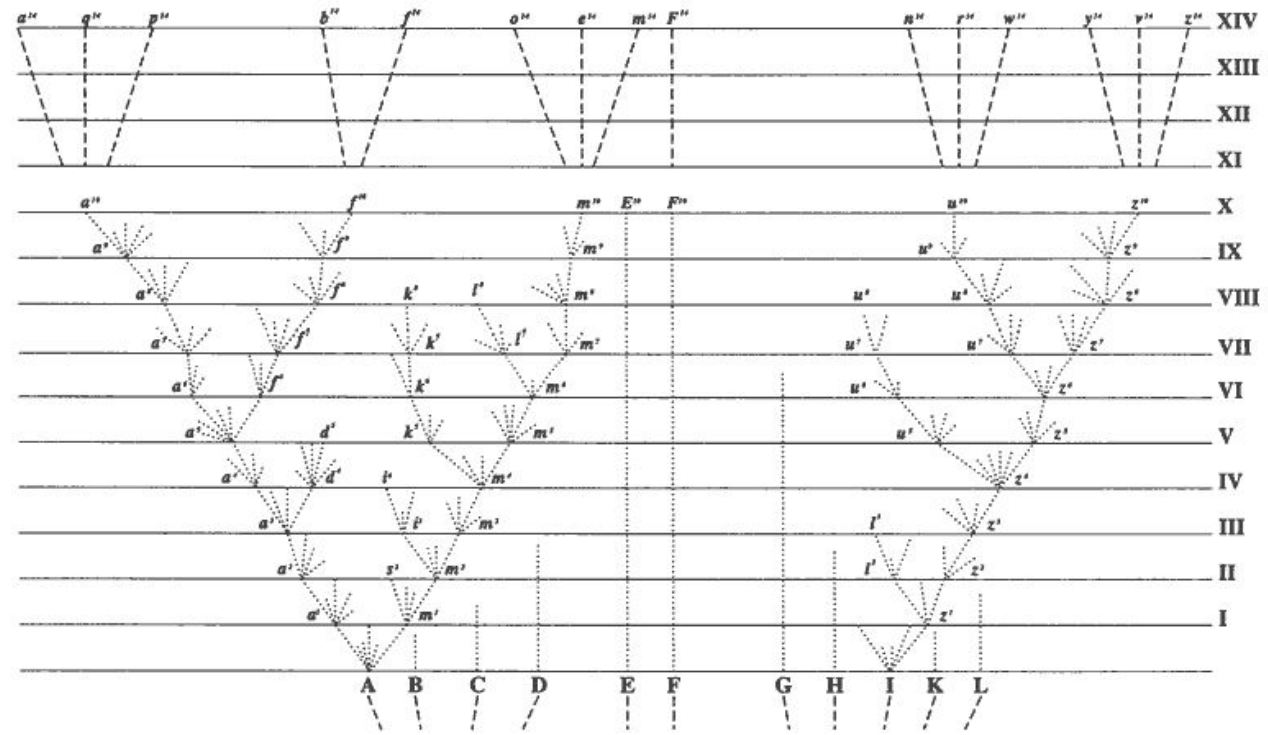
Co-occurrence des sites de restriction qui encadrent le RBD

	EcoRI	BstEII
HuCoV2_WH01_2019_215	ACAGGCTGCGTTATAGCTTGGAAATCTAACAA	GGTGTGTTGTTACCAACCATACAGAA
BtRaTG13_2013_Yunnan_215	ACTGGTTGTGTTATAGCTTGGAAATCTAAGCA	GGTGTGGGCACCAACCTTATAGGA
PnMP789_214	ACAGGTTGTGTAATAGCTTGGAAATCTAACAA	GGTGTGGTTACCAACCTTATAGAA
PnGX-P1E_2017_215	ACTGGTTGTGTTATTGCTTGGAAACACAGTTAA	PnGX-P1E_2017_215 GGTGTTAACCTACCAACCTTTTAGAA
PnGX-P2V_2018_215	ACTGGTTGTGTTATTGCTTGGAACTCAGTTAA	PnGX-P2V_2018_215 GGTGTTAACCTACCAACCTTTTAGAA
BtZC45_215	ACAGGTTGTGTCATAGCTTGGAAACACAGCCAA	BtZC45_215 CCACTTGAATACCAAGCTACAAGG
BtZXC21_214	ACAGGTTGTGTCATAGCTTGGAAACACAGCAA	BtZXC21_214 CCGCTTGAATATCAAGCTACAAGG
BtLYRa11_215	ATGGGTTGTGTCCTTGCTTGGAAACACAGGAA	BtLYRa11_215 GGCATGGTTACCAACCTTATAGAA
BtRs4874_214	ACGGGTTGTGTCCTTGCTTGGAAACTAGGAA	BtRs4874_214 GGCATGGCTACCAACCTTATAGAA
BtYN2018B_215	ATGGGTTGTGTCCTTGCTTGGAAACTAGGAA	BtYN2018B_215 GGCATGGCTATCAACCTTATAGAA
HuSARS-Frankfurt-1_2003_215	ATGGGTTGTGTCCTTGCTTGGAAACTAGGAA	HuSARS-Frankfurt-1_2003_215 GGCATGGCTACCAACCTTACAGAA
Cv007-2004_214	ATGGGTTGTGTCCTTGCTTGGAAACTAGGAA	Cv007-2004_214 GGCATGGCTACCAACCTTACAGAA
BtHKU3-12_214	ACTGGCTGTGTAATTGCTTGGAAACTAGCTAA	BtHKU3-12_214 CCAGTAGCATATCAGGCTACTAGGA
Btrec-SARSg_2008_215	ACTGGCTGTGTAATTGCTTGGAAACTAGCTAA	Btrec-SARSg_2008_215 CCAGTAGCATATCAGGCTACTAGGA
BtYN2013_212	ACAGGCTGTGTCATAGCTTGGAAACTAGCTAA	BtYN2013_212 CCCTTGATATCAAGCCACCAGAA
BtRm1/2004_215	ACAGGCTGTGTAATAGCTTGGAAACTAGCTAA	BtRm1/2004_215 CCAGTTGAATACCAAGCCACTAGGA
BtRp3-2004_214	ACTGGTTGCGTAATAGCTTGGAAACTAGCTAA	BtRp3-2004_214 CCGGTTGCTTATCAGGCTACTAGGA
BtGX2013_212	ACTGGCTGTGTAATAGCTTGGAAACTAGCTAA	BtGX2013_212 CCAGTGGCATATCAGGCTACTAGGA
BtRp-Shaanxi2011_213	ACAGGCTGTGTAATAGCTTGGAAACTAGCTAA	BtRp-Shaanxi2011_213 CCACTTGAATATCAGGCTACTAGGA
BtSC2018_214	ACTGGCTGTGTAATAGCTTGGAAACTAGCTAA	BtSC2018_214 CCGGTGGCATATCAGGCTACTAGAA
BtCp-Yun_2011_214	ACAGGTTGCGTAATAGCTTGGAAACTAGCTAA	BtCp-Yun_2011_214 CCACTTGAATACCAAGCTACTAGAA
BtRs_672-2006_208	ACAGGCTGTGTCATAGCTTGGAAACTAGCTAA	BtRs_672-2006_208 CCATTTGAATATCAGGCTACTAGGA
BtYN2018C_215	ACAGGCTGTGTTATTGCTTGGAAACTAGCTAA	BtYN2018C_215 CCATTTGAATATCAGGCTACTAGGA
BtYNLF_31C_215	ACAGGCTGTGTTATTGCTTGGAAACTAGCTAA	BtYNLF_31C_215 CCCGTTGAATATCAAGCCACTAGAA
BtJL2012_212	ATAGGTTGTGTTATAGCTTGGAAACTAGCTAA	BtJL2012_212 CCCTTGAGTACCAAGCCACTAGAA
BtBtKY72_214	ACTGGCTGTGTTTATAGCTTGGAAACTAGCTAA	BtBtKY72_214 GGTGTGGTTACCAACCATACAGAA
BtBM48-31_213	ACAGGTTGTGTAATAGCTTGGAAACTAGCTAA	BtBM48-31_213 GGAAATGGCTTTCAACCATACAGAA
BtHKU5_219	TCT---TCACCAAATGCAATTACAACCTAACAA	BtHKU5_219 GTTACTCTTCTTACAGTGGACT
HuOC43_241	ACAAGTTGTGTCAGTTGATTTATAATTTACCTGC	HuOC43_241 AAGTGCCCCCAAACCTAAATCTTTA
CmMERS_218	CCCACATGTTTGATTTTAGCGACTGTTCCCTCA	CmMERS_218 CCACTTGAAGGTTGGTGGCTGGCTT
HuMERS_172-06_2015_218	CCCACATGTTTGATTTTAGCGACTGTTCCCTCA	HuMERS_172-06_2015_218 CCACTTGAAGGTTGGTGGCTGGCTT
HuTGEV_210	TATTGATTGTATATCTTTTAAATTTGACCACATG	HuTGEV_210 AAACACAGCTATTACAAAGGTGAC
BtHKU9-1_211	TACGGTTGTTTGCATGCATTCTATTTGAATTC	BtHKU9-1_211 CCCTTTCTTAT---GTTTATGGTT
Hu229E_205	-----GCTAGTATTAACACGGGAAA	Hu229E_205 ATACCCGGTGGTTGCGCAATGCC
HuNL63_210	---TCATGGCACATTTATTTAAAGAGTGGCAC	HuNL63_210 GTGCCTGGTAGTTGTAATTTCCG
PiSADS_205	CCTTATGAATGTTTTGGTGGTTCATGGAATGA	PiSADS_205 -----GGTGG
PiPRCV_205	-----GCCACCGCTGTATAAAAACCTGGTAC	PiPRCV_205 TTTGGTGTAAATGTAAGTTTGA

Arbres de la vie

La divergence des caractères

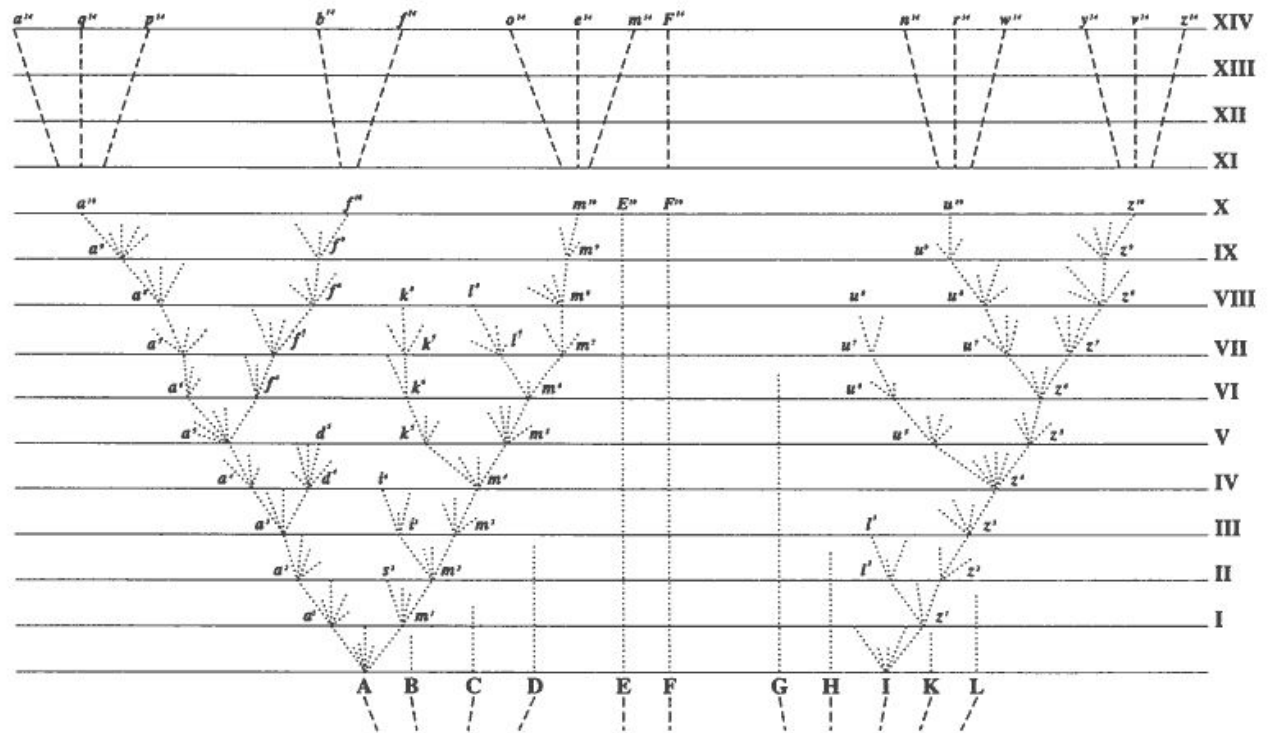
La seule figure de l'Origine des Espèces (C. Darwin, 1859) est une représentation conceptuelle de l'arbre de la vie.



La divergence des caractères

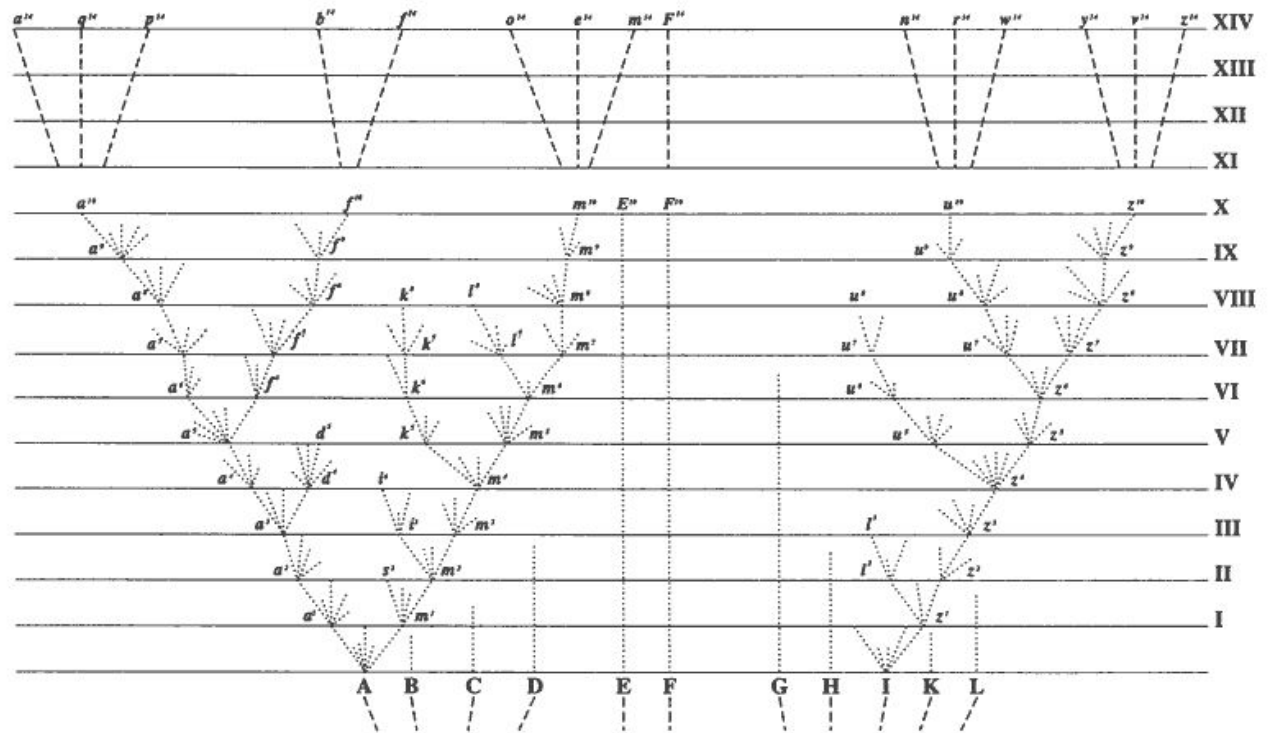
Il s'agit d'un arbre synchrone : chaque niveau horizontal représente un moment donné.

