Correction des TP1 et TP2 :

Jeux de données

Analyser le tableau 1
Nous pouvons observer que le nombre de systèmes homologues à ComED est très variable d'un génome à l'autre.

S. dysgalactiae	S. equi	S. gallolyticus	S. gordonii
5	5	6	2
S. mitis	S. mutans	S. oralis	S. parauberis
4	2	2	4
S. pneumoniae	S. pyogenes	S. salivarius	S. sanguinis
4	3	9	2
S. thermophilus	S. uberis		
2	6		

	HK	RR
S. dysgalactiae	3	2
S. equi	3	2
S. gallolyticus	3	3
S. gordonii	1	1
S. mitis	2	2
S. mutans	1	1
S. oralis	1	1
S. parauberis	3	1
S. pneumoniae	2	2
S. pyogenes	2	1
S. salivarius	5	4
S. sanguinis	1	1
S. thermophilus	1	1
S. uberis	4	2

Ces systèmes sont absents de *S. agalactiae* et *S. suis*. Ils sont présents à un nombre variable d'exemplaires même dans le même groupe taxonomique. Cette versatilité suggère des évènements de gains/pertes de gènes récents au cours de l'évolution. Si nous utilisons la proximité chromosomique pour reconstruire les systèmes, nous observons que le nombre de partenaires histidine kinase (HK) peut varier de un à trois (*S. uberis*), un et deux partenaire HK étant ce qu'il est trouvé de plus fréquent. Nous pouvons observer que l'annotation fonctionnelle des séquences donne peu d'information sur leur fonction biologique, de même les noms de gènes/protéines utilisés sont très peu fiables.

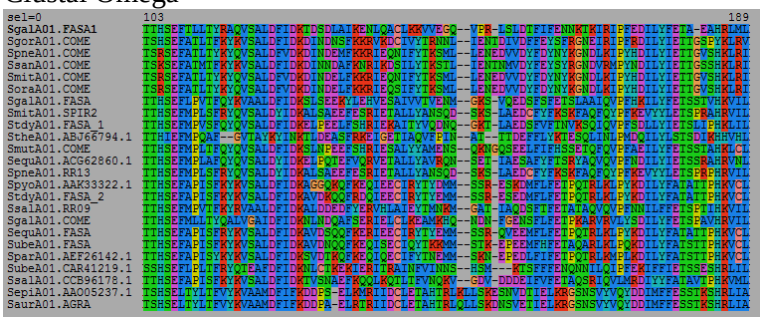
Alignement multiples des séquences homologues a ComE de *S. pneumoniae*

L'alignement obtenu est de très bonne qualité avec très peu d'insertions/délétions (indels) en dehors des régions Nter et Cter. Une variabilité au niveau de la partie Nter des protéines est souvent observée en raison de la difficulté de prédire correctement les débuts des gènes. Ces régions peuvent être éditées pour supprimer les indels mais cela aura peu de répercussions sur les reconstructions d'arbres (les changements topologiques observés sont associés à des branches faiblement supportées). En effet, par défaut les méthodes basées sur une distance éliminent les colonnes comportant au moins une délétion et les méthodes basées sur le maximum de vraisemblance prennent en compte efficacement les délétions.

Si vous réalisez l'alignement des séquences ComE avec clustalo (Clustal Omega), vous observerez quelques différences dans la localisation des indels. Sur cet exemple, le choix du logiciel pour réaliser l'alignement aura peu d'impact sur la reconstruction des arbres.

Muscle

Clustal Omega



Construction des arbres en utilisant la méthode de distance NJ

Remarques : seule la longueur des branches horizontales est significative et chaque valeur de bootstrap est associée à une bi-partition de l'arbre.
La rotation des branches autour des nœuds (swap) ne change pas les bi-partitions et les longueurs de branches, l'arbre conserve sa topologie. Les arbres obtenus ne sont pas enraciné. Par défaut, le logiciel utilise la méthode du point médian. Comme nous disposons d'un groupe externe (*Staphylococcus epidermidis* et *Staphylococcus aureus*), nous allons l'utiliser pour enraciner tous nos arbres. Le nœud ancêtre sera sur la branche reliant ce groupe externe aux autres séquences. Cet enracinement permet d'orienter l'arbre (distinguer les nœuds pères des nœuds fils) et donc les différents événements qui se sont produits au cours de l'évolution.

Format Newick

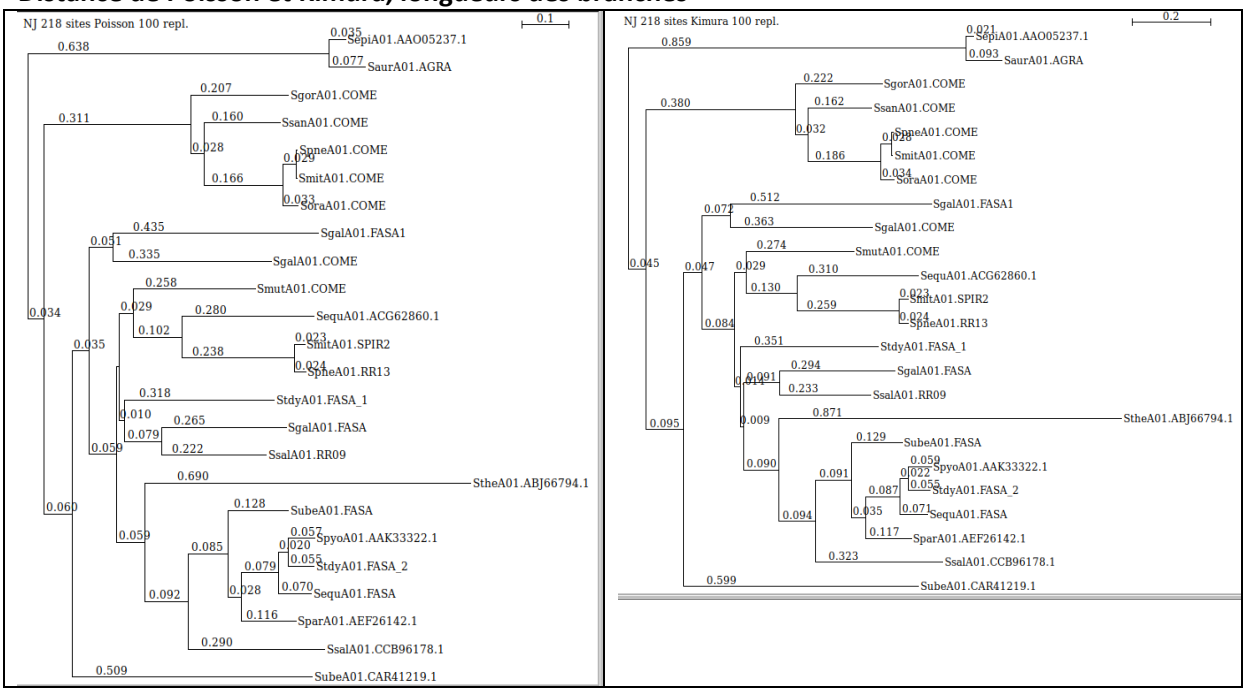
((SaurA01.AGRA:0.07724,SepiA01.AAO05237.1:0.03460):0.63758,((((SoraA01.COME:0.03309,(SmitA01.COME:0.00189,SpneA01.COME:0.00271)100:0.02865)100:0.16585,SsanA01.COME:0.15999)63:0.02846,SgorA01.COME:0.20679)100:0.31145,(SubeA01.CAR41219.1:0.50854,(((((SsalA01.RR09:0.22188,SgalA01.FASA:0.26463)94:0.07901,StdyA01.FASA_1:0.31770)27:0.00995,(((SpneA01.RR13:0.02410,SmitA01.SPIR2:0.02292)100:0.23804,SequA01.ACG62860.1:0.28026)100:0.10241,SmutA01.COME:0.25764)56:0.02946)30:0.00579,((SsalA01.CCB96178.1:0.28979,((SparA01.AEF26142.1:0.11551,(SequA01.FASA:0.06998,(StdyA01.FASA_2:0.05534,SpyoA01.AAK33322.1:0.05650)84:0.02009)100:0.07856)80:0.02830,SubeA01.FASA:0.12754)100:0.08488)93:0.09193,StheA01.ABJ66794.1:0.69047)75:0.05891)62:0.05919,(SgalA01.COME:0.33494,SgalA01.FASA1:0.43487)65:0.05123)37:0.03511)55:0.06021):0.03356)100;

Exemple : le sous arbre (SgalA01.COME:0.33494,SgalA01.FASA1:0.43487)65:0.05123
0.33494, 0.43487 longueurs des branches des feuilles au dernier ancêtre commun et 0.05123 longueur de la branche de cet ancêtre au nœud suivant. Cette branche, qui sépare SgalA01.COME et SgalA01.FASA1 de toutes les autres feuilles, a un support de bootstrap de 65%.

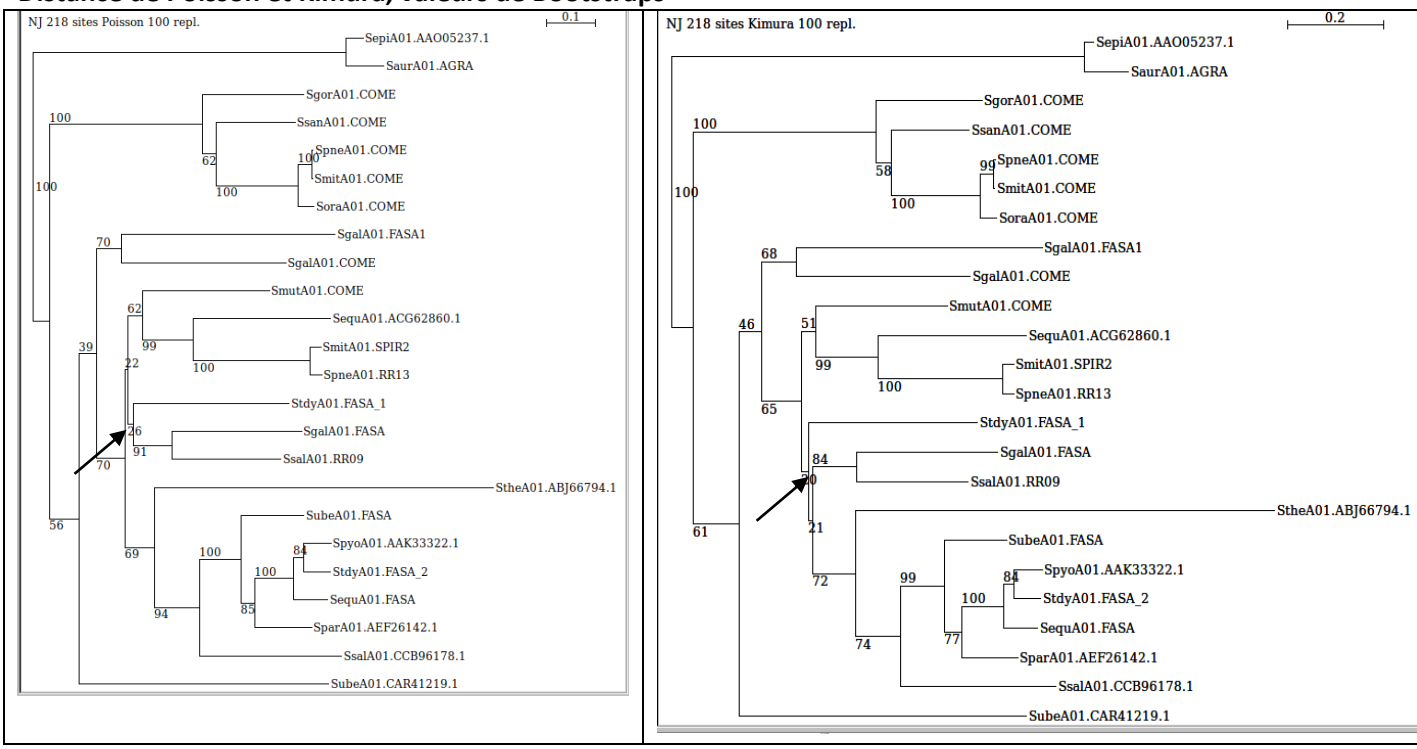
Comparaison topologie obtenue avec distance de Poisson et Kimura (approximation de la distance PAM)

Nous pouvons remarquer que les branches ont des longueurs plus grandes avec la méthode Kimura en particulier pour les branches les plus profondes (noter la différence d'échelle entre les deux topologies).

