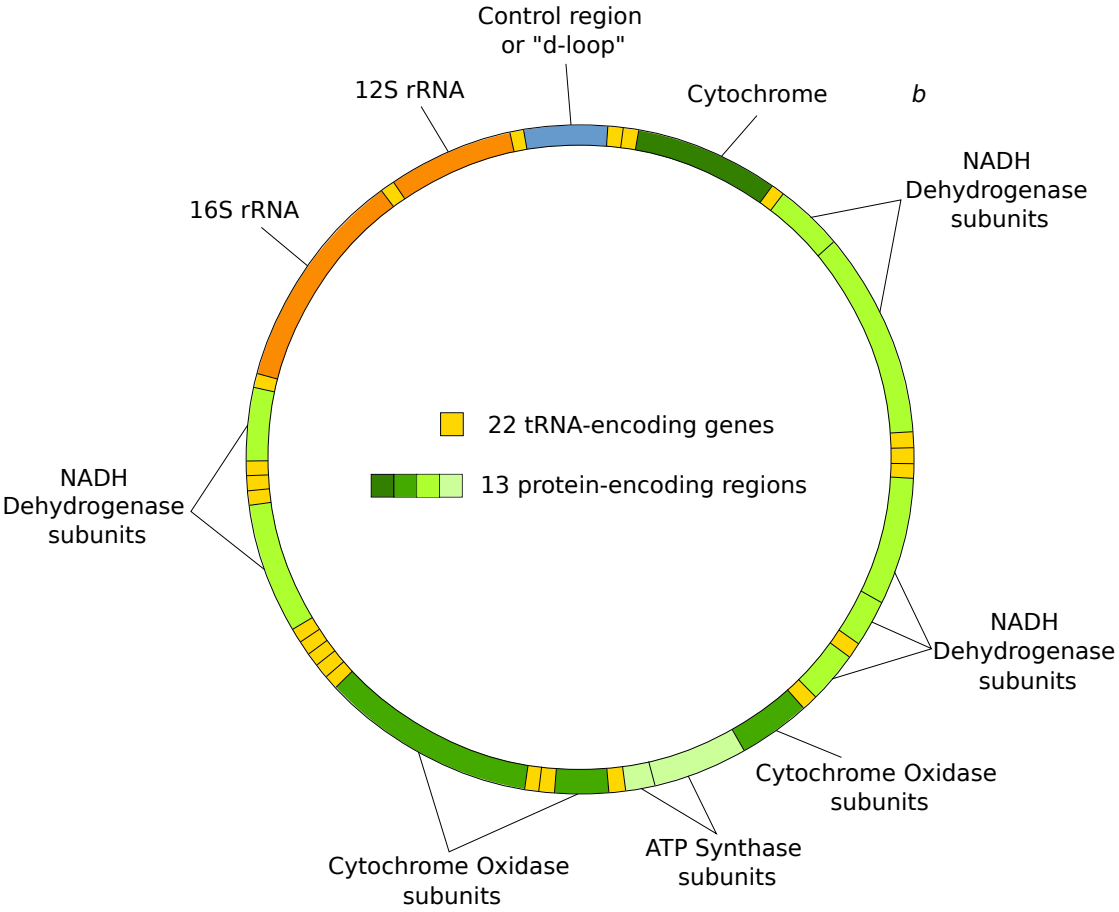
Énoncé du TP Phylogénomique du mercredi 25 octobre

Au cours de ce TP nous allons nous intéresser à la phylogénie des primates.
Pour cela nous utiliserons le génome mitochondrial. Il contient des gènes codants des protéines et des gènes d'ARN non codants (tRNA et rRNA).



http://en.wikipedia.org/wiki/Mitochondrial_DNA#mediaviewer/File:Mitochondrial_DNA_en.svg

Première partie :
Les questions sont là pour vous guider et pour mettre le focus sur les points importants de l'analyse.
Attention : sur le cluster toujours utiliser qsub ou qrsh et configurer les jobs pour qu'ils écrivent dans le /work de votre compte fleur.
anemone arome aster bleuet camelia capucine chardon clematite cobee coquelicot cosmos cyclamen dahlia digitale geranium gerbera glaieul hortensia iris jacinthe. Mdp : f1o2r3 !

