TP Bioinformatique - M1 SPGF

Phylogénie moléculaire

Janvier 2008

Introduction

La **phylogénie** (ou *phylogénèse*) est une reconstruction de l'histoire évolutive des êtres vivants. La phylogénie moléculaire correspond à de la phylogénie par comparaison de gènes ou de protéines.

Un **arbre phylogénétique** est une représentation graphique de la phylogénèse d'un groupe de taxa réalisé à partir de l'étude d'un ou plusieurs caractères.

L'homologie implique une origine évolutive commune mais les caractères étudiés peuvent avoir des formes ou des fonctions différentes.
L'analogie (ou *homoplasie*) implique une similitude de forme ou de fonction sans qu'il n'y ait d'origine évolutive commune.
Enfin, la **xénologie** correspond à l'acquisition d'un caractère sans héritage (cas des transferts horizontaux).

Alignement de séquences

Etape essentielle à toutes les méthodes de phylogénie moléculaire.

La Reconstruction Phylogénétique (*RP*)

On distingue deux grands groupes de méthodes de RP :
- les méthodes de distances (ou *phénétiques*)
- les méthodes de caractères (ou *cladistiques*)

Les modèles de calcul de distances

A cause des substitutions multiples, la distance observée entre deux caractères sous-estime la distance réelle (ou *distance évolutive*). Plusieurs modèles de calcul de distances ont été développés pour corriger ce biais.

Le modèle de Jukes et Cantor (1969) est un modèle à un seul paramètre. Ce modèle assume que les quatre bases ont les mêmes fréquences et que les substitutions sont équiprobables.

$$d_{evo} = -\frac{3}{4}\, ln(1 - \frac{4}{3}\, d_{obs}) \qquad (1)$$

(d_{obs} *correspond à la fréquence de substitutions observées.*)

Le modèle de Kimura (1982) est un modèle à un deux paramètres. Ce modèle considère également que les quatre bases ont les mêmes fréquences mais il tient compte de la proportion entre le nombre de transitions[1] et transversions[2].

$$d_{evo} = -\frac{1}{2}\, ln\Big((1-2P-Q)\,\sqrt{1-2Q}\Big) \tag{2}$$

(*P et Q correspondent respectivement aux fréquences de transitions et transversions observées.*)

Remarque : Les modèles précédents permettent de calculer des distances entre des séquences nucléiques, en tenant compte des propriétés physico-chimiques des nucléotides, mais en omettant complètement le code génétique. Pour des séquences non codantes, c'est effectivement ce qu'il convient de faire.
Par contre, pour des séquences codantes, la survenue d'une mutation dépend des acides aminés correspondants. Pour des séquences protéiques, le calcul de distance se résume généralement au score d'alignement en utilisant une matrice de subsitution type BLOSUM ou PAM. Des modèles de calcul de distances peuvent aussi être utilisés (cf. option *Distance Model* de `protdist`).

Les méthodes phénétiques de RP

La méthode UPGMA (*Unweight Pair Group Method with Arithmetic mean*) (1958) est un algorithme de clusterisation séquentiel qui consiste à regrouper les deux unités taxonomiques (OTU) les plus proches, puis recalculer les distances moyennes avec les autres groupes et ainsi de suite. Elle impose que les distances soient *ultramétriques* (hypothèse d'horloge moléculaire).

La méthode NJ (*Neighbor Joining*) (1987) est la méthode de distances la plus utilisée. Elle considère que les distances sont proches de l'additivité (donc n'implique pas l'hypothèse d'horloge moléculaire). C'est aussi un algorithme de clusterisation séquentiel qui consiste à regrouper les deux OTU dont le regroupement va minimiser la longueur totale de l'arbre.

[1] purine ↔ purine ou pyrimidine ↔ pyrimidine
[2] purine ↔ pyrimidine

Exercices

Exercice 1 :

A partir de la matrice de distances suivante, reconstruire les arbres phylogénétiques par les méthodes UPGMA et NJ. Comparer les deux arbres obtenus.

	A	B	C	D	E	F
A	-					
B	5	-				
C	4	7	-			
D	7	10	7	-		
E	6	9	6	5	-	
F	8	11	8	9	8	-

En vous rendant sur le site http ://mobyle.pasteur.fr, utilisez le programme `neighbor` [3] pour reconstruire les arbres selon les deux méthodes. Les arbres pourront être générés à l'aide des logiciels `drawtree` ou `NjPlot` en local sur vos machines.

Exercice 2

D'après un travail de Jean-Stéphane Varré au LIFL

Nous allons maintenant construire une phylogénie sur les ongulés. Le matériel génétique qui nous servira sera la séquence nucléique du cytochrome B. Nous avons déjà réalisé l'alignement multiple des séquences par `ClustalW` et enregistré le résultat au format `PHYLIP`. Cet alignement est diponible sur http ://julien.maupetit.free.fr.

Réalisez un arbre à l'aide des programmes `dnadist` et `neighbor` en faisant varier les modèles de calcul de distance ainsi que l'algorithme de RP. Comparez les arbres réalisés.

Recommencez ce travail en enlevant la séquence du cochon. Observez vous des différences ?

Vous pouvez maintenant tester la robustesse de vos noeuds à l'aide des méthodes de *bootstrap* ou de *jacknife*. D'une manière générale, évaluez toujours votre arbre à l'aide des modules `seqboot` et `consense`.

Recherchez une séquence du cytochrome B chez une espèce qui ne soit pas un ongulé de manière à enraciner l'arbre (méthode du *outgroup*). Vous pouvez vous rendre sur le site *The Ultimate Ungulate Page* pour vérifiez les phylogénies que vous avez obtenues.

[3] Les noms de programmes notés `programme` appartiennent au package `PHYLIP`.