

PARTIE VI : TP

PARTIE VI : TP

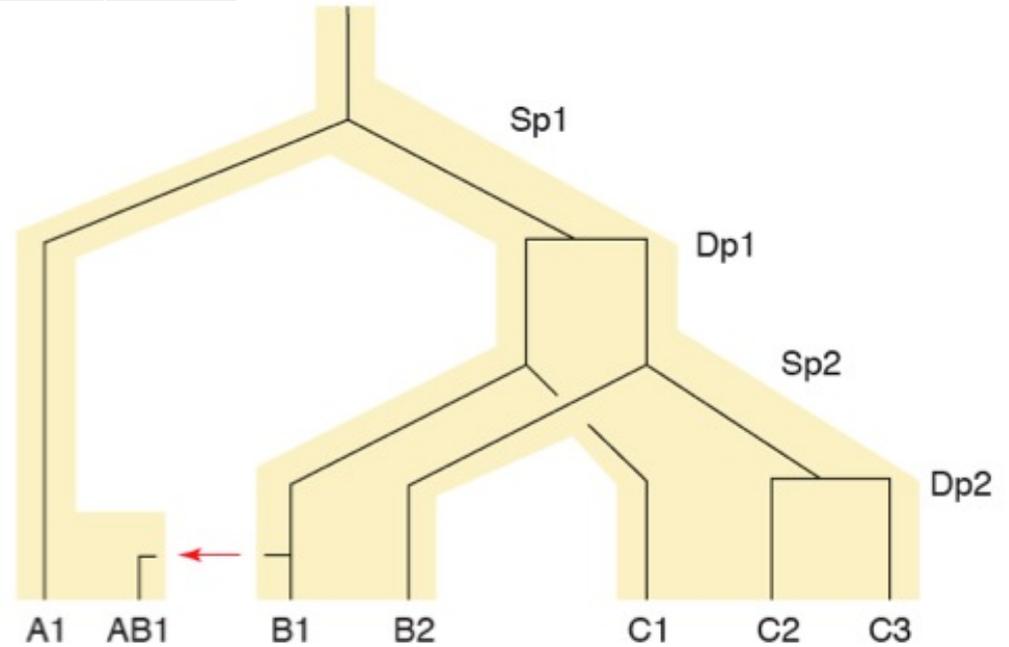
TP 0 / ORTHOLOGIE, PARALOGIE

Qualifiez les relations entre chaque paire de gènes représentés dans l'arbre ci dessous.

P paralogue A analogue
O orthologue X xenologue

Orthologie, Paralogie

	A1	AB1	B1	B2	C1	C2	C3
A1							
AB1							
B1							
B2							
C1							
C2							
C3							



PARTIE VI : TP

TP 1 / ANALYSE DES THIOREDOXINES

Le but du TP est d'étudier la phylogénie du gène de la thiorédoxine pour les 10 organismes suivants:

1. *Helicobacter pylori* (P66928)
2. *Bacillus subtilis* (P14949)
3. *Homo sapiens* (P10599)
4. *Penicillium chrysogenum* (P34723)
5. *Listeria monocytogenes* (POA4L3)
6. *Escherichia coli* (POAA25)
7. *Gallus gallus* (P08629)
8. *Mus musculus* (P10639)
9. *Neurospora crassa* (P42115)
10. *Drosophila melanogaster* (P47938)

Acquisition des données

Q 1. Allez chercher les séquences en faisant une recherche par mots-clés dans la banque de données Uniprot (<http://www.uniprot.org>). Pour que la recherche soit plus rapide, vous pouvez faire une recherche multi-critères, sur la description (thioredoxin) et l'organisme. Pour chaque organisme, vérifiez qu'il s'agit de la bonne protéine. Sauvegardez la séquence au format FASTA, et notez la taxonomie associée.

Q 2. A la main, faites une classification rapide à partir des informations taxonomiques fournies par Swissprot.

Q 3. Charger le fichier fasta des séquences de thiorédoxine dans Galaxy. Réalisez un alignement multiple avec Muscle. Le résultat se trouve dans le fichier Muscle Alignment.

Q 4. Avant de construire une phylogénie, il faut s'assurer de la correction de l'alignement multiple. Pour cela, il est possible de tirer parti de connaissances extérieures sur la fonction des séquences étudiées, en vérifiant par exemple que les domaines connus sont bien alignés. Prosite (<http://prosite.expasy.org>) permet de localiser les sites connus. Lancez-le sur une des séquences de thioredoxine. Vous devez trouver un domaine caractéristique de la famille. Quel est le motif associé à ce domaine ? Vérifiez sur l'alignement que le motif est présent et bien aligné dans toutes les séquences. Cela garantit que l'alignement est pertinent au moins dans cette région.

Une petite remarque : en général, il est également souhaitable de "nettoyer" l'alignement multiple, en supprimant les régions non informatives, celles qui sont mal conservées. Sur cet exemple, comme les séquences sont relativement bien conservées, cela n'est pas nécessaire.

Méthode de reconstruction

Il existe plusieurs techniques pour reconstruire un arbre phylogénétique à partir de données moléculaires.

Pour ce TP, vous allez appliquer une méthode de distance, appelée Neighbor Joining.. Elle regroupe les séquences deux par deux progressivement à partir de la matrice de distances.

Q 5. Reprenez la page de résultat de Muscle dans Galaxy. Pour calculer la matrice des distances à partir de l'alignement multiple, sélectionnez l'outil Phylip protdist. Choisir Muscle Alignement dans le champ Sequences alignment file. Choisissez 'PAM' comme modèle et lancez le calcul. La matrice de distances est obtenue dans le fichier Distance matrix. A vue d'œil, quels sont les organismes les plus proches, les plus éloignés ?

Construction de l'arbre

Q 6. Avec la matrice, nous allons calculer un arbre. Pour cela, sélectionnez « BioNJ » dans la liste des outils.

Q 7. Visualiser l'arbre avec Figtree.