Distance de Poisson et Kimura, longueurs des branches



Distance de Poisson et Kimura, valeurs de Bootstraps

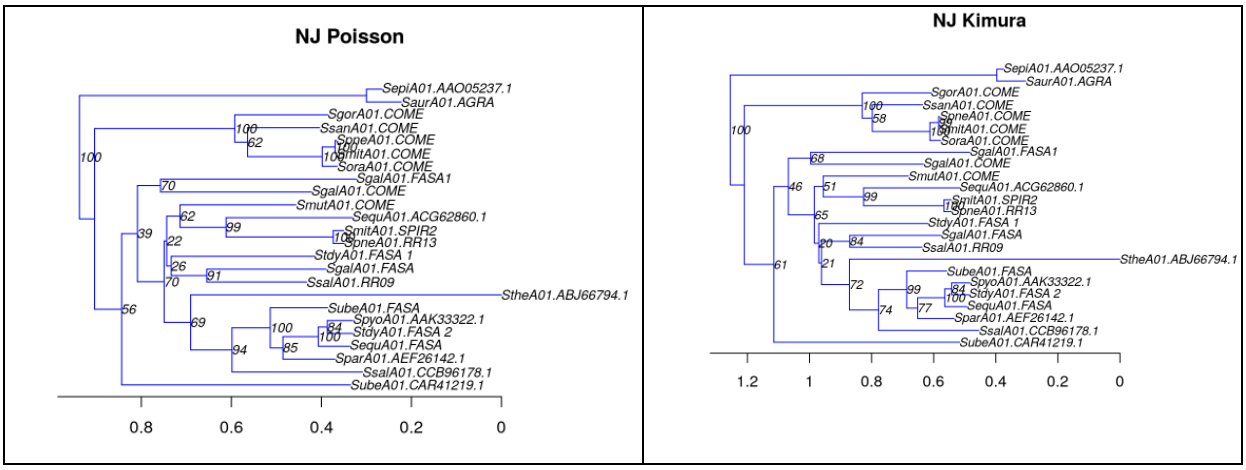


Les valeurs de bootstraps sont globalement un peu meilleures avec Kimura.

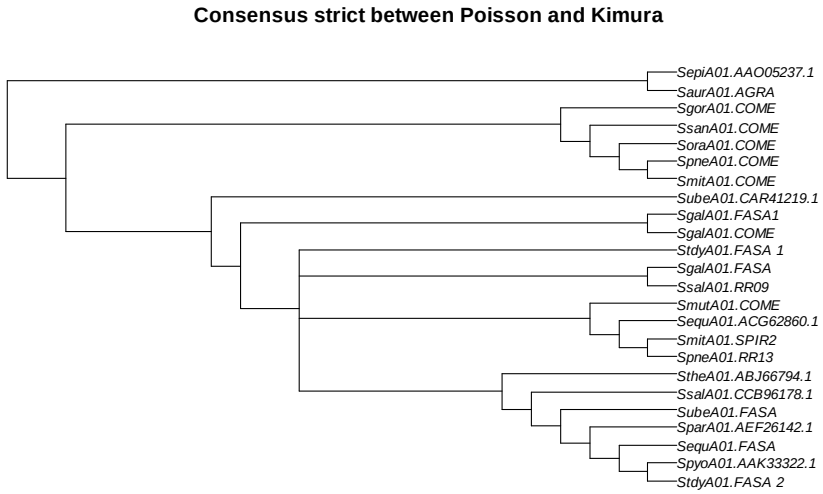
Nous pouvons remarquer que les séquences semblent évoluer à peu près à la même vitesse (elles sont alignées verticalement) sauf la séquence StheA01.ABJ66794.1 qui montre clairement une accélération.

Il y a peu d'incongruences entre ces deux arbres, elles sont généralement associées à des branches courtes supportées par de faibles valeurs de bootstrap (indiquées par une flèche sur les arbres ci-dessous)

Comparaison des bipartitions des arbres obtenues avec les distances de Poisson et Kimura. Construction de l'arbre consensus.

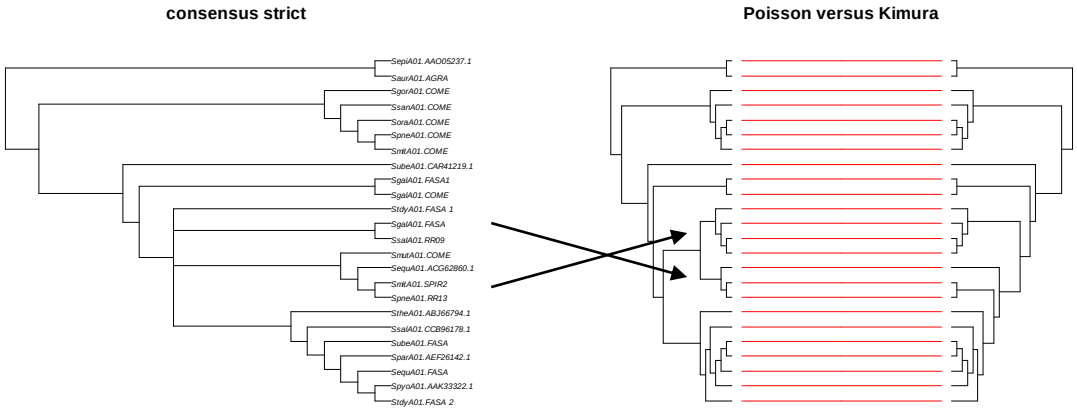


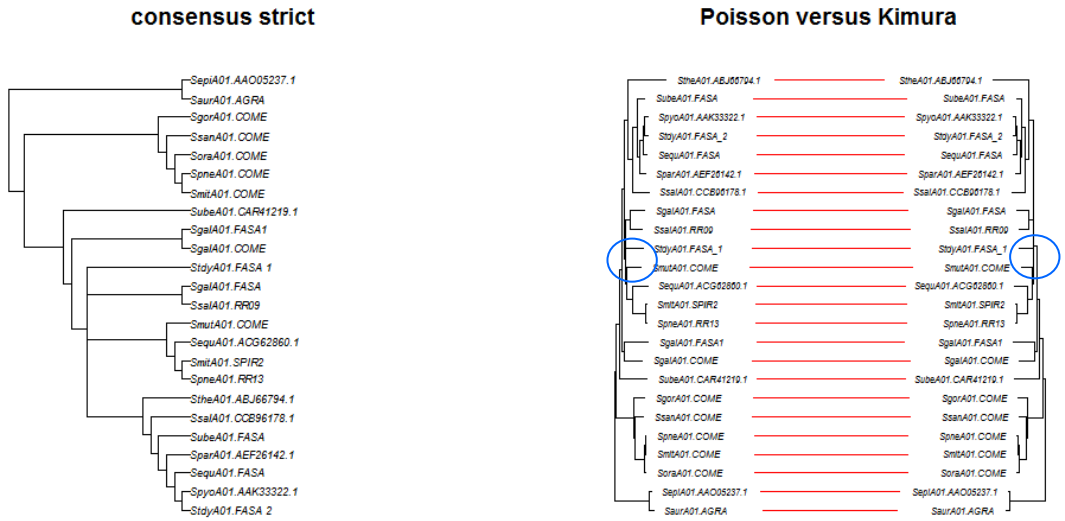
Arbre consensus



Attention, les longueurs de branches associées à l'arbre consensus n'ont pas de signification phylogénétique. Nous observons une bonne résolution de l'arbre consensus ce qui traduit une très grande majorité de bipartitions communes entre les deux arbres. Il y a une seule région ou les bi-partitions ont été fusionnées (multifurcation ou polytomie, nœud dans un arbre qui connecte plus de trois branches).

Congruence des arbres obtenus avec les distances de Poisson et Kimura

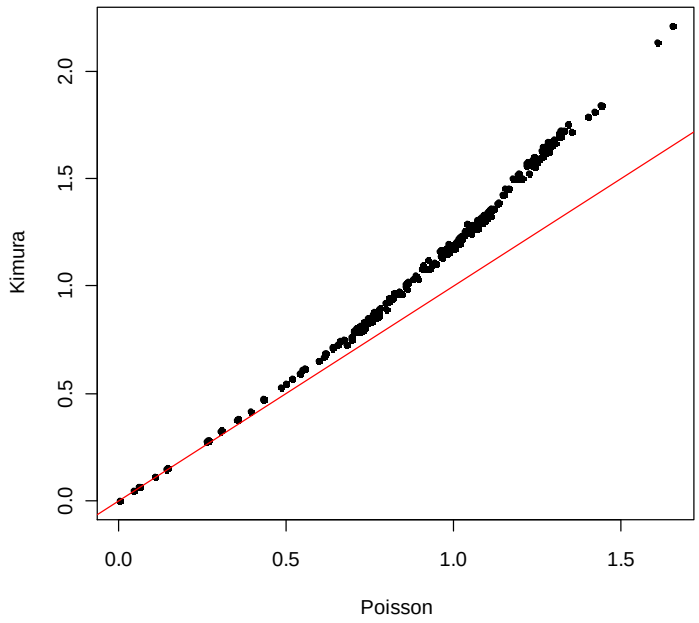




Attention, les deux arbres ne sont pas exactement congruents (cercles bleus) ! Comme précédemment, les longueurs de branches n'ont pas de signification phylogénétique.

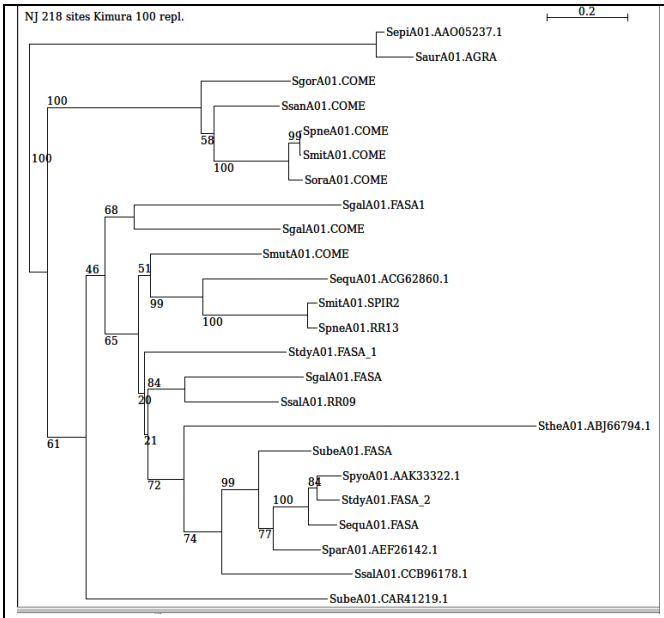
Relation entre les distances d'arbres Poisson/ Kimura

On observe une bonne corrélation entre les distances obtenues avec les deux modèles pour les petites distances. Par contre, il y a un décrochage très net pour les distances > 0.5. Cela montre que la distance de Poisson sous-estime les distances 'réelles' par rapport à la distance de Kimura quand la divergence augmente entre les paires de séquences.

