I. Alignement de séquences

Exercice I. Recherche des homologues d'une séquence protéique d'intérêt.

1. Recherchez la séquence ayant comme identifiant P04118 dans la banque Swiss-Prot
2. *De quel organisme provient-elle? Quelle est sa taille?*
3. *Quand a-t-elle été déposée dans la banque de séquences?*
4. *Quelle est sa fonction? Où est-elle exprimée ? Quelle est sa localisation cellulaire ? Forme-t-elle un complexe protéique ?*
5. *Quelle est la localisation chromosomique du gène qui la code?*
6. *A-t-elle des homologues connus?*

Précisez la stratégie de recherche que vous allez privilégier? Justifiez vos choix.

Exercice II. Recherche de séquences divergentes.

1. Recherchez la séquence protéique de la globine alpha humaine (identifiant swissprot P69905)

2. Cette séquence possède-t-elle des homologues chez l'homme?
Précisez la stratégie de recherche que vous allez privilégier? Justifiez vos choix.

3. Il existe en réalité 13 homologues de cette protéine chez l'homme. *Pourquoi ne les avez-vous pas tous détectés?*

4. Refaites l'analyse en utilisant le logiciel PSI-BLAST.

5. Refaites l'analyse réalisée en 2., mais en ajustant les paramètres du programme utilisé afin de prendre en compte la forte divergence des globines humaines.

6. Téléchargez le fichier HSglobin_NA.fasta contenant les globines humaines.

7. Ouvrez le fichier avec SeaView. Alignez les séquences avec MUSCLE. Ouvrez le fichier avec un deuxième SeaView. Alignez les séquences avec CLUSTAL0.
Comparez les alignements obtenus.

8. Utilisez Gblocks pour éliminer les régions où l'alignement est ambigu. *Combien de positions sont gardées si on se base sur l'alignement obtenu avec MUSCLE? Et si on se base sur l'alignement obtenu avec CLUSTAL0?*

II. Une bactérie âgée de 250 millions d'années

En 2000, Vreeland et collaborateurs ont annoncés qu'ils avaient isolés une bactérie âgée de 250 millions d'années à partir d'un cristal salin.
La séquence de l'ARNr 16S de cette bactérie (notée unknown293), alignée avec d'autres séquences provenant d'organismes actuels est disponible dans le fichier : permians.nxs.

Une chose importante à noter est que les séquences intitulées BACSUCG.* proviennent toutes de Bacillus subtilis 168 et qu'elles correspondent à différentes copies paralogues de l'ARNr 16S dans cette bactérie.

1. Sauvegardez ce fichier au format texte sur votre ordinateur.

2. Chargez-le dans SeaView.

3. Refaites la phylogénie en utilisant tout d'abord la parcimonie puis comparez l'arbre reconstruit avec celui obtenu avec le Neighbour-Joining.

4. *Quelle est l'information importante apportée par les longueurs de branches dans le cas de l'analyse effectuée par Neighbour-Joining ?*

5. *Que peut-on en conclure quant aux résultats de Vreeland et al. (2000) ?*

Vous pouvez consulter l'article de Graur et Pupko (2001) démontrant pourquoi cette bactérie est probablement d'origine beaucoup plus récente.

6. Ouvrez un terminal et placez-vous (en utilisant la commande cd) dans le dossier où se trouve le fichier permians.nxs.

7. Lancez MrBayes en tapant la commande mb. Pour voir la liste des options disponibles, tapez help.

8. Ouvrez le fichier permians.nxs en tapant exe permians.nxs.

En tapant help lset, vous pourrez observer les paramètres de l'analyse par défaut. Reportez-vous à la section correspondante du manuel de MrBayes pour voir ce qu'ils signifient.

9. Lancez une analyse en tapant mcmc ngen=100000.
Quelle est la signification du paramètre ngen ? Combien de chaînes sont-elles lancées par défaut ?
Observez l'évolution des différentes chaînes, et estimez le temps attendu pour l'analyse.

