Phylogénie moléculaire
durée 2h

Tous documents et moyens de calculs (calculatrice, ordinateur) autorisés

Question 1 (2 points) :

Donnez la formule (*i.e.*, la fonction de vraisemblance) permettant d'estimer au maximum de vraisemblance la distance évolutive entre deux séquences protéiques, ceci sous l'hypothèse du modèle de Poisson.

Réponse :

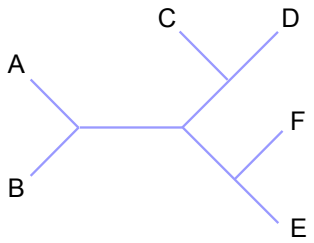
Sachant que, sous l'hypothèse du modèle de Poisson, on a $p = 1 - e^{-d}$, on en déduit la fonction de vraisemblance $L(d)$, telle que :

$$\begin{aligned} L(d) = f(p|d) &= \binom{\ell}{n} p^n (1-p)^{\ell-n} \\ &= \frac{\ell!}{n!(\ell-n)!} \left(1 - e^{-d}\right)^n \left(e^{-d}\right)^{\ell-n} \end{aligned}$$

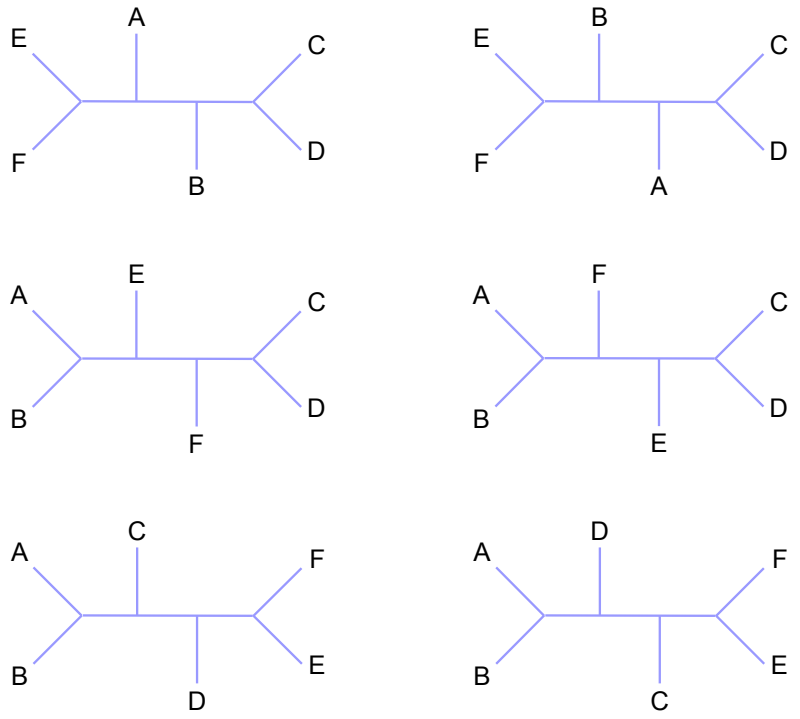
avec ℓ le nombre de sites dans l'alignement et n le nombre de sites pour lesquels il y a une substitution entre les deux séquences.

Question 2 (6 points) :

Dessinez l'ensemble des topologies qu'il est possible d'obtenir par réarrangement NNI à partir de l'arbre non raciné à six UTO ci-dessous :



Réponse :



Question 3 (4 points) :

Ci-dessous figure un alignement de deux séquences nucléotidiques, A et B :

```
A:  AGCTGGCATTGCTTGCATGCATCTTGTCATATGGCAGCGCGCTTGCAGTA
B:  AGCAGAGTCTCCATGTTTCCATCTTGCCTTACAGTAGCAGTGTTGGAGAA
```

Calculez la distance de Kimura à deux paramètres entre ces séquences ainsi que leur ratio transition/transversion. Le détail des calculs devra être donné et toute réponse ne contenant que les résultats numériques sera considérée comme nulle et non avenue.