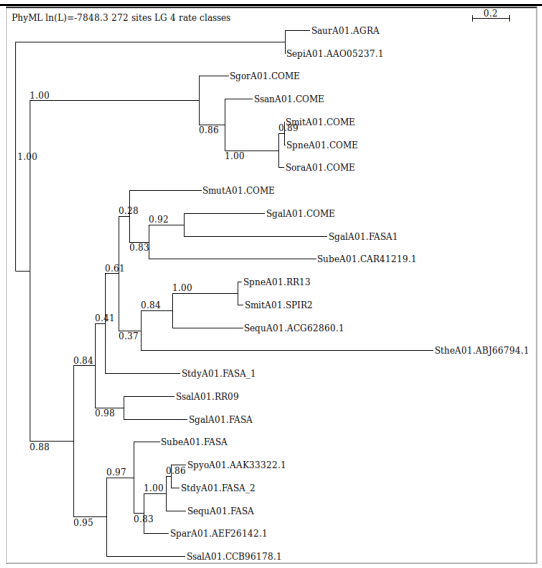


Construction des arbres en utilisant une méthode du maximum de vraisemblance

Nous observons des changements topologiques important entre ces deux arbres. Le plus important concerne la position de la séquence StheA01.ABJ66794.1 qui est radicalement différente. Nous pouvons également observer une perturbation générale qui conduit à un décalage des groupes de séquences par rapport à la verticale, ce qui traduit des vitesses relatives d'évolution différentes pour ces groupes. La topologie obtenue avec la méthode PhyML LG suggère l'existence de quatre clades. Les valeurs de bootstrap sont un peu meilleures pour la méthode PhyML LG mais restent faibles pour les régions incompatibles entre les deux arbres.



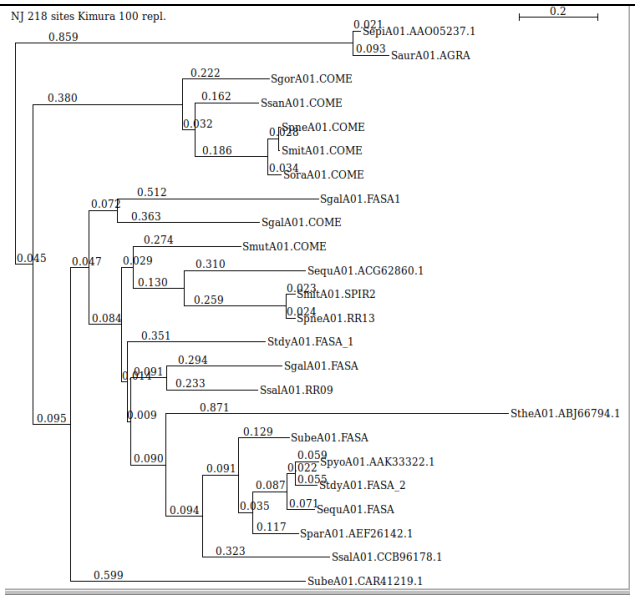
NJ distance de Kimura



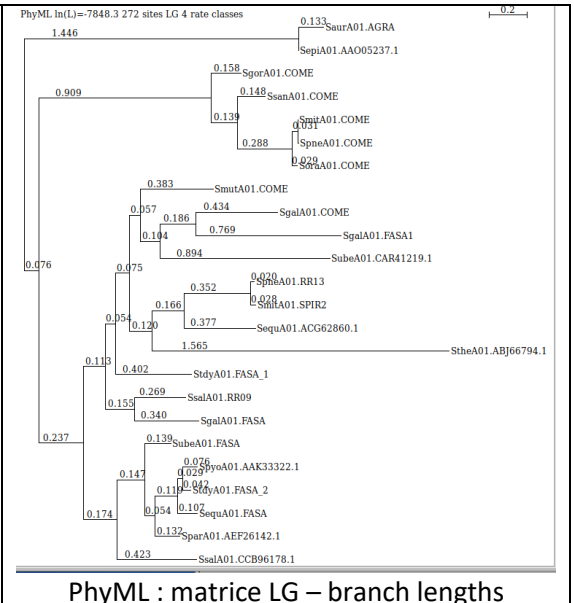
PhyML : matrice LG

Informations données lors du déroulement du programme PhyML :

- 269 patterns found (out of a total of 272 sites) : 269 sites ont été utilisés pour le calcul de l'arbre
- 30 sites without polymorphism (11.03%) : 30 positions invariantes dans l'alignement
- **Log likelihood of PhyML LG 4 tree:** -7848.341886.

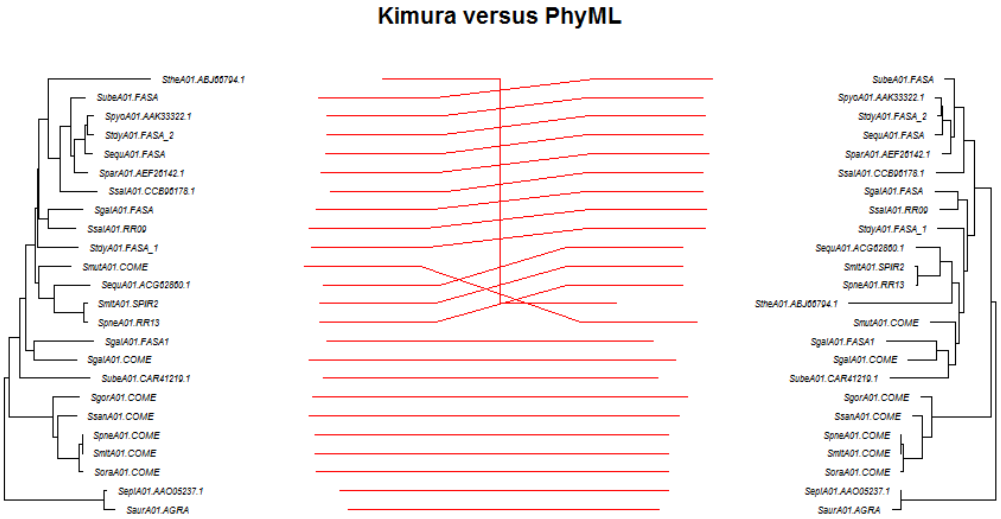


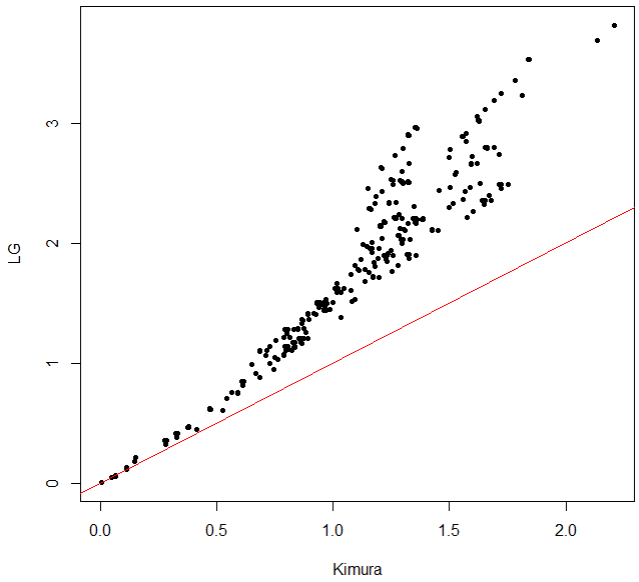
NJ distance de Kimura – branch lengths



PhyML : matrice LG – branch lengths

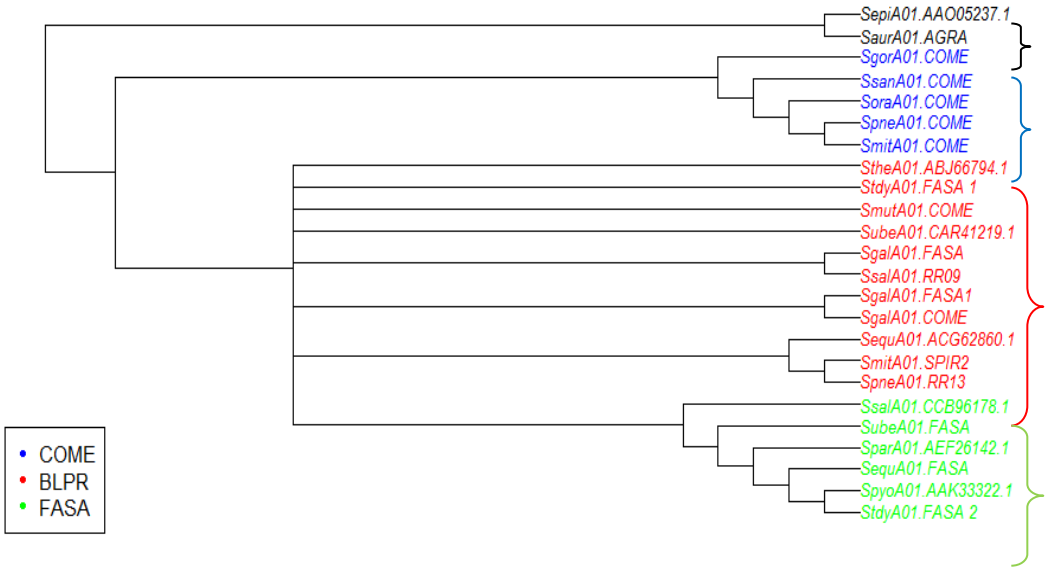
Relation entre les distances d'arbres NJ Kimura / PhyML LG





La corrélation est beaucoup moins bonne que celle observée précédemment. Nous observons une sous-estimation des distances par la méthode NJ Kimura en regard de la méthode PhyML LG. De plus, il y a une dispersion importante des points suggérant un traitement différent des substitutions observées entre les paires de séquences.

Comparaison des bipartitions des trois arbres



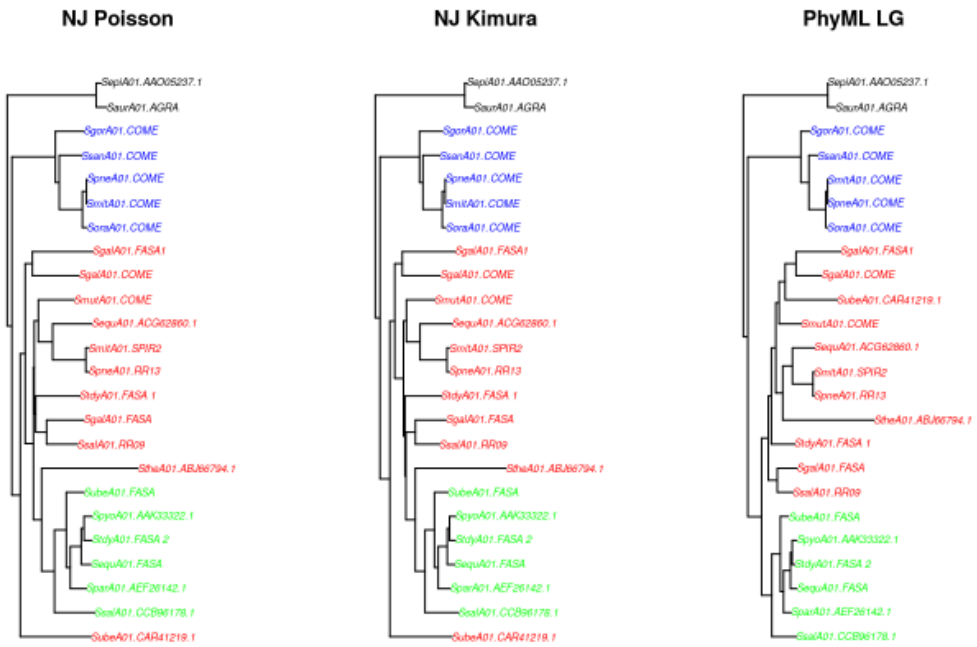
Nous observons trois sous arbres stables selon toutes les méthodes (le noir, outgroup; le bleu correspondant aux séquences de ComE; le vert correspondant aux séquences des régulateurs de réponse (RR) du système Fas). Le quatrième groupe de séquences (en rouge) qui se retrouvent entre le groupe bleu (Com) et le groupe vert (Fas) présente plus de variabilité topologique en fonction des

méthodes car nous observons beaucoup de multifurcations. De même, la localisation du groupe fas (vert) par rapport à ce groupe rouge est différente suivant les topologies.