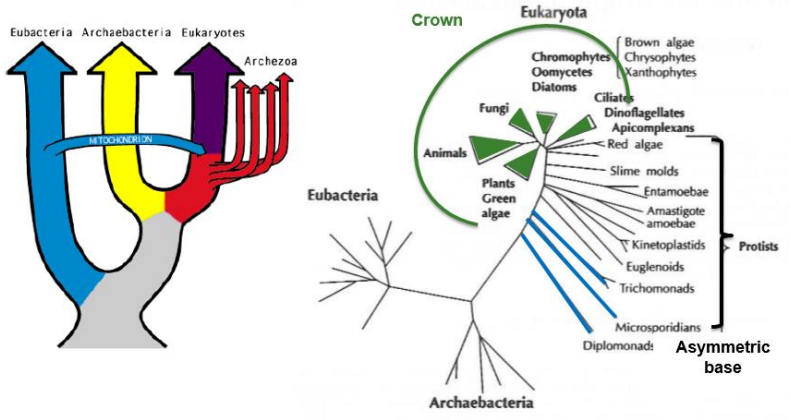
10. Lorsque l'analyse est terminée (ou quand vous l'aurez interrompue faute de temps par Ctrl-C), résumez les résultats (sumt burnin=250 et sump burnin=250). *Que signifie ce paramètre de burnin ?*

11. Observez les arbres (par exemple avec showtree ou en utilisant SeaView).
Que signifient les indices compris entre 0 et 1 pour chacune des branches internes ?

12. *D'une façon générale, les résultats produits par l'analyse bayésienne corroborent-ils ceux obtenus avec le Neighbor-Joining ?*

Exercice II. L'hypothèse « Archezoa » - de l'importance du choix des modèles d'évolution en phylogénie moléculaire.

Les *Archezoa* est un taxon proposé par Thomas Cavalier-Smith en 1989. Il regroupe diverses lignées de protistes supposés primitifs car dépourvus de mitochondries, telles que les *Microsporidia*, les *Trichomonada* et les *Diplomonada*. D'après cette hypothèse, l'endosymbiose mitochondriale aurait eu lieu après la divergence des archezoa (cf. schéma ci-dessous).



1) Pour tester cette hypothèse, téléchargez le fichier 28S_rRNA.fasta (les séquences sont déjà alignées) et visualisez l'alignement avec SeaView.

-Éliminez les régions où l'alignement est de faible qualité avec Gblocks avec les paramètres par défaut.

-Construisez l'arbre correspondant à votre jeu de données par la méthode du Maximum de Vraisemblance en utilisant le modèle d'évolution TN93 (sans distribution gamma) et utilisant les NNI pour l'exploration de l'espace des arbres. N'oubliez pas de permettre l'optimisation du taux de transitions/transversions.

Quelle est la valeur de vraisemblance associée à l'arbre reconstruit ?

En supposant que l'enracinement au point-moyen est correct, quelles hypothèses pouvez-vous faire concernant la position phylogénétique des Microsporidia ? Est-elle en accord avec l'hypothèse Archezoa proposée par T.C. Smith ?

Qu'en est-il de l'origine de la mitochondrie chez les eucaryotes.

2) Refaites l'analyse phylogénétique en utilisant cette fois-ci le modèle d'évolution suggéré par le serveur IQ-TREE avec le critère BIC et l'approche couplant les NNI et le SPR pour l'exploration de l'espace des arbres. N'oubliez pas de permettre l'optimisation du taux de transitions/transversions.

Quelle est la valeur de vraisemblance associée à l'arbre reconstruit ?

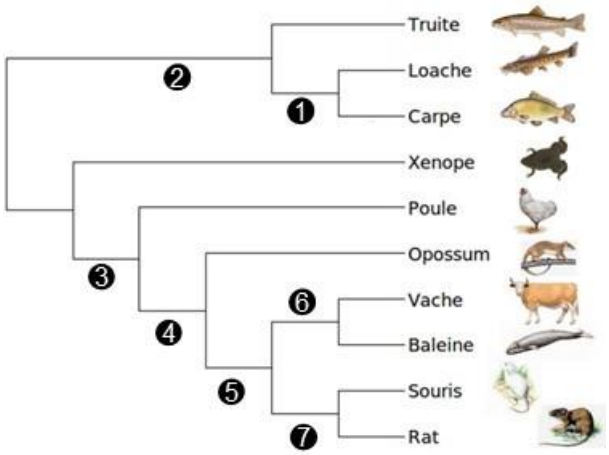
Quelles différences majeures présente l'arbre obtenu avec le précédent ? Cela vous amène-t-il à réviser votre hypothèse sur l'endosymbiose mitochondriale ?

