Applications de la phylogénie

UE BNF104

Partie 1 : Les arbres

Il existe un format de fichier particulier qui permet de décrire les relations phylogénétiques entre des espèces sans avoir à dessiner l'arbre correspondant : le format NEWICK. Certains programmes permettant de réaliser des phylogénies donnent le résultat sous cette forme. Voici une description de ce format :

http://fr.wikipedia.org/wiki/Newick

Dessiner les arbres suivants donnés au format Newick

- *(A,(B,C),(D,(E,F))) ;*
- *(B:6,(A:5,C:3,E:4):5,D:11) ;*

Plusieurs logiciels de dessin d'arbres phylogénétiques écrits au format Newick sont disponibles sur Internet (http://mobyle.pasteur.fr/cgi-bin/portal.py#welcome).
Les programmes permettant de représenter des arbres phylogénétiques se trouvent dans la partie « Programs », section « Phylogeny » à la catégorie « Display ».

Comparez les arbres que vous avez dessinés avec les résultats obtenus pour chacun des 3 programmes disponibles.
Note : Dans les options avancées, choisissez le format de sortie « MS WindowsBitmap ».
Les arbres en sortie de newicktops sont au format GhostScript (extension .ps). Vous les trouverez ici au format pdf.

Partie 2 : Les primates

L'objectif de cette première partie est de reconstruire une phylogénie crédible des primates suivants :

- Lémur (http://fr.wikipedia.org/wiki/L%C3%A9mur)
- Singe hurleur (http://fr.wikipedia.org/wiki/Singe_hurleur)
- Macaque (http://fr.wikipedia.org/wiki/Macaque)
- Gibbon (http://fr.wikipedia.org/wiki/Gibbon)
- Orang outan (http://fr.wikipedia.org/wiki/Orang-outan)
- Chimpanzé (http://fr.wikipedia.org/wiki/Chimpanz%C3%A9)
- Bonobo (http://fr.wikipedia.org/wiki/Bonobo_%28primate%29)
- Homme (http://fr.wikipedia.org/wiki/George_W._Bush)

A. Les données à utiliser

Cette étude sera faite à partir des données moléculaires concernant le cytochrome oxydase, gène mitochondrial codant pour une enzyme intervenant dans le cycle énergétique de la cellule.

Pourquoi baser notre étude sur l'ADN mitochondrial (ADNmt) ?

Pour rapatrier les séquences protéiques de ce gène pour chacune des espèces, il faut aller sur NCBI (http://www.ncbi.nlm.nih.gov/) dans la section « Protein ». Le moteur de recherche permet de retrouver ces séquences en indiquant le nom du gène (cytochrome c oxidase subunit II) et le nom de l'espèce en latin ou en anglais. Puis, il vous suffit de mettre la combo box « Display » sur « fasta » pour récupérer la séquence dans le bon format.

Pour vous aider un peu dans ce travail laborieux et pas forcément très intéressant, vous trouverez un fichier fasta contenant les 8 séquences nécessaires à l'adresse :
http://www.griv.org/ED_BNF104/

Compléter ce fichier avec la séquence du Gorille (http://fr.wikipedia.org/wiki/Gorille).

B. Alignement des séquences

Maintenant que l'on a récupéré nos séquences, il va falloir les aligner de façon à appliquer les algorithmes de reconstruction phylogénétique sur des données homogènes.
Afin d'aligner les séquences, utiliser le logiciel ClustalW disponible sur le site :
http://mobyle.pasteur.fr/cgi-bin/portal.py#welcome

Dans la liste « Programs » à gauche, sélectionnez la catégorie « Alignment », puis la section « multiple ». Cliquez sur « clustalw-multialign ».

Alignez les séquences et sauvegardez-les dans un fichier fasta. Quels changements constatez- vous ?

Sauvegadez les séquences alignées dans un nouveau fichier.

C. Reconstruction de la phylogénie par UPGMA

Une fois les séquences alignées, nous allons appliquer la méthode de reconstruction la plus simple : UPGMA (méthode des distances + respect de l'hypothèse d'horloge moléculaire).

Avant d'appliquer la méthode UPGMA, il est nécessaire de générer la matrice de distance. Pour ce faire, dans la liste des programmes, choisissez dans la sous-section « phylogeny » le programme *protdist*. Entrez les séquences alignées dans le cadre réservée aux données d'entrée. Sauvegardez la matrice de distance générée dans un nouveau fichier.

Vous pouvez désormais reconstruire la phylogénie des espèces étudiées à l'aide de la

méthode UPGMA. Dans la liste des programmes, allez dans la section « phylogeny », cliquez sur la section « distance » et choisissez le programme *neighbor*. En haut de la page, sélectionnez UPGMA dans la combo-box « Distance method ». Entrez ensuite la matrice de distance dans le cadre réservé.

Sur la page des résultats, l'arbre phylogénétique est donné dans le cadre « Neighbor output tree file ». Afin de visualiser cet arbre, dans la section « phylogeny », à la sous-section « display », cliquez sur le programme *newicktops*. Cliquez sur le bouton « advanced options » en haut à droite. Activez l'option d'affichage de la longueur des branches.

Commentez l'impact de l'hypothèse d'horloge moléculaire sur l'arbre obtenu et l'enracinement de l'arbre ?

Comment feriez-vous pour enraciner l'arbre plus proprement ?

D. Enracinement de l'arbre

Pour faire un enracinement crédible d'un point de vue de l'évolution, il faut rajouter à notre phylogénie une autre espèce dont on sait qu'elle a divergé des primates il y a longtemps.

Quelle espèce peut-on prendre par exemple ? Justifiez.

Allez chercher la séquence correspondante à votre choix dans NCBI et refaite toute la démarche de la partie C pour obtenir un nouvel arbre. Commentez le résultat.

E. Reconstruction de la phylogénie par NJ et mesure de robustesse de l'arbre (Bootstrap)

Reconstruisez l'arbre phylogénétique à l'aide de la méthode du neighbor-joining. Pour cela, remplacez UPGMA par Neighbor-Joining dans la méthode employée par le logiciel neighbor. Comparez les arbres obtenus par la méthode UPGMA et la méthode Neighbor-Joining.

Afin d'évaluer la robustesse de l'arbre, il faut générer de nouveaux jeux de données, et donc de nouvelles matrices de distances, à l'aide de permutation (méthode du Bootstrap). Etant donné que ce processus prend du temps, vous trouverez 100 nouvelles matrices de distances dans le fichier « boostrap_distmat » à l'adresse http://www.griv.org/ED_BNF104. Pour information, la génération de ces nouvelles données s'effectue en activant l'option « Perform a bootstrap analysis » des options avancées du logiciel *protdist*.

A l'aide du logiciel neighbor précédemment utilisé, reconstruisez la phylogénie des primates par NJ (changez de méthode dans le logiciel « neighbor »). Dans les options avancées, mettez l'option « Analyze multiple data sets » sur Yes, et la valeur 100 pour l'option « How many data sets ». Choisissez un nombre impair pour le Random seed number. Activez ensuite l'option « Compute a consensus tree ».
Le format Newick de cet arbre consensus doit être légérement modifié afin de pouvoir faire

apparaître les valeurs de bootstrap. Dans le logiciel newicktops, dessinez l'arbre suivant : (((((((Homme,(Bonobo,Chimpanze)92)48,(Gorille,Orang)52)74,Gibbon)55,Macaque)100,Singe)93,Chien),Lemurien)100) ;
Commentez les degrés de confiance dans les divergences évolutives successives.

F. Réflexion complémentaire

Quelle stratégie adopteriez-vous pour confirmer vos résultats ?

Partie 3 : Le virus du SIDA

Le virus de l'immunodéficience humaine (VIH) est un rétrovirus infectant l'homme et conduisant à plus ou moins long terme au syndrome d'immunodéficience acquise (SIDA), qui se caractérise par un affaiblissement du système immunitaire et donc, une vulnérabilité à de multiples infections opportunistes.
(http://fr.wikipedia.org/wiki/VIH#Variants_g.C3.A9n.C3.A9tiques).

Étant donné que le génome du VIH fait de l'ordre de 10 kb (10000 nucléotides), on peut se permettre de travailler sur une grande partie son génome (ici ~3kb) afin d'établir une phylogénie des différentes souches.
Note: Le génome des virus est de l'ordre de 3kb à 200kb.

Voici les différentes souches que l'on va étudier :

Accession Number	Sub Type	Organism	Country
L20571	O	HIV-1	Cameroon
X52154	CPZ	SIV	Gabon
U09127	A1	HIV-1	Uganda
U27426	G	HIV-1	Uganda
U27445	G	HIV-1	Russian Federation
AF067158	C	HIV-1	India
U09126	C	HIV-1	Brazil
U27399	D	HIV-1	Uganda
U43386	D	HIV-1	Uganda
L02317	B	HIV-1	USA
AF025763	B	HIV-1	USA
U08443	B	HIV-1	Haiti
AF042106	B	HIV-1	Australia

Pourquoi intégrer dans l'étude une souche provenant du singe (SIV-CPZ) ?

Les données nucléotidiques de chaque souche sont disponibles à l'adresse suivante :
http://www.griv.org/ED_BNF104/

Reconstruisez une phylogénie des différentes souches par NJ avec bootstrap

Commentez les résultats ?

Comparez ces résultats à une classification connue.