Edition et annotation des arbres

Nous pouvons faire les mêmes remarques que précédemment. Les quatre groupes de séquences apparaissent très clairement. Nous avons confirmation que les incongruences entre les arbres sont imputables aux feuilles du groupe rouge. Il est à noter que ce groupe est monophylétique avec la méthode PhyML matrice LG (support de aLTR > 0.80).

```
layout(matrix(1:3, 1, 3));
plot.phylo(tP, cex=0.6, tip.color=col);
title('NJ Poisson);
plot.phylo(tK, cex=0.6, tip.color=col);
title('NJ Kimura');
plot.phylo(tLG, cex=0.6, tip.color=col);
title('PhyML LG');
layout(1);
```



En rapport avec la question biologique que nous sommes posée, à savoir si la différence entre les temps de latence observée chez *S. pneumoniae* et *S. mutans* entre le moment où le CSP a été ajouté et celui où la transcription des gènes précoces est observée peut s'expliquer en analysant les protéines impliquées dans la régulation du processus, nous pouvons observer que la séquence de ComE de *S. mutans* appartient au groupe rouge, comme les séquences BlpR de *S. pneumoniae* (SpneA01.RR13, Blp : **Bacteriocin-like peptide**) et non au groupe bleu renfermant la séquence de ComE de *S. pneumoniae*. Les gènes *comE* de *S. mutans* et S. *pneumoniae* sont donc paralogues et non pas orthologues, ce qui suggère des différences fonctionnelles.

Fas (fibronectin/fibrinogen binding/haemolytic activity/streptokinase regulator)

Recherche du modèle évolutif le plus adapté à l'alignement ComE

Résultats obtenus eu utilisant l'onglet "Model selection" du site "http://iqtree.cibiv.univie.ac.at/ »

```
IQ-TREE 1.6.12 built Aug 15 2019

Input file name: SpneA01.COME_Muscle.fst
Type of analysis: ModelFinder + tree reconstruction
Random seed number: 647569

REFERENCES
----------

To cite ModelFinder please use:

Subha Kalyaanamoorthy, Bui Quang Minh, Thomas KF Wong, Arndt von Haeseler,
and Lars S Jermiin (2017) ModelFinder: Fast model selection for
accurate phylogenetic estimates. Nature Methods, 14:587–589.
https://doi.org/10.1038/nmeth.4285

SEQUENCE ALIGNMENT
------------------

Input data: 24 sequences with 272 amino-acid sites
Number of constant sites: 30 (= 11.0294% of all sites)
Number of invariant (constant or ambiguous constant) sites: 30 (= 11.0294% of all
sites)
Number of parsimony informative sites: 224
Number of distinct site patterns: 269

ModelFinder
-----------

Best-fit model according to AIC: LG+F+I+G4

List of models sorted by AIC scores:

Model             LogL          AIC      w-AIC      AICc     w-AICc       BIC      w-BIC
LG+F+I+G4       -7832.8847  15797.7695 + 1.0000  15840.9109 + 1.0000  16035.7524 + 1.0000
WAG+F+I+G4      -7869.6406  15871.2812 - 0.0000  15914.4227 - 0.0000  16109.2642 - 0.0000
JTT+F+I+G4      -7880.2750  15892.5500 - 0.0000  15935.6915 - 0.0000  16130.5329 - 0.0000
Dayhoff+F+I+G4  -7921.6654  15975.3307 - 0.0000  16018.4722 - 0.0000  16213.3137 - 0.0000
Blosum62+F+I+G4 -7936.4700  16004.9400 - 0.0000  16048.0815 - 0.0000  16242.9229 - 0.0000

AIC, w-AIC   : Akaike information criterion scores and weights.
AICc, w-AICc : Corrected AIC scores and weights.
BIC, w-BIC   : Bayesian information criterion scores and weights.

Plus signs denote the 95% confidence sets.
Minus signs denote significant exclusion.

SUBSTITUTION PROCESS
--------------------

Model of substitution: LG+F+I+G4

State frequencies: (empirical counts from alignment)

  pi(A) = 0.0438
  pi(R) = 0.0519
  pi(N) = 0.0465
  pi(D) = 0.0619
  pi(C) = 0.0132
  pi(Q) = 0.0445
  pi(E) = 0.0887
  pi(G) = 0.0282
  pi(H) = 0.0277
  pi(I) = 0.0920
```

```
  pi(L) = 0.0898
  pi(K) = 0.0852
  pi(M) = 0.0210
  pi(F) = 0.0669
  pi(P) = 0.0247
  pi(S) = 0.0603
  pi(T) = 0.0507
  pi(W) = 0.0024
  pi(Y) = 0.0392
  pi(V) = 0.0615

Model of rate heterogeneity: Invar+Gamma with 4 categories
Proportion of invariable sites: 0.0228
Gamma shape alpha: 1.1491

 Category  Relative_rate  Proportion
  0         0              0.0228
  1         0.1696         0.2443
  2         0.5294         0.2443
  3         1.0447         0.2443
  4         2.3495         0.2443
Relative rates are computed as MEAN of the portion of the Gamma distribution
falling in the category.
```

Le modèle +G+F+I avec la matrice LG est le plus vraisemblable.

Alignement multiples des séquences homologues a ComD de *S. pneumoniae*

Recherche du modèle évolutif le plus adapté à l'alignement ComD

```
IQ-TREE 1.6.12 built Aug 15 2019

Input file name: SpneA01.COMD_CleanUp_muscle.fst
Type of analysis: ModelFinder + tree reconstruction
Random seed number: 349582

REFERENCES
----------

To cite ModelFinder please use:

Subha Kalyaanamoorthy, Bui Quang Minh, Thomas KF Wong, Arndt von Haeseler,
and Lars S Jermiin (2017) ModelFinder: Fast model selection for
accurate phylogenetic estimates. Nature Methods, 14:587–589.
https://doi.org/10.1038/nmeth.4285

SEQUENCE ALIGNMENT
------------------

Input data: 30 sequences with 483 amino-acid sites
Number of constant sites: 21 (= 4.34783% of all sites)
Number of invariant (constant or ambiguous constant) sites: 21 (= 4.34783%
of all sites)
Number of parsimony informative sites: 437
Number of distinct site patterns: 476

ModelFinder
-----------

Best-fit model according to AIC: LG+F+I+G4

List of models sorted by AIC scores:
```

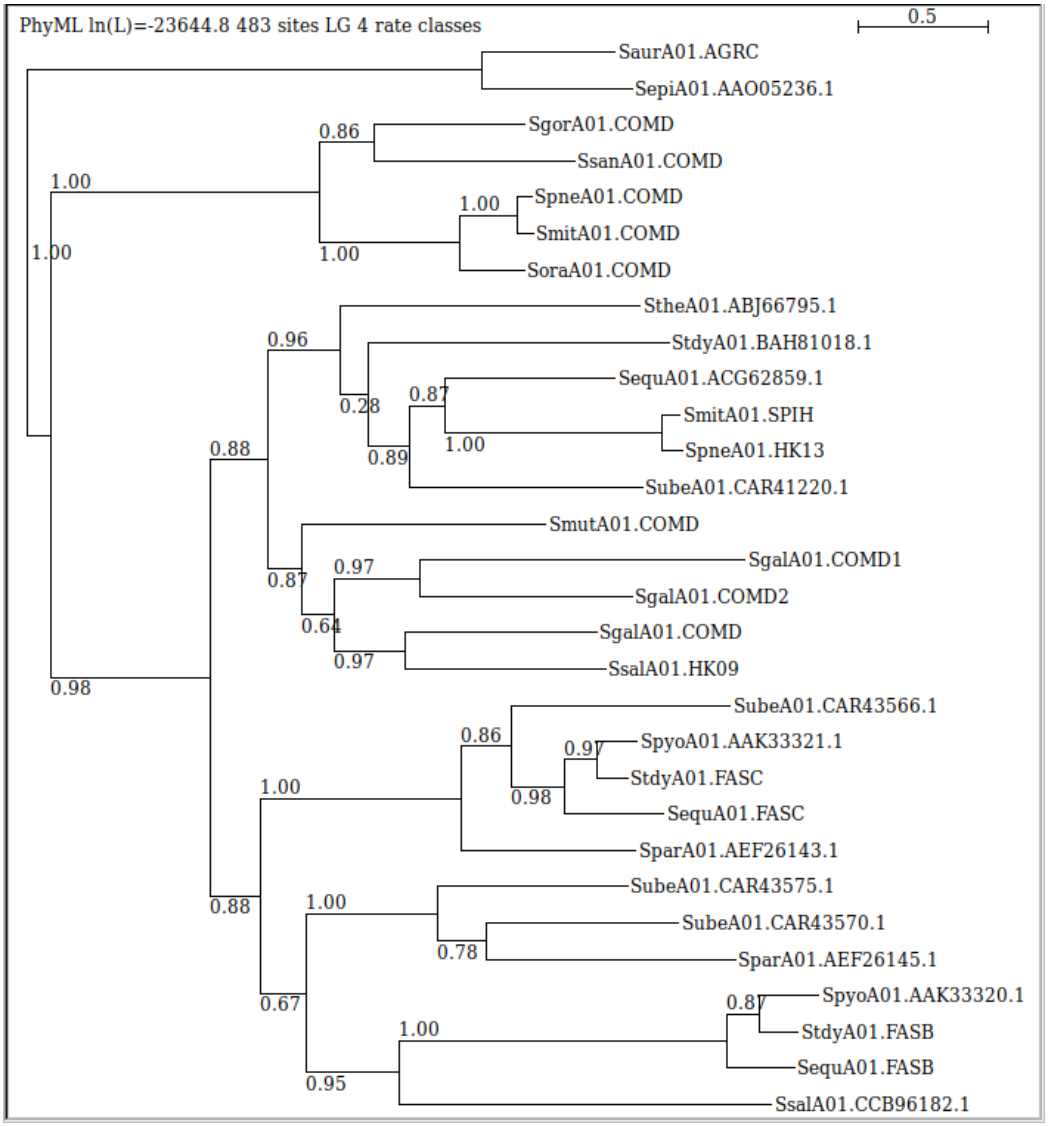
```
Model             LogL          AIC      w-AIC      AICc     w-AICc
BIC      w-BIC
LG+F+I+G4       -23685.9971  47527.9941 + 1.0000  47558.4991 + 1.0000
47854.0354 + 1.0000
JTT+F+I+G4      -23737.1234  47630.2468 - 0.0000  47660.7518 - 0.0000
47956.2881 - 0.0000
WAG+F+I+G4      -23771.8532  47699.7065 - 0.0000  47730.2114 - 0.0000
48025.7478 - 0.0000
Blosum62+F+I+G4 -23881.8277  47919.6555 - 0.0000  47950.1604 - 0.0000
48245.6968 - 0.0000
Dayhoff+F+I+G4  -23988.3952  48132.7904 - 0.0000  48163.2954 - 0.0000
48458.8317 - 0.0000

AIC, w-AIC   : Akaike information criterion scores and weights.
AICc, w-AICc : Corrected AIC scores and weights.
BIC, w-BIC   : Bayesian information criterion scores and weights.

Plus signs denote the 95% confidence sets.
Minus signs denote significant exclusion.

SUBSTITUTION PROCESS
--------------------

Model of substitution: LG+F+I+G4

State frequencies: (empirical counts from alignment)

  pi(A) = 0.0410
  pi(R) = 0.0389
  pi(N) = 0.0509
  pi(D) = 0.0449
  pi(C) = 0.0060
  pi(Q) = 0.0354
  pi(E) = 0.0572
  pi(G) = 0.0362
  pi(H) = 0.0194
  pi(I) = 0.1093
  pi(L) = 0.1351
  pi(K) = 0.0601
  pi(M) = 0.0252
  pi(F) = 0.0751
  pi(P) = 0.0200
  pi(S) = 0.0779
  pi(T) = 0.0443
  pi(W) = 0.0062
  pi(Y) = 0.0519
  pi(V) = 0.0651

Model of rate heterogeneity: Invar+Gamma with 4 categories
Proportion of invariable sites: 0.0086
Gamma shape alpha: 1.5285


 Category  Relative_rate  Proportion
  0         0              0.0086
  1         0.2317         0.2479
  2         0.5984         0.2479
  3         1.0611         0.2479
  4         2.1434         0.2479
```

```
Relative rates are computed as MEAN of the portion of the Gamma
distribution falling in the category.
```