- 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"



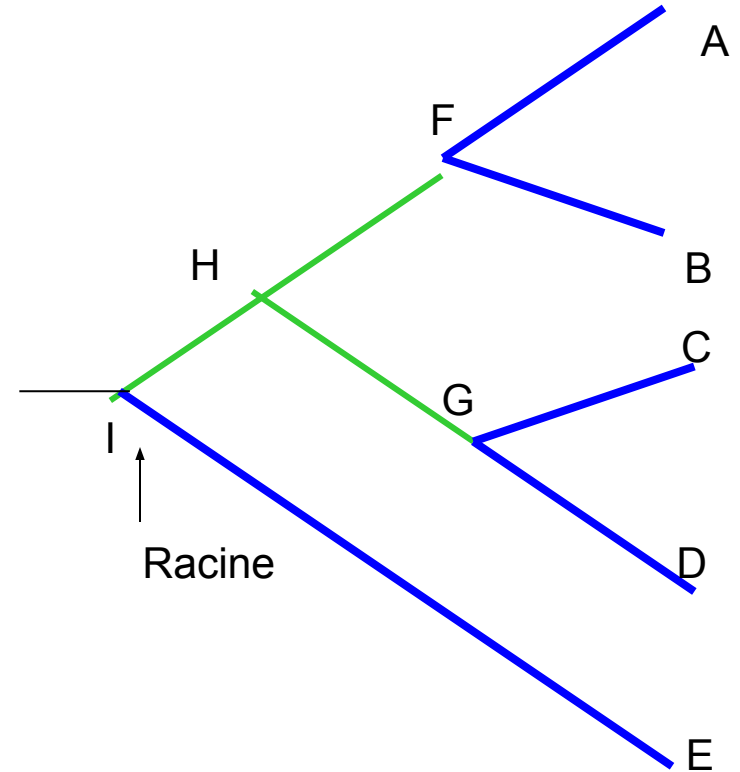
La divergence des caractères

- 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.



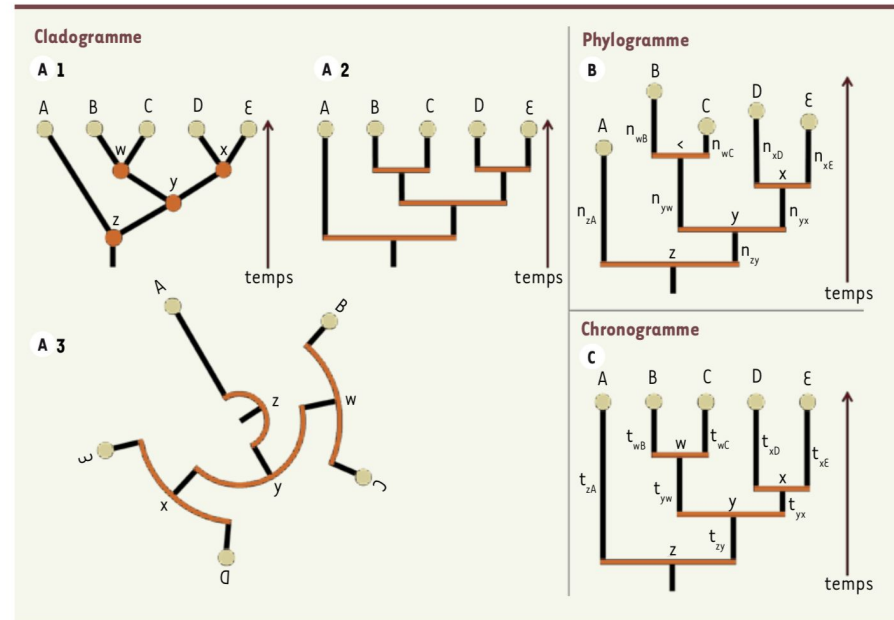
Unités taxonomiques opérationnelles (OTU) et hypothétiques (HTU)

- 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.



Représentations arborescentes des histoires évolutives

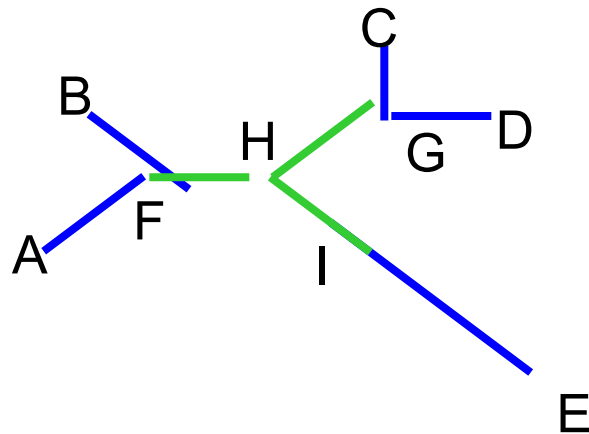
- On représente les histoires évolutives sous forme d'arbre
- Différents types de représentation peuvent être utilisés selon les cas.
 - Bifurcations triangulaires ou rectangulaires
 - Disposition radiale
- Selon les cas, les longueurs des branches représentent
 - Le nombre de divergences (cladogramme)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



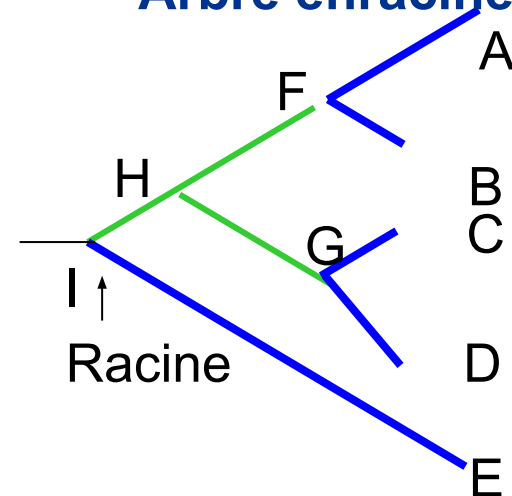
Arbres enracinés ou non enracinés

- 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 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).

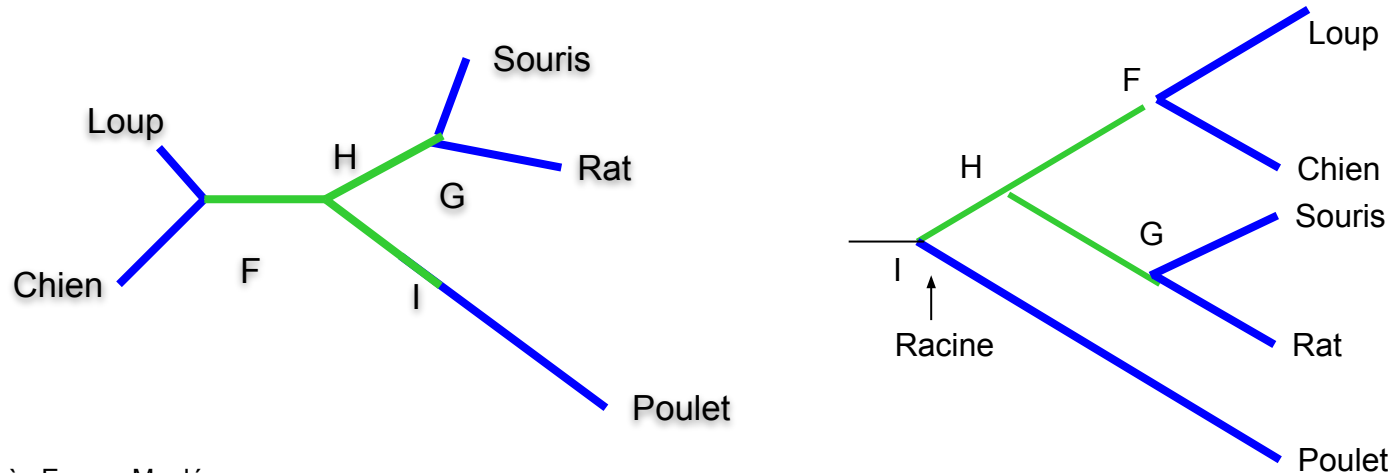
n	Nb arbres enracinés	Nb arbres non-enracinés
2	1	1
3	3	1
4	15	3
5	105	15
6	945	105
7	10,395	945
8	135,135	10,395
9	2,027,025	135,135
10	3.45E+07	2,027,025
11	6.55E+08	3.45E+07
12	1.37E+10	6.55E+08
13	3.16E+11	1.37E+10
14	7.91E+12	3.16E+11
15	2.13E+14	7.91E+12
16	6.19E+15	2.13E+14
17	1.92E+17	6.19E+15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

$$N_R = \frac{(2n! 3)!}{2^{n!2} (n! 2)!}$$

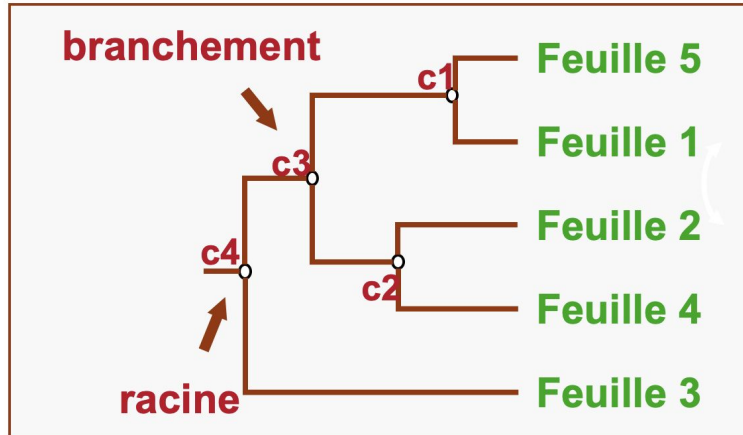
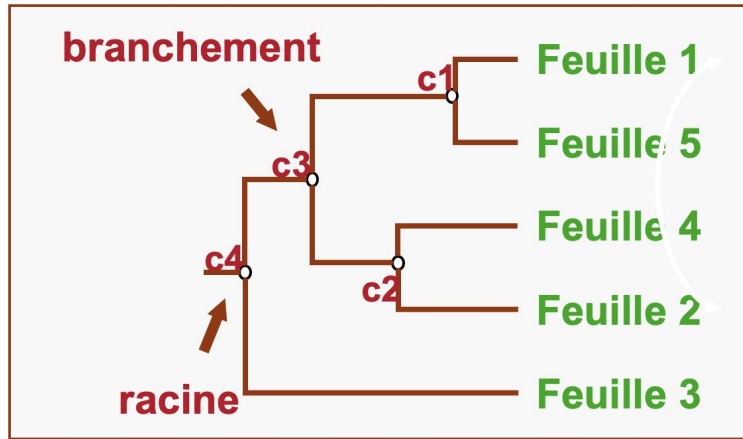
$$N_U = \frac{(2n! 5)!}{2^{n!3} (n! 3)!}$$

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.



