

Formation phylogénie : Travaux Pratiques

Guy Perrière - guy.perriere@univ-lyon1.fr
Héloïse Philippon - heloise.philippon@univ-lyon1.fr

9-10-30 Novembre et 1er Décembre 2017

Table des matières

1	Alignement de séquences	2
1	Recherche des homologues d'une séquence protéique	2
2	Recherche de séquences divergentes	2
2	Une bactérie âgée de 250 millions d'années	3
1	Première partie : arbre en parcimonie	3
2	Deuxième partie : arbre en distance	3
3	Troisième partie : arbre en bayésien	3
3	Tests de comparaison de phylogénies	4
4	Grippe de type C en Inde	5
5	Endosymbiose mitochondriale	6
1	Phylogénie à partir de l'ARNr 18S	7
2	Phylogénie à partir du gène codant pour la protéine IscS	7
6	Phylogénie des mammifères – de l'importance du choix des marqueurs en phylogénie moléculaire	8

AVANT DE DÉBUTER

Tout au long du TP vous allez obtenir des résultats intermédiaires (fichiers d'homologues, alignements, alignements filtrés, etc). A chaque étape il sera important de sauvegarder ces résultats dans des fichiers aux noms explicites comme par exemple :

```
permians_muscle_Gblocks.fasta
```

pour l'alignement effectué par Muscle puis filtré par Gblocks dans la question 3 de la partie 2.

Conseil : Créez des répertoires pour chaque partie et chaque exercice !

Pour ces TP vous aurez également besoin de quatre logiciels :

1. **Seaview** [1] téléchargeable ici : <http://pbil.univ-lyon1.fr/software/seaview.html/>
2. **BEAST** [2] : le **.zip** est téléchargeable ici : <https://github.com/beast-dev/beast-mcmc/releases/tag/v1.8.4>. Lancez l'exécutable de BEAST, si Java 8 n'est pas installé sur votre ordinateur, suivez les instructions proposées.
3. **Tracer** [3] téléchargeable ici : <http://tree.bio.ed.ac.uk/software/tracer/>
4. **FigTree** téléchargeable ici : <http://tree.bio.ed.ac.uk/software/figtree/>

1 Alignement de séquences

1/ Recherche des homologues d'une séquence protéique

1. Recherchez la séquence ayant comme identifiant P04118 dans la banque UniProt.
2. De quel organisme provient-elle ? Quelle est sa taille ?
3. Quand a-t-elle été déposée dans la banque ?
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.

2/ Recherche de séquences divergentes

1. Recherchez la séquence protéique de la globine alpha humaine (identifiant UniProt 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. Aligned les séquences avec Muscle. Ouvrez le fichier avec un deuxième SeaView. Aligned les séquences avec ClustalO.
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 ClustalO ?

2 Une bactérie âgée de 250 millions d'années

En 2000, [Vreeland et al.](#) ont annoncé qu'ils avaient isolés une bactérie âgée de 250 millions d'années à partir d'un cristal salin [5]. La séquence de l'ARNr 16S de cette bactérie (nommée 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 BAC-SUCG.* proviennent toutes de *Bacillus subtilis 168* et qu'elles correspondent à différentes copies paralogues de l'ARNr 16S dans cette bactérie.

1/ Première partie : arbre en parcimonie

1. Sauvegardez ce fichier au format texte sur votre ordinateur.
2. Chargez-le dans SeaView.
3. Aligned les séquences avec Muscle puis sauvez l'alignement dans un autre fichier (par exemple `permians_aln.nxs`) en conservant bien le format Nexus.
4. Refaites la phylogénie en utilisant la parcimonie avec les paramètres par défaut proposés par le programme.
Concernant les paramètres, en quoi l'option `Randomize seq. order` a-t-elle un intérêt pour la parcimonie ? Quelles sont les informations importantes figurant dans la ligne de commentaire située en haut de la fenêtre contenant l'arbre ?
5. Racinez l'arbre au moyen de la séquence de *L. casei*.
6. Sauvegardez l'arbre construit, tout d'abord dans le menu **Trees**, puis dans un fichier pour une utilisation ultérieure.