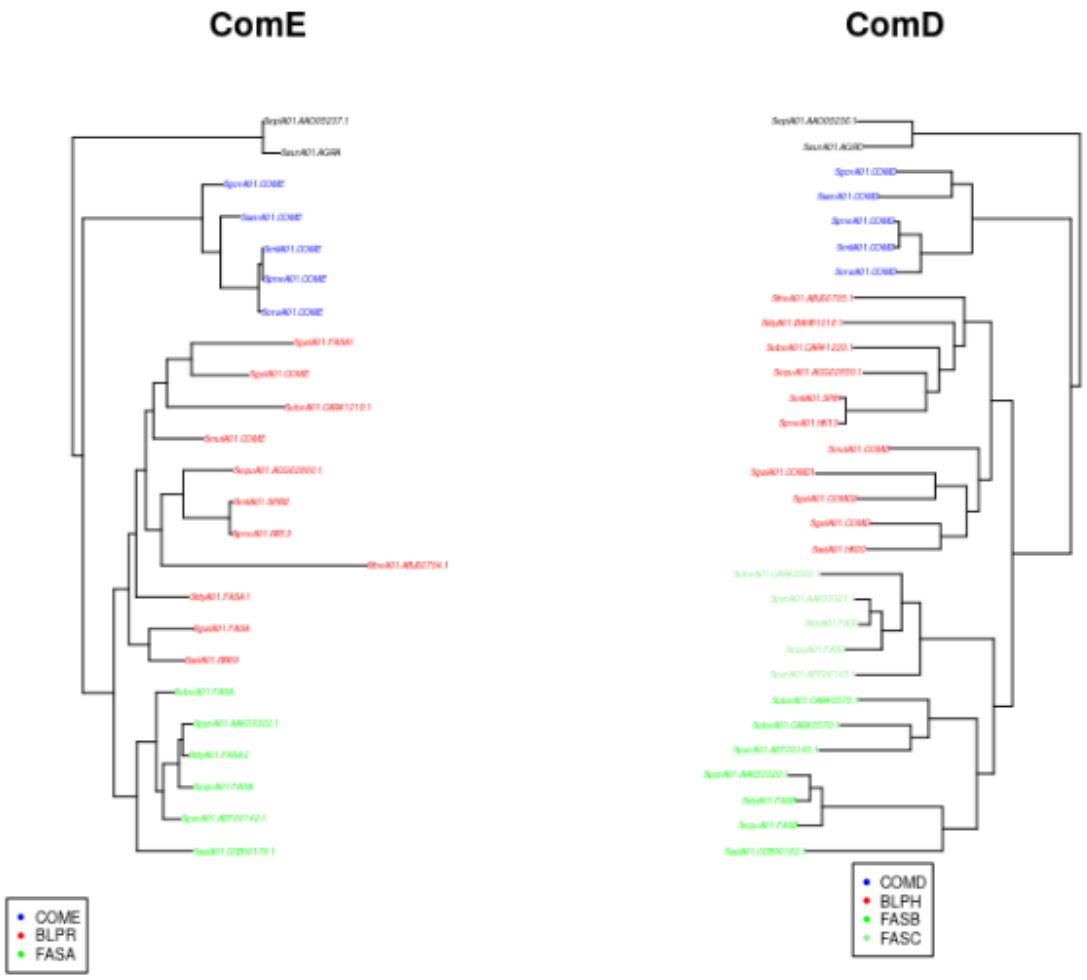
Le modèle LG+G+F+I est le plus vraisemblable comme dans le cas de ComE. Par contre, les modèles suivants ne changent que la matrice, ainsi pour chaque matrice les modèles +G+F+I devancent les modèles plus simples.

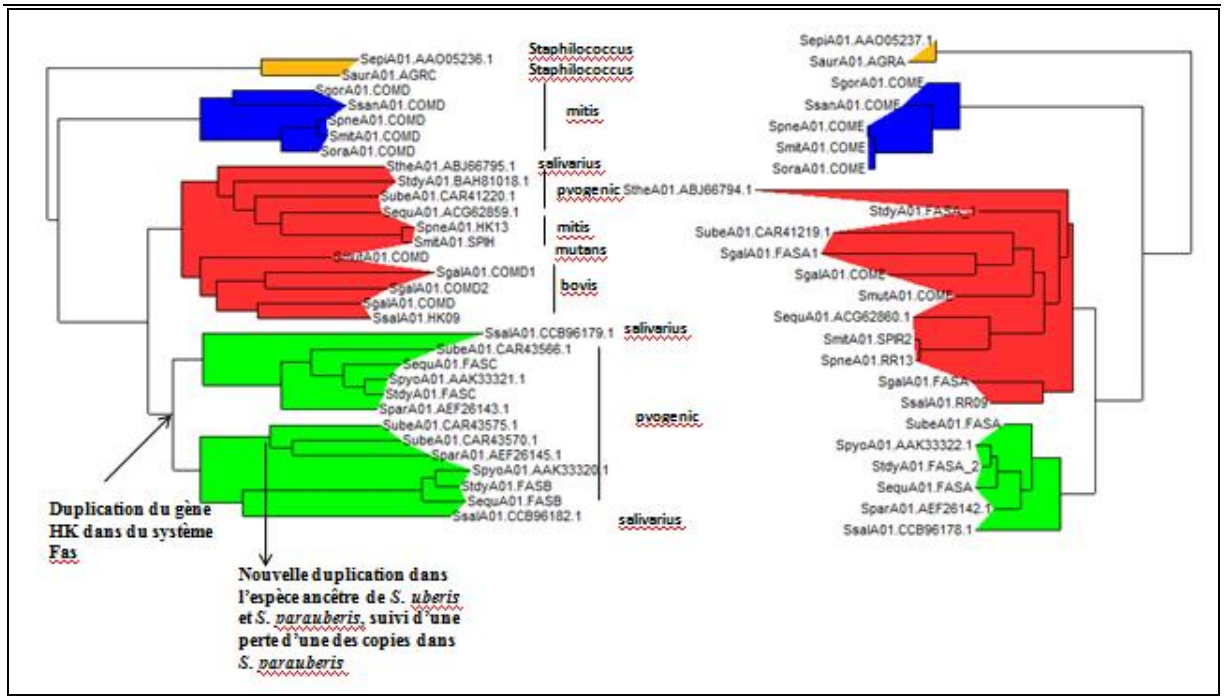
Construction de l'arbre des protéines ComD en utilisant le modèle le plus adapté et une méthode du maximum de vraisemblance

476 patterns found. (out of a total of 483 sites)
21 sites without polymorphism (4.35%)
Log likelihood of the current tree: -23644.844815



Comparaison des arbres obtenus sur les séquences homologues à ComD et à ComE: interprétation biologique





Les séquences de Staphylococcus (Saur et Sepi) servent de groupe externe et permettent donc de connaître le nœud correspondant au nœud ancêtre hypothétique de l'ensemble de nos séquences de streptocoques. A partir de ce nœud, nous remarquons qu'une branche conduit à un nœud interne regroupant un certain nombre de séquences (coloriées en bleu sur chacun des deux arbres) parmi lesquelles ComD et ComE de *S. pneumoniae*. Parmi ces séquences il n'y a pas de paralogie (pas deux séquences appartenant à la même espèce) donc nous avons un groupe de séquences orthologues. Tous les génomes appartiennent au groupe mitis. Le système ComDE de *S. mutans* n'appartient pas à ce sous-arbre, il ne forme donc pas un système orthologue au système ComDE de *S. pneumoniae* et n'a donc probablement pas la même fonction. Ceci pourrait expliquer les différences de temps de latence observées lors de l'ajout du CSP avant le déclenchement de l'état de compétence chez *S. mutans*. Son système ComDE ne doit pas intervenir de la même manière dans la régulation de la compétence que le système ComDE de *S. pneumoniae*. Il est également possible qu'il ne soit pas impliqué dans cette régulation.

Pour les deux sous-arbres bleus, les topologies sont identiques. On dit que les deux sous-arbres sont **congruents.** En termes évolutifs, cela indique que les deux partenaires du système, ComD et ComE, ont **coévolué**.

La deuxième partie des arbres est plus complexes à analyser. Première remarque : des séquences homologues à BlpR et BlpH de *S. pneumoniae* (groupe rouge) présentent la plus grande distribution taxonomique avec 9 espèces représentées appartenant à 5 groupes taxonomiques (*salivarius*, *mitis*, *pyogenic*, *mutans*, *bovis*). Ceci suggère que les gènes codant pour ce système étaient présents dans l'ancêtre commun aux streptocoques et certaines espèces les auraient perdus. Le génome de *S. gallolyticus* se distingue par l'occurrence de trois copies du système (paralogues). On remarquera aussi que les séquences de *S. thermophilus* (StheA01) et *S. salivarius* (SsalA01) (groupe salivarius) ne sont pas regroupées ce qui suggère des transferts horizontaux de gènes.

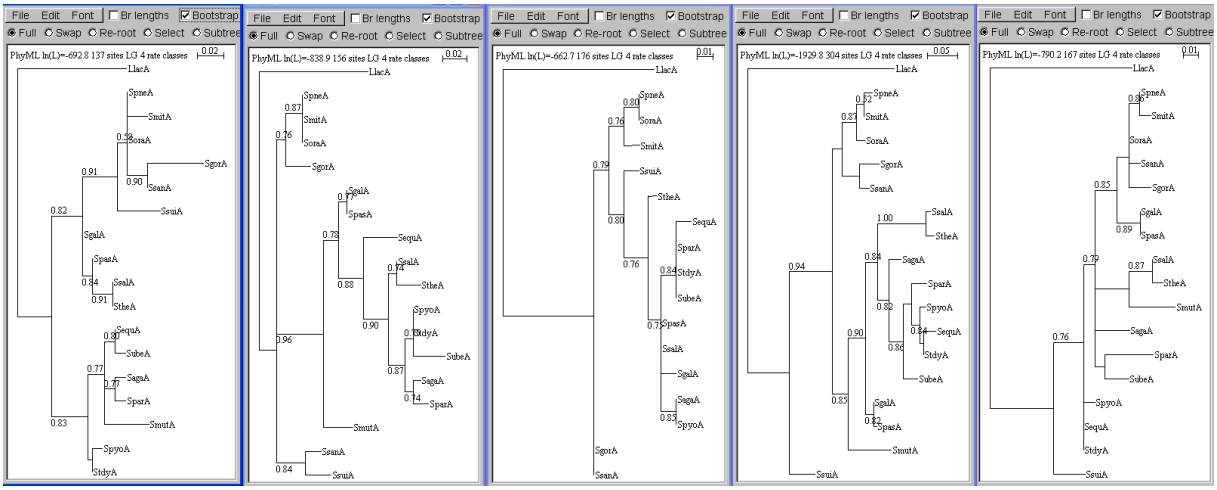
Le groupe vert est décomposé en deux sous arbres sur les ComD et un seul sous-arbre sur les ComE. Ce groupe renferme les séquences du système Fas qui posséderait deux HK par RR. Des signaux différents pourraient être "sentis" par chacun des senseurs et activer le même régulateur et donc activer les mêmes gènes. Ils sont trouvés majoritairement dans le groupe pyogenic.