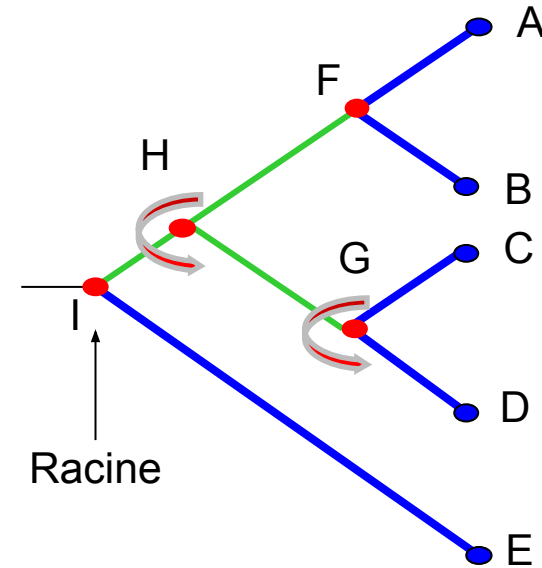
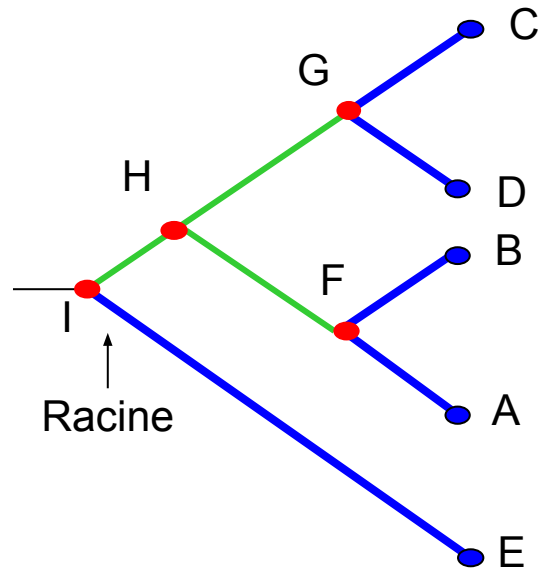
Isomorphisme sur un arbre



- Dans un arbre, les deux enfants de chaque branche peuvent être interchangés.
- Le résultat est un arbre **isomorphe**, considéré équivalent à l'arbre initial.
- Les deux arbres de gauche 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

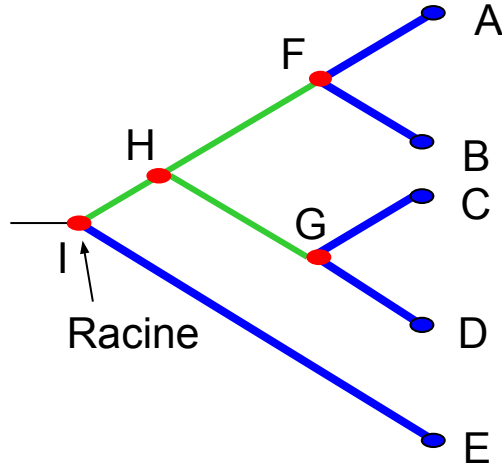
- Il faut éviter le piège d'évaluer les distance entre feuilles sur base de leur proximité verticale.
 - Les structures ci-dessous sont absolument identiques.
 - Pourtant les feuilles B et D semblent voisines sur le graphe de gauche, et éloignées sur celui de droite.
- Pour évaluer la distance entre 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).



d'un arbre phylogénétique

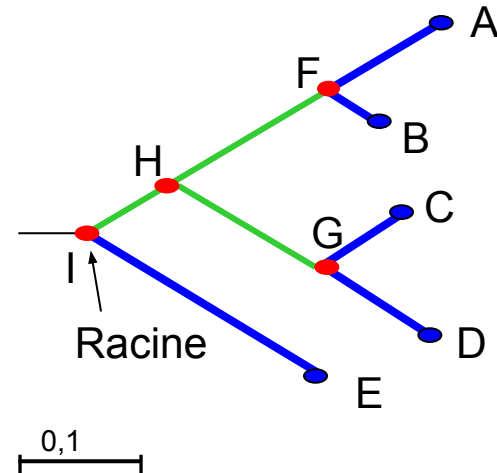
■ Cladogramme

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



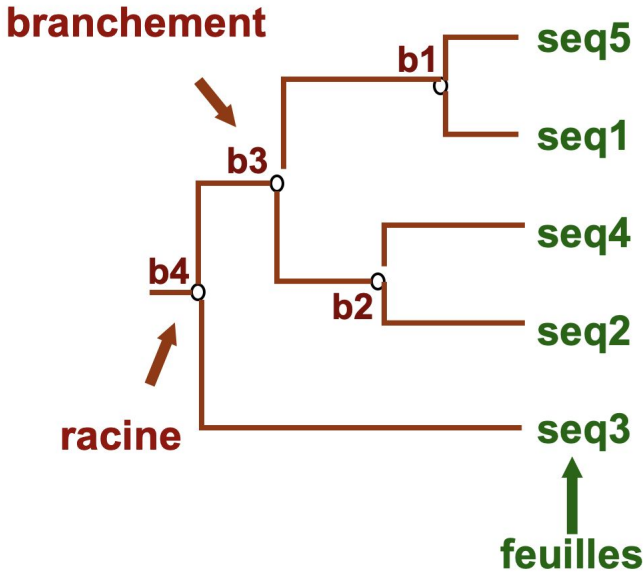
■ Phylogramme

- 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).

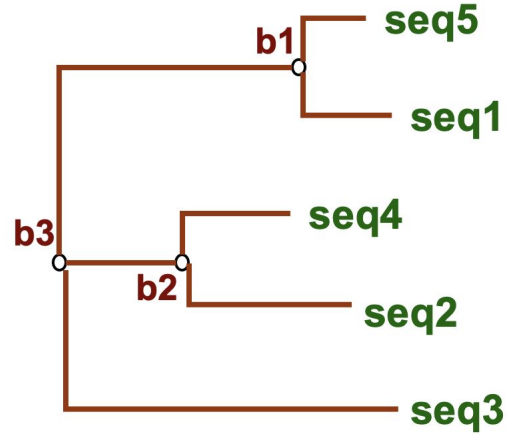


Calcul de la distance sur un arbre

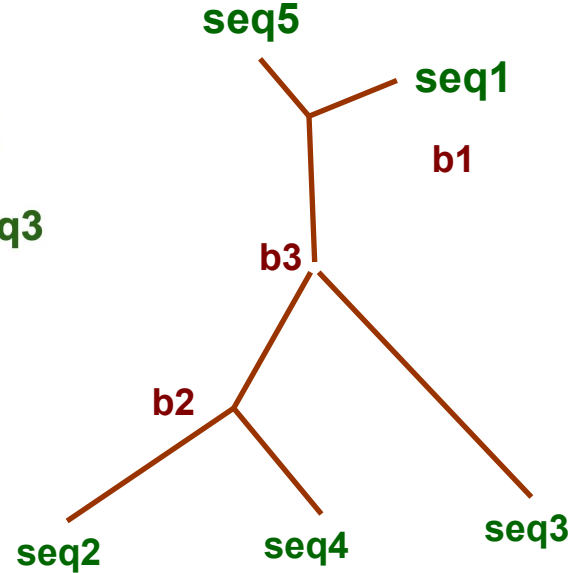
Arbre enraciné



Arbre non-enraciné



Arbre non-enraciné



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

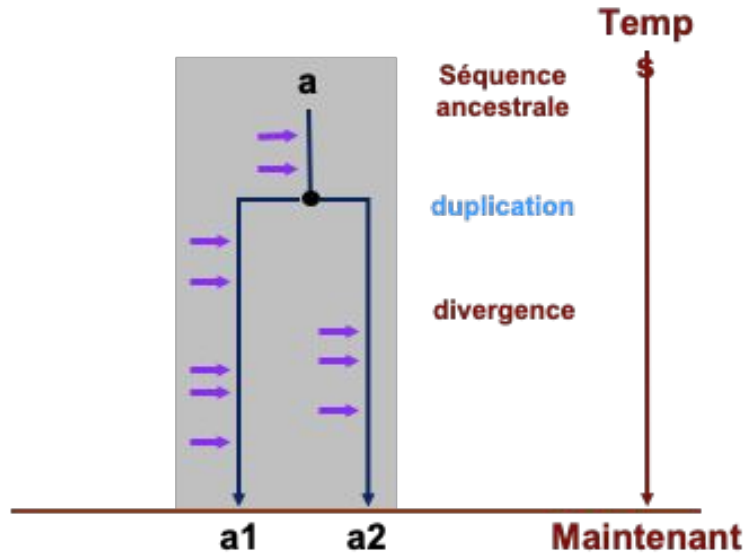
Concepts: homologie, paralogie, orthologie

- 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**

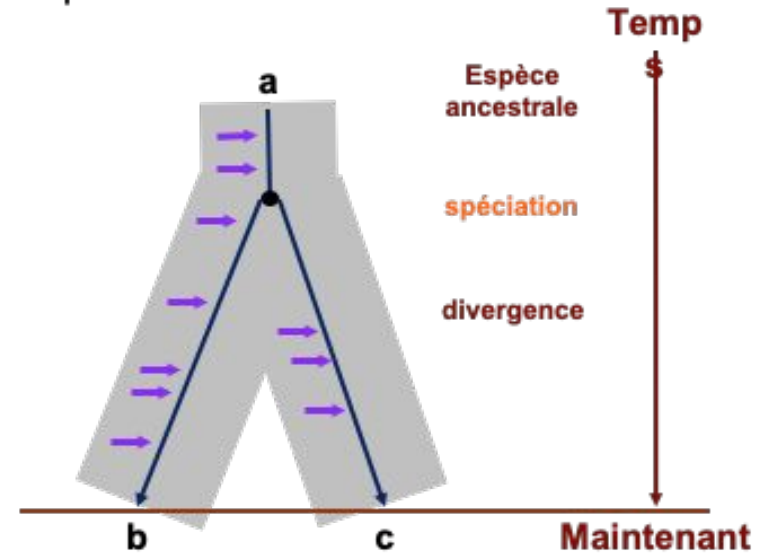
Scénarios évolutifs

- 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.

Duplication



Spéciation



- 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
 - 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.

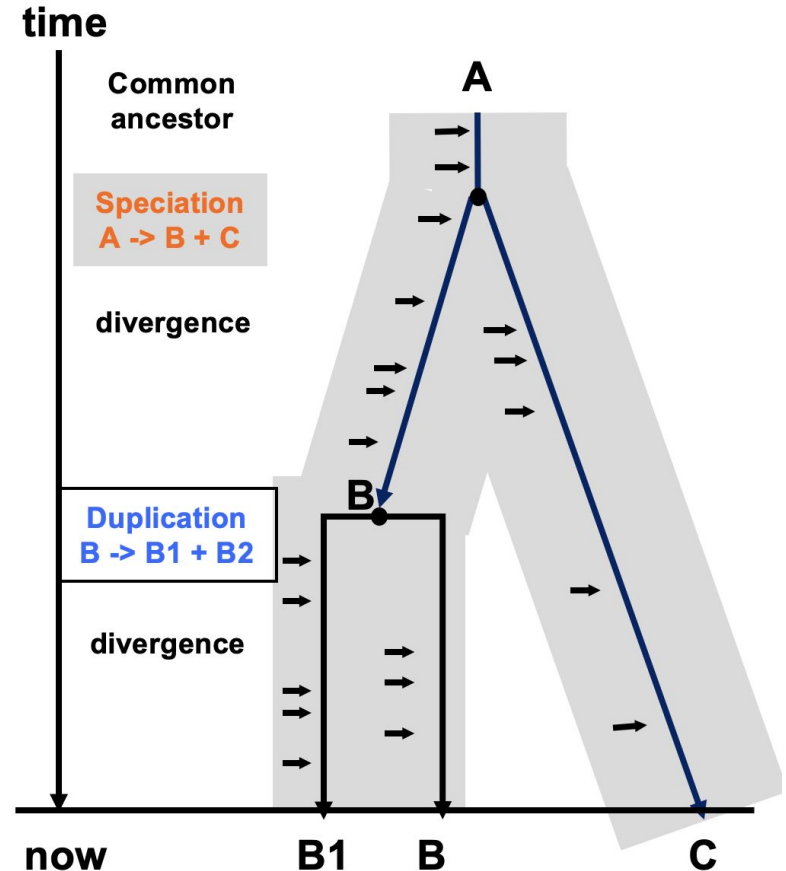


L'homologie est une relation logique (soit vraie, soit fausse).

- 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

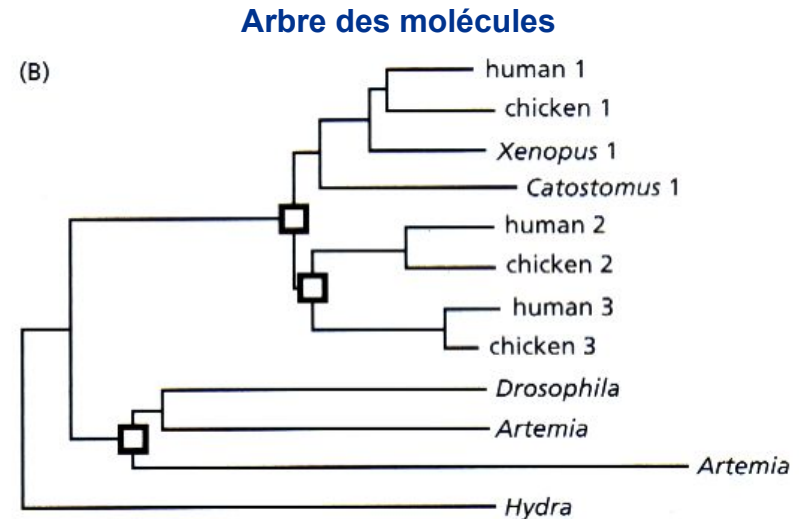
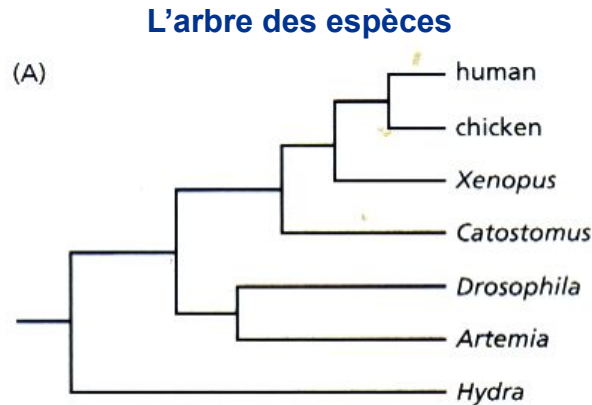
- Zvelebil & Baum (2000) fournissent une définition claire et opérationnelle des concepts d'orthologie et paralogie.
 - **Orthologues**: séquences dont le dernier ancêtre commun précède immédiatement un événement de spéciation.
 - **Paralogues** séquences dont le dernier ancêtre commun précède immédiatement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

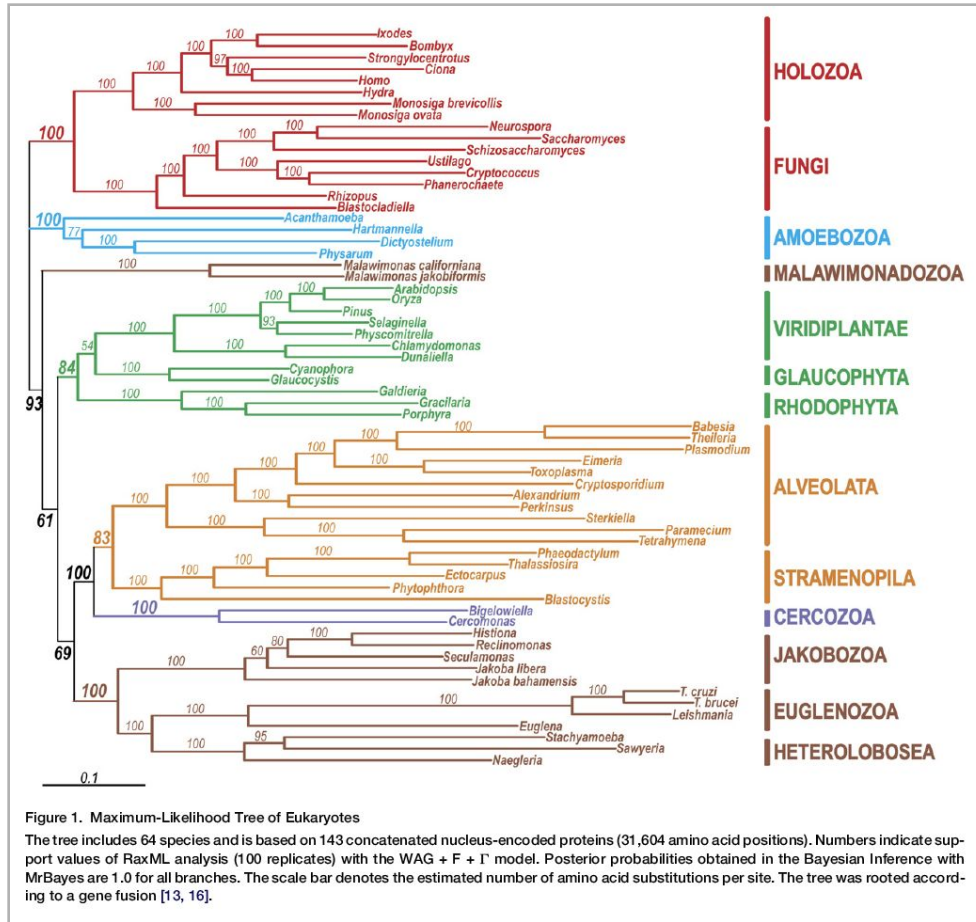


Approches d'inférence phylogénétique

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.