Les fichiers sont disponibles ici :
http://genoweb.toulouse.inra.fr/~formation/M2_Phylogenomique/data/MitoPrimates/
ou en ligne de commandes : /home/formation/public_html/M2_Phylogenomique/data/MitoPrimates/
Les scripts pour ce TP sont :
/home/formation/public_html/M2_Phylogenomique/scripts/MitoPrimates

Nous avons pour vous récupéré les séquences à partir des fichiers Genbank spécifiés dans la table supplémentaire I de l'article de Menezes et al. Il manque donc par rapport à leur papier les 4 souches qu'ils ont séquencées. Nous les avons téléchargé du ncbi en format fasta puis pour produire les superalignements nous avons procédé comme suit :

1) Les superalignements

Le superalignement des 13 gènes de protéines a été fait en nucléotides (*prot_nt.concat.phy*) et en acides aminés (*prot_aa.concat.phy*).
Pour l'alignement nucléotidique : nous avons utilisé : Transeq, Clustalo pour chaque protéine indépendamment avec les paramètres par défaut, catfasta2phyml.pl (pour concaténer les gènes ensemble), Gblocks avec les paramètres par défaut, un script pour recoder l'alignement en codon (alAA2AN.pl) et fasta2phylip.pl
Pour l'alignement protéique : Transeq, Clustalo pour chaque protéine avec les paramètres par défaut, catfasta2phyml.pl, Gblocks avec les paramètres par défaut, et fasta2phylip.pl

Question 1 : à quoi sert chacune des étapes ?

Le superalignement des 25 gènes d'ARN (*RNA.concat.phy)* a été fait de la façon suivante :
Mafft with -qinsi --reorder, catfasta2phyml.pl, Gblocks

Question 2 : à quoi sert chacune des étapes ?

2) Les arbres

Pour trouver le meilleur modèle sur l'alignement des gènes d'ARN non codants, il vous faudra lancer modelgenerator. Pour cela préparer un fichier (cmd2.txt dans la ligne de commande ci-dessous) contenant le nom du fichier d'alignement nettoyé : RNA.concat.fa-gb, un retour à la ligne et 4 (pour le nombre de catégorie de la loi gamma) suivi de deux retours à la ligne.

Lancez modelgenerator de la façon suivante : qsub -V -b Y -l h_vmem=50G -l mem=40G "java -jar /usr/local/bioinfo/bin/modelgenerator.jar < cmd2.txt"

Utilisez le modèle sélectionné pour lancer phyML sur l'alignement phylip RNA.concat.phy avec un support de branches aLRT, en estimant le paramètre gamma et la proportion des sites invariants et enfin en utilisant la méthode SPR pour l'amélioration de la topologie de l'arbre.

Pensez à demander un peu plus de mémoire : qsub -V -b Y -N phymlRNA -l h_vmem=10G -l mem=8G

Nous avons fait pour vous les arbres des superalignements protéiques et nucléiques sur les gènes codant des protéines mitochondriales. Regarder les fichiers *prot_aa.concat.phy_phyml_stats.txt* et *prot_nt.concat.phy_phyml_stats.txt* pour savoir ce qui a été lancé. Les arbres sont dans prot_aa.concat.phy_phyml_tree.txt et prot_nt.concat.phy_phyml_tree.txt.

Question 3 : comparez les trois arbres visuellement avec figtree par exemple (ou avec des outils en ligne tels que : phylo.io, PhyloD3...).
Pour utiliser figtree sur le cluster pensez à se loguer en -XY sur le cluster. Les noms des espèces sont disponibles dans le fichiers SpeciesNames.txt. Quelles sont les principales différences ? Que pensez-vous des ces arbres ? Leurs supports ? Leurs congruences ? Les grands groupes sont-ils retrouvés ? Comparer avec les arbres des papiers (Perelman 2001 et Menezes 2013) que j'ai mis dans le dossier du TP. Sur la dernière page de ce document j'ai reporté sur la figure du papier de Menezes et al. les numéros d'accession du NCBI correspondant aux labels de vos arbres (SpeciesNames.txt).