Arbre obtenus sur les protéines de la petite et la grande sous-unité du ribosome

Que pensez-vous du choix des protéines ribosomiques pour la reconstruction de l'arbre des espèces?

On s'attend à ce que les gènes codant pour les protéines ribosomiques soient conservés dans tous les génomes avec une très faible fréquence de duplication de gènes ce qui permet d'identifier sans ambigüité les gènes orthologues. D'autre part, les produits de ces gènes sont impliqués dans un processus vital pour la bactérie, la traduction, qui implique le maintien d'interactions fortes entre protéines. On s'attend donc à ce que ces gènes soient rarement impliqués dans des transferts horizontaux. Ils seraient donc un bon marqueur de l'évolution des génomes (désavantages : petite taille et forte conservation de séquence !).

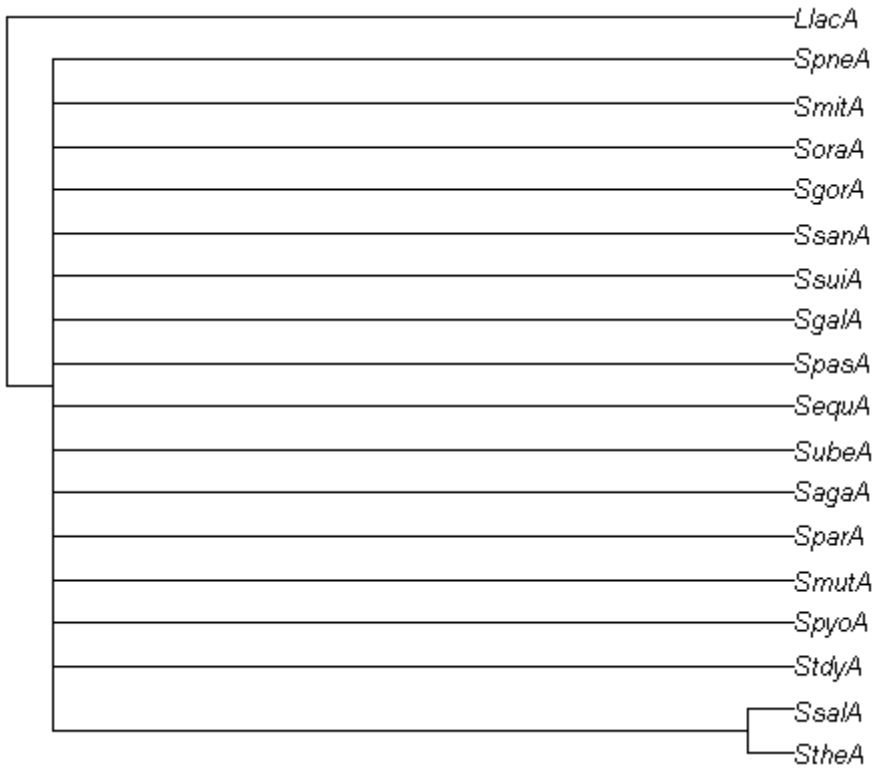
```
43 clusters retained
LlacA   11111111111111111111111111111111111111111111   100.00
SagaA   11111111111111111111111111111111111111111111   100.00
SequA   11111111111111111111111111111111111111111111   100.00
SgalA   11111111111111111111111111111111111111111111   100.00
SgorA   11111111111111111111111111111111111111111111   100.00
SmitA   11111111111111111111111111111111111111111111   100.00
SmutA   11011110011111111111101111111111111111111111   90.70
SoraA   11111111111111111111111111111111111111111111   100.00
SparA   11111111111111111111111111111111111111111111   100.00
SpasA   11111111111111111111111111111111111111111111   100.00
SpneA   11111111111111111111111111111111111111111111   100.00
SpyoA   11111111111111111111111110111111111111111111   97.67
SsalA   11111111111111111111111111111111111111111111   100.00
SsanA   11111111111111111111111111111111111111111111   100.00
SsuiA   11111111111111110101111110111111111111101110   88.37
StdyA   11111111111111111111111111111111111111111111   100.00
StheA   11111111111111111111111111111111111111111111   100.00
SubeA   11111111111111111111111111111111111111111111   100.00
```



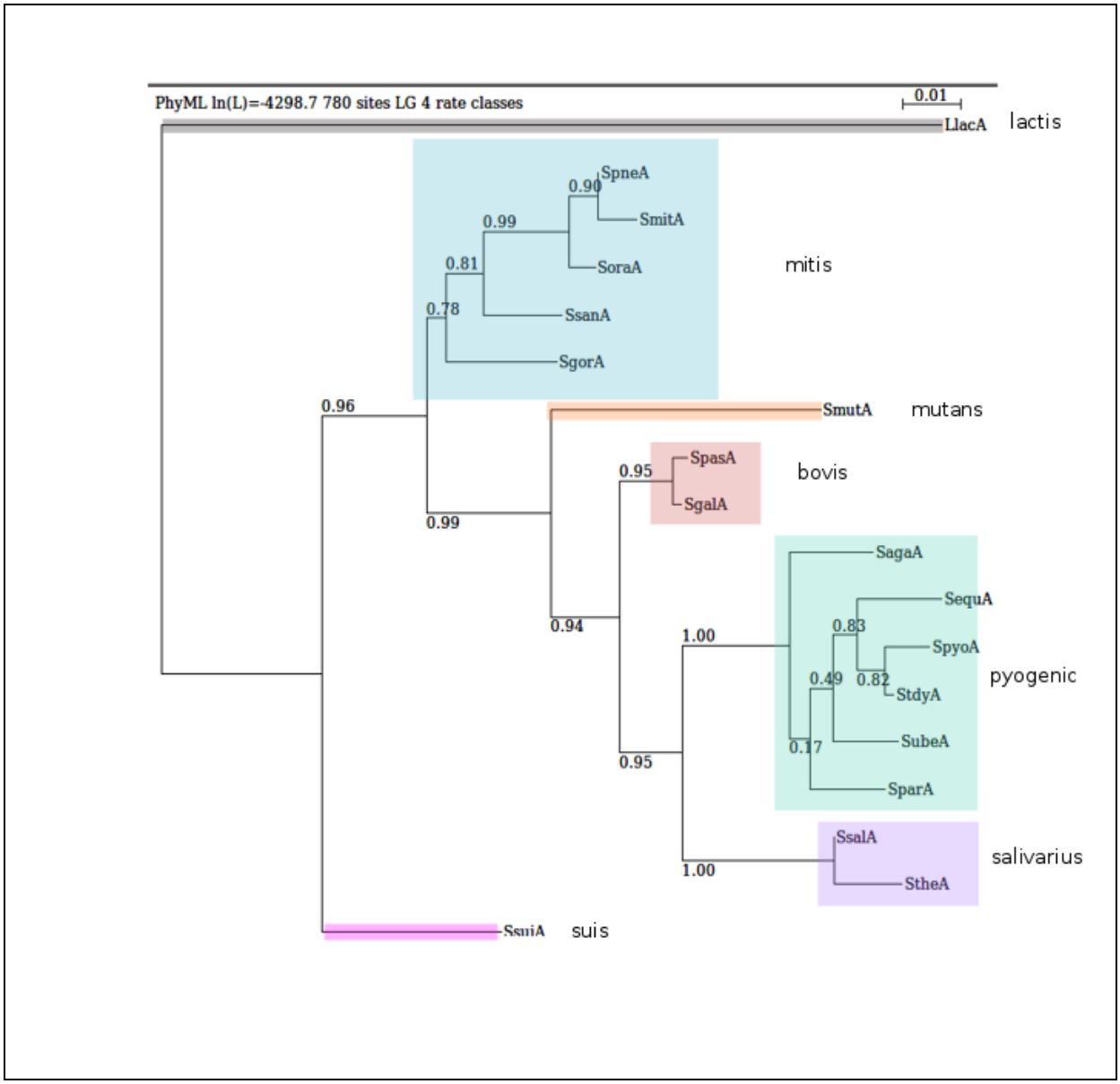
Arbres obtenus avec la méthode PhyML et la matrice LG (paramètres par défauts) pour les 5 premiers fichiers. Nous ne pouvons observer aucune congruence entre ces arbres et de nombreuses branches non résolues.

Arbre consensus

Un des cinq arbres ne possède que 17 feuilles (tips) alors que les autres en possèdent 18. On ne peut calculer un arbre consensus que si les différents arbres ont exactement les mêmes feuilles, d'où le refus de la méthode quand nous demandons le consensus avec les 5 arbres. En supprimant l'arbre incriminé, nous obtenons l'arbre suivant qui confirme une totale incongruence entre les différents arbres (que des multifurcations) car aucune bifurcation commune, excepté pour SsalA et StheA.

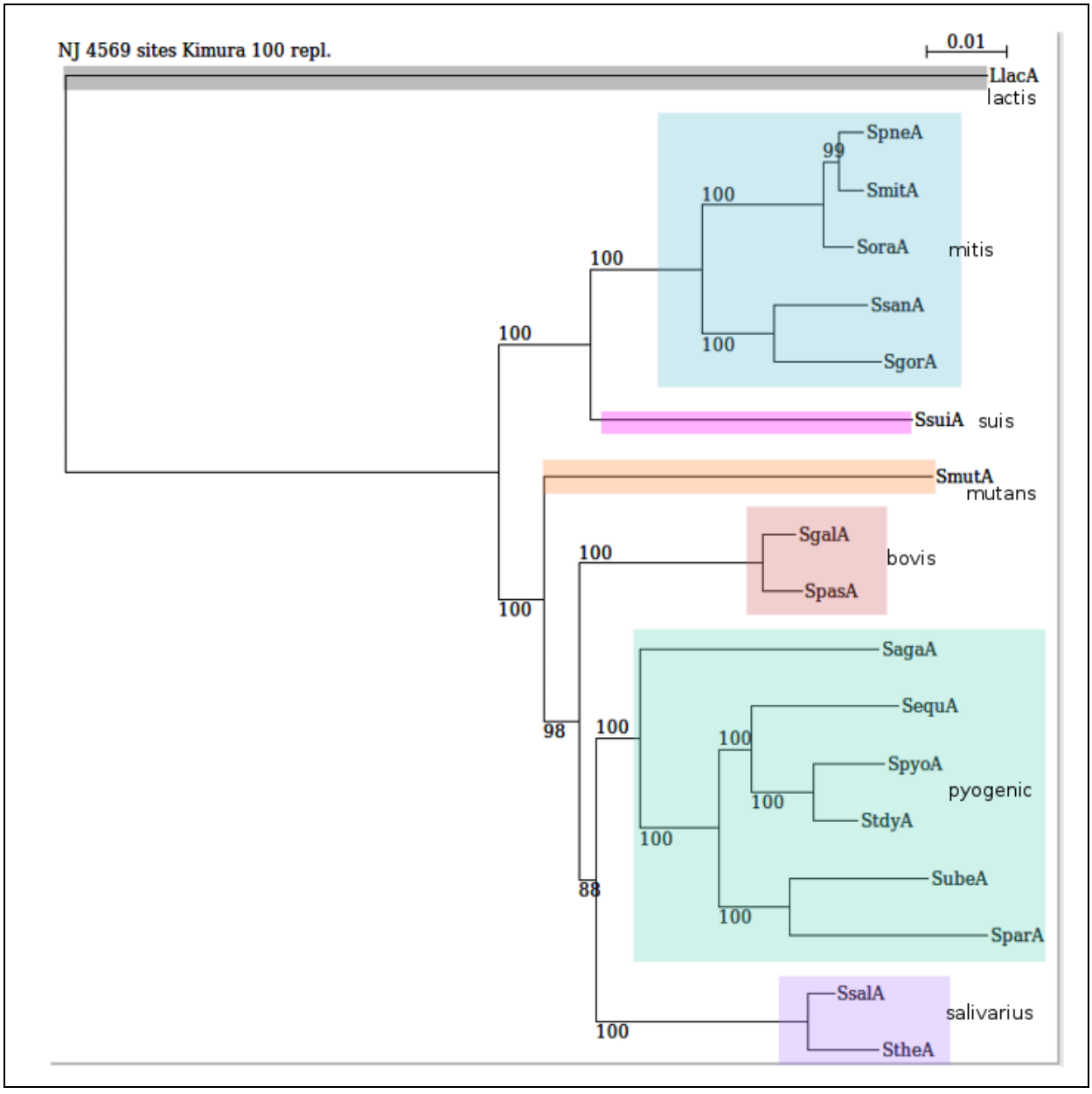


Arbre obtenu en concaténant les 5 alignements (méthode PhyML, matrice LG paramètres par défaut).