Réponses :

Sachant que le nombre de sites dans l'alignement est $\ell = 50$, que le nombre de transitions est $n_r = 8$ et que le nombre de transversions est $n_v = 13$, on en déduit la distance :

$$d = -\frac{1}{2}\ln\left(1 - 2 \times 8/50 - 13/50\right) - \frac{1}{4}\ln\left(1 - 2 \times 13/50\right) \simeq 0.6172$$

De même, le ratio transition/transversion est égal à :

$$\kappa = \frac{2\ln\left(1 - 2 \times 8/50 - 13/50\right)}{\ln\left(1 - 2 \times 13/50\right)} - 1 \simeq 1.364$$

Question 4 (4 points) :

Dans le tableau ci-dessous figurent les comptages des conservations et des différentes substitutions observées entre deux séquences nucléotidiques A et B :

A/B	A	C	T	G
A	90	3	2	3
C	2	96	5	4
T	5	3	94	1
G	3	8	2	79

1. Calculez la distance de Jukes et Cantor (détail des calculs exigé) entre ces deux séquences ainsi que la variance de cette estimation.
2. Calculez la distance de Kimura à deux paramètres (détail des calculs exigé) entre ces deux séquences ainsi que la variance de cette estimation.
3. Comparez les résultats. Qu'en concluez-vous ?

Réponses :

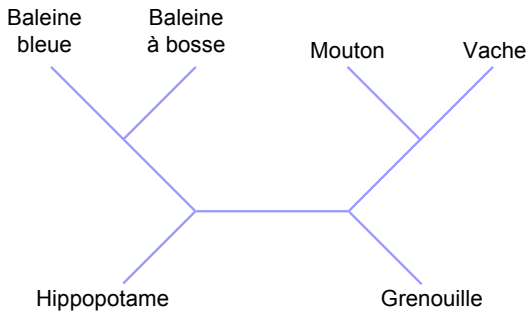
1. Le nombre de sites dans l'alignement est égal au grand total tu tableau, soit $\ell = 400$. Par ailleurs, le nombre de substitutions observées est égal à la somme des éléments non diagonaux, soit $n = 41$. On en déduit que la distance de Jukes et Cantor entre A et B est égale à :
$$d = -\frac{3}{4}\ln\left(1 - \frac{4}{3} \times \frac{41}{400}\right) \simeq 0.1102$$
De même, la variance de cette estimation est égale à :
$$\mathbb{V}(d) = \frac{9 \times 41/400 \times (1 - 41/400)}{400 \times (3 - 4 \times 41/400)^2} \simeq 3.086 \times 10^{-4}$$

2. Le nombre de transitions est $n_r = 14$ tandis que le nombre de transversions est $n_v = 27$. On en déduit que la distance de Kimura à deux paramètres entre A et B est égale à :
$$d = -\frac{1}{2}\ln\left(1 - 2 \times 14/400 - 27/400\right) - \frac{1}{4}\ln\left(1 - 2 \times 27/400\right) \simeq 0.1102$$
Pour faciliter le calcul de la variance, on pose :
$$\begin{aligned} c_1 &= 1/(1 - 2 \times 14/400 - 27/400) \simeq 1.159 \\ c_2 &= 1/(1 - 2 \times 27/400) \simeq 1.156 \\ c_3 &= (1.159 + 1.156)/2 \simeq 1.158 \end{aligned}$$
On en déduit l'estimation de la variance :
$$\begin{aligned} \mathbb{V}(d) &= \frac{1.159^2 \times 14/400 + 1.158^2 \times 27/400 - (1.159 \times 14/400 + 1.158 \times 27/400)^2}{400} \\ &\simeq 3.086 \times 10^{-4} \end{aligned}$$

3. On constate que les estimations des distances ainsi que de leurs variances respectives sont égales. Ceci signifie que les taux instantanés des transitions et des transversions inférés par le modèle de Kimura sont égaux entre eux ($\alpha = \beta$), ce qui nous ramène au modèle de Jukes et Cantor.

Question 5 (4 points) :

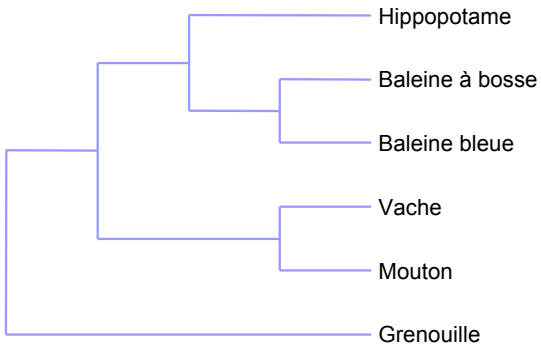
Ci-dessous figure un arbre non raciné avec six espèces de vertébrés :



1. Dessinez la version racinée de cet arbre en utilisant l'organisme que vous estimez comme étant le plus approprié pour servir de groupe externe.
2. Sur cet arbre raciné, de quelle espèce l'Hippopotame est-il le plus proche, la Baleine ou le Mouton ?