2/ Deuxième partie : arbre en distance

1. Refaites la phylogénie sur les séquences alignées obtenues en 1.3 en utilisant le *Neighbor-Joining*. Utilisez les paramètres par défaut proposés par le programme.
2. Comparez l'arbre obtenu avec celui de parcimonie.
Quelle est l'information importante apportée par les longueurs de branches dans le cas de l'analyse effectuée par Neighbor-Joining ? Que peut-on en conclure quant aux résultats de Vreeland et al. (2000) ?
Vous pouvez consulter l'article de [Graur et Pupko \(2001\)](#) [6] démontrant pourquoi cette bactérie est probablement d'origine beaucoup plus récente.

3/ Troisième partie : arbre en bayésien

1. Avant de reconstruire l'arbre en bayésien, nous allons sélectionner le modèle évolutif le plus adapté au jeu de donnée en utilisant le serveur d'**IQTREE**. Pour cela, chargez le le fichier aligné et filtré obtenu en question 3) et gardez les paramètres par défaut.
Quel modèle est privilégié selon le critère BIC ?

2. Ouvrez un terminal et placez-vous dans le dossier où se trouve le fichier contenant les séquences que vous avez alignées en 1.3.
3. Lancez MrBayes en tapant la commande `mb`. Pour voir la liste des options disponibles, tapez `help`.
4. Ouvrez le fichier de séquences alignées en tapant `exe nom_du_fichier`. 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.
5. Lancez une analyse en tapant `mcmc ngen=100000`. Observez l'évolution des différentes chaînes, et estimez le temps attendu pour l'analyse.
Quelle est la signification du paramètre `ngen` ? Combien de chaînes sont-elles lancées par défaut ?
6. 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` ?
7. Observez les arbres (par exemple avec la commande `showtree` ou en utilisant SeaView).
Que signifient les indices compris entre 0 et 1 pour chacune des branches internes ? Les résultats produits par l'analyse bayésienne corroborent-ils ceux obtenus avec le Neighbor-Joining ?

3 Tests de comparaison de phylogénies

Rokas *et al.* (2003) ont publié une analyse phylogénomique d'un groupe de levures proches de *Saccharomyces cerevisiae*. Leur analyse portait sur un jeu de 106 gènes orthologues présent en une copie chez chacune des huit espèces considérées. Les alignements correspondant à cinq de ces 106 gènes sont disponibles à partir des liens ci-dessous :

[YGL001Cnuc.fst](#)
[YOL097Cnuc.fst](#)
[YBL091Cnuc.fst](#)
[YER005Wnuc.fst](#)
[YNL155Wnuc.fst](#)

1. Reconstituez les arbres de maximum de vraisemblance pour ces cinq gènes en utilisant le modèle $GTR+I+\Gamma_4$.
Que pensez-vous de ces arbres ?
2. Nous allons concaténer les cinq alignements. Pour ce faire, ouvrez le premier d'entre eux avec SeaView, puis en utilisant le menu **Open**, ouvrez l'alignement suivant. Retournez ensuite à la première fenêtre et utilisez l'onglet **Concatenate**, choisissez l'alignement que vous venez d'ouvrir et cliquez sur **OK**. Une fois tous les alignements concaténés ensemble, utilisez le menu **Save as** pour enregistrer ce concaténat. Vérifiez que sa longueur est bien 3993 puis construisez l'arbre de ce concaténat.
Comparez l'arbre du concaténat aux autres arbres obtenus, observez notamment le soutien de ses branches.
3. Nous allons maintenant tester la significativité des différences observées entre les arbres, ceci au moyen de tests de comparaison de topologies. Comme il y a au total six tests à effectuer (un pour chaque alignement individuel plus un pour le concaténat), nous allons répartir la tâche entre les différentes personnes présentes.
 - (a) Tout d'abord il faut que chaque groupe mette dans un fichier l'ensemble des six arbres qui ont été calculés, avec un arbre par ligne. *Cette opération doit impérativement être effectuée avec Notepad. N'utilisez surtout pas Word qui ne sauvegardera pas en format*

texte par défaut! Au cas où vous n'y arriveriez pas, le résultat de cette opération est disponible [ici](#).

