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 III. L'hypothèse « Archezoa ».

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*.

1) Pour tester cette hypothèse, téléchargez le fichier 28S_rRNA.fasta (les séquences sont déjà alignées).
-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.

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 ?

Que pouvez-vous en déduire concernant le moment où a eu lieu l'endosymbiose mitochondriale 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 présentent 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.

3) 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 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].

-Téléchargez le jeu de données IscS.fasta contenant un échantillon de séquences eucaryotes et procaryotes.
-Alignez les séquences avec Clustal0.
-Éliminez les régions mal alignées avec Gblocks (paramètres par défaut).
-Construisez des arbres phylogénétiques par la méthode du Maximum de vraisemblance (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 ?

IV. Origine évolutive de la thésaurine du Xénope

Les oocytes prévitellogéniques contiennent deux types de complexes ribonucléoprotéiques, appelés thésaurisomes, dont la fonction est le stockage des ARN 5S et ARNt-chargés.
Le thésaurisome 42S est constitué de 4 sous-unités, chacune d'elle étant composée de : 3 ARNt, 1 ARN 5S, 2 thésaurines a liant les ARNt, et 1 thésaurine b liant l'ARN 5S.
Il serait aussi impliqué dans la synthèse des protéines en fournissant des ARNt aux ribosomes.
La question de l'origine évolutive des thésaurisomes est importante car elle renvoie à celle de la formation des réserves des oocytes chez le Xénope.

1) Des recherches basées sur la similarité de séquences dans les bases de données ont montré que la thésaurine a était homologue au facteur d'élongation EF-1a (appelé EF-Tu chez les bactéries).
-Téléchargez le fichier thesauORI.fasta qui contient un échantillon représentatif de séquences d'EF-1a et EF-Tu.
-Alignez les séquences avec Muscle.
-Eliminez les régions mal alignées avec Gblocks en utilisant le critère le plus stringent.
-Construisez un arbre phylogénétique par la méthode de distances BioNJ (modèle d'évolution Poisson, 100 réplicats de bootstrap).

Quelles informations vous apportent ces analyses quant à l'origine évolutive de la thésaurine a chez le Xénope ? Est-ce qu'une origine mitochondriale semble plausible ?
Quelle hypothèse forte implique la position phylogénétique de la thésaurine a sur l'histoire évolutive du facteur d'élongation EF-1a chez les eucaryotes ?

2) Vous allez réaliser l'analyse phylogénétique de votre jeu de données en utilisant une méthode plus sophistiquée, le maximum de vraisemblance avec PhyML implémenté dans SeaView avec les paramètres suivants (tous les autres paramètres sont laissés par défaut) :
- modèle LG+G4, NNI+SPR
- modèle LG+G4, NNI
- modèle LG, NNI+SPR
- modèle LG, NNI

Notes : La distribution Gamma est choisie lorsque que la case « optimized » du cadre « Across Site Rate Variation » est cochée. Pour ne pas intégrer de distribution Gamma, il faut cocher la case « None ».

Relevez les valeurs de vraisemblance (Ln L) associées à chaque arbre reconstruit.
Quels sont les facteurs qui influencent le plus la vraisemblance des arbres reconstruits ?

3) Vous allez maintenant tester l'influence du choix du modèle.
Reconstruisez la phylogénie de vos séquences en utilisant PhyML avec les paramètres suivants (tous les autres paramètres sont laissés par défaut) :
- modèle WAG+G4, NNI+SPR
- modèle JTT+G4, NNI+SPR
- modèle Dayhoff+G4, NNI+SPR

- modèle WAG, NNI+SPR
- modèle JTT, NNI+SPR
- modèle Dayhoff, NNI+SPR

Comment évoluent les valeurs de vraisemblance des arbres reconstruits en fonction du modèle ?
Quelles différences notables présentent les topologies reconstruites ? Ce résultat vous amène-t-il à réviser votre scénario ?

Exercice V. Identification d'une bactérie inconnue.

Une souche bactérienne PhosAc3 a été isolée récemment à partir d'un mélange de sédiments marins contaminés par du phosphogypse en Tunisie (phosphogypse = gypse non naturel issu du traitement industriel des minerais calciques fluorophosphatés). L'ARNr 16S de cette souche a été séquencé (identifiant : FN611033).

1) Connectez-vous sur le site de la RDP (Ribosomal data base project II, http://rdp.cme.msu.edu/index.jsp).

À l'aide de l'outil « Classifier » identifiez la position taxonomique probable de la souche PhosAC3.

2) Les seules souches cultivées représentant le taxon auquel appartiendrait votre souche sont thermophiles ou hyperthermophiles (i.e. vivant à des températures ~ 80°C).

En quoi l'affiliation taxonomique proposée par la RDP est donc surprenante ?

3) Vous allez vérifier l'affiliation taxonomique proposée par la RDP à l'aide d'une analyse phylogénétique.

-Téléchargez le jeu de données meso16S.fasta. Il contient les séquences d'ARNr 16S de toutes les souches cultivées appartenant au même taxon que PhosAc3. Le groupe extérieur est composé de séquences d'ARNr 16S provenant d'autres grands groupes bactériens.
-Ouvrez le jeu de données avec SeaView.
-Alignez les séquences avec le logiciel Muscle implémenté dans SeaView.
-Éliminez les régions où la qualité de l'alignement est faible avec Gblocks (paramètres par défaut).
-Construisez l'arbre correspondant à votre jeu de données par la méthode de distances BioNJ. Seaview vous offre la possibilité d'utiliser trois modèles d'évolution différents pour l'analyse de séquences nucléiques (JC, K2P et HKY) avec cette méthode. Déterminez quel est le modèle d'évolution le plus adapté à vos données grâce au serveur IQ-TREE (http://iqtree.cibiv.univie.ac.at/) en utilisant le critère BIC.
-Utilisez le maximum de parcimonie pour confirmer le résultat précédent (Randomize seq. order 5 times, 10 réplicats de bootstrap).

Les arbres obtenus à partir de chacune méthodes sont-ils congruents ? Confirmez-vous l'affiliation proposée par la RDP ?

4) Des ARNr 16S de ce phylum bactérien ont été séquencés à partir d'environnements variés. Téléchargez l'arbre phylogénétique construit en incluant les séquences environnementales (fichier meso16S.pdf, extrait de *Ben Hania et al. 2011 Systematic Applied Microbiology*).

Quelles hypothèses pouvez-vous faire sur notre connaissance de la biodiversité de ce groupe ? Sur l'adaptation à la température au sein de ce groupe ?