Travaux dirigés de Phylogénie moléculaire
M1 Master EcoSciences/Microbiologie – UE Biologie Evolutive

Liste des logiciels utilisés
- SeaView : http://pbil.univ-lyon1.fr/software/seaview.html
- Mafft : http://mafft.cbrc.jp/alignment/server/
- Clustal0, Muscle (implémenté dans SeaView)
- Gblocks (implémenté dans SeaView)

Exercice I. 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 ?

Exercice II. 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 ?

Exercice III. Origine évolutive de la thésaurine a chez le 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, distribution Gamma, NNI+SPR
- modèle LG, distribution Gamma, NNI
- modèle LG, pas de distribution Gamma, NNI+SPR
- modèle LG, pas de distribution Gamma, 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 ». NNI et SPR sont des heuristiques permettant d'explorer l'espace des arbres.

Relevez les valeurs de vraisemblance (Ln L) associées à chaque arbre reconstruit. La valeur de vraisemblance est indiquée en haut à gauche dans la fenêtre contenant l'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, distribution Gamma, NNI+SPR
- modèle JTT distribution Gamma, NNI+SPR
- modèle Dayhoff, distribution Gamma, NNI+SPR

Comment évoluent les valeurs de vraisemblance des arbres reconstruits en fonction du modèle ?
Quelles différences notables présente la topologie associée à la meilleure vraisemblance ? Ce résultat vous amène-t-il à réviser votre scénario ?

3) Dans les cellules somatiques, l'insertion des ARNt chargés au niveau du site A des ribosomes requière l'action d'un facteur d'élongation spécifique, l'EF-1a.
Quelle hypothèse fonctionnelle pourriez-vous avancer concernant l'apparition des thésaurisomes ?

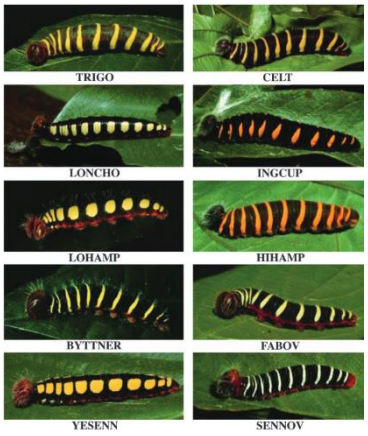
Exercice IV. Phylogénie et classification de l'espèce *Astrapes fulgerator*

La phylogénie moléculaire est de plus en plus utilisée pour l'identification et la classification des êtres vivants. Si l'ARNr 16S s'est imposé comme marqueur de référence chez les procaryotes, ce sont plutôt les gènes mitochondriaux, et en particulier le gène codant pour la sous-unité I de la cytochrome c oxydase (CoxI), qui sont utilisés chez les eucaryotes. Cette approche moléculaire de l'identification et de la classification a été popularisée sous l'appellation « Barcoding ».

Dans le cadre de cet exercice, vous allez utiliser le gène CoxI pour positionner une nouvelle variété de de papillons appartenant à l'espèce *Astrapes fulgerator*.

Cette espèce tropicale commune de Lépidoptères a été décrite pour la première fois en 1775. Depuis plusieurs années un certain nombre d'entomologistes se sont interrogés sur la possibilité que l'espère *A. fulgerator* soit en fait un complexe d'espèces. En effet, bien que les représentants de cette espèce ne présentent aucune différence au niveau des appareils reproducteurs mâles et femelles (caractère fréquemment variables au sein des espèces tropicales), les chenilles présentent des différences marquées au niveau des taches colorées qui les ornent, ainsi qu'au niveau de leurs habitats et leurs régimes alimentaires.
Des chenilles ornées de disques jaunes, se nourrissant principalement de Fabaceae et habitant dans des forêts humides ou pluvieuses ont été découvertes récemment (elles apparaissent sous le nom YESENN dans le tableau et sur les photos ci-après).

