Correction du Contrôle continu octobre 2012

Questions de cours

Question 1

Nous utilisons un modèle évolutif pour réaliser une reconstruction phylogénétique car la distance observée qui correspond aux nombres de substitutions observées entre deux séquences sur le nombre de sites alignés sous-estime la distance évolutive quand les séquences sont issues d'organismes éloignés dans l'évolution. Ceci à pour cause l'existence de substitutions multiples qui ont pu se produire au même site mais qui ne sont pas observables. Ce phénomène est plus critique dans le cas des séquences d'acides nucléiques car elles possèdent un alphabet plus pauvre que les séquences protéiques : quatre lettres au lieu de 20. Pour tenter de corriger le biais du aux mutations multiples, des hypothèses sont faites sur la façon dont les bases se sont substituées à un locus donné conduisant à la construction d'un modèle évolutif.

Modèle de Jukes et Cantor et modèle de Tamura

La différence entre les modèles d'évolution est liée à la définition des μ_{ij} correspondant au taux de substitution instantané d'une base d'un état *i* vers un état *j* ($i \neq j$).

Dans le cas du modèle de Jukes et Cantor, toutes les substitutions sont équiprobables donc un seul taux de substitution instantané α pour chacun des changements possibles (tous les $\mu_{ij} = \alpha$). C'est le modèle évolutif le plus simple mais qui correspond à une vision très simplificatrice de l'évolution.

Dans le cas du modèle de Tamura, les substitutions se produisent suivant deux taux distincts, l'un pour les transitions, l'autre pour les transversions, les transitions étant plus fréquentes (transition = A<->G ou T<->C). De plus, le modèle précédent de Jukes et Cantor, impose que les fréquences des bases à l'équilibre soient toutes égales à ¼, donc que le taux global de GC soit égal à ½. Or ceci est rarement vérifié sur les séquences réelles. Des modèles alternatifs ont été proposés pour rendre compte de cette réalité biologique. Le modèle de Tamura intègre un paramètre supplémentaire θ représentant la fréquence de GC de la ou des séquence(s) considérée(s).

Le modèle de Tamura représentant mieux la réalité biologique, il est plus approprié pour la construction d'un arbre phylogénétique à partir de séquences d'ARNr 16S.

Question 2

Les modèles évolutifs sont établis dans un cadre conceptuel : le modèle de Markov en temps continu. Dans les modèles markoviens, l'information utile pour la prédiction du futur est contenue dans l'état présent du processus. Donc, l'état futur d'un site de dépendra que de son état présents et pas des états passés. Plusieurs hypothèses sont liées au modèle markovien. L'une d'entre-elles est l'uniformité du processus **:** tous les sites d'une séquence suivent le même processus, c'est-à-dire que les probabilités et taux de substitutions sont applicables à tous les sites. La conséquence est que l'on suppose que les sites évoluent à la même vitesse. On sait que cette hypothèse est fausse mais elle utilisée dans la plupart des modèles

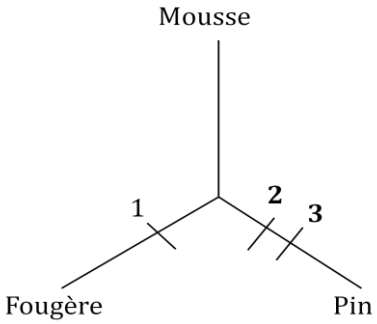
d'évolution. En effet, les contraintes fonctionnelles engendrent des taux d'évolution (*r*) différents selon les sites. Il a été démontré que ce taux *r* est modélisable par une loi Gamma (séquences nucléiques ou protéiques). L'utilisation de la distribution Gamma permet donc de prendre en compte l'existence de vitesses d'évolution différentes. Si nous utilisons une correction Gamma 4 catégories, nous considérerons que les sites évoluent suivant quatre taux d'évolution différents.