- (b) Allez sur le site web d'[IQ-TREE](#). Pour effectuer le test, chaque groupe doit envoyer sur le site le fichier d'alignement qui lui a été assigné, ceci au moyen du bouton **Browse** de l'option **Alignment file** (menu **Input Data**). Concernant le modèle utilisé pour le calcul de l'arbre, nous allons bien sûr utiliser le même que celui employé précédemment, à savoir GTR plus la correction par la loi Gamma ainsi que la prise en compte des invariants (menu **Substitution Model Options**).
- (c) Faites ensuite dérouler le menu **Tree Topology Evaluation and Test**. A ce niveau, chaque groupe doit sélectionner au moyen du bouton **Browse** associé le fichier contenant les six arbres créé à l'étape 3.a. Laissez les options par défaut qui sont sélectionnées.
- (d) Soumettez votre analyse au moyen du bouton **SUBMIT JOB**.

*Sur la page de résultats, notez quels sont les arbres qui ne rejettent pas l'hypothèse nulle au seuil $\alpha = 0.05$ avec le test *ELW*.*

Que pensez-vous de ces résultats? Quelle est selon vous la source des incongruences observées par Rokas et al. ?

4 Grippe de type C en Inde

Nous allons étudier l'origine évolutive de trois échantillons de virus de la grippe de type C isolés en Inde en 2011, 2012 et 2013 nommés respectivement : C/India/P119564/2011, C/India/P121719/2012 et C/India/P135047/2013. Les données sont issues de l'article de [Potdar et al.](#) (2017) [4]. Tout d'abord, téléchargez le fichier [influenza.india.fasta](#) contenant 32 séquences du gène de l'hémagglutinin-esterase (HE) du virus de grippe C isolés en différents endroits de la planète entre 1947 et 2013.

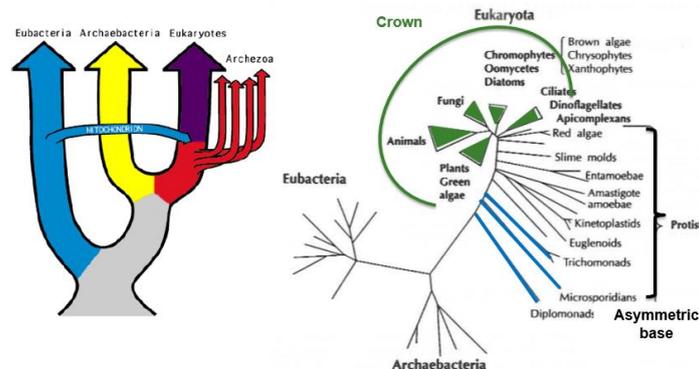
1. Chargez le fichier sous Seaview et regardez les séquences en protéiques grâce à **Props** → **View as proteins**. Aligned-les en utilisant l'algorithme MUSCLE (**Align** → **Alignment options** → **Muscle** puis **Align** → **Align all**).
Que pensez-vous de la conservation du gène HE dans ces différentes souches? Pensez-vous que reconstruire une phylogénie à l'aide des séquences protéiques soit une bonne idée?
2. Repassez en nucléotides et faites une sélection de sites conservés à l'aide de l'algorithme GBLOCKS en mode permissif (**Sites** → **Create set** → **Gblocks**. Sélectionnez les trois premières options). Enregistrer la sélection dans un nouveau fichier au format Nexus : **File** → **Save selection**.
3. Sélectionnez le modèle évolutif le plus adapté au jeu de donnée en utilisant le serveur d'[IQTREE](#) et le critère d'AIC corrigé (AICc).
Quel est le modèle sélectionné?
4. Nous allons maintenant utiliser le logiciel BEAST. Tout d'abord, lancez l'exécutable **Beauti** (disponible dans la suite de BEAST) et ajoutez le fichier aligné et filtré obtenu en 2. (**File** → **Import data** → **Filter: all Files** ou en cliquant sur + en bas de la fenêtre).
5. Nous allons utiliser les dates fournies dans les noms des séquences. Pour cela, dans l'onglet **Tips**, choisissez d'utiliser les dates et cliquez sur **Guess date**. L'année étant la dernière donnée dans le nom, choisissez **Order: last**.
6. Regardez dans les différents onglets les paramètres par défaut (modèles et priors utilisés, etc). Changez le modèle utilisé si-besoin en fonction résultat obtenu en 3.
Retrouvez-vous dans BEAST tous les modèles testés par IQTREE?