Nom	Ornements			Régime alimentaire							Forêts		
	Rayures	Disques	Tâches	Trigoniaceae	Celtidaceae	Fabaceae	Malvaceae	Styracaceae	Sterculiaceae	Sapindaceae	Pluvieuses	Humides	Sèches
TRIGO	x			x							x		x
CELT	x				x						x		
LONCHO		x				x						x	
LOHAMP		x					x				x	x	
HIHAMP	x						x					x	
BYTTNER	x								x				x
SENNOV	x					x						x	x
FABOV	x					x							x
YESENN		x				x					x	x	
INGCUP			x			x		x		x	x	x	x



Sur la base des caractères morphologiques et écologiques de quelle(s) autre(s) population(s) de chenilles les YESENN pourraient être rapprochées ?

Pour confirmer ou infirmer cette/ces hypothèse(s) vous allez réaliser l'analyse phylogénétique du gène Cox I.

-Téléchargez le fichier coxI.fasta contenant les 441 séquences obtenues à partir de 465 individus prélevés au nord-ouest du Costa Rica par Hebert et al en 2004.
-Alignez les séquences avec Clustal0.
-Eliminez les régions mal alignées avec Gblocks (paramètres par défaut).
-Construisez des arbres phylogénétiques par la méthode du BioNJ (modèle d'évolution divergence observée, 100 réplicats de bootstrap).

Analysez votre arbre ? Analysez la distribution taxonomique des ornementations, des régimes alimentaires et des habitats pour ces organismes ? Quel semble être le critère (ou les critères) le plus discriminant entre les différents groupes ?

Exercice V. Origine de la DNA polymérase mitochondriale γ eucaryote.

Les ADN polymérases γ permettent la réplication de l'ADN mitochondrial chez les Eucaryotes. Elles forment une sous-famille au sein des ADN polymérases de type A.

Quelle hypothèse pourriez-vous émettre a priori sur l'origine évolutive des ADN polymerases γ mitochondriales ?

1) Pour tester cette hypothèse, téléchargez le fichier DNAgam.fasta.
-Alignez les séquences avec Clustal0.
-Éliminez les régions où l'alignement est de faible qualité avec Gblocks d'abord avec les paramètres par défaut, puis en utilisant les critères les moins stringents.

Combien de positions sont conservés par Gblocks avec les paramètres par défaut. A quoi cela est-ce du ? Ce nombre de positions vous semble-t-il suffisant pour réaliser une phylogénie ?
Combien de positions sont conservées par Gblocks avec les paramètres les moins stringents ?

Observez bien les régions conservées, pouvez-vous identifier des signatures (résidus conservés) qui permettent de discriminer des groupes de séquences ? Si oui quels types de regroupements proposeriez-vous ?

2) Construisez des arbres phylogénétiques par la méthode de distances BioNJ (modèle d'évolution Poisson, 100 réplicats de bootstrap) et par la méthode du maximum de parcimonie (Randomize seq. order 5 times, 10 réplicats de bootstrap).

Sur la base des arbres obtenus, quelles conclusions pouvez-vous émettre concernant l'origine des séquences d'ADN polymérases mitochondriales γ ?

3) Rendez-vous sur le site web du « Tree of life project ». Explorez l'arbre afin de trouver la page dédiée aux eucaryotes dans leur ensemble.

Quels sont les grands groupes d'eucaryotes actuels ?
Sachant que le gène codant pour l'ADN polymérase γ n'est présent que dans les génomes de Fungi (Champignons), Métazoaires (Animaux), Choanoflagellates et Amoebozoa, dans quels grands groupes d'eucaryotes est-il absent ?
Sur la base de la phylogénie des organismes présentée sur cette page web, à quel moment de l'histoire évolutive des eucaryotes feriez-vous apparaître les ADN polymérases γ ?
Qu'est-ce que cela implique au niveau de la réplication de l'ADN mitochondrial pour les autres groupes eucaryotes ?