Trois paramètres ont été changés entre la première et la seconde analyse. Testez l'influence de chacun d'eux séparément. Pour ce faire regardez l'évolution de la valeur de vraisemblance associée à chaque reconstruction. Concluez.

Exercice III. Phylogénie des mammifères – de l'importance du choix des marqueurs en phylogénie moléculaire.

L'objectif de cet exercice est d'étudier la qualité et la variabilité de l'information contenue dans les marqueurs moléculaires et leur pouvoir résolutif.

Pour ce faire, vous allez étudier et comparer l'information portée par 15 gènes codés par la mitochondrie de 10 espèces de vertébrés (la baleine, la vache, la souris, le rat, l'opossum, la poule, le xénope, la truite, la carpe et la loche) dont la phylogénie est bien établie et non remise en cause.



Les fichiers des gènes étudiés sont accessibles en cliquant sur les liens suivant : ARN 12S, ARN 16S, le cytochrome b, les sous-unités 6 et 8 de l'ATPase, les sous-unités 1, 2 et 3 de l'oxydase du cytochrome *c* et les sous-unités 1, 2, 3, 4, 4L, 5 et 6 de la déshydrogénase du NADH.

Chaque étudiant travaillera sur un ou deux de ces gènes.

1) Réalisez l'analyse phylogénétique de chaque gène. Pour ce faire :

- Téléchargez le ou les fichiers d'intérêt. Ouvrez le fichier avec SeaView.

- Alignez les séquences avec le programme d'alignement MUSCLE.

Nota bene pour les gènes codant pour des protéines, l'alignement doit se faire à partir des séquences protéiques déduites des séquences des gènes. Pour cela, cochez la case « view as proteins » dans le menu « props ».

- Éliminez les régions où l'alignement est de faible qualité avec Gblocks (paramètres par défaut).

- Pour les gènes codant pour des protéines, construisez l'arbre correspondant aux séquences protéiques par la méthode des distances BioNJ avec le modèle d'évolution Poisson + 100 réplicats de bootstrap.

- Construisez l'arbre correspondant aux séquences nucléiques par la méthode des distances BioNJ avec le modèle d'évolution HKY + 100 réplicats de bootstrap.

Nota bene pour les gènes codant pour des protéines, faire l'analyse une première fois en conservant les trois bases des codons. Refaire l'analyse en conservant uniquement les deux premières bases des codons (menu sites -> create set -> 1st + 2nd codon pos).

- En vous inspirant du document accessible sur spiral connect, pour chaque gène et pour chacune des analyses phylogénétiques réalisées (au niveau nucléique et au niveau protéique), indiquez la valeur de bootstrap associée à chaque branche interne des arbres obtenus (B1 à B8). Si une branche n'apparaît pas, indiquez NO (Non observée). Indiquez pour chaque analyse le nombre de positions conservées par Gblocks.

Pour les gènes codant pour des protéines, comparez les arbres obtenus par l'analyse des séquences protéiques et des séquences nucléiques. Les topologies sont-elles identiques ?

Sont-elles en accord avec la phylogénie de référence des espèces ?

Proposez une ou plusieurs explications ?

2) Réalisez l'analyse phylogénétique du génome mitochondrial complet. Pour ce faire :

- Téléchargez le fichier mito_complet.fst.

- Alignez les séquences avec le programme d'alignement MUSCLE.

- Éliminez les régions où l'alignement est de faible qualité avec Gblocks (paramètres par défaut).

- Construisez l'arbre correspondant aux séquences nucléiques par la méthode des distances BioNJ avec le modèle d'évolution HKY + 100 réplicats de bootstrap.

La topologie obtenue est-elle en accord avec la phylogénie de référence des espèces ?