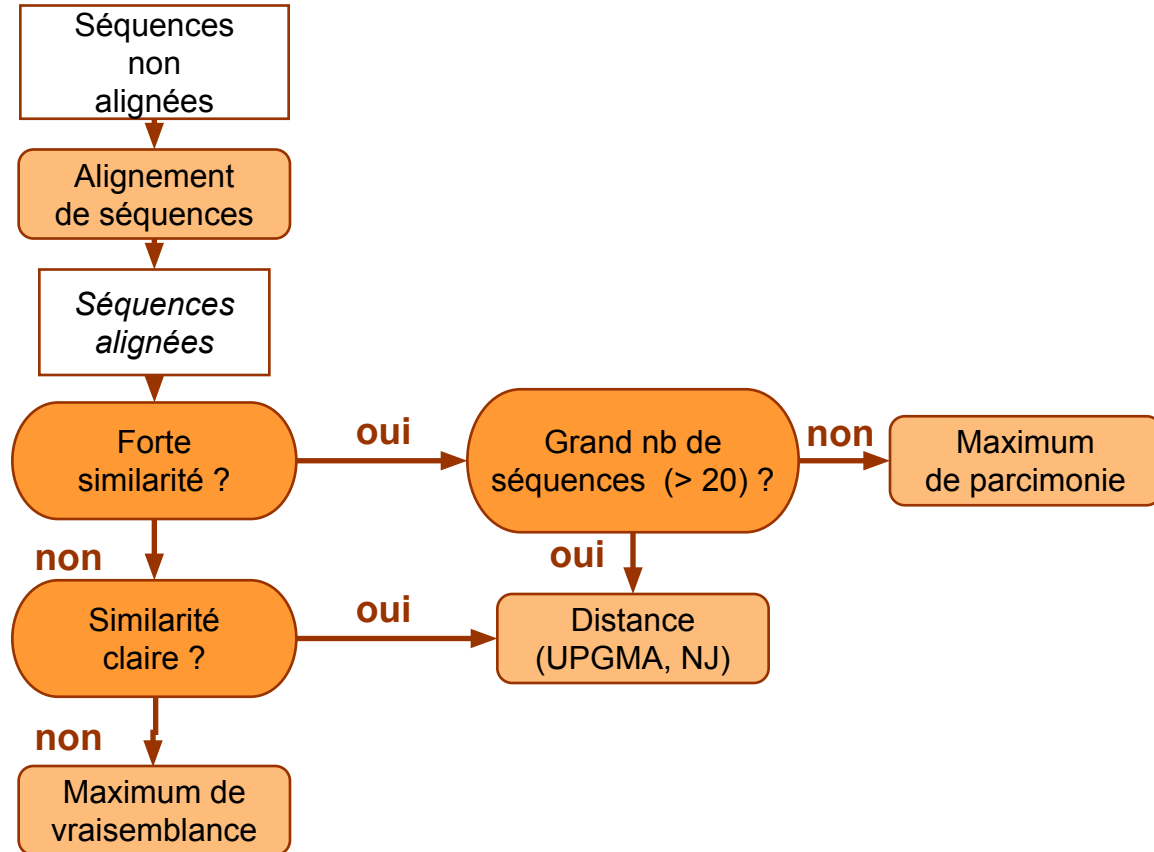


- En phylogénie moléculaire, une approche classique consiste à se concentrer sur un gène considéré comme représentatif, et à construire un arbre sur base de la divergence de séquence de ce gène.
- Ces approches peuvent maintenant être généralisées en comparant les séquences de plusieurs centaines de 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.



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)

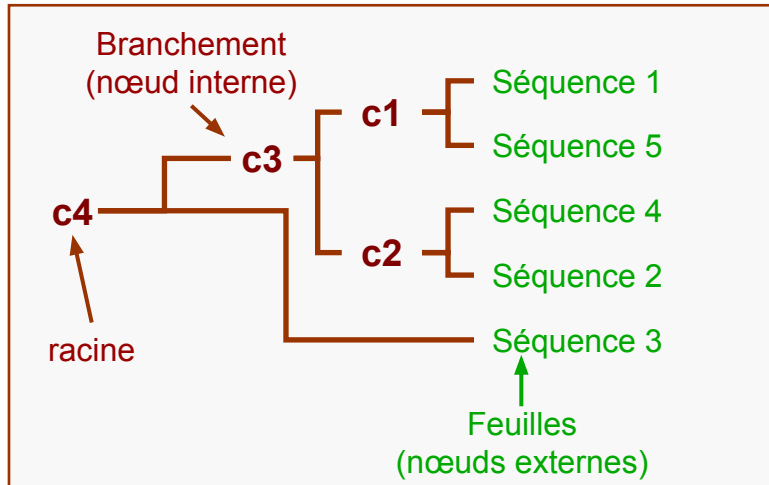


Principe de la construction de l'arbre

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

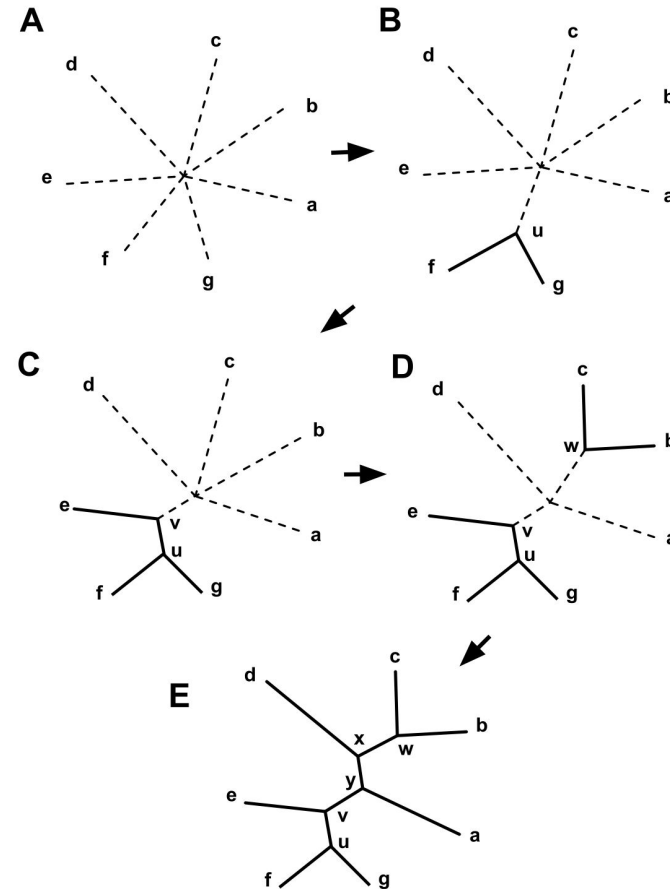
Arbre



- 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.
- Algorithmes
 - 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

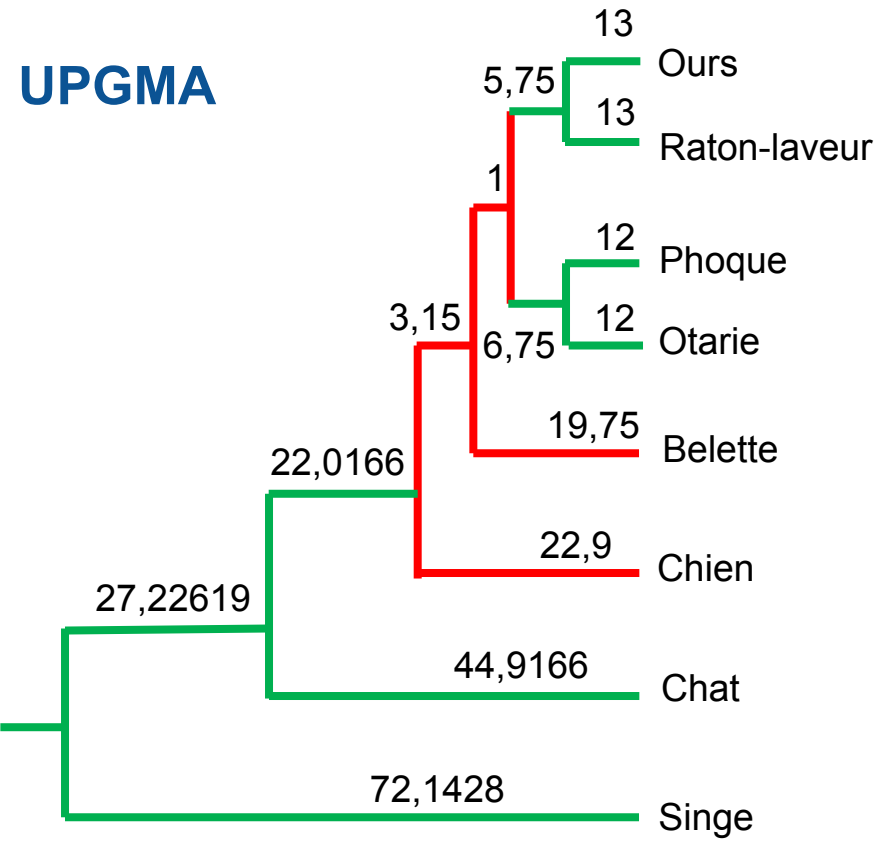


Propriétés de la méthode NJ

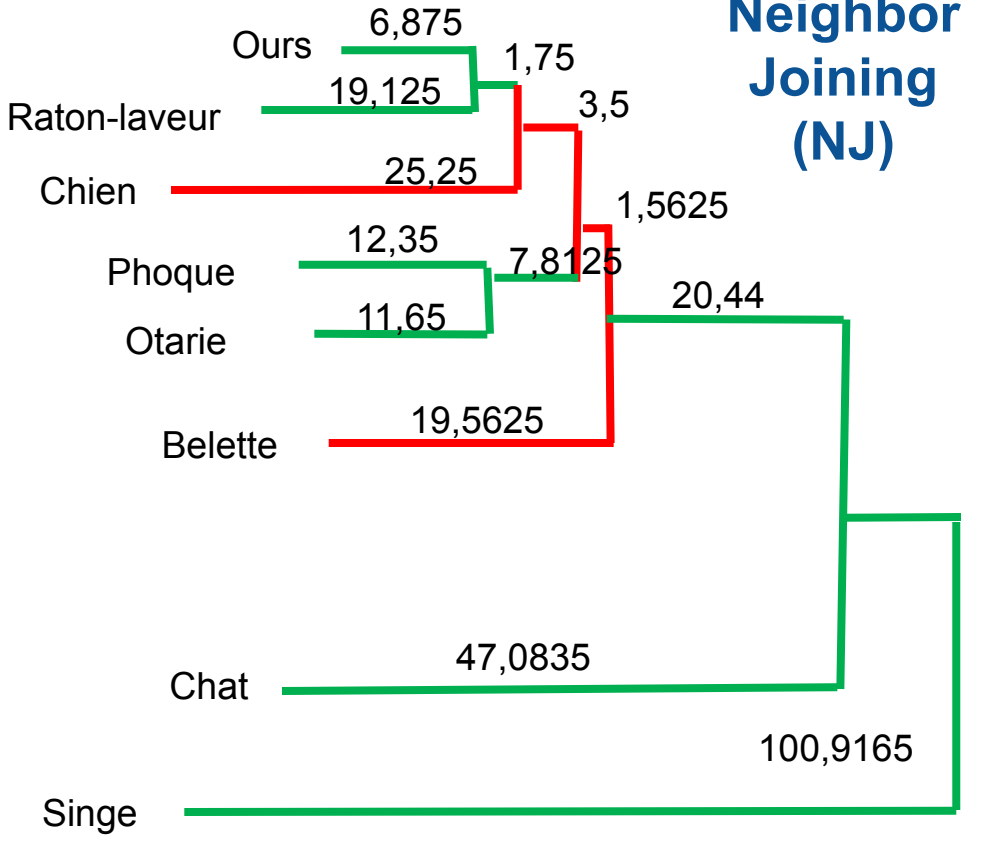
- Méthode rapide et simple qui permet de travailler avec un très grand nombre de taxons.
- Les arbres ne sont pas enracinés.
- Les longueurs des branches sont informatives (phylogramme).
- Bonne approximation de la méthode du minimum d'évolution (l'arbre le plus court).
- Retrouve l'arbre vrai si la 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.