7. Dans l'onglet **Clocks** choisissez l'horloge relaxée non-corrélée et une distribution LogNormal.
8. Enfin, fixez le nombre de pas de la chaîne à 100000, le nombre d'états intermédiaires à 100 et demandez un **log** tous les 100 pas. Générez le fichier BEAST correspondant (format **xml**).
9. Générez un deuxième fichier **xml** avec les mêmes paramètres qu'en 6. et 7. mais avec 2000000 de pas, un nombre d'états intermédiaires fixé à 2000 et un **log** tous les 2000 pas. **Pensez bien à modifier le nom de sortie du fichier pour ne pas écraser celui créé en 8.**
10. Lancez BEAST et chargez le fichier **.xml** créé en 8. Désélectionnez l'utilisation de la librairie BEAGLE et lancez l'analyse. En parallèle, faites de même avec le fichier **.xml** généré en 9. (*i.e.*, avec 2000000 pas). Cette analyse va prendre du temps, en attendant continuez le TP jusqu'à la question 12. avec les résultats obtenus avec 100000 pas.
Le fichier .log contient les valeurs des différents paramètres à la fréquence demandée en 8. Combien de lignes de valeurs de paramètres va-t-il contenir ? Combien d'arbres vont être générés ?
11. Lancez le logiciel Tracer et chargez le fichier **.log** obtenu en 10. (**File** → **Import Trace file** ou en cliquant sur le +).
Qu'appelle-t-on la phase de burn-in ? A combien est-elle fixée par défaut dans Tracer ?
12. Fixez le paramètre de *burn-in* à 25000 et regardez la trace.
Que pensez-vous des valeurs d'ESS (Effective Sample Size) ? Pensez-vous avoir atteint la convergence ? Comparez avec la trace obtenue avec 2000000 pas.
13. Le fichier [influenza_india_10_millions.log](#) contient la trace de l'arbre obtenu avec le même jeu de données et les mêmes paramètres excepté que 10000000 de pas ont été faits.
Pensez-vous que la convergence a été atteinte ? A part augmenter le nombre de pas, que peut-on faire pour espérer atteindre la convergence ?
14. Lancez maintenant le programme TreeAnnotator (distribué avec BEAST). Fixez le paramètre de *burn-in* à 25000 et la probabilité postérieure minimale à 0.5. Enregistrez l'arbre consensus dans votre répertoire de travail.
15. Ouvrez l'arbre obtenu avec le logiciel FigTree. Pour son racinement, aidez-vous de l'arbre phylogénétique [influenza_C_D.pdf](#).
Selon vous, pourquoi est-il raciné par des séquences de virus de la grippe de type D ? Comment ont-elles été détectées ? Quelle(s) séquences de virus de type C allez-vous donc utiliser en groupe externe pour votre arbre ? Pourquoi ?
16. Nous allons modifier les paramètres de FigTree pour obtenir un arbre semblable à la figure 2.A de Potdar *et al.* Dans **Node labels** choisissez d'afficher les probabilités postérieures. Dans **Scale Axis** choisissez un axe inverse, fixez les paramètres **Label spacing** et **Tick spacing** respectivement à 10 et à 1. Enfin, dans **Time Scale**, fixez le paramètre **Scale Factor** à -1.
*Analysez l'arbre obtenu. Les patients dont les échantillons ont été prélevés par Potdar *et al.* étaient-ils infectés de la même souche de grippe C ? Selon cet arbre, combien de souches différentes de virus de la grippe de type C circulaient au Japon en 2004 ?*
17. Dans leur article, Potdar *et al.* reconstruisent également les phylogénies de six autres gènes du virus de la grippe de type C.
Pourquoi ? A quoi doit-on faire particulièrement attention avec les phylogénies de virus ?