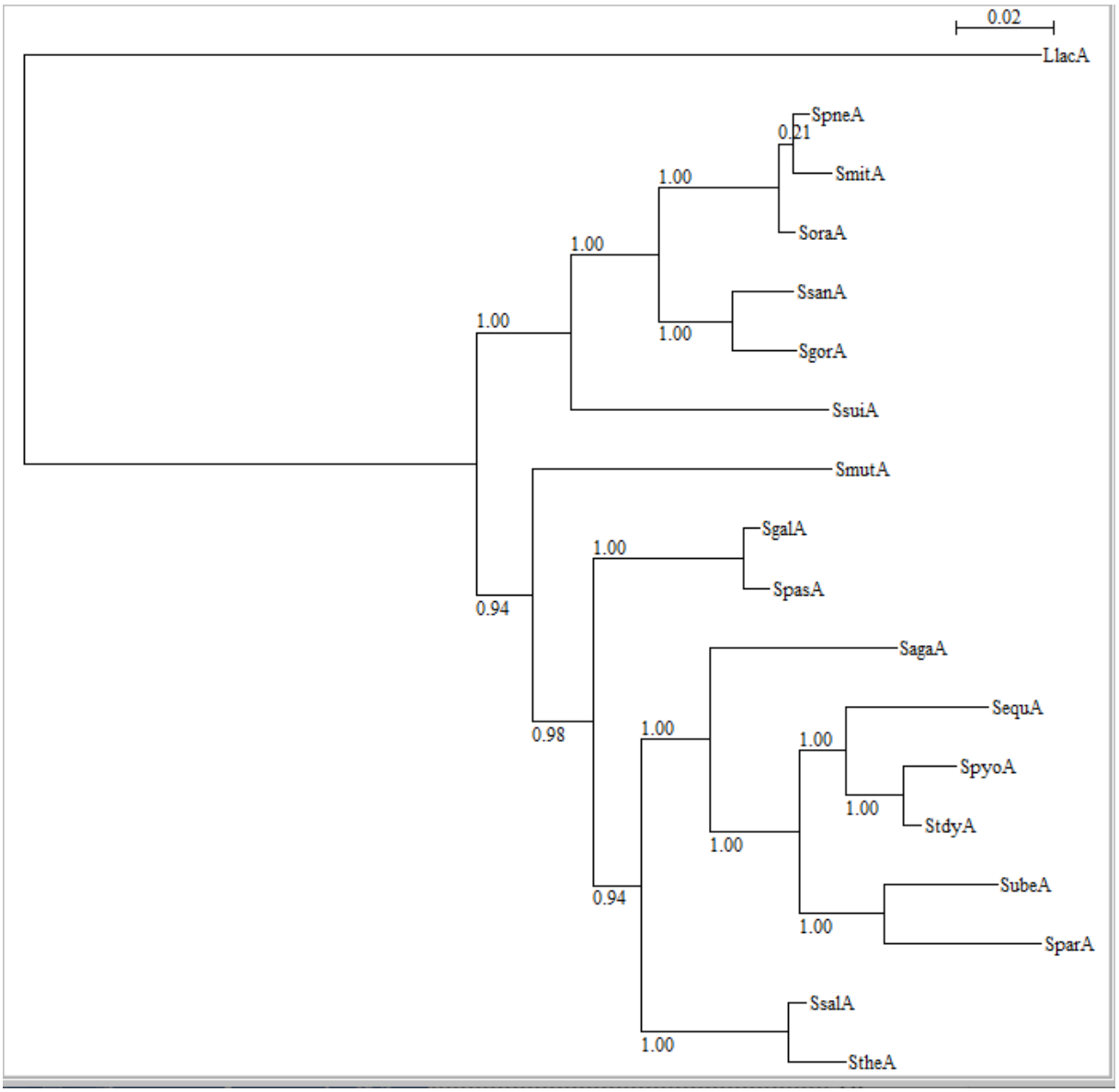


L'arbre obtenu en concaténant 5 fichiers est remarquablement cohérent avec la classification en groupe des Streptocoques. Nous pouvons également remarquer de bonnes valeurs de support des branches (aLRT).

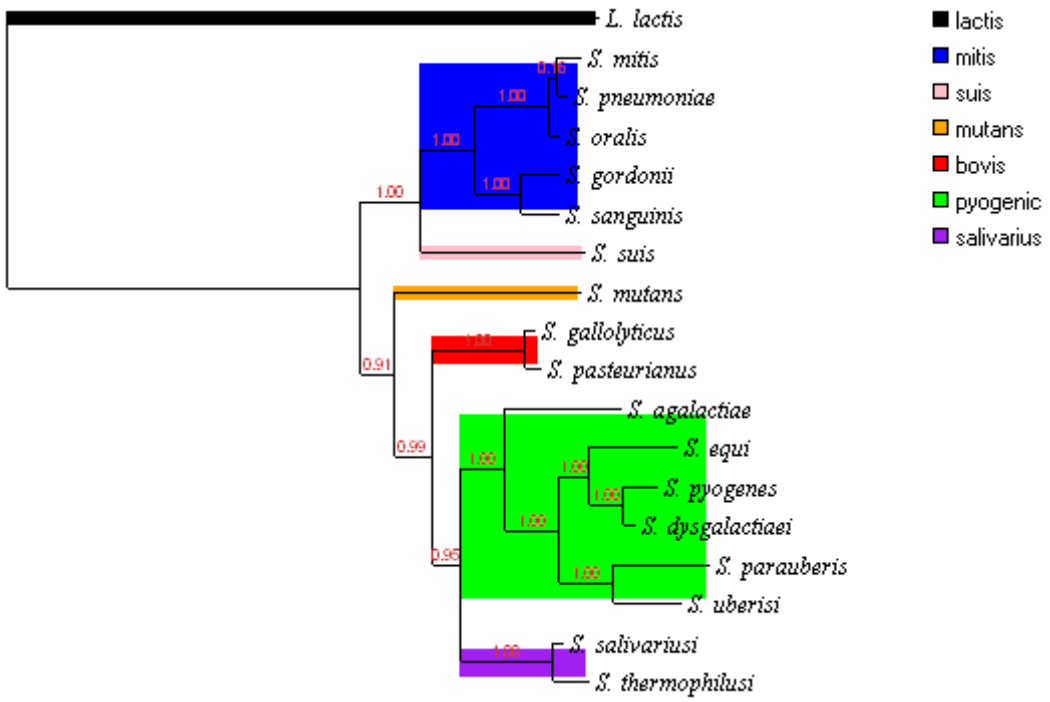
Arbre obtenu sur l'alignement concaténé des 43 familles de protéines avec la méthode NJ distance de Kimura



Arbre obtenu sur l'alignement concaténé des 43 familles de protéines avec PhyML

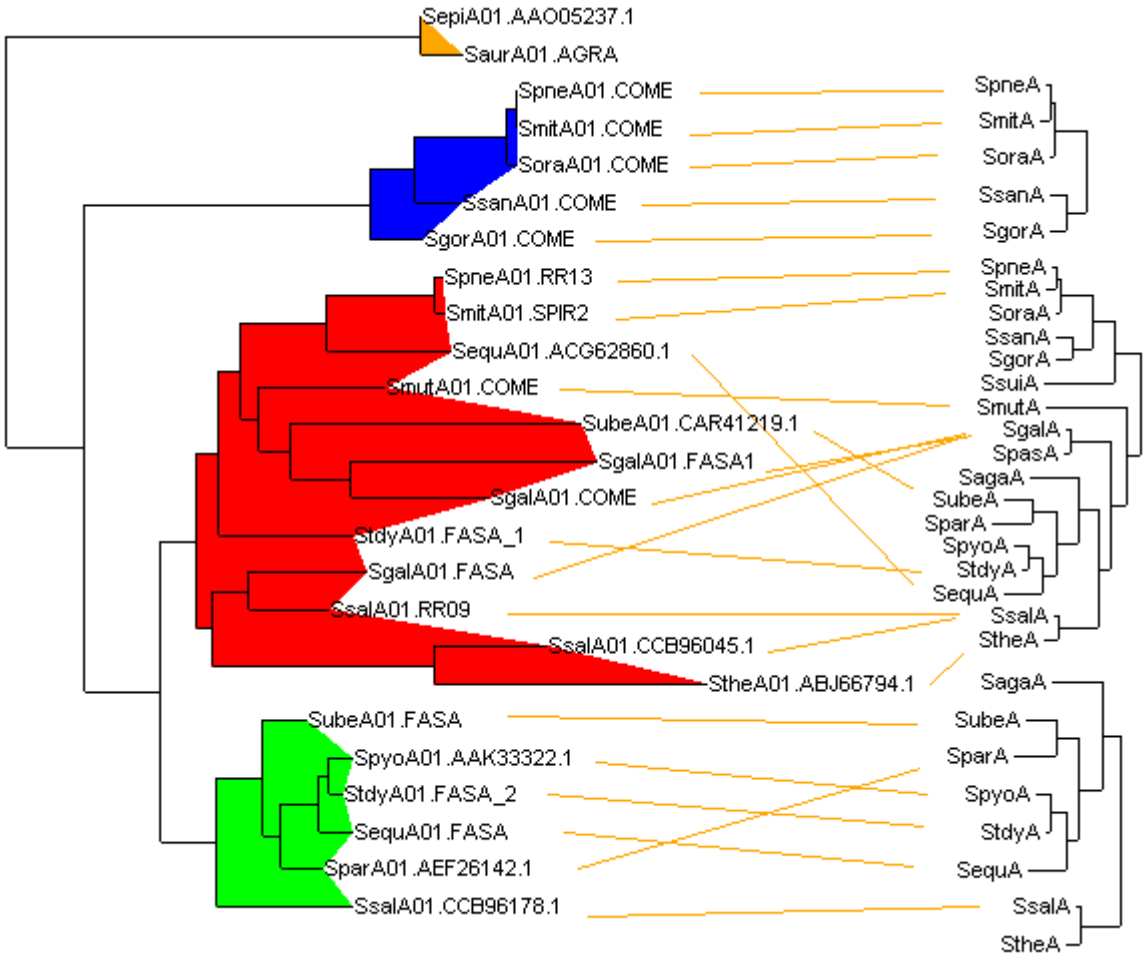


Quelle que soit la méthode utilisée, les arbres ont exactement la même topologie avec de très bonnes valeurs de bootstrap, ce qui suggère que l'alignement des 43 protéines ribosomiques contient suffisamment d'information phylogénétique pour résoudre les relations phylogénétiques entre ces espèces. Nous pouvons néanmoins observer que la branche menant à *S. pneumoniae*, et *S. mitis* a une très faible valeur de aLRT avec PhyML. Ce résultat pourrait être dû à la faible divergence des protéines ribosomiques qui n'apporteraient pas suffisamment d'information pour les espèces qui auraient divergées récemment.