Comparison UPGMA - Neighbour Joining

UPGMA



Neighbor Joining (NJ)

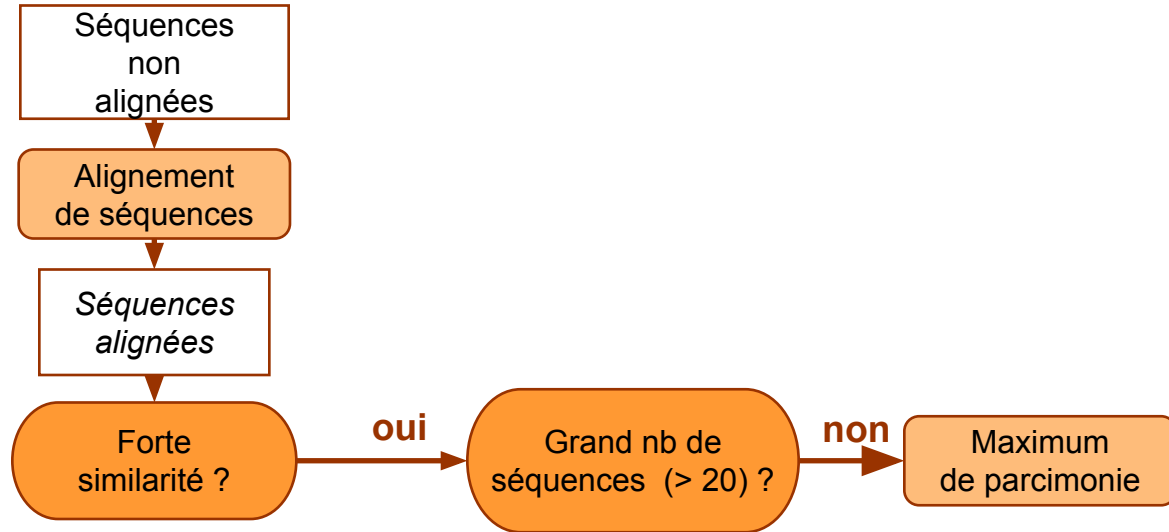


*Inférence phylogénétique par la méthode
du maximum de parcimonie*

Inférence phylogénétique par comparaison de séquences

■ Approches alternatives

- Maximum de parcimonie
- Distance
- Maximum de vraisemblance

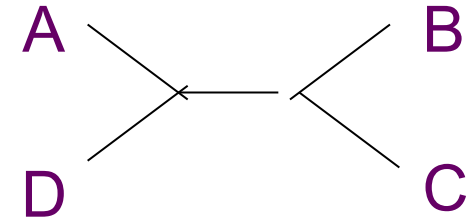
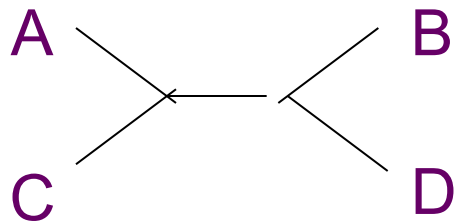
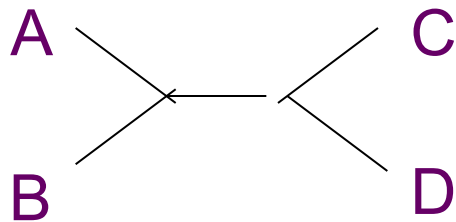


- Principe:
 - ❑ Identifier la topologie T qui implique le plus petit nombre de changements évolutifs suffisant à rendre compte des différences observées entre les OTU étudiées.
 - ❑ Utilise des états de caractères discrets => L'arbre le plus parcimonieux => plus court chemin conduisant aux états de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

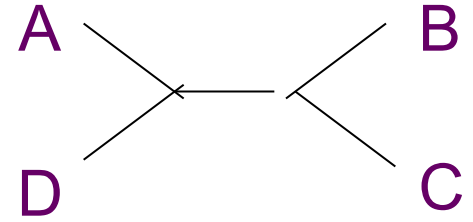
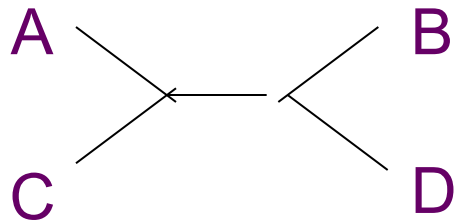
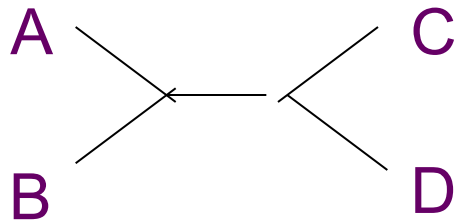
Espèces	Séquences								
	1	2	3	4	5	6	7	8	9
A	A	A	G	A	G	T	T	C	A
B	A	G	C	C	G	T	T	C	T
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	T



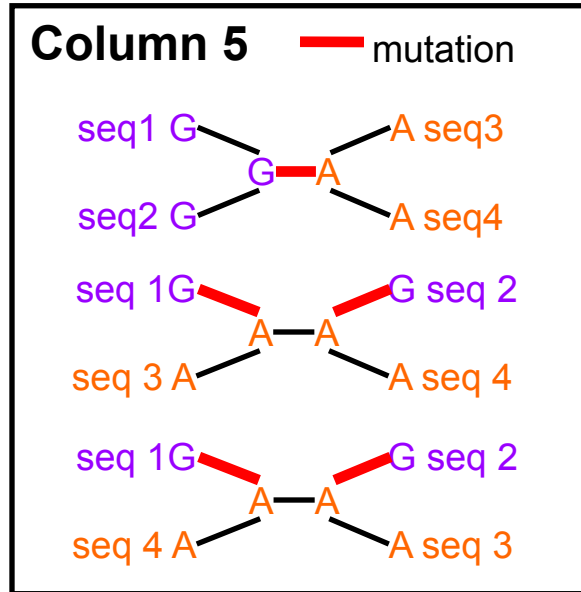
Maximum de parcimonie – classification des sites

- 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



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

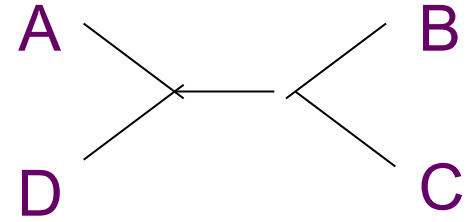
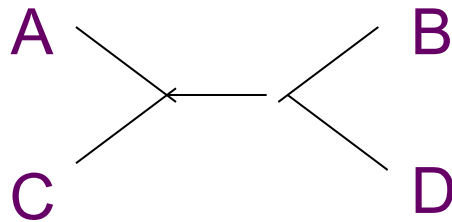
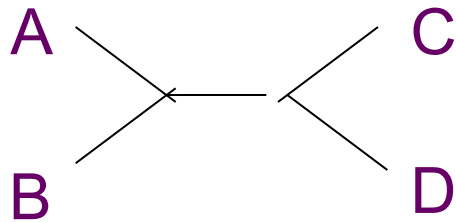


Adapted from Mount (2000)

- 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

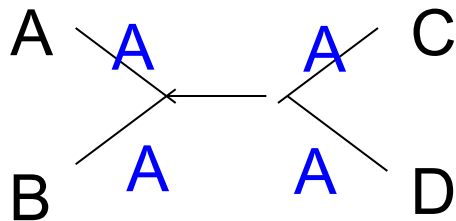
Pour un caractère donné, on compte le nombre de changements évolutifs (CE) pour chaque topologie possible.

Étude du caractère n°1

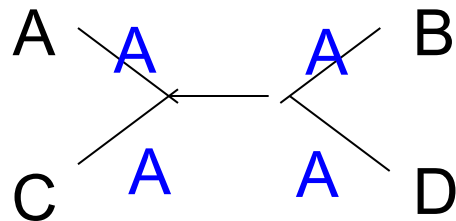
	1	2	3	4	5	6	7	8	9
A	A	A	G	A	G	T	T	C	A
B	A	G	C	C	G	T	T	C	T
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	T

Caractère constant (même état de caractère à tous les sites).

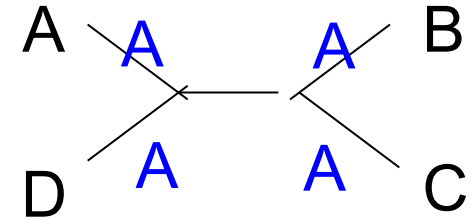
Caractère **non informatif** : ne favorise aucune topologie par rapport à une autre.



Nb CE= 0



Nb CE= 0



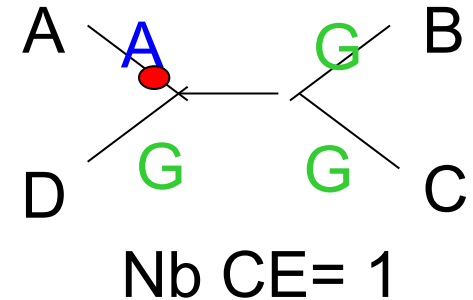
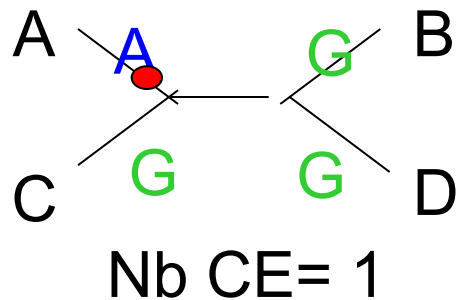
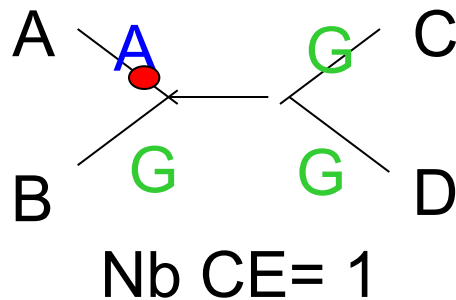
Nb CE= 0

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

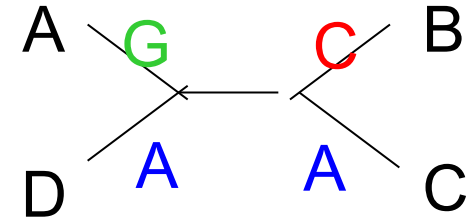
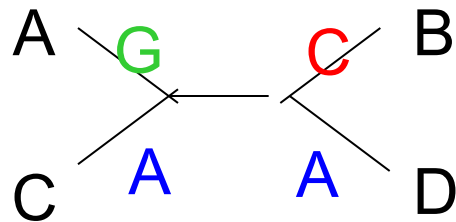
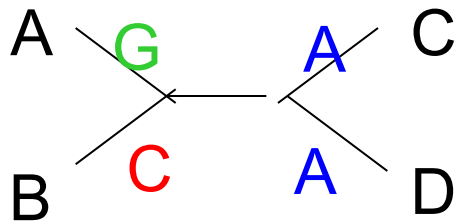
Caractère **variable** mais **non informatif**.

Caractère ne favorisant aucune topologie par rapport à une autre.

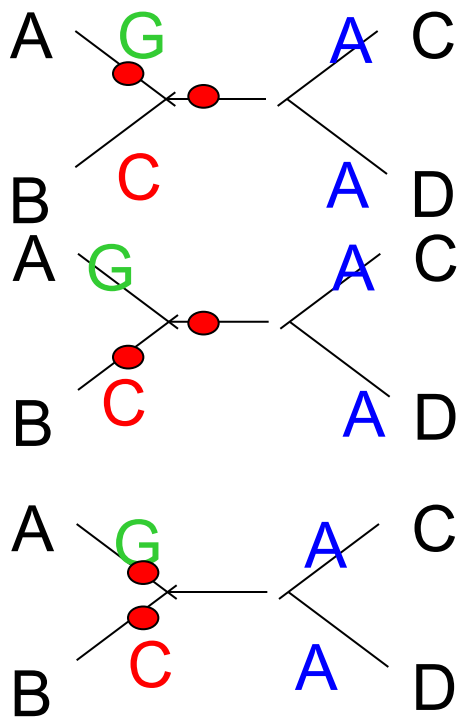


Étude du caractère n°3

	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



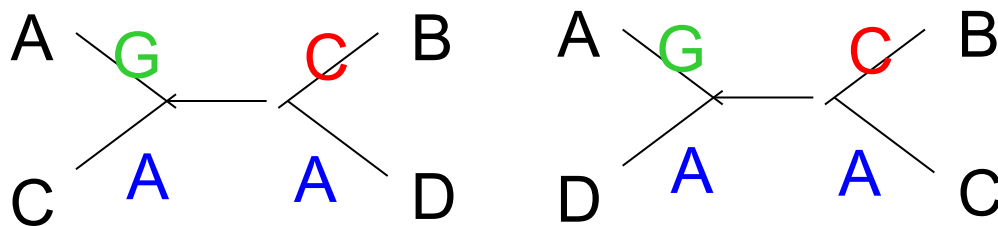
Étude du caractère n°3



Nb CE = 2

Arbre 1

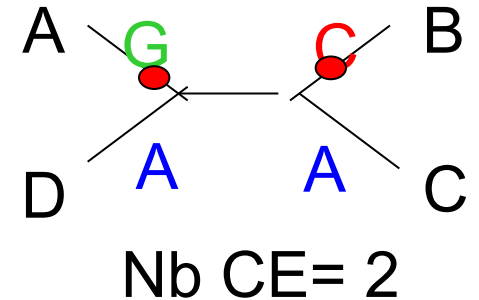
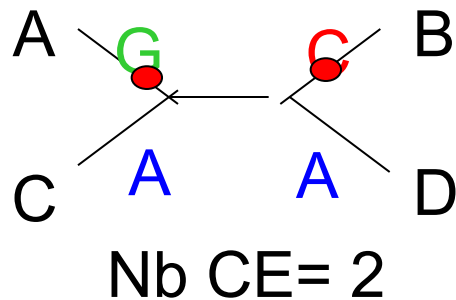
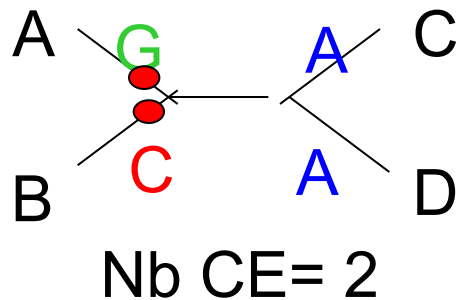
	1	2	3	4	5	6	7	8	9
A	A	A	G	A	G	T	T	C	A
B	A	G	C	C	G	T	T	C	T
C	A	G	A	T	A	T	C	C	A
D	A	G	A	G	A	T	C	C	T



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.