5 Endosymbiose mitochondriale

L'acquisition des mitochondries a été un événement clé dans l'histoire évolutive des eucaryotes. Tous les eucaryotes actuels connus possèdent une mitochondrie, à l'exception de quelques lignées, comme les Microsporidies, les Diplomonadines, les Trichomonadines, ou *Entamoeba*. Dans

les années 1980, ces lignées dépourvues de mitochondries furent regroupées sous l'appellation *Archezoa* (voir schémas ci-dessous). Pour étudier cette acquisition, vous allez dans un premier temps reconstruire une phylogénie des eucaryotes basée sur l'ARNr 18S.



1/ Phylogénie à partir de l'ARNr 18S

1. Téléchargez le fichier [euca18S.fasta](#) contenant des séquences d'ARNr 18S eucaryotes représentatives de la diversité de ce domaine. Ouvrez le fichier avec SeaView puis alignez les séquences avec Muscle. Éliminez les régions mal alignées avec Gblocks.
2. Réalisez l'analyse phylogénétique de ces séquences par la méthode du Maximum de vraisemblance (paramètres par défaut).

Analysez la phylogénie obtenue, semble-t-elle cohérente ? En particulier, retrouvez-vous la monophylie de chacun des grands groupes eucaryotes ? Quelle est la position phylogénétique des Archezoa ? Forment-ils un groupe monophylétique ? Quelle hypothèse évolutive pouvez-vous formuler concernant l'absence de mitochondries chez Entamoeba ? Et pour les autres Archezoa ? Quelle hypothèse pouvez-vous faire concernant le moment où a eu lieu l'endosymbiose mitochondriale chez les Eucaryotes ?

2/ Phylogénie à partir du gène codant pour la protéine IscS

La formation des clusters [Fe/S] est une fonction primordiale pour toutes les cellules, qu'elles soient bactériennes, archéennes ou eucaryotes. En effet, de nombreuses activités cellulaires (*e.g.*, la photosynthèse, la réparation et la réplication de l'ADN, le contrôle de l'expression des gènes, etc.) dépendent de protéines porteuses de clusters [Fe/S]. Des dysfonctionnements au niveau des systèmes permettant de former ou de réparer les clusters [Fe/S] sont associés à de nombreuses maladies.