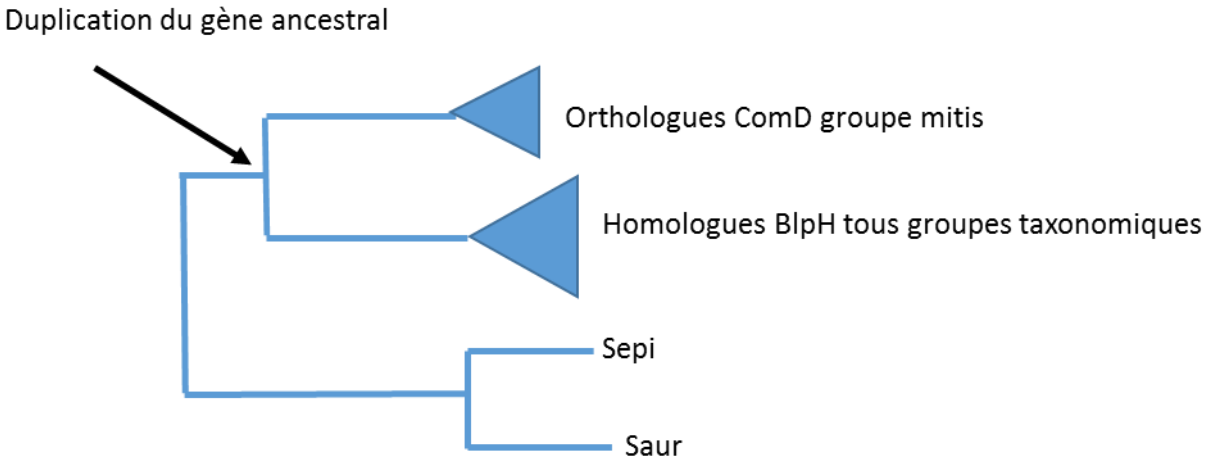
Comparaison de la topologie de l'arbre obtenu sur ComE avec l'arbre des espèces

Quand nécessaire le sous-arbre des espèces a été extrait pour le mettre en face de la topologie des sous-groupes de ComE (bleu et vert). Il y a une très bonne congruence entre les deux pour le sous-arbre bleu des séquences ComE (et même meilleur avec ComD !). Nous voyons cependant pour les autres sous-groupes, quelques différences avec l'arbre des espèces.
Pour le sous-arbre rouge, *S. equi* est à une place non attendue par rapport à la phylogénie des espèces. Sgal possède 3 paralogues dont deux sont correctement placé avec Sube et pourraient correspondre à une duplication dans Sgal. Par contre, le sous-groupe formé par la troisième copie de Sgal et SsalA1.RR9 pose un problème.
Dans le sous-arbre vert, seule la séquence de Spar pose un petit problème car elle ne possède pas un ancêtre commun avec Sube mais cependant se branche juste après Sube en groupe externe de Spyo, Stdy et Sequi.



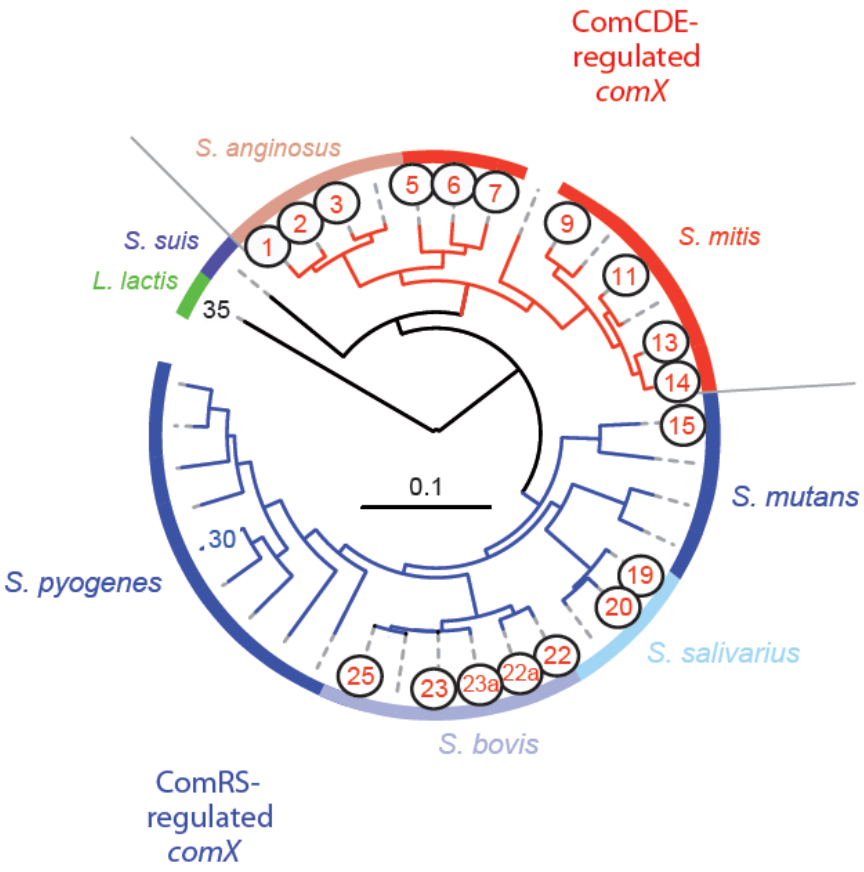
Comme nous l'avons déjà remarqué, les séquences orthologues à ComE et ComD de *S. pneumoniae* ne sont présentes que dans le groupe *mitis* (groupe bleu). L'arbre obtenu avec les 43 protéines ribosomiques supposé représenter la phylogénie des espèces de Streptocoques montre que les différentes espèces du groupe *mitis* descendent bien d'une espèce ancêtre commune. Nous pouvons donc émettre l'hypothèse que le système ComDE a été acquis par cette espèce ancêtre et hérité ensuite par spéciation par les espèces actuelles.

Cependant, que ce soit dans les arbres obtenus avec une méthode de distance (NJ) ou avec une méthode de maximum de vraisemblance (PhyML), les sous-arbres correspondant aux systèmes ComDE branchent à l'extérieur du sous-arbre comportant les autres séquences. Ceci indique que ce système n'a probablement pas été acquis par duplication mais par transfert horizontal par l'ancêtre commun. En effet une duplication aurait dû se traduire par la topologie d'arbre suivante (idem pour ComE) :

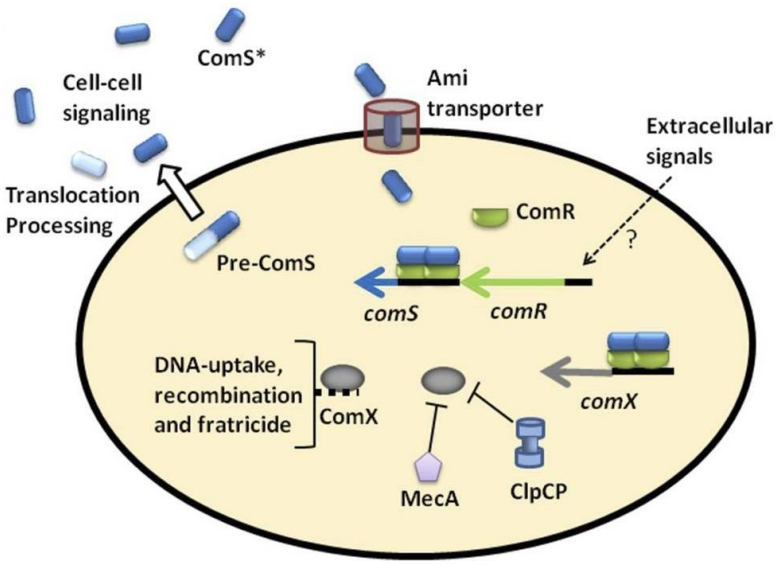


Cette hypothèse d'acquisition par HGT est soutenue par l'observation que l'opéron *comCDE* possède une composition en G+C plus faible de celle de l'ensemble du génome (~30%) et qu'il est bordé par deux gènes codant pour des ARNt, deux caractéristiques associées aux ilots de pathogénies et plus généralement aux transferts horizontaux de gènes.

La cascade de régulation identifiée chez S. *pneumoniae* serait partagée par les autres membres du groupe *mitis*. En effet, chez ces espèces, l'opéron *comCDE* est localisé à proximité de l'origine de réplication du chromosome, elles codent pour un transporteur orthologue à ComAB et possèdent deux copies du gène *comX*.



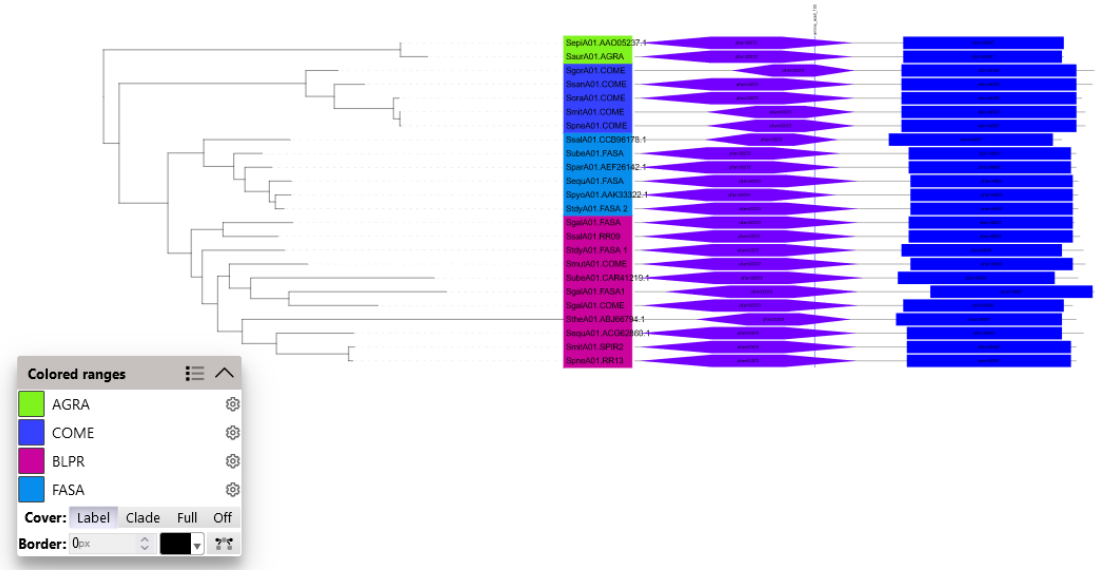
La compétence est donc régulée chez *S. mutans* par un autre système ComRS.



Régulation de la transformation par le système ComRS. On ne connait pas la régulation du niveau basal de l'opéron *comRS*, ce pourrait être un signal extracellulaire. Le produit du gène *comS*, le Pre-Coms, est exporté et maturé (ComS*) par un transporteur qui n'a pas encore été identifié. ComS* est importé dans la cellule par le transporteur d'oligopeptides Ami. Dans le cytoplasme, il se fixe à ComR et l'active. ComR activé se fixe sur les boites ECom au niveau des promoteurs de *comS* et *comX*, conduisant à une amplification du signal (boucle auto catalytique) et à l'expression des gènes tardifs. Les protéines ClpC and MecA préviennent l'accumulation de ComX dans les conditions qui ne sont pas optimum pour le développement de la compétence.

Exemple de représentation et annotation des arbres avec iTOL.

Ici l'arbre obtenu avec PhyML sur ComE où les feuilles sont coloriées en fonction de la famille de protéines et les domaines Pfam présents dans ces protéines sont annotés.



Arbre des espèces annoté avec iTOL avec la distribution des différentes familles de protéines dans chacun des génomes