Question 3

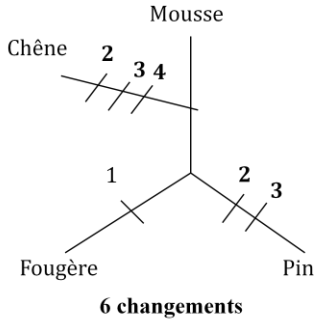
Nous avons ici des données morphologiques représentées par la présence/absence d'un caractère dérivé. L'approche que nous utiliserons sera donc une approche cladistique utilisant la méthode de parcimonie pour la construction de l'arbre phylogénétique.

Reconstruction de l'arbre :

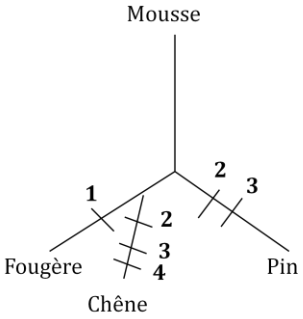
Première étape : On construit un arbre avec les 3 premières espèces et on reporte sur les branches le numéro du caractère transformé. Ici la mousse sert de groupe externe car elle ne possède aucun des caractères dérivés analysés. Nous obtenons l'arbre présenté ci-dessous.



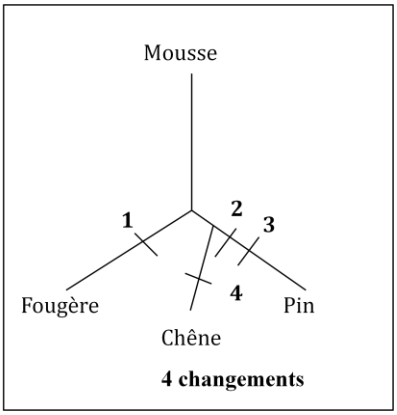
Seconde étape : Nous allons rajouter la quatrième espèce sur cet arbre. Il y a trois possibilités car trois branches internes. Pour chaque arbre, nous allons placer sur ces branches l'apparition des caractères dérivés permettant d'expliquer la topologie. Nous conserverons l'arbre le plus parcimonieux, c'est-à-dire celui dont la topologie s'explique par le minimum de changements. Dans note cas, il s'agit de l'arbre encadré dont la topologie s'explique par 4 changements. Pour les deux autres arbres, nous avons supposé que les caractères 2 et 3 étaient apparus indépendamment sur les branches menant au chêne et au pin. Nous aurions pu faire l'hypothèse qu'ils étaient apparus avant la séparation chêne/pin et auraient subi une réversion sur la branche menant à la fougère. Le nombre de changement aurait été équivalent.



6 changements



6 changements



4 changements

Problème

1) Démarche bioinformatique

- Utiliser chacun des produits protéiques codés par les gènes de la voie de biosynthèse des stérols comme sonde pour une recherche par similarité avec BlastP sur l'ensemble des protéines codées par les génomes complets eucaryotes disponibles.

- Réaliser ensuite les alignements multiples de chaque famille protéique. Vérifier ces alignements de manière à identifier la présence de séquences partielles qui seront alors éliminées de l'analyse. Vérifier également que les séquences sont correctement alignées et qu'il n'est pas nécessaire d'intervenir sur l'alignement pour l'améliorer.

- Rechercher le modèle évolutif le plus approprié pour construire l'arbre et ceci pour chaque famille de protéines à l'aide du logiciel Protest.

- Construire les arbres correspondant en choisissant de préférence la méthode du maximum de vraisemblance PhyML ou alors une méthode de distance (NJ ou bioNJ) et en utilisant le modèle évolutif sélectionné à l'étape précédente. Pour vérifier la robustesse de l'arbre, nous effectuerons 100 (ou 1000) bootstrap.

2) Arbre consensus

Un arbre consensus permet de comparer et d'évaluer les topologies de plusieurs arbres. Dans le cas du TP nous avons utilisé un consensus strict. Dans ce cas, la comparaison des branches (appelées aussi bipartitions) des arbres conduit à garder la branche dans l'arbre consensus si elle est présente dans l'ensemble des arbres comparés, sinon cela se traduira par une multifurcation sur l'arbre consensus. A noter que les longueurs de branches associées à l'arbre consensus n'ont pas de signification phylogénétique.

3) nombre figurant sur les branches des arbres ERG7 et ERG5

Les nombres figurant sur les branches des arbres correspondent aux valeurs de bootstrap.
Le calcul des valeurs de bootstrap sert à évaluer la robustesse de l'arbre, c'est-à-dire s'il est bien supporté par les données. C'est une méthode de rééchantillonage (tirage aléatoire avec remise des positions de l'alignement pour construire un nouvel alignement de même longueur qui servira à construire un arbre). Elle teste individuellement la validité de chaque branche interne de l'arbre en calculant le pourcentage de fois où une branche de départ se retrouve dans les arbres construits par rééchantillonage. De manière générale, une faible valeur de bootstrap indique que la quantité d'information supportant la bipartition induite par une branche interne est faible. Si la valeur du boostrap est inférieure à 70%, la branche n'est pas bien supportée par les données. Il faut alors interpréter la topologie avec précaution.

4) Séquences protéiques versus séquences d'acides nucléiques

Les arbres ont été construits à partir de séquences protéiques et non nucléiques car lorsque les espèces sont distantes dans l'évolution, les séquences nucléiques peuvent avoir subi des

substitutions multiples qui conduiront à une sous-estimation de leurs distances évolutives. On peut même dans certains cas avoir perdu le signal phylogénétique. Ceci est du au petit alphabet de ces séquences (4 lettres, les 4 bases). On préfère donc travailler au niveau protéique.

5) Séquences SHC

Les séquences SHC bactériennes ont été ajoutées à l'analyse pour la construction de l'arbre sur les protéines ERG7 de manière à fournir un groupe externe. Ceci permet donc de connaître l'ancêtre commun (racine) des séquences ERG7 qui est alors localisé sur la branche qui relie le groupe externe aux autres séquences. D'autre part, ici le groupe externe est constitué de plusieurs espèces présentant des distances évolutives étalonnées, ce qui permet de « casser » les longues branches et donc de limiter l'impact de l'artéfact de l'attraction des longues branches auquel sont sujettes les méthodes de reconstruction d'arbre.
Sur les quatre génomes bactériens, trois possèdent un exemplaire de cette séquence SHC. Le génome de *Stigmatella aurantiaca* en est dépourvu.

6) Comparaison arbre ERG7 et arbre des espèces.

La protéine ERG7 étant indispensable à la synthèse des stérols, les organismes qui en sont dépourvus ne peuvent donc pas réaliser cette biosynthèse. En comparant l'arbre obtenu sur ERG7 et celui donnant la phylogénie des eucaryotes, il est donc possible d'identifier les génomes dépourvus de ce gène. Si plusieurs génomes eucaryotes (feuilles de l'arbre des espèces) issus d'un ancêtre commun ne possèdent plus ce gène, c'est que celui-ci a été perdu sur la branche de l'arbre conduisant à cet ancêtre (lui-même ne possédait plus le gène). Les croix marron sur les branches de l'arbre des espèces ci-dessous indiquent les positions dans l'arbre des espèces où la capacité de synthèse des stérols a été perdue.

7) Analyse de l'arbre des protéines ERG7

Des duplications sont visibles sur l'arbre. Les génomes sont : *Trypanosoma cruzi*, *Oriza sativa*, *Arabidopsis thaliana* et *Aspergillus fumigatus*. Dans ce dernier génome, il y a trois exemplaires de la protéine, on peut donc supposer qu'après une première duplication, un des deux gènes s'est de nouveau dupliqué.