Q 8. Comparez l'arbre avec la classification communément admise que vous avez retracée en début de TP.

Q 9. On observe trois espèces qui se distinguent bien des autres. En observant l'alignement multiple, expliquez pourquoi. Expliquez pourquoi les mammifères et le poulet sont regroupés. Qu'est-ce qui différencie le poulet des mammifères ?

Evaluation des résultats

L'arbre que vous venez d'obtenir semble globalement réaliste. Mais il se peut que localement, ou concernant les longueurs des branches, celui-ci ne soit pas correct (la longueur des branches est indicative de l'évolution). Il est possible de tester la robustesse d'un arbre phylogénétique avec des techniques de **bootstrap**. L'idée du bootstrap est que si l'on effectue des petits changements sur les données on doit être capable de retrouver le même arbre. Concrètement, à partir de l'alignement multiple initial obtenu avec Muscle, on construit des nouveaux alignements qui sont obtenus en échangeant des colonnes. C'est légitime car les méthodes de reconstruction phylogénétique supposent que les sites évoluent de manière indépendante. Pour chaque nouvel alignement, on construit une matrice de distance, puis l'arbre correspondant.

Q 10. Il est possible de faire une analyse avec bootstrap à partir de Phylip prodist. Tester la robustesse avec un bootstrap de 50. Consulter le fichier Distance matrix généré.

Q 11. Il est possible de calculer les 50 arbres correspondants avec BioNJ. Par défaut, BioNJ va générer 50 arbres s'il y a 50 matrices en entrée ; il est donc habituellement nécessaire d'enchaîner par la création d'un arbre consensus (étape qui peut se faire avec l'outil consense ou bien directement avec BioNJ en cochant l'option Generate consensus tree, qui va générer un arbre consensus majoritaire). Dans Galaxy, générez les 50 arbres avec BioNJ.

Q 12. A partir des 50 arbres du bootstrap, il faut en obtenir un seul. En faisant un consensus strict, on peut faire l'union de ces 50 arbres, ce qui permet de mettre en évidence les nœuds mal résolus dans les arbres. Utilisez le programme consense dans Galaxy pour générer le consensus strict.

Q 13. Que pensez-vous de l'arbre obtenu ? Pouvons-nous réellement tirer des conclusions de l'arbre obtenu sans bootstrap ?

Q 14. L'utilisation du consensus peut aussi se faire d'une autre manière en affectant à chaque nœud des valeurs de bootstrap. A chaque nœud, on indique le degré de confiance, en comptant le nombre d'arbres pour lesquels les deux groupes issus du nœud sont séparés. C'est ce qu'on appelle le consensus majoritaire. Revenez au calcul du consensus. Choisissez cette fois l'option consensus majoritaire et générez l'arbre.

Q 15. Visualiser l'arbre dans Figtree.

Q 16. Combien de fois le groupe poulet-mammifères est-il retrouvé ?

Q 17. Les nœuds auxquels on peut avoir confiance sont ceux ayant une valeur de bootstrap supérieure à 80%. Affichez dans Figtree les valeurs de bootstrap et identifiez ces nœuds.

(Attention, consense ne renvoie pas les vraies valeurs de bootstrap, mais le nombre de fois que chaque groupe apparaît dans les arbres en entrée. Donc si on a demandé 50 itérations de bootstrap, les valeurs des branches de l'arbre consensus seront comprises entre 0 et 50.)

Répéter l'opération avec l'outil SEAVIEW 4

<http://www.phylogeny.fr/>

PARTIE VI : TP

TP 2 / DE L'ORIGINE DES CICHLIDÉS

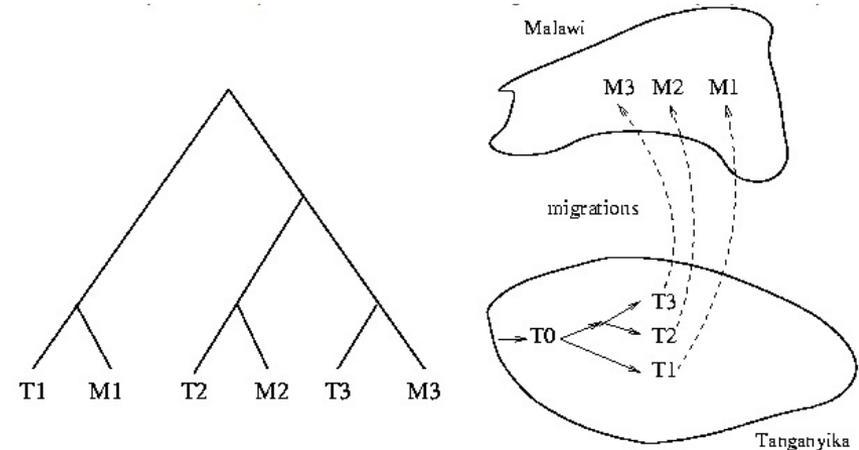
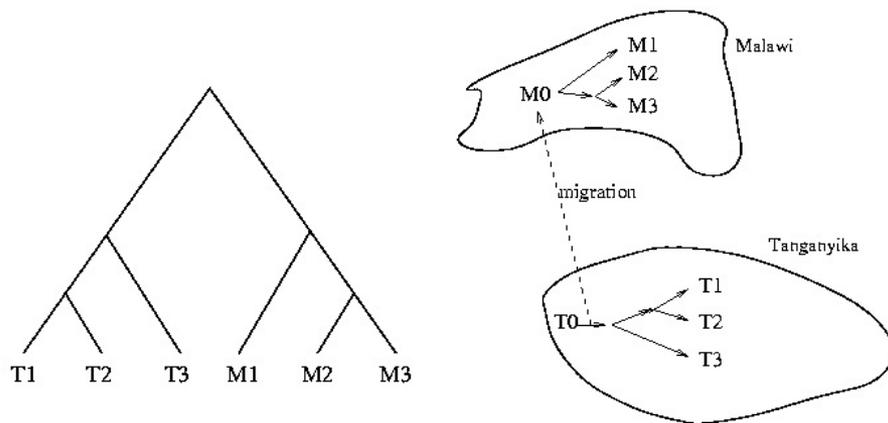
De l'origine des cichlidés

- Les *Cichlidés* sont une famille de poissons d'eau douce. Nous nous intéressons à leur évolution dans deux grands lacs d'Afrique de l'Est: le lac Malawi et le lac Tanganyika.
- Nous travaillerons plus spécialement sur douze espèces, six vivant dans le lac Tanganyika, et six dans le lac Malawi.
- *Petrochromis sp*, *Bathybates ferox*, *Iobochilotes labiatus*, *Tropheus brichardi*, *Cyphotilapia frontosa* et *Julidochromis ornatus* pour le lac Tanganyika ;
- *Petrotilapia sp.*, *Rhamphochromis sp.*, *Placidochromis milomo*, *Pseudotropheus microstoma*, *Cyrtocara moori* et *Melanochromis auratus* pour le lac Malawi.
- Ce qui est mystérieux, et que nous tacherons d'élucider, c'est que ces douze espèces se ressemblent deux à deux. A chaque espèce du lac Tanganyika (T), on peut associer une espèce du lac Malawi (M) sur des critères morphologiques:

Tanganyika	Malawi	Ressemblances
<i>Petrochromis sp</i>	<i>Petrotilapia sp.</i>	Herbivore très efficace, adapté au raclage des algues, même habitat
<i>Bathybates ferox</i>	<i>Rhamphochromis sp.</i>	Gros prédateurs pélagiques, même morphologie hydrodynamique
<i>lobochilotes labiatus</i>	<i>Placidochromis milomo</i>	Prédateurs pétricoles, grosses lèvres charnues et molles
<i>Tropheus brichardi</i>	<i>Pseudotropheus microstoma</i>	Herbivore, bouche infère et dents bicuspidés, petite taille
<i>Cyphotilapia frontosa</i>	<i>Cyrtocara moori</i>	Les mâles portent une bosse frontale
<i>Julidochromis sp.</i>	<i>Melanochromis auratus</i>	Rayures horizontales sur le corps

Convergence des caractères: Il y a eu évolution indépendante des poissons dans les deux lacs. Les caractères similaires sont dus au milieu similaire dans lequel les poissons vivent. Ces organismes ont subi indépendamment les mêmes pressions évolutives. Le lac Tanganyika est plus ancien que le lac Malawi. On peut supposer qu'il y a d'abord eu une migration de T vers M, puis évolution indépendante dans les deux lacs.

Homologie: toutes les ressemblances sont dues à l'existence d'un ancêtre commun. Il y a donc eu plusieurs événements de migration de T vers M (un pour chaque espèce).



De l'origine des cichlidés

Votre but est de déterminer quelle hypothèse est la plus vraisemblable en vous basant sur une analyse phylogénétique.

Prendre le fichier contenant les 12 séquences : all.fas.

Q 1. Comme précédemment, faites l'analyse avec Galaxy. Pour le calcul de distance (Phylip dnadist), utiliser une matrice Kimura-2-parameters.

Q 2. Pour enraciner l'arbre précédent, et avoir davantage de confiance dans le scénario évolutif proposé, il faut utiliser un groupe externe (outgroup). On vous propose deux jeux de séquences de la famille des *Cichlidés*, l'un provenant de l'espèce *Geophagus brasiliensis* d'Amérique du Sud (**newworld.fas**), l'autre de l'espèce *Cyprichromis Leptosoma* de l'île de Malasa située dans le lac Tanganyika. (**malasaisland.fas**). Lequel des deux jeux utiliser pour l'enracinement? Pour quelle raison? Concaténer le fichier all.fas et le fichier fasta de l'outgroup choisi avec l'outil Concatenate datasets. Construire l'arbre phylogénétique de l'ensemble, et trouver l'enracinement de l'arbre des 12 espèces.

Q 3. Depuis environ quelle période les espèces de *Cichlidés* d'Amérique du Sud ont-elles divergé de leurs cousins éloignés d'Afrique ? Vous pouvez vous aider de ce site. (<http://www2.ggl.ulaval.ca/personnel/bourque/s4/pangee.auj.html>)

PARTIE VI : TP

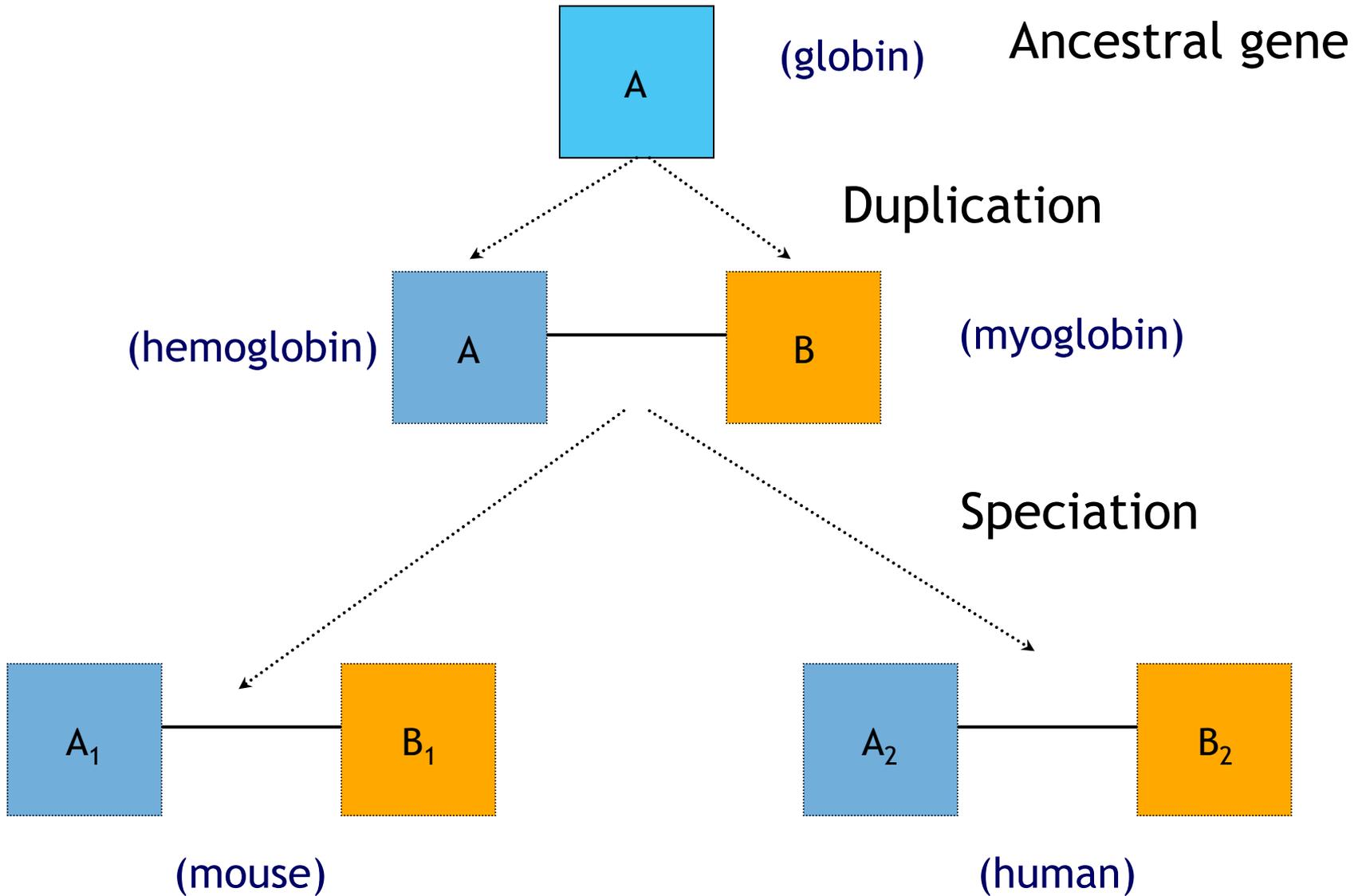
TP 3 / PHYLOGENIE DES HÉMOGLOBINES

Ce jeu de données a été constitué à partir de l'ensemble des séquences d'hémoglobines de la base de données non redondante SWISSPROT. Pour les espèces animales, seules celles ayant au moins 3 chaînes de disponibles ont été conservées

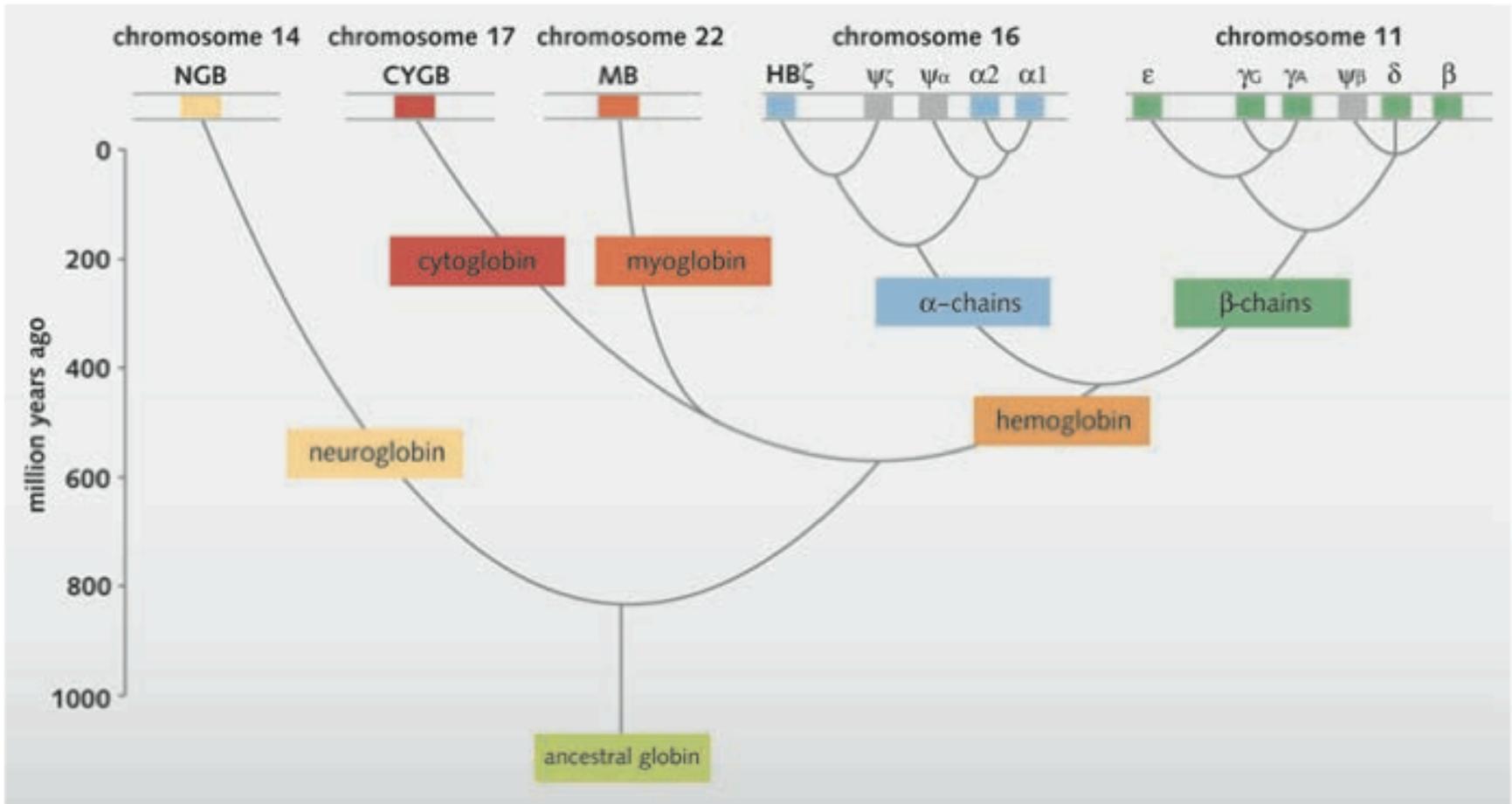
- Trouver les événements de spéciation et de duplication.
- Quels gènes sont orthologues/paralogues?

- ADEL : *Aldabrachelys elephantina* : tortue géante des seychelles
- CHRPI : *Chrysemys picta bellii* : tortue peinte de l'ouest
- PASMO : *Passer montanus* : Moineau friquet
- PHACO : *Phasianus colchicus* : Faisan de Colchide
- CTEGU : *Ctenodactylus gundi* : Goundi de l'Atlas : rongeur
- PIG
- HUMAN
- NOTCO : *Notothenia coriiceps* : Poisson antarctique
- TRICR : *Triturus cristatus* : triton
- LUCPE : *Lucina pectinata* : palourde (manque un heme-binding histidine résidu)
- MEDSA : *Medicago sativa* : Luzerne (plante globine : capteur O₂ : pas transport)
- CASGL : *Casuarina glauca* : arbre tropical des organes dérivés suivants : les nodules fixateurs d'azote, les mycorhizes et les racines touffes qui facilitent l'assimilation d'éléments minéraux.

Phylogénie des hémoglobines



Phylogénie des hémoglobines



PARTIE IV : TP

TP 4 / SUPER ALIGNEMENTS

Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae;
 Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae;
 Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
 Bacteria; Proteobacteria; Gammaproteobacteria; Alteromonadales;

>BapSg *Buchnera aphidicola* Sg

- >CFT073 *Escherichia coli* CFT073 (UPEC)
- >EcoRIMD Enterohemorrhagic *Escherichia coli* (EHEC)
- >MG1655 *Escherichia coli* K-12 MG1655
- >PstDC3000 *Pseudomonas syringae* DC3000
- >SCRI1043 *Pectobacterium atrosepticum* SCRI1043
- >SenChB67 *Salmonella Choleraesuis* SC-B67
- >SenLT2 *Salmonella Typhimurium* LT2
- >SenTy2 *Salmonella Typhi* Ty2
- >Sfl301 *Shigella flexneri* 301
- >Wbrevi *Wigglesworthia brevialpis*
- >YPCO92 *Yersinia pestis* CO92
- >Ypt32953 *Yersinia pseudotuberculosis* IP32953
- >EccWPP14 *Pectobacterium carotovorum* subsp. *carotovorum* WPP14

>BspAPS *Buchnera aphidicola* APS

- >ECH3937 *Erwinia chrysanthemi* 3937
- >EDL933 *Escherichia coli* EDL933 (EHEC)
- >PluTTO1 *Photobacterium luminescens* TTO1
- >Sbo227 *Shigella boydii* 227
- >Sdy197 *Shigella dysenteriae* 197
- >SenCT18 *Salmonella Typhi* CT18
- >SenPA9150 *Salmonella Paratyphi_A* ATCC9150
- >Sfl2457T *Shigella flexneri* 2457T
- >Sso046 *Shigella sonnei* 046
- >YP91001 *Yersinia pestis* 91001
- >YPKIM *Yersinia pestis* KIM
- >SonMR-1 *Shewanella oneidensis* MR-1

Q1 Nous allons construire un arbre des espèces suivantes en récupérant les séquences d'ARN16S dans genbank.

Q2 Nous disposons également d'autres séquences d'autres gènes :

- *atpD (ATP synthase subunit beta)*
- *bioB (biotin synthase)*
- *carA (carbamoyl phosphate synthetase small subunit)*
- *dnaJ (chaperone protein DnaJ)*
- *murA (UDP-N-acetylglucosamine 1-carboxyvinyltransferase (peptidoglycane synthase)),*
- *rpoB (RNA polymerase, subunit beta)*

Générez et comparez 3 arbres d'espèces à partir de l'alignement des 27 gènes orthologues. Lequel vous semble le plus proche de la phylogénie 16S?

- **Q3** Une technique plus fine et plus robuste pour faire des études de phylogénie est d'utiliser les séquences de plusieurs gènes en même temps: on parle de superarbre ou d'arbre concaténés.

- Une technique plus fine et plus robuste pour faire des études de phylogénies est d'utiliser les séquences de plusieurs gènes en même temps: on parle de superarbre ou d'arbre concaténés.
 - Clann software for inferring phylogenetic supertrees :
<http://bioinf.nuim.ie/clann/>
 - <http://genome.cs.iastate.edu/CBL/RFsupertrees/>
 - http://genome.cs.iastate.edu/supertree/userdata_analysis/userdata_analysis.html

Nous disposons du fichier contenant les 6 gènes concaténés.
 Recommencez l'analyse avec l'ensemble des gènes concaténés,
 comparez et interprétez...

PARTIE VI : TP

TP 5 / TRICHOTOMIE

HUMAIN / CHIMPANZÉ/GORILLE

- A partir de 3 jeux de données
 - ARN 16S mitochondrial
 - NADH4L gène mitochondrial
 - BRCA1 gène nucléaire sensibilité au cancer

- *Ateles sp* (Atèle sp.) Famille : CEBIDAE
- *Lagothrix lagotricha* (Lagothriche de Humboldt) Famille : CEBIDAE
- *Alouatta palliata* (Hurleur à pèlerine) Famille : CEBIDAE
- *Aotus trivirgatus* (Douroucouli) Famille : CEBIDAE
- *Cebus apella* (Sajou apelle) Famille : CEBIDAE
- *Pithecia pithecia* (Saki à tête pâle) Famille : CEBIDAE
- *Callicebus moloch* (Titi moloch) Famille : CEBIDAE
- *Saimiri sciureus* (Singe écureil) Famille : CEBIDAE
- *Leontopithecus rosal* (Singe Lion) Famille : CALLITRICHIDAE
- *Callithrix jacchus* (Ouistiti; marmouset commun) Famille : CALLITRICHIDAE
- *Cebuella pygmaea* (Ouistiti mignon) Famille : CALLITRICHIDAE
- *Callimico goeldii* (Tamarin de Goeldi) Famille : CALLITRICHIDAE
- *Saguinus geoffroyi* (Tamarin de Geoffroy) Famille : CALLITRICHIDAE
- *Nasalis larvatus* (Nasique) Famille : CERCOPITHECIDAE
- *Papio hamadryas* (Babouin hamadryas) Famille : CERCOPITHECIDAE
- *Hylobates lar* (Gibbon à mains blanches) Famille : HYLOBATIDAE
- *Pongo pygmaeus* (Orang-outan) Famille : HOMINIDAE
- *Pan troglodytes* (Chimpanzé) Famille : HOMINIDAE
- *Pan paniscus* (Bonobo) Famille : HOMINIDAE
- *Gorilla gorilla* (Gorille) Famille : HOMINIDAE
- *Homo sapiens* (Homme) Famille : HOMINIDAE

Q1 16S :

- Jouer sur la sélection de sites et d'espèces
- 21 espèces 7 espèces (gap ou -global gap removal)

Q2 NADH4L (! Code mitochondrial):

- NJ
- Comparer les modèles Ka et Ks
- diminuer le nombre d'espèces (17 à 5) en Ks et observer la variation de % bootstrap, association H/G/C et distance
- Utiliser uniquement les codons 1 et/ou 2 et/ou 3. Comparer les distances

Q3 BRCA1 :

- même chose

PARTIE VI : TP

TP 6 / PHYLOGÉNIE

DES PROTÉOBACTÉRIES ALPHA

Généralités.

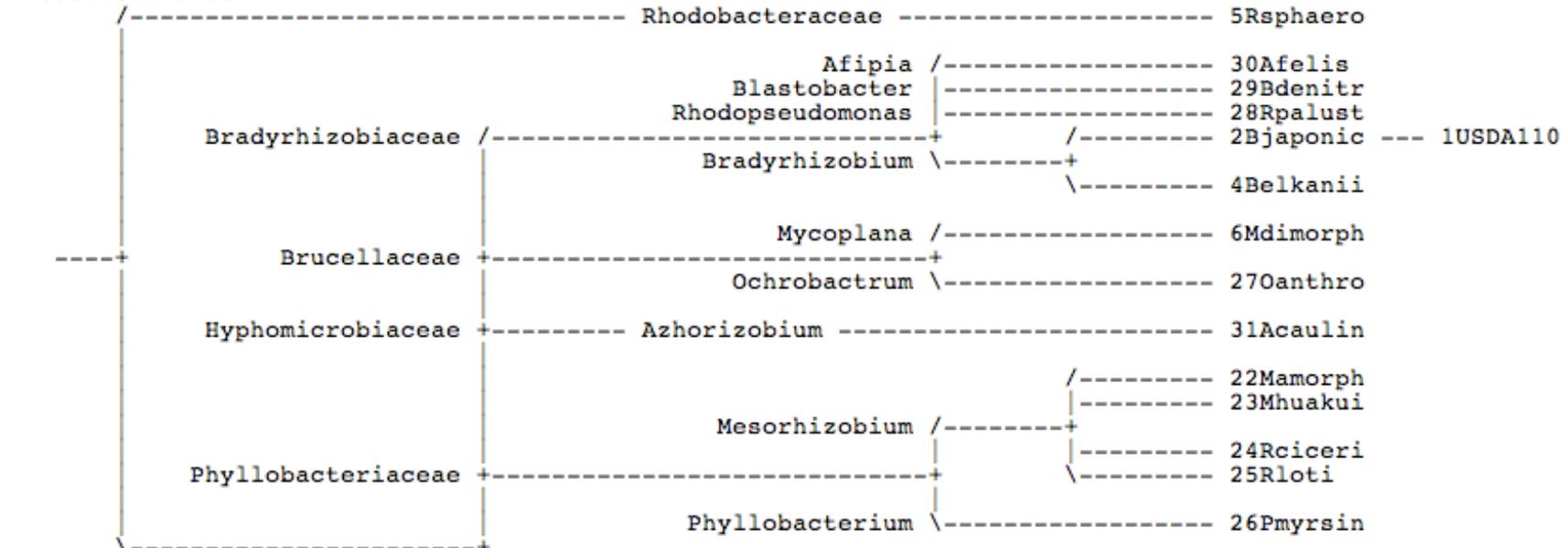
Le but de ce TP est d'étudier la classification phylogénétique d'un groupe de Rhizobia (alpha-protéobactéries) en se basant sur deux gènes issus de l'opéron *rrn*, les gènes des ARNr 16S et 23S.

L'hypothèse à tester concerne la qualité de la classification connue actuellement, qui repose essentiellement sur l'utilisation du gène 16S.

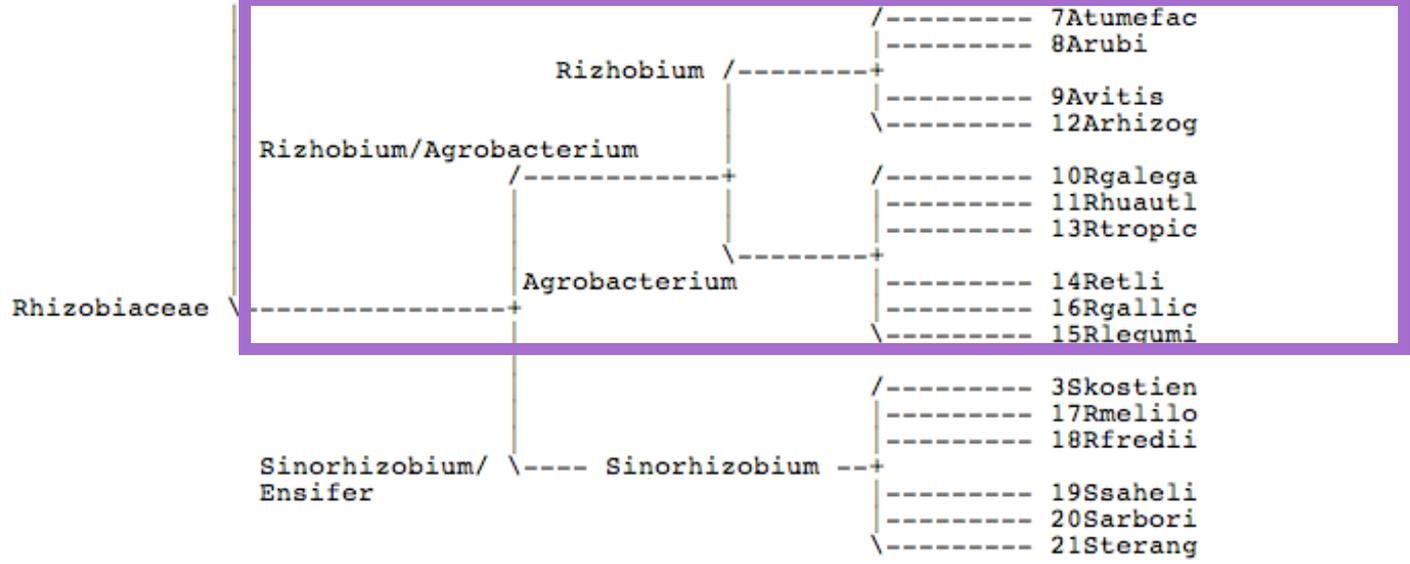
Jeu de données

- *Afipia felis*
- *Agrobacterium rhizogenes*,
- *Agrobacterium rubi*,
- *Agrobacterium tumefaciens*,
- *Agrobacterium vitis*,
- *Azorhizobium caulinodans*,
- *Blastobacter denitrificans*,
- *Bradyrhizobium elkanii*,
- *Bradyrhizobium japonicum*,
- *Bradyrhizobium japonicum USDA 110*,
- *Mesorhizobium amorphae*,
- *Mesorhizobium ciceri*,
- *Mesorhizobium huakuii*,
- *Mesorhizobium loti*,
- *Mycoplana dimorpha*,
- *Ochrobactrum anthropi*,
- *Phyllobacterium myrsinacearum*,
- *Rhizobium etli*,
- *Rhizobium galegae*,
- *Rhizobium gallicum*,
- *Rhizobium huautlense*,
- *Rhizobium leguminosarum*,
- *Rhizobium tropici*,
- *Rhodobacter sphaeroides*,
- *Rhodopseudomonas palustris*,
- *Sinorhizobium arboris*,
- *Sinorhizobium fredii*,
- *Sinorhizobium kostiense*,
- *Sinorhizobium meliloti*,
- *Sinorhizobium saheli*,
- *Sinorhizobium terangaiae*.

Rhodobacterales



Rhizobiales

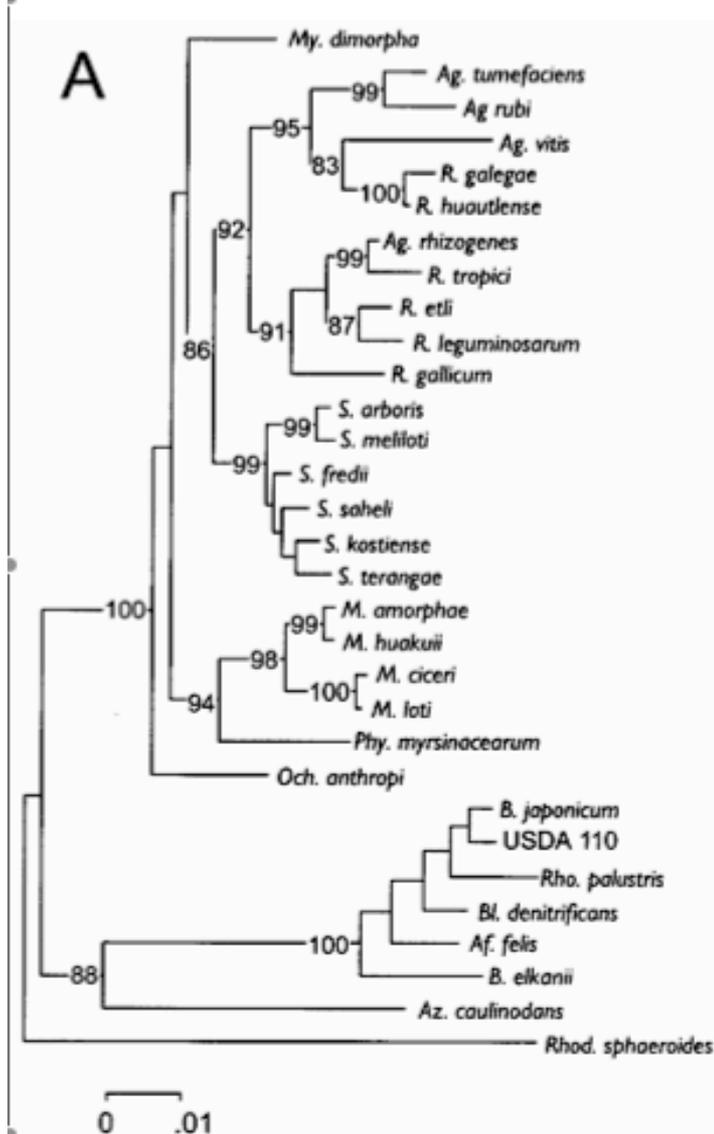


Q1 : A partir de 2 jeux de données 16S et 23S évaluer les différentes phylogénies obtenues