Deuxième partie :

1) Préparation des arbres en entrée

Pour préparer le super-arbre nous allons d'abord découper le problème en élément cohérents. Je vous propose pour cela de partager les gènes comme suit : chaque gène protéique sera analysé indépendamment (ATP6, ATP8, COX1, COX2, COX3, CYTB, ND1, ND2, ND3, ND4, ND4L, ND5, ND6), les gènes d'ARN ribosomiques seront analysés séparément également (s-rRNA et l-rRNA) mais par contre je propose de concaténer ensemble les gènes de tRNA (fichiers tRNA*.idx.2.fa.align.rename).
Utilisez les fichiers fasta ayant cette forme : NAME.idx.2.fa.align.rename.
Partagez-vous le travail entre vous et mettez à disposition des autres votre arbre et le fichier .stats de phyML pour qu'ils aient les paramètres utilisés. Pour cela nommez vos fichiers résultats avec un nom explicité dans le /tmp. Vous devriez avoir 16 arbres différents.
Pour les gènes protéiques : Utilisez perl fasta2phylip.pl pour transformer l'alignement au format phylip puis utiliser modelgenerator pour trouver le meilleur modèle et finir en lançant phyML pour contruire l'arbre.
Pour les gènes d'ARNr : procédez de la même façon.
Pour les gènes d'ARNt : utilisez catfasta2phyml.pl, fasta2phylip.pl puis modelgenerator et phyML.
A la fin du process les 16 arbres devraient être dans le /tmp.
Chacun les concatène (commande cat) dans un même fichier dans son work.

Remarquez que nous ne faisons pas le nettoyage des alignements.

2) Les super-arbres

Nous allons commencer par utiliser la méthode la plus répandue : MRP.
Pour cela : tapez qrsh dans un autre terminal connecté à genotoul. Remettez vous dans le bon répertoire. Puis appelez R.
Si le résultat de votre concaténation d'arbres est le fichier : MonPath/allTree.txt.
Lancez ensuite les commandes suivantes :
library(phytools)
trees=read.tree(« MonPath/allTree.txt »)
supertrees<-mrp.supertree(trees,rearrangements= « SPR », start= « NJ »)
Vous avez obtenu les super-arbres les plus parcimonieux. Sauvez-les en utilisant la fonction write.tree de R.
Par exemple : write.tree(supertrees, file = "/home/choede/work/formation_phylo/superTree")

Sortez de R mais restez sur le nœud et faites un consensus (avec consense) de ces super-arbres.
Voici l'aide en ligne : https://faculty.cs.byu.edu/~clement/phylip/doc/consense.html.
Utilisez le type de consensus par défaut.
Renommez la sortie : mv outtree supertree.cons

Faites également le consensus (restez en qrsh) avec consense des 16 arbres obtenus dans la partie intitulée **1)Préparation des arbres en entrée**
Renommez la sortie : mv outtree consensus.tree

3) Autre méthodes : PhySIC

Lancez PhySIC via l'interface web sur la fichier contenant l'ensemble de vos arbres en entrée en baissant le seuil de support des nœuds pour avoir un compromis entre le nombre de nœud PI et PC.
Voir http://www.atgc-montpellier.fr/physic/usersguide.php.
Vous pourrez visualiser ces labels (PI et PC) avec PhyD3.

Question 4 : Que pensez-vous de cet arbre ?

4) Comparer les arbres

Nous allons utiliser une méthode de distance topologique très simple pour comparer les arbres. Pour cela concaténez dans un même fichier tous les arbres que vous souhaitez comparer. Se souvenir de l'ordre.
cat phySicTree consensus.tree supertree.cons prot_aa.concat.phy_phyml_tree.txt prot_nt.concat.phy_phyml_tree.txt RNA.concat.phy_phyml_tree.txt > trees.final

Utilisez treedist (symmetric difference) entre toutes les paires d'arbres.

Question 5 : comment cette distance topologique fonctionne-t-elle avec des arbres contenant des polytomies ?

Quels sont les arbres les plus proches topologiquement ? Les plus éloignés ? Et si on enlève le super arbre PhySIC ? Est-ce attendu ? Que pensez-vous de l'arbre PhySIC ? Proposez des explications.