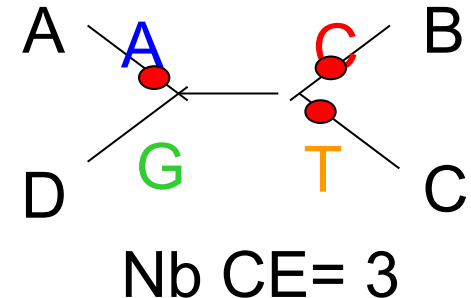
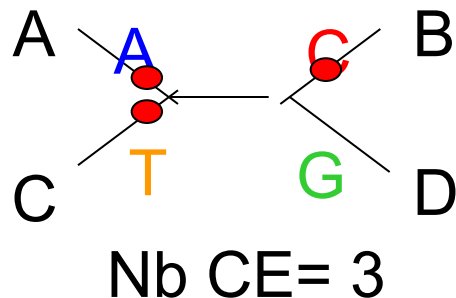
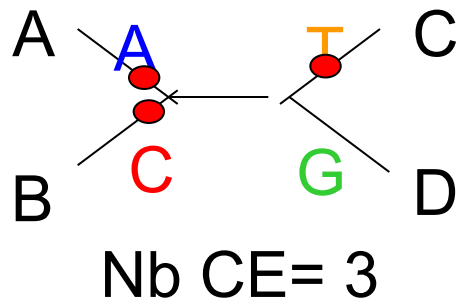


Étude du caractère n°4

	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

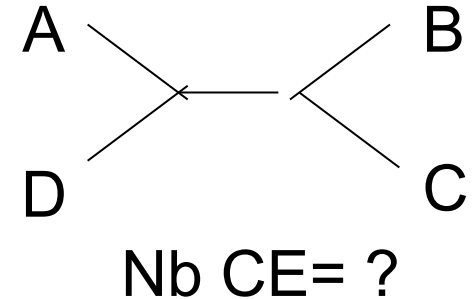
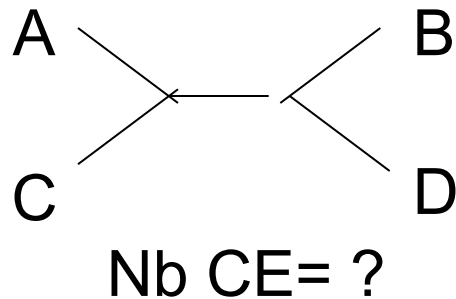
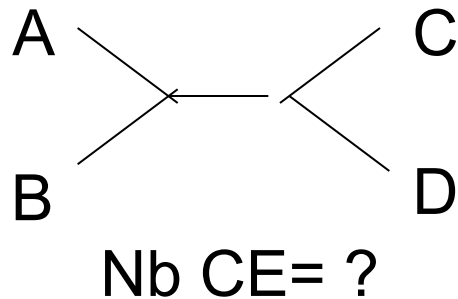
Caractère **variable** mais **non informatif**.

Caractère ne favorisant aucune topologie par rapport à une autre.



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

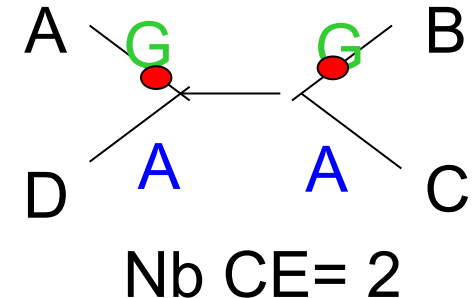
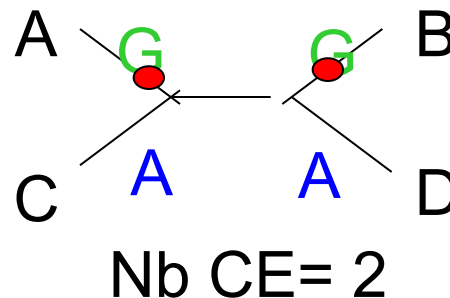
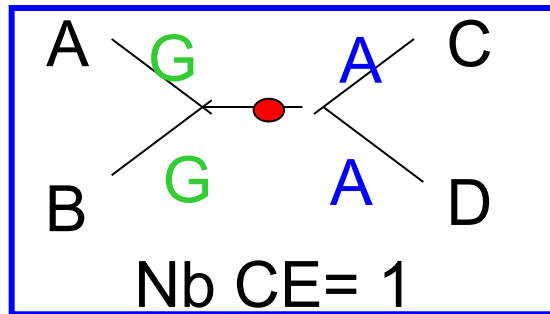


Étude du caractère n°5

	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.



- 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:

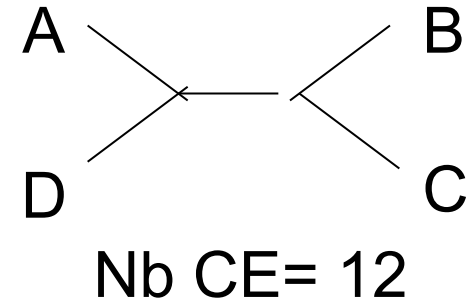
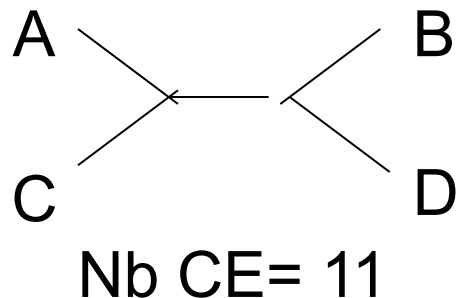
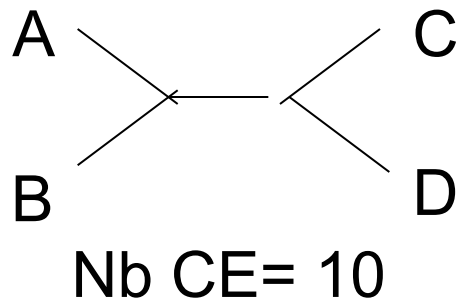
$$T1 = 0+1+2+3+1+0+1+0+2=10$$

$$T2 = 0+1+2+3+2+0+2+0+1=11$$

$$T3 = 0+1+2+3+2+0+2+0+2=12$$

	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



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

$$N_R = \frac{(2n! 3)!}{2^{n!2} (n! 2)!}$$

$$N_U = \frac{(2n! 5)!}{2^{n!3} (n! 3)!}$$

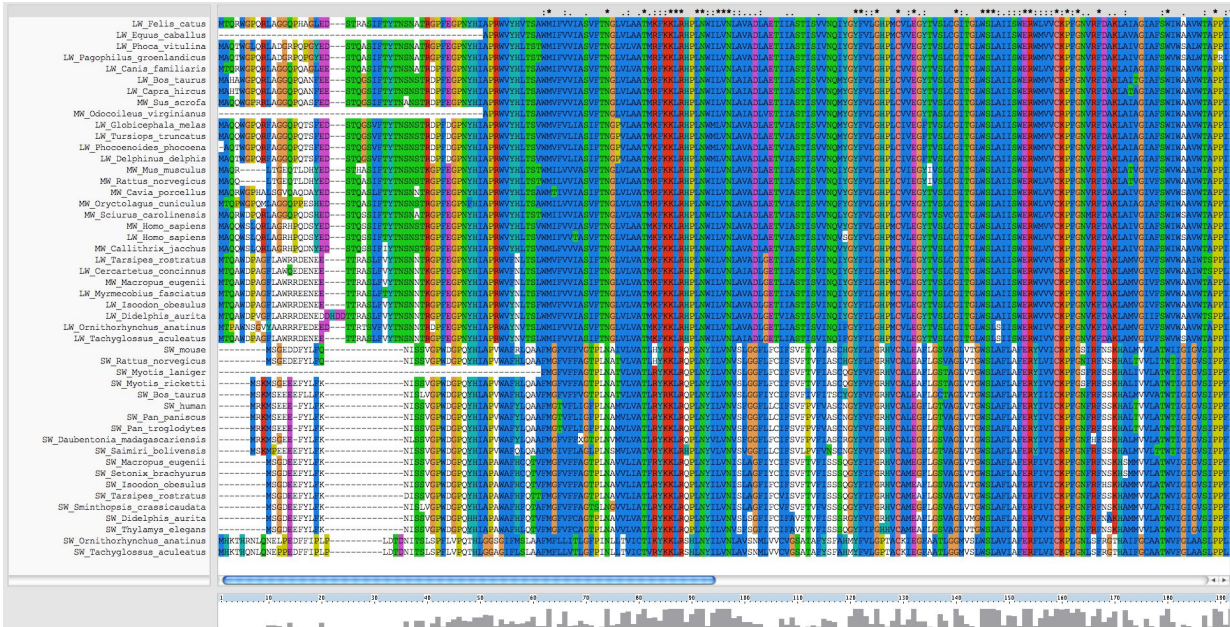
- Résumé des méthodes d'inférence d'arbre implémentées dans PHYLIP.
- Note: le temps de calcul augmente drastiquement quand on passe de méthodes de voisinage (NJ, UPGMA: temps quadratique) aux méthodes de kitch ou fitch (puissance 4 de la longueur des séquences).

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 branches

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

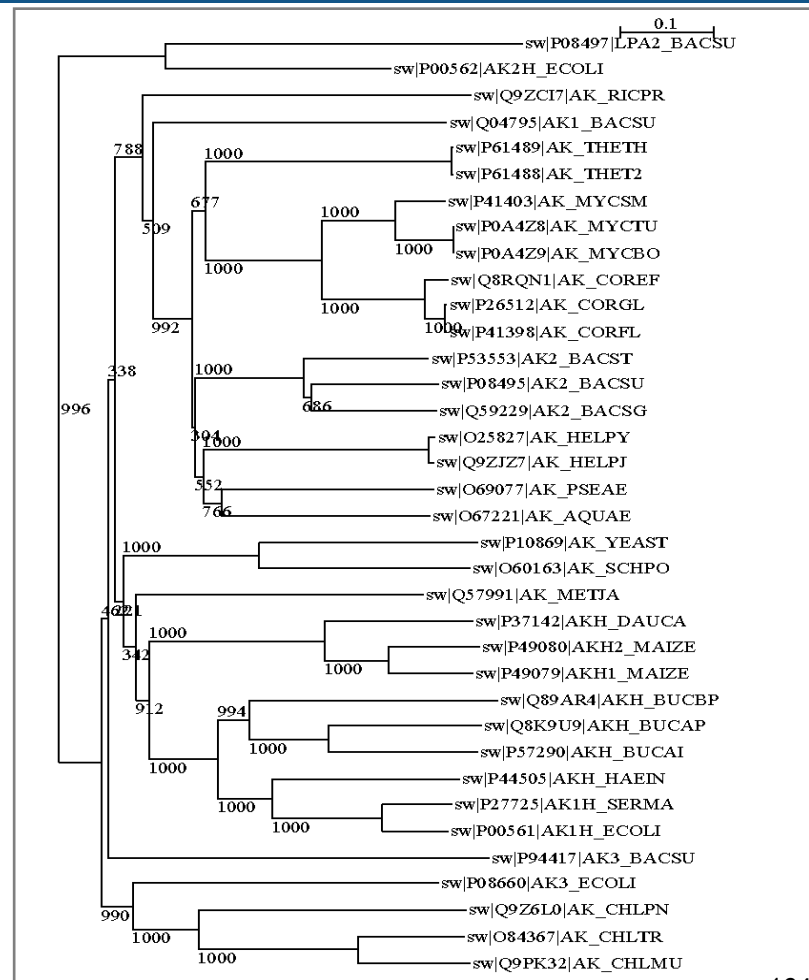
Quelle est la fiabilité d'un arbre inféré ?

- 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 ?



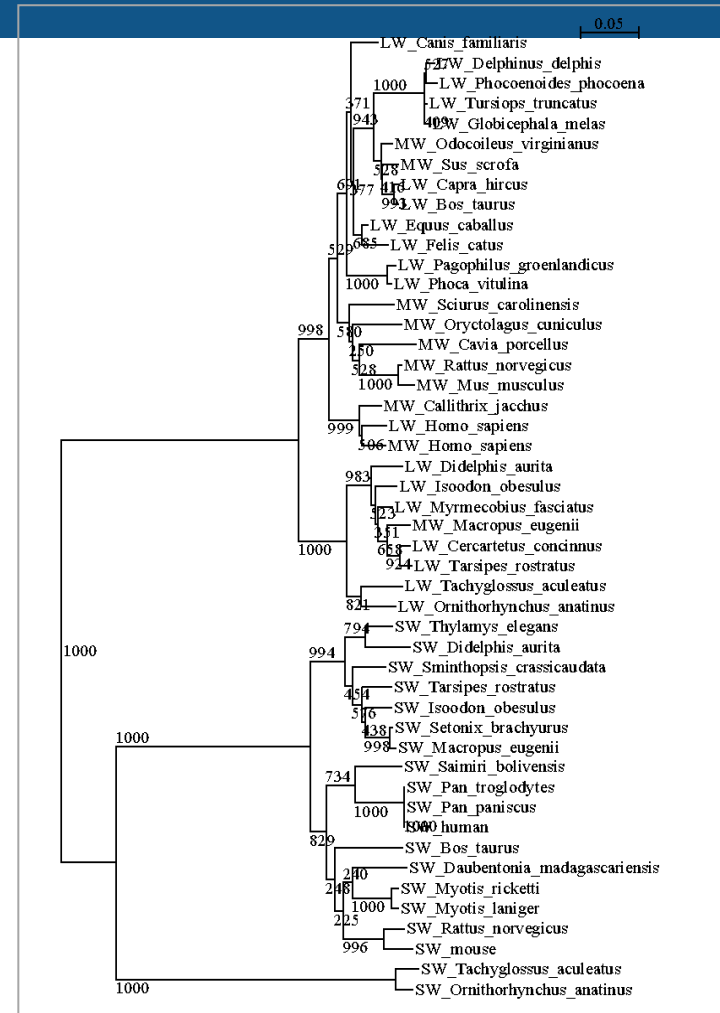
Bootstrapping

- 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.