Réponses :

1. Le groupe le plus approprié pour enraciner l'arbre est la grenouille, qui est un batracien, alors que toutes les autres UTO correspondent à des mammifères :



2. Sur cet arbre, l'Hippopotame est plus proche de la Baleine que du Mouton.

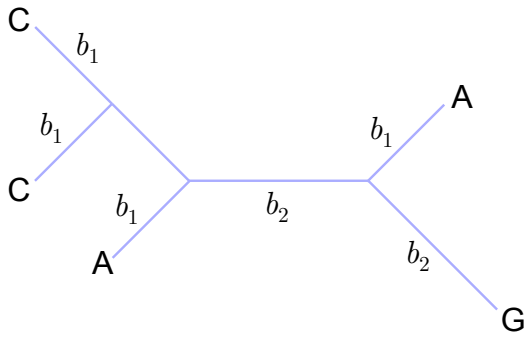
Phylogénie moléculaire
durée 3h

29 novembre 2013

Tous documents et moyens de calculs (calculatrice, ordinateur) autorisés
La partie pratique nécessite l'utilisation du programme SeaView

1 Partie théorique : vraisemblance d'un site (5 points) :

Ci-dessous figure un arbre phylogénétique construit à partir de cinq séquences d'ADN. On considère un seul site de l'alignement et les nucléotides figurant dans ce site sont placés aux feuilles correspondantes dans l'arbre. Par ailleurs, les différentes longueurs de branches sont égales à b_1 ou à b_2. A côté de l'arbre figurent les deux matrices de probabilités de substitutions correspondant à ces longueurs de branches :



Vers

De ↗	A	C	G	T
A	0.896	0.026	0.052	0.026
C	0.156	0.767	0.052	0.026
G	0.156	0.026	0.793	0.026
T	0.156	0.026	0.052	0.767

b_1

Vers

De ↗	A	C	G	T
A	0.780	0.055	0.110	0.055
C	0.330	0.504	0.110	0.055
G	0.330	0.055	0.559	0.055
T	0.330	0.055	0.110	0.504

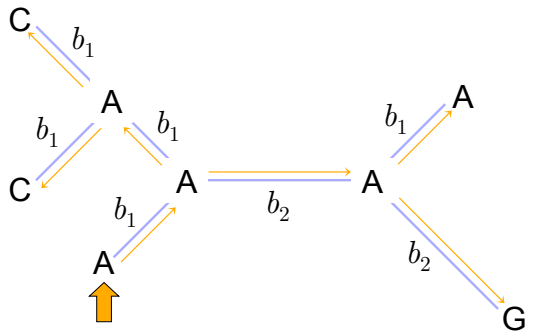
b_2

Considérant que les fréquences à l'équilibre des quatre nucléotides sont $\pi_A = 0.6$, $\pi_C = 0.1$, $\pi_G = 0.2$ et $\pi_T = 0.1$:

1. Si l'on fait l'hypothèse que les nucléotides ancestraux présents dans l'arbre sont tous des A, calculez la vraisemblance de cet arbre étant donné les informations fournies.

2. Maintenant, si l'on fait l'hypothèse que les nucléotides ancestraux sont tous des C, refaites le calcul de la vraisemblance. Des deux hypothèses, laquelle est la plus vraisemblable ?

Du fait que les nucléotides aux nœuds internes sont tous fixés, il n'est pas nécessaire d'utiliser les vecteurs de vraisemblances partielles et le calcul peut être initialisé partir de n'importe quelle feuille (le résultat final sera identique, quelle que soit la feuille choisie). Dans l'exemple donné ci-dessous, l'initialisation a été effectuée au niveau de la feuille indiquée par une flèche pleine et le flux des calculs est donné par les flèches figurant le long des branches :



Dans ce cas, le calcul de la vraisemblance est donné par la formule :

$$\begin{aligned} L &= \pi_{\mathrm{A}}\mathbb{P}(\mathrm{A}|\mathrm{A},b_1)\mathbb{P}(\mathrm{A}|\mathrm{A},b_1)\mathbb{P}(\mathrm{A}|\mathrm{C},b_1)\mathbb{P}(\mathrm{A}|\mathrm{C},b_1)\mathbb{P}(\mathrm{A}|\mathrm{A},b_2)\mathbb{P}(\mathrm{A}|\mathrm{A},b_1)\mathbb{P}(\mathrm{A}|\mathrm{G},b_2) \\ &= 0.6 \times 0.896 \times 0.896 \times 0.026 \times 0.026 \times 0.780 \times 0.896 \times 0.110 \\ &\simeq 2.503 \times 10^{-5} \end{aligned}$$

Soit, sous forme logarithmique : $\ln L \simeq -10.6$.