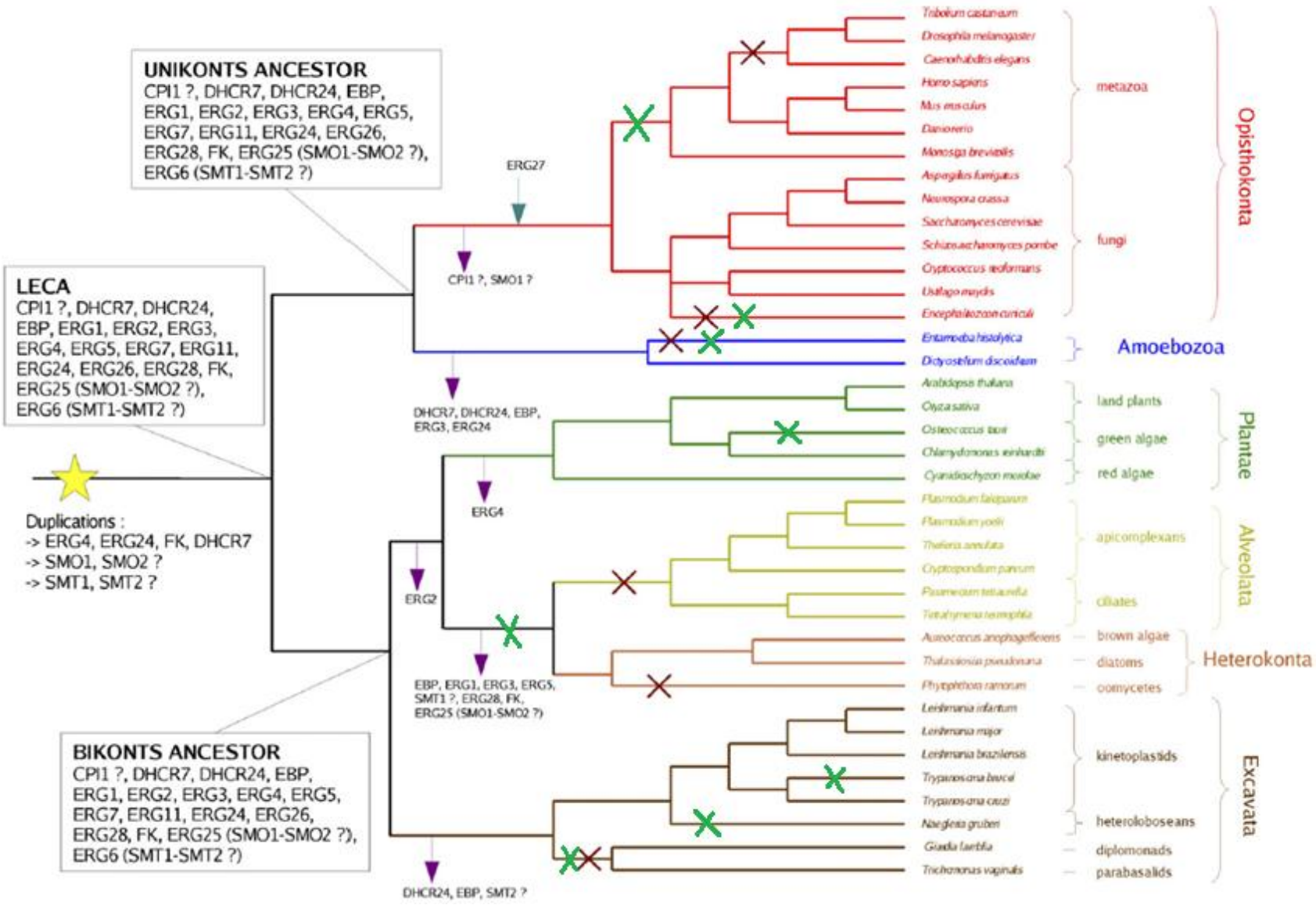
Transferts horizontaux :
La localisation de la séquence bactérienne de *Stigmatella aurantiaca* avec les séquences eucaryotes est une indication de l'acquisition de cette séquence par la bactérie au travers d'un transfert horizontal d'une séquence provenant d'un génome eucaryote. De plus, l'identification de la séquence ERG7 dans un petit nombre de génomes bactériens (quatre) est aussi en faveur de l'acquisition de cette séquence par ces génomes via un transfert horizontal dont la source serait un génome eucaryote. En effet, autrement il faudrait supposer que la séquence du gène ERG7 était présente dans l'ensemble des génomes procaryotes, au moins ceux possédant SHC et qu'ensuite elle ait été perdue par la majorité de ces génomes excepté les quatre génomes en question. Ceci nécessiterait d'envisager un grand nombre de pertes indépendantes, ce qui n'est pas l'hypothèse la plus parcimonieuse et la position de la séquence de *S. aurantiaca* indique clairement l'acquisition horizontale du gène.

8) Les gènes eucaryotes codant pour la protéine ERG7 sont homologues car ils possèdent un ancêtre commun. Ils sont orthologues car ils ont été acquis par héritage vertical donc suite à la

spéciation. Cependant dans les cas où des duplications ont eu lieu, ces gènes sont paralogues entre-eux (dans le même génome) et ils forment des groupes d'orthologues quand comparés aux autres espèces.

9) Protéines ERG5

La distribution taxonomique de la protéine ERG5 montre que son gène est identifié dans des génomes UNIKONTS comme par exemple *Saccharomyces cerevisiae* mais aussi dans des génomes BIKONTS comme *Arabidopsis Thaliana*. Le gène codant pour cette enzyme devait donc être présent dans le génome ancestral avant la séparation BIKONTS et UNIKONTS, c'est-à-dire dans le génome de LECA.
Ce gène a subi plusieurs pertes indiquées par des croix vertes sur les branches de la phylogénie des eucaryotes. Nous pouvons également observer des évènements de duplication récents dans certains génomes car les feuilles des arbres correspondant aux gènes dupliqués sont regroupées sous un même nœud ancêtre. Un évènement de duplication a eu lieu dans le génome d'*Oryza sativa* et dans celui de *Cyanidioschyzon merolae*. Le génome d'*Arabidopsis thaliana* a subi trois évènements de duplication. Après le premier évènement, chacun des deux gènes s'est de nouveau dupliqué.


UNIKONTS ANCESTOR
CPI1 ?, DHCR7, DHCR24, EBP, ERG1, ERG2, ERG3, ERG4, ERG5, ERG7, ERG11, ERG24, ERG26, ERG28, FK, ERG25 (SMO1-SMO2 ?), ERG6 (SMT1-SMT2 ?)
LECA
CPI1 ?, DHCR7, DHCR24, EBP, ERG1, ERG2, ERG3, ERG4, ERG5, ERG7, ERG11, ERG24, ERG26, ERG28, FK, ERG25 (SMO1-SMO2 ?), ERG6 (SMT1-SMT2 ?)
Duplications :
-> ERG4, ERG24, FK, DHCR7
-> SMO1, SMO2 ?
-> SMT1, SMT2 ?
BIKONTS ANCESTOR
CPI1 ?, DHCR7, DHCR24, EBP, ERG1, ERG2, ERG3, ERG4, ERG5, ERG7, ERG11, ERG24, ERG26, ERG28, FK, ERG25 (SMO1-SMO2 ?), ERG6 (SMT1-SMT2 ?)
ERG27
CPI1 ?, SMO1 ?
DHCR7, DHCR24, EBP, ERG3, ERG24
ERG4
ERG2
EBP, ERG1, ERG3, ERG5, SMT1 ?, ERG28, FK, ERG25 (SMO1-SMO2 ?)
DHCR24, EBP, SMT2 ?
Tribolium castaneum
Drosophila melanogaster
Caenorhabditis elegans
Homo sapiens
Mus musculus
Danio rerio
Monosiga brevicollis
Aspergillus fumigatus
Neurospora crassa
Saccharomyces cerevisiae
Schizosaccharomyces pombe
Cryptococcus neoformans
Ustilago maydis
Encephalitozoon cuniculi
Entamoeba histolytica
Dictyostelium discoideum
Arabidopsis thaliana
Oryza sativa
Ostreococcus tauri
Chlamydomonas reinhardtii
Cyanidioschyzon merolae
Plasmodium falciparum
Plasmodium yoelii
Theileria annulata
Cryptosporidium parvum
Paramecium tetraurelia
Tetrahymena thermophila
Aureococcus anophagefferens
Thalassiosira pseudonana
Phytophthora ramorum
Leishmania infantum
Leishmania major
Leishmania braziliensis
Trypanosoma brucei
Trypanosoma cruzi
Naegleria gruberi
Giardia lamblia
Trichomonas vaginalis
metazoa
fungi
Opisthokonta
Amoebozoa
land plants
green algae
red algae
Plantae
apicomplexans
ciliates
Alveolata
brown algae
diatoms
oomycetes
Heterokonta
kinetoplastids
heteroloboseans
diplomonads
parabasalids
Excavata