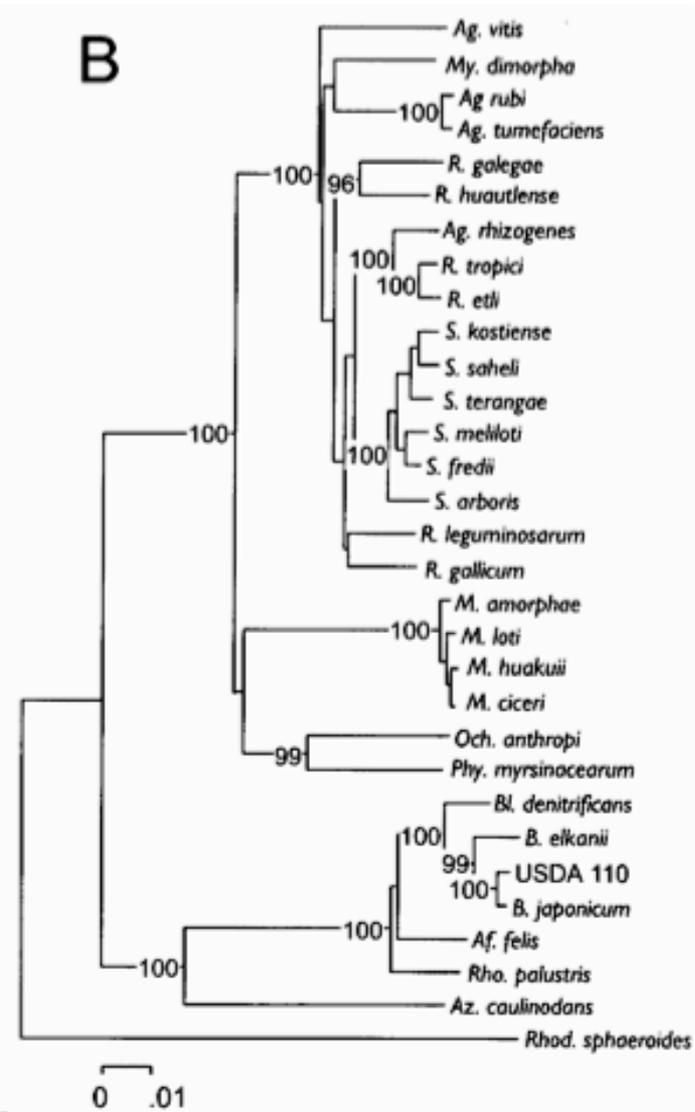
Aligner les séquences dans Galaxy avec le logiciel M-coffee.

- M-Coffee lance tous les logiciels d'alignement ci-dessous et sélectionne le meilleur :
- T-Coffee <http://www.tcoffee.org>
- ClustalW <ftp://www.ebi.ac.uk/pub/clusterw>
- MAFFT <http://www.biophys.kyoto-u.ac.jp/~katoh/programs/align/mafft/>
- Dialign-tx <http://dialign-tx.gobics.de/>
- POA <http://www.bioinformatics.ucla.edu/poa/>
- ProbCons <http://probcons.stanford.edu/>
- Muscle <http://www.drive5.com/muscle/>
- PCMA <ftp://iole.swmed.edu/pub/PCMA/>
- Kalign <http://msa.cgb.ki.se>
- AMAP <http://bio.math.berkeley.edu/amap/>
- PRODA <http://bio.math.berkeley.edu/proda/>
- PRANK <http://www.ebi.ac.uk/goldman-srv/prank/>

Phylogénie des protéobactéries alpha



16S



23S

PARTIE VI : TP

TP7 / TROUVER L'ANIMAL LE PLUS PROCHE DE FRED

Trouver l'animal le plus proche de Fred



Écureuil



Tamarin



Colobus



Mammouth laineux



Saki



Chimpanzé



AGM singe vert



Fred



Singe hurleur



Titi



Macaque rhesus



Ouisiti



Patas singe rouge



Chouette



Entelle



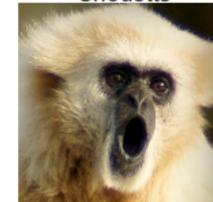
Araignée



Babouin



Orang-outan



Gibbon

Le but de ce TP est d'inférer une phylogénie complète en suivant la méthode du maximum de vraisemblance et de créer un workflow dans Galaxy.

Rappel sur les différentes étapes pour générer une phylogénie :

1. Sélectionner un lot de séquences homologues (Utiliser le fichier de séquences ADN primatesNuc.fasta)
2. Réaliser l'alignement multiple de ces séquences
3. Sélectionner les sites d'intérêt présents dans l'alignement
4. Trouver le modèle qui explique le mieux vos données
5. Inférer l'arbre par la méthode du maximum de vraisemblance

Q1 Générer le workflow précédent à l'aide des outils présents dans Galaxy jusqu'au choix du modèle.

Q2 Exécuter le workflow. Combien de blocks sont conservés dans l'alignement ? Les différents critères de sélection des modèles s'accordent-ils ?

Q3 Générer l'arbre en maximum de vraisemblance en fonction du modèle choisi précédemment en utilisant un test de robustesse aLRT-SH-like puis avec un bootstrap de 100.

Q4 Visualiser les arbres avec FigTree. Pour l'arbre avec bootstrap, afficher les valeurs de robustesse sur les branches. Qui est le plus proche parent de Fred ?

Q5 Créer un deuxième workflow pour traiter des données protéiques.

Q6 Exécuter le workflow avec le fichier primatesAA.fasta.

Q7 Choisissez un modèle et inférer l'arbre avec la méthode bayésienne.

Q8 Visualiser l'arbre avec Figtree. Les deux méthodes permettent-elles d'obtenir le même parent proche de Fred ?