Pour ce jeu de données avait-on besoin de faire un super-arbre ? Pourquoi n'as-tu pas nettoyé les alignements pour effectuer les arbres en entrée du super arbre ?

Dans le consensus, comme dans le super arbre, est-ce que NC_010299 Daubentonia madagascariensis est bien placé ? (Il doit être plus proche des Indriidae que des Lorisidae). Dans quel arbre est-il mal placé ?

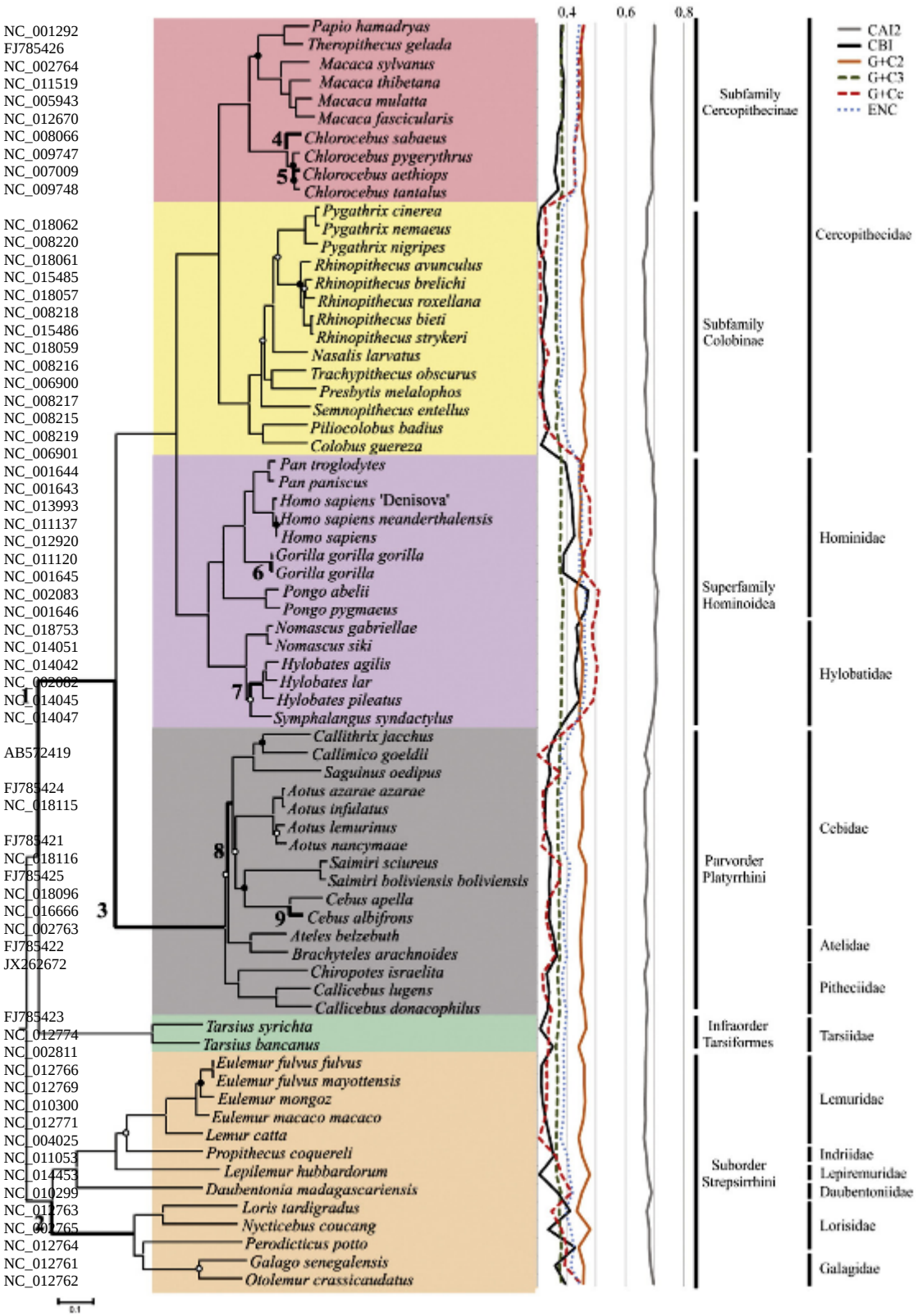
Qu'en pensez-vous ?