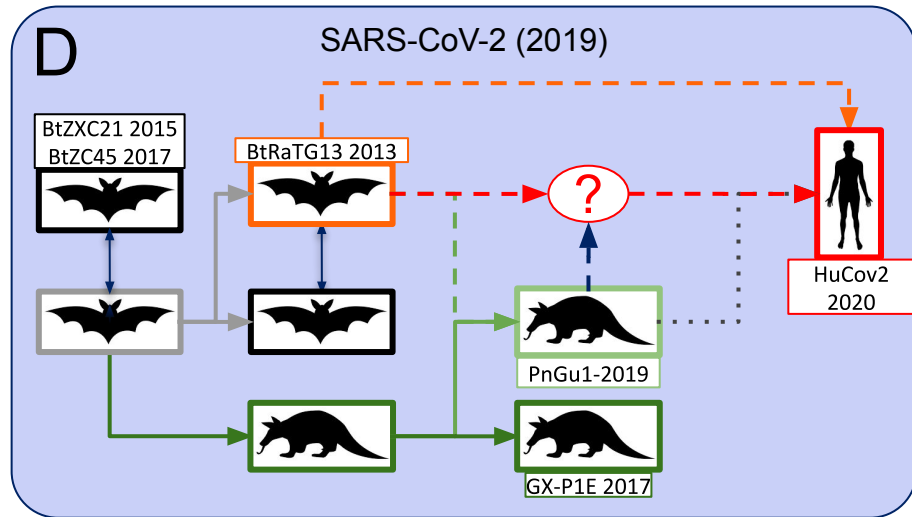
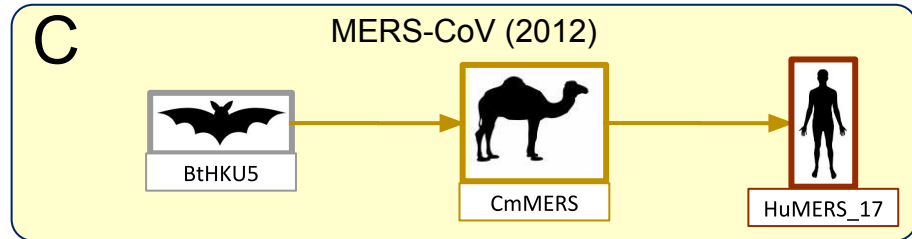
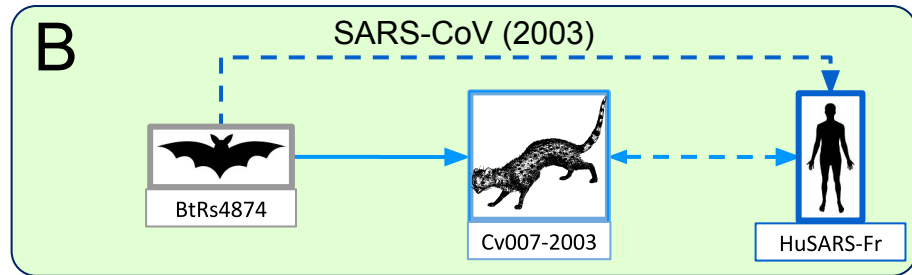
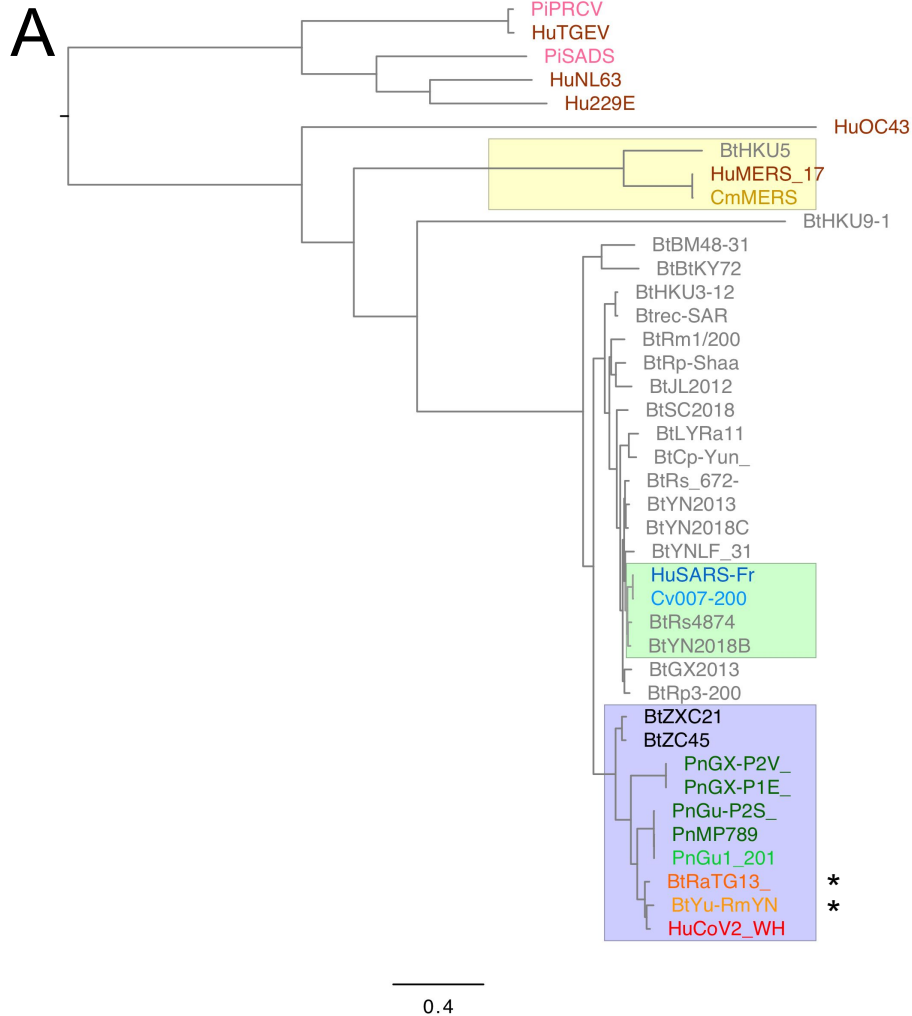


Bootstrapping

- 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).

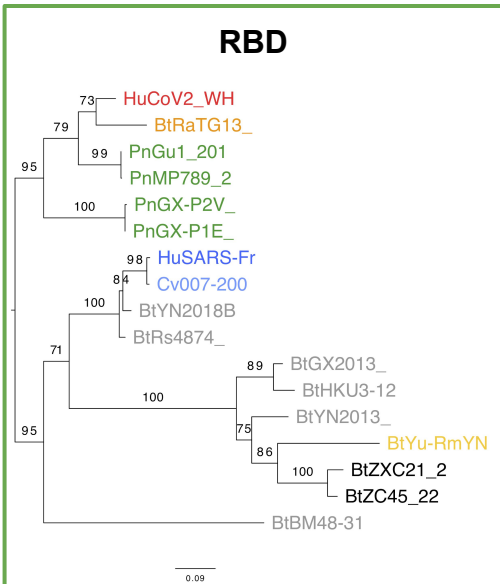
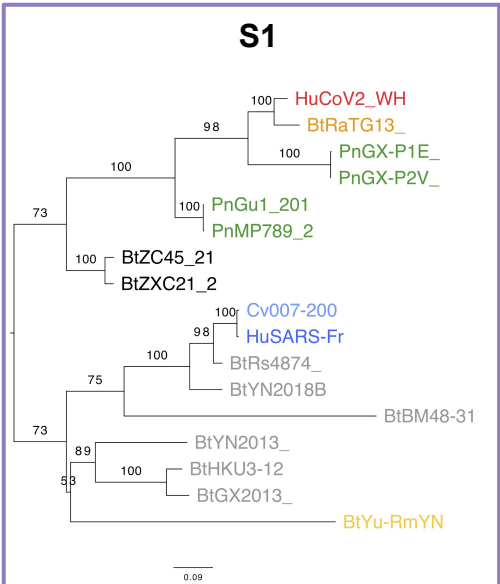
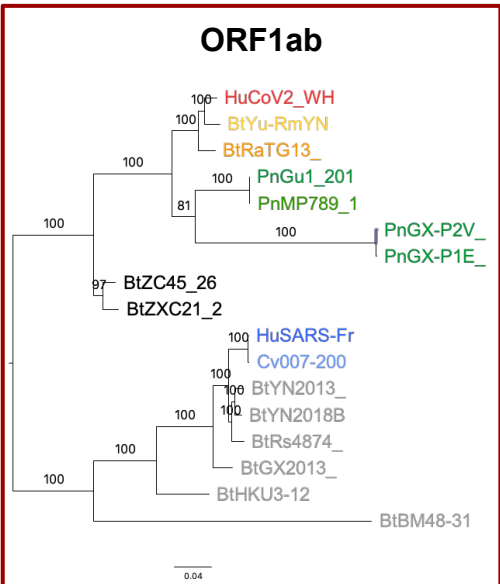
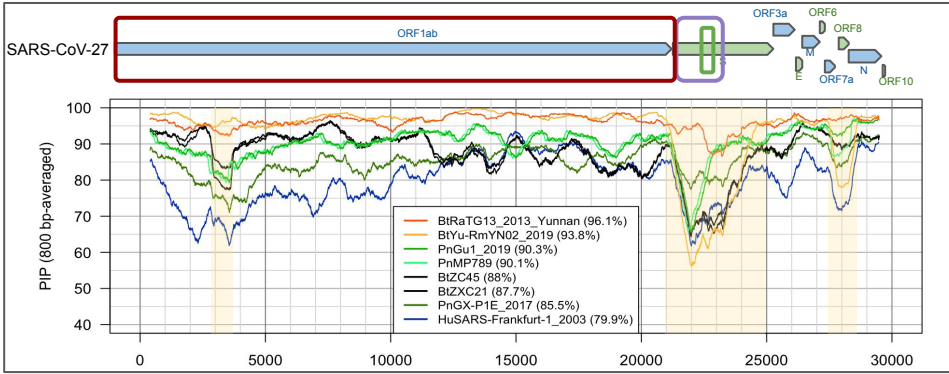


Phylogénie des coronavirus

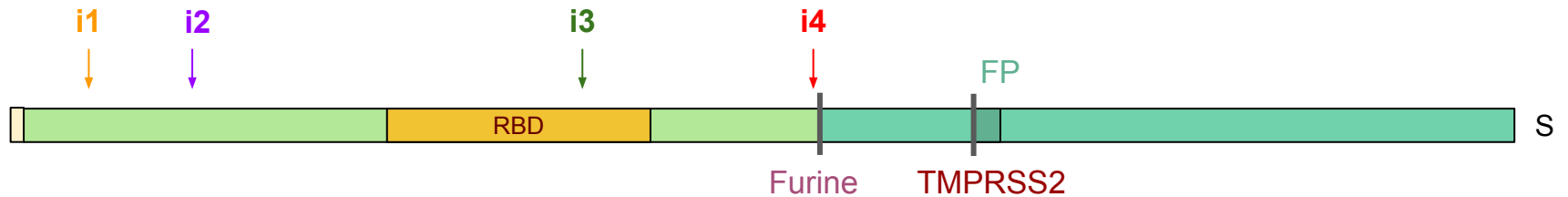


Comparaison entre coronavirus - gène S

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).

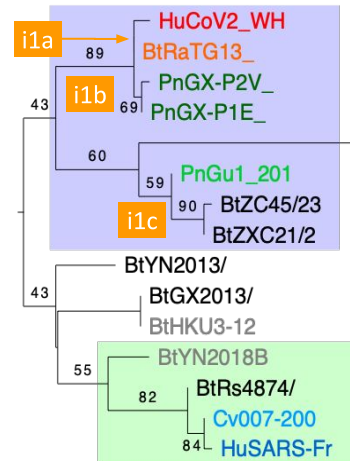


Insertion 1



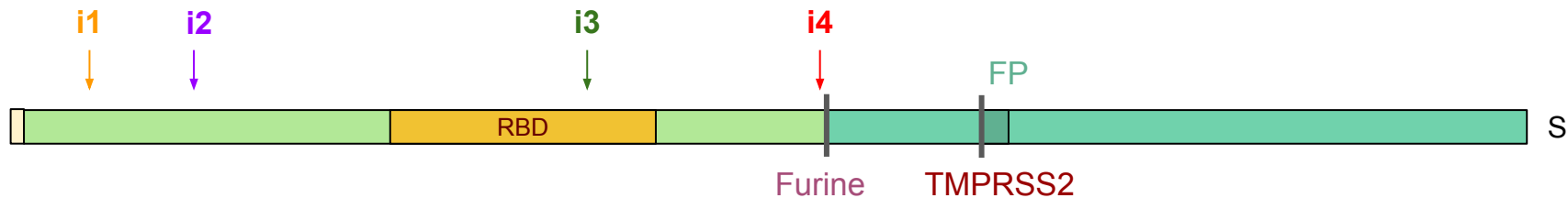
Strain	Sequence
HuCoV2_WH01_2019	WFHAIHVSGTNGTKRFDNP
BtRaTG13_2013_Yunnan	WFHAIHVSGTNGIKRFDNP
PnGX-P1E_2017	WFNTI--NYQGGFKKFDNP
PnGX-P2V_2018	WFNTIHLNYQGGFKKFDNP
PnGu1_2019	WYYAL-TKTNSAEKRVDNP
BtZC45	WYYSL-TTNNAATKRFDNP
BtZXC21	WYYSL-TTNNAATKRFDNP
BtYu-RmYN02_2019	WYNFW-----NQAYTSR
BtYN2013	QFFTQ-----GTNIDNP
BtHKU3-12	QYFSL-NVSDRYTYFDNP
BtGX2013	QYFSL-NVSDRYTYFDNP
BtYN2018B	RFITF-----GLNFDNP
BtRs4874	GFHTI-----NHRFDNP
Cv007-2004	GFHTI-----NHTFDNP
HuSARS-Frankfurt-1_2003	GFHTI-----NHTFGNP

.....80.....90

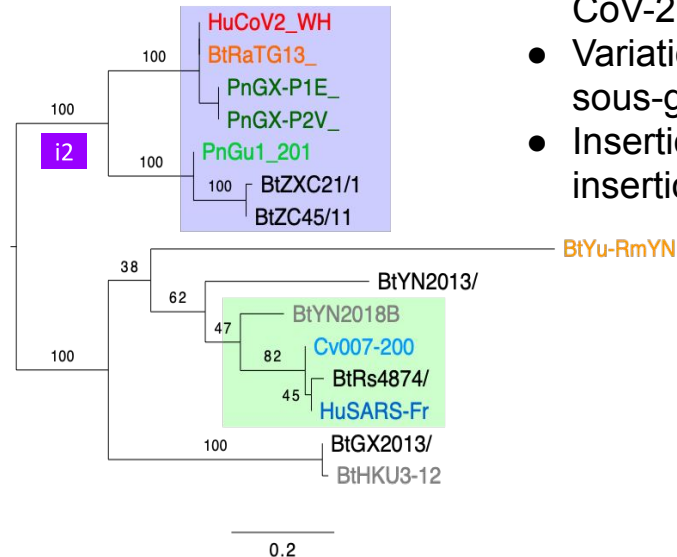


- Insertion retrouvée dans tous les génomes de la branche CoV-2.
- Variations selon les sous-groupes.
- Insertions indépendantes ou insertion suivie de mutations ?

Insertion 2

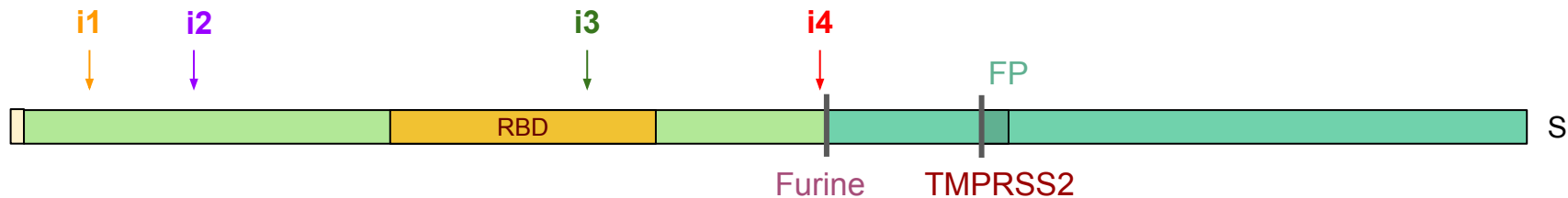


HuCoV2_WH01_2019	YYHKNNKSWMESEFRVYSS	.	::
BtRaTG13_2013_Yunnan	YYHKNNKSWMESEFRVYSS		
PnGX-P1E_2017	YYHNNKNTWVENEFRVYSS		
PnGX-P2V_2018	YYHNNKNTWVENEFRVYSS		
PnGu1_2019	YYH-NNKTWSTREFAVYSS	i2	
BtZC45	YYH-NNKTWSIREFAVYSS		
BtZXC21	YYH-NNKTWSIREFAVYSF		
BtYu-RmYN02_2019	AGGQOTSAA-----VYIS		
BtYN2013	FKSNNSQLSH-----LFS		
BtHKU3-12	SRGTQONAW-----VYQS		
BtGX2013	SRGTQONSW-----VYQS		
BtYN2018B	LRSNNTQIPSY----IFNN		
BtRs4874	SKPTGTQTHM----IFDN		
Cv007-2004	SKPMGTQTHM----IFDN		
HuSARS-Frankfurt-1_2003	SKPMGTQTHM----IFDN		
160.....170		

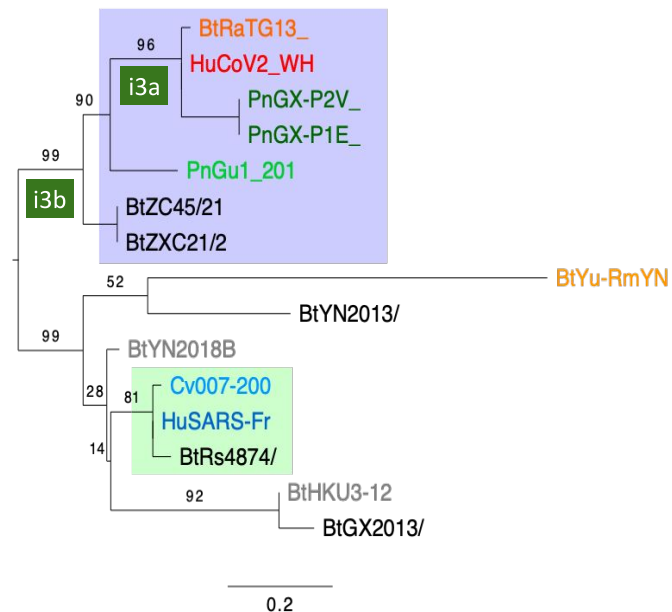


- Insertion retrouvée dans tous les génomes de la branche CoV-2.
- Variations selon les sous-groupes.
- Insertions indépendantes ou insertion suivie de mutations ?

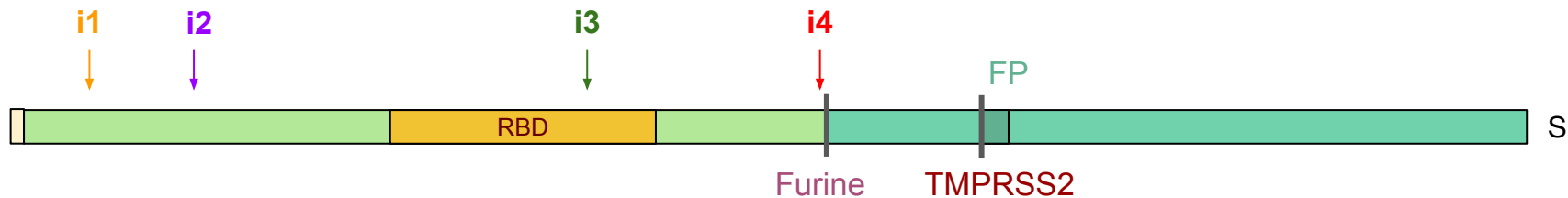
Insertion 3



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

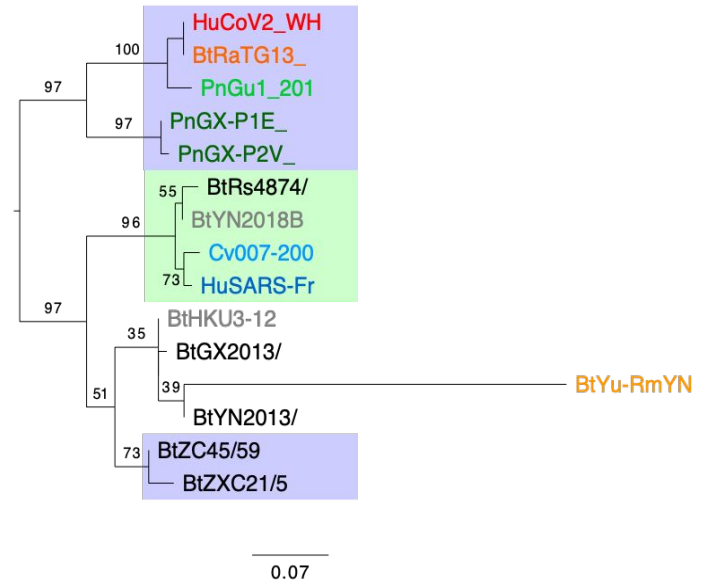


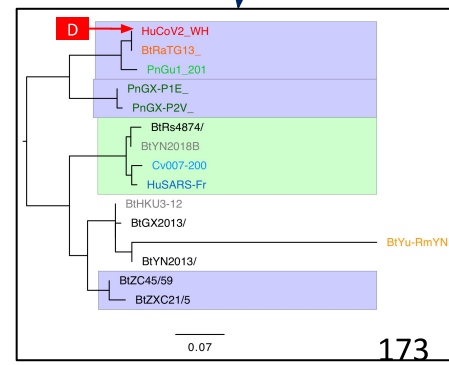
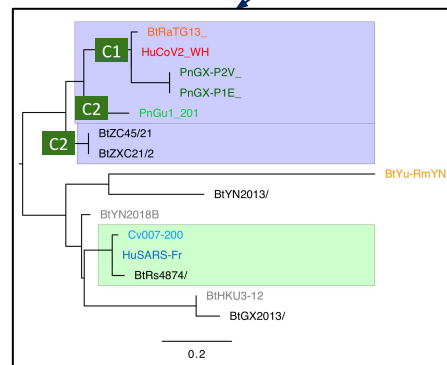
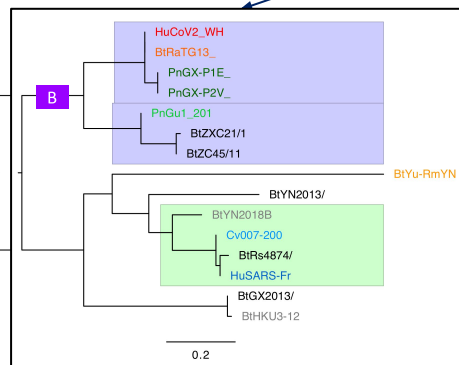
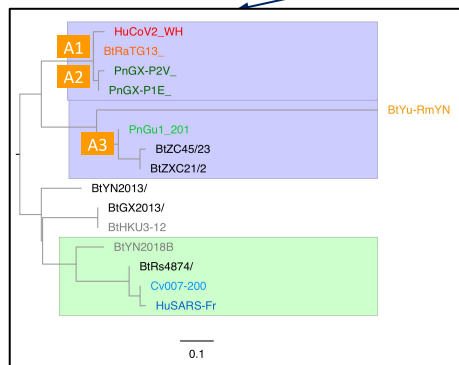
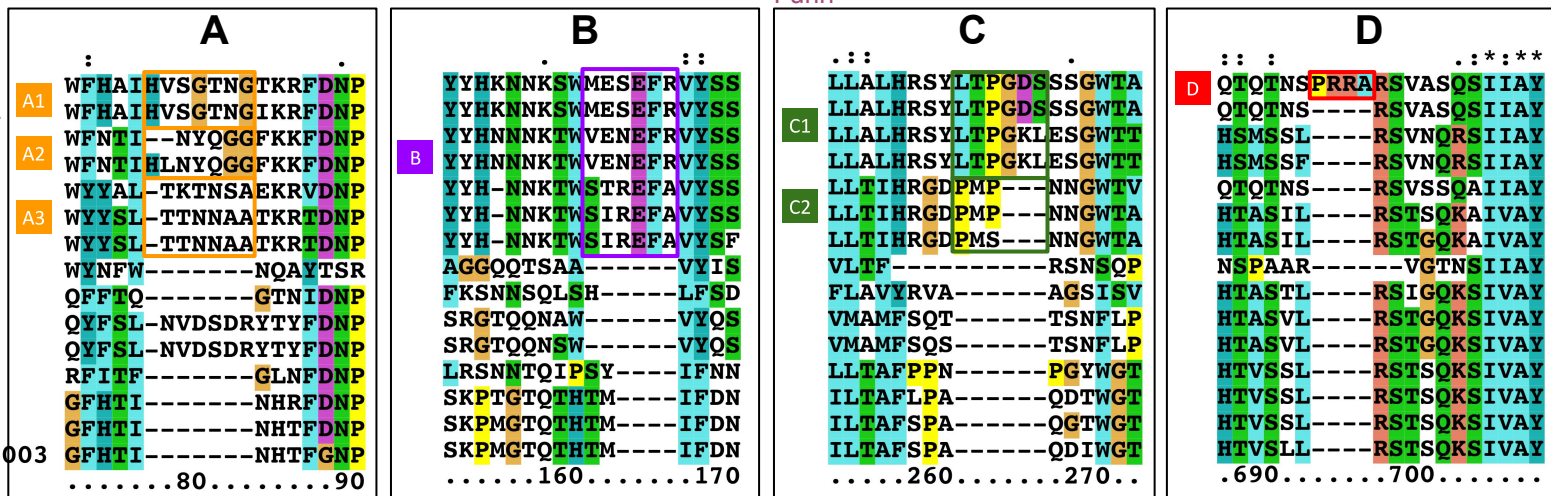
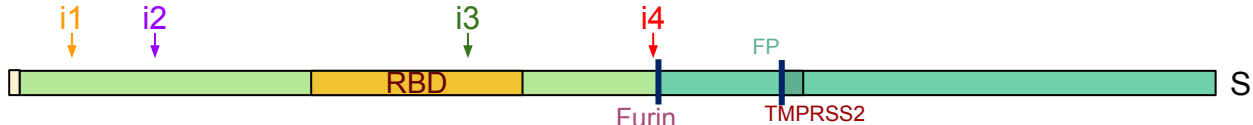
Insertion 4



Sequence	Conservation	Insertion Site	Residues
HuCoV2_WH01_2019	:: :	i4	QTQTNSPRRRARSVASQSIIAY
BtRaTG13_2013_Yunnan			QTQTNS----RSVASQSIIAY
PnGX-P1E_2017			HSMSSL----RSVNORSIIAY
PnGX-P2V_2018			HSMSSF----RSVNORSIIAY
PnGu1_2019			QTQTNS----RSVSSQAI IAY
BtZC45			HTASIL----RSTSOKAIVAY
BtZXC21			HTASIL----RSTGOKAIVAY
BtYu-RmYN02_2019			NSPAAR-----VGTNSIIAY
BtYN2013			HTASTL----RSIGOKSIVAY
BtHKU3-12			HTASVL----RSTGOKSIVAY
BtGX2013			HTASVL----RSTGOKSIVAY
BtYN2018B			HTVSLL----RSTSOKSIVAY
BtRs4874			HTVSLL----RSTSOKSIVAY
Cv007-2004			HTVSLL----RSTSOKSIVAY
HuSARS-Frankfurt-1_2003			HTVSLL----RSTSOKSIVAY

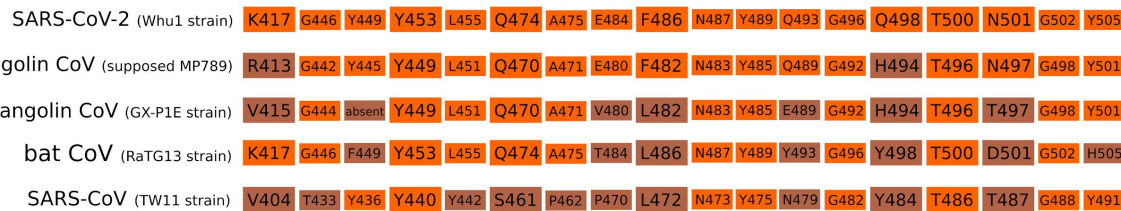
.690.....700.....



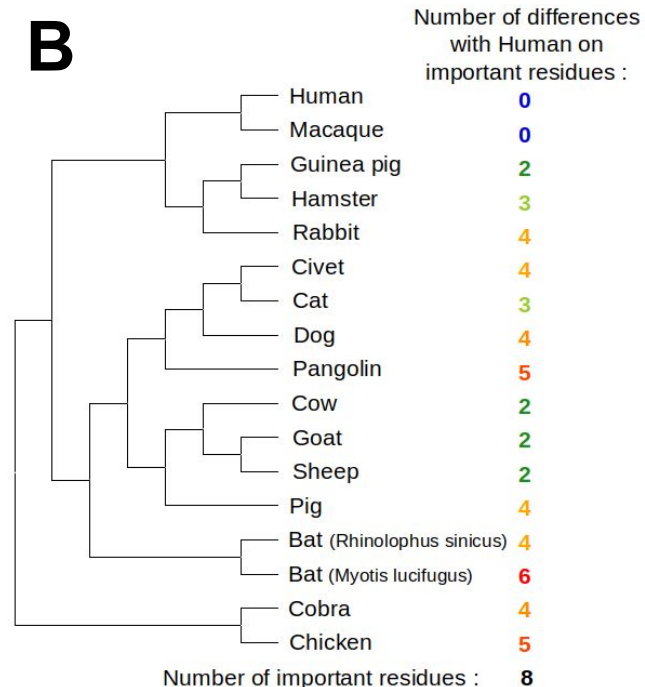


A

ACE2 receptor



S protein

B

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