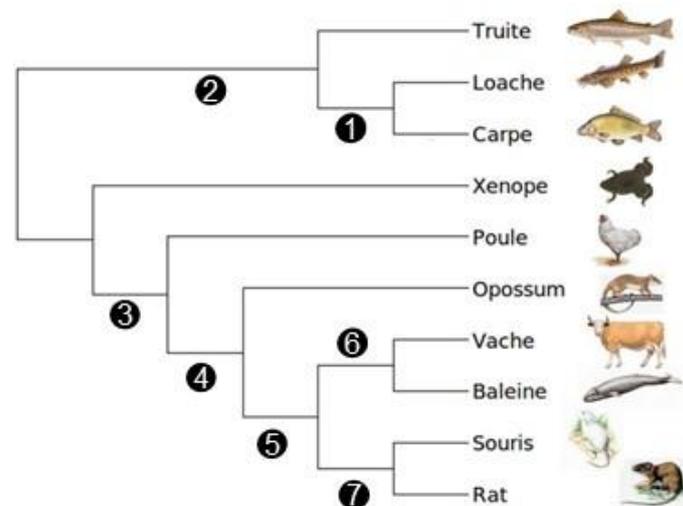
Vous allez maintenant vous intéresser à l'origine évolutive de processus chez les eucaryotes au travers de l'étude de la protéine IscS, une cystéine désulfurase qui utilise la L-Cystéine pour former de la L-Alanine et du Soufre élémentaire. Ce dernier sera ensuite utilisé pour la formation ou la régénération de clusters [Fe/S].

1. Téléchargez le fichier [IscS.fasta](#) contenant un échantillon de séquences eucaryotes et pro-caryotes. Ouvrez le fichier avec SeaView puis alignez les séquences avec Muscle. Éliminez les régions mal alignées avec Gblocks.
2. Réalisez l'analyse phylogénétique de ces séquences avec PhyML (paramètres par défaut). *Analysez la phylogénie obtenue. Que pouvez-vous dire de l'origine du gène codant pour la protéine IscS chez les Eucaryotes ? Observez attentivement la distribution taxonomique du gène IscS chez les Eucaryotes. Quelle information importante vous apporte-t-elle concernant l'origine des Archezoa et de l'endosymbiose mitochondriale ?*

6 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 groupe travaillera sur un ou deux de ces gènes.



1. Réalisez l'analyse phylogénétique de chaque gène. Pour ce faire :
 - (a) Téléchargez le ou les fichiers d'intérêt. Ouvrez le fichier avec SeaView.
 - (b) Aligned 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**.
 - (c) Éliminez les régions où l'alignement est de faible qualité avec Gblocks (paramètres par défaut).
 - (d) 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 répliquats de *bootstrap*.
 - (e) Construisez l'arbre correspondant aux séquences nucléiques par la méthode des distances BioNJ avec le modèle d'évolution HKY + 100 répliquats 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**).
 - (f) 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 :
 - (a) Téléchargez le fichier [mito_complet.fst](#).
 - (b) Alignez les séquences avec le programme d'alignement Muscle.
 - (c) Éliminez les régions où l'alignement est de faible qualité avec Gblocks (paramètres par défaut).
 - (d) Construisez l'arbre correspondant aux séquences nucléiques par la méthode des distances BioNJ avec le modèle HKY + 100 réplicats de bootstrap.
La topologie obtenue est-elle en accord avec la phylogénie de référence des espèces?

Références

- [1] **Gouy M, Guindon S and Gascuel O.** 2010. SeaView version 4 : A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol.* 27 :221-224
- [2] **Drummond AJ, Suchard MA, Xie D and Rambaut A .** 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol.* 29 : 1969-1973
- [3] **Rambaut A, Suchard MA, Xie D and Drummond AJ.** 2014. Tracer v1.6, Available from <http://tree.bio.ed.ac.uk/software/tracer/>
- [4] **Potdar V. A., Hinge D. D., Dakhave M. R., Manchanda A., Jadhav N., Kulkarni P. B., and Chadha M. S. .** 2017. Molecular detection and characterization of Influenza 'C' viruses from western India. *Infect. Genet. Evol.* 54 :466-477
- [5] **Vreeland, RH, Rosenzweig, WD, Powers, DW.** 2000. Isolation of a 250 million-year-old halotolerant bacterium from a primary salt crystal. *Nature*, 407, 6806 :897-900.
- [6] **Graur D, Pupko T..** 2001. The Permian bacterium that isn't. *Mol Biol Evol.* 18(6) :1143-6.