Consignes pour le compte-rendu de TP :

Faites chacun un document reprenant les parties principales d'un article scientifique relatif à ce TP.

Il n'a pas besoin d'être trop long, mais le matériel et les méthodes utilisés pour chacune des étapes doivent être spécifiés, les résultats présentés sous la forme de figures ou de tableau commentés, puis vous devez discuter les résultats et les méthodes utilisées par rapport aux deux papiers proposés (Perelman 2001 et Menezes 2013) au minimum.
Essayez d'être concis mais précis.

Bon courage ;-)


0.4
0.6
0.8
CAI2
CBI
G+C2
G+C3
G+Ce
ENC
NC_001292
FJ785426
NC_002764
NC_011519
NC_005943
NC_012670
NC_008066
NC_009747
NC_007009
NC_009748
NC_018062
NC_008220
NC_018061
NC_015485
NC_018057
NC_008218
NC_015486
NC_018059
NC_008216
NC_006900
NC_008217
NC_008215
NC_008219
NC_006901
NC_001644
NC_001643
NC_013993
NC_011137
NC_012920
NC_011120
NC_001645
NC_002083
NC_001646
NC_018753
NC_014051
NC_014042
NC_002082
NC_014045
NC_014047
AB572419
FJ785424
NC_018115
FJ785421
NC_018116
FJ785425
NC_018096
NC_016666
NC_002763
FJ785422
JX262672
FJ785423
NC_012774
NC_002811
NC_012766
NC_012769
NC_010300
NC_012771
NC_004025
NC_011053
NC_014453
NC_010299
NC_011763
NC_002765
NC_012764
NC_012761
NC_012762
Papio hamadryas
Theropithecus gelada
Macaca sylvanus
Macaca thibetana
Macaca mulatta
Macaca fascicularis
Chlorocebus sabaeus
Chlorocebus pygerythrus
Chlorocebus aethiops
Chlorocebus tantalus
Pygathrix cinerea
Pygathrix nemaeus
Pygathrix nigripes
Rhinopithecus avunculus
Rhinopithecus brelichi
Rhinopithecus roxellana
Rhinopithecus bieti
Rhinopithecus strykeri
Nasalis larvatus
Trachypithecus obscurus
Presbytis melalophos
Semnopithecus entellus
Piliocolobus badius
Colobus guereza
Pan troglodytes
Pan paniscus
Homo sapiens 'Denisova'
Homo sapiens neanderthalensis
Homo sapiens
Gorilla gorilla gorilla
Gorilla gorilla
Pongo abelii
Pongo pygmaeus
Nomascus gabriellae
Nomascus siki
Hylobates agilis
Hylobates lar
Hylobates pileatus
Symphalangus syndactylus
Callithrix jacchus
Callimico goeldii
Saguinus oedipus
Aotus azarae azarae
Aotus infulatus
Aotus lemurinus
Aotus nancymaae
Saimiri sciureus
Saimiri boliviensis boliviensis
Cebus apella
Cebus albifrons
Ateles belzebuth
Brachyteles arachnoides
Chiropotes israelita
Callicebus lugens
Callicebus donacophilus
Tarsius syrichta
Tarsius bancanus
Eulemur fulvus fulvus
Eulemur fulvus mayottensis
Eulemur mongoz
Eulemur macaco macaco
Lemur catta
Propithecus coquereli
Lepilemur hubbardorum
Daubentonia madagascariensis
Loris tardigradus
Nycticebus coucang
Perodicticus potto
Galago senegalensis
Otolemur crassicaudatus
1
2
3
4
5
6
7
8
9
0.1
Subfamily Cercopithecinae
Subfamily Colobinae
Superfamily Hominoidea
Parvorder Platyrrhini
Infraorder Tarsiiformes
Suborder Strepsirrhini
Cercopithecidae
Hominidae
Hylobatidae
Cebidae
Atelidae
Pitheciidae
Tarsiidae
Lemuridae
Indriidae
Lepiremuridae
Daubentoniidae
Lorisidae
Galagidae