Projet de Bioinformatique

Phylogénie moléculaire

M1 SPGF/IMVI

28 Février 2008

A propos ...

La *dead line* pour me rendre le compte-rendu de votre projet est fixée au **6 mars 2008 minuit**. Vous devrez me l'envoyer par courriel au format **PDF**. Un point sera retiré par heure dépassée.

Parcimonie de Fitch (1971)

Dix sites variables ont été observés sur des séquences (fragments de 50 bases de la protéine enveloppe du virus HIV1) de cinq isolats notés T1 à T5).

Isolats \ Sites	4	13	14	24	25	26	28	32	34	41
T1	T	C	A	A	C	G	C	T	T	A
T2	T	C	A	A	C	C	T	C	A	G
T3	A	T	G	A	T	A	G	C	A	T
T4	G	T	G	T	T	G	G	C	G	T
T5	T	C	A	C	T	C	C	A	T	A

TAB. 1 – Description des sites pour les cinq isolats.

Questions

1. Quels sont les sites **informatifs** à considérer pour des analyses ultérieures ?

   Pour rappel, un site est considéré comme informatif si au moins deux bases différentes y sont observées au moins deux fois.

2. Nous supposons que seuls trois arbres peuvent être considérés acceptables. D'après leurs représentations parenthèsées, donnez la topologie de chaque arbre sous forme de dendogramme.

$$\textbf{Arbre A} \rightarrow 4 :\Big(3 :\big(1 :(\text{T1}, \text{T2}), \text{T5}\big), 2 :(\text{T3}, \text{T4})\Big)$$
$$\textbf{Arbre B} \rightarrow 4 :\Big(3 :\big(1 :(\text{T2}, \text{T5}), \text{T1}\big), 2 :(\text{T4}, \text{T3})\Big)$$
$$\textbf{Arbre C} \rightarrow 4 :\Big(3 :\big(1 :(\text{T1}, \text{T5}), \text{T4}\big), 2 :(\text{T2}, \text{T3})\Big)$$

3. Déterminez lequel de ces trois arbres est le plus optimal au sens du *critère de parcimonie*. On vous demande de préciser la base (la plus adéquate) pour chacun des sites et ceci pour chaque noeud de l'arbre.

4. Evaluez pour chaque arbre le nombre de transitions et de transversions qu'il a fallu réaliser. En prenant comme critère le nombre minimal de transitions, dire lequel des trois arbres doit être choisi ?

5. Comme l'ont souligné *Brown et al. (1982)*, la fréquence des transitions est beaucoup plus élevée que celle des transversions. Dans ce cas, il semble raisonnable de donner un poids plus élevé aux transversions qu'aux transitions. En effet, les transversions, plus rares, apportent une information phylogénétique plus solide que les transitions dont la trop grande fréquence d'apparition finit par ne produire que du "bruit" (*Darlu et Tassy, 1993*). Si vous considérez un poids $P_{ts} = 1$ pour une transition et $P_{tv} = 3$ pour une transversion, lequel des trois arbres est optimal ?

6. Refaire l'analyse en considérant maintenant qu'il n'existe que deux états possibles R (les purines : `A` et `G`) et Y (les pyrimidines `C` et `T`). Cette approche revient à choisir un arbre au sens de la "parcimonie des transversions" (*Swoffort et Olsen, 1990*) et donc à ignorer les transitions.

Remarques

- Dans chacun des cas, comparez les arbres obtenus, et faites une conclusion générale de vos analyses.

Utilisation du package PHYLIP

En utilisant les outils du package `PHYLIP` vu en TP, refaites l'exercice proposé précédemment. Les séquences au format `fasta` sont disponibles sur :

`http://julien.maupetit.free.fr/teachingFiles/2007/HIV1-EnvPSequences.fasta`

Pour rappel, tous les outils du package PHYLIP sont accessibles sur la plateforme **Mobyle** à l'url suivante :

`http://mobyle.pasteur.fr`

Questions

1. Est-ce que ces séquences sont alignées ? Si oui, convertissez cet alignement au format `PHYLIP`, sinon réalisez l'alignement avec l'outil `ClustalW` en prenant soin de préciser `-output=phylip`.

2. Réalisez dans un premier temps une recherche du meilleur arbre par l'algorithme de parcimonie (`dnapars`), évaluez ensuite les poids des trois arbres proposés dans la première partie de l'énoncé. Comparez.

3. Utilisez maintenant l'option `Transversion Parcimony`, est-ce que cela modifie vos résultats ?

Remarques :

- Il est fortement conseillé d'utiliser les différents outils du package PHYLIP vus en TP pour cette partie (`consense`, `drawtree`, etc ...)
- N'hésitez pas à me contacter si vous avez un problème.
- Attention aux pièges ...