Si maintenant les A aux nœuds internes sont remplacés par des C, la formule de calcul de la vraisemblance devient :

$$\begin{aligned} L &= \pi_{\mathrm{A}}\mathbb{P}(\mathrm{A}|\mathrm{C},b_1)\mathbb{P}(\mathrm{C}|\mathrm{C},b_1)\mathbb{P}(\mathrm{C}|\mathrm{C},b_1)\mathbb{P}(\mathrm{C}|\mathrm{C},b_1)\mathbb{P}(\mathrm{C}|\mathrm{C},b_2)\mathbb{P}(\mathrm{C}|\mathrm{A},b_1)\mathbb{P}(\mathrm{C}|\mathrm{G},b_2) \\ &= 0.6 \times 0.026 \times 0.767 \times 0.767 \times 0.767 \times 0.504 \times 0.156 \times 0.110 \\ &\simeq 6.088 \times 10^{-5} \end{aligned}$$

Soit, sous forme logarithmique : $\ln L \simeq -9.707$.

L'hypothèse où les nucléotides ancestraux sont tous des C est donc plus vraisemblable que celle où ils sont tous des A.

2 Partie pratique : phylogénie et origine de HIV (15 points) :

L'Homme n'est pas le réservoir naturel des virus HIV (*Human Immuno-deficience Virus*). Au lieu de cela, ces virus ont infecté la population humaine par une ou plusieurs transmissions zoonotiques, c'est-à-dire d'espèce à espèce. Deux souches de HIV circulent actuellement au sein de la population humaine : HIV-1 – qui cause la majorité des cas déclarés du SIDA – et HIV-2. Par ailleurs, il a été montré que les primates étaient très probablement le réservoir naturel de HIV. En effet, des virus très similaires à HIV, d'un point de vue de la structure même

du virus ou de son génome, ont été retrouvés chez les primates. Ces virus ont été nommés SIV pour *Simian Immuno-deficience Virus*. Des virus SIV ont été observés chez deux sous-espèces de chimpanzés, *Pan troglodytes troglodytes* (Ptt) et *Pan troglodytes schweinfurthii* (Pts) alors qu'ils n'ont pour le moment pas été détectés chez les deux autres sous-espèces que sont *Pan troglodytes verus* et *Pan troglodytes vellerosus*. La répartition géographique de ces quatre sous-espèces est donnée dans la Figure 1.

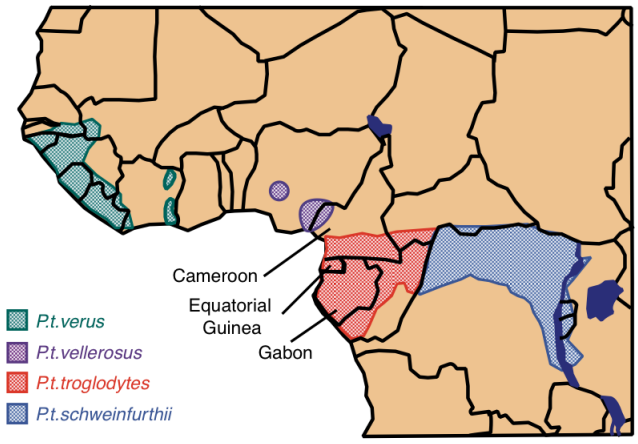


FIGURE 1 – Répartition géographique des quatre sous-espèces de chimpanzés communs.

2.1 Origine de HIV

Les jeux de données que vous allez manipuler contiennent à la fois des séquences d'HIV-1, d'HIV-2 et de SIV de nombreux primates, ainsi qu'une séquence de FIV (pour *Feline Immuno-deficience Virus*) qui servira de groupe externe à nos arbres. Le fichier utilisé (`HIVpol.fst`) contient un ensemble de séquences d'ADN correspondant au gène *pol*.

Lancez le programme `SeaView` puis chargez le fichier `HIVpol.fst`. Comme vous pouvez le constater, les données présentes dans ce fichier sont des séquences nucléiques. L'analyse phylogénétique va, quant à elle, utiliser les séquences protéiques correspondantes. Pour ce faire, sélectionnez dans le menu `Props` l'option `View as proteins`. Une fois ceci fait, il est nécessaire d'aligner les séquences. Vérifiez tout d'abord que `Muscle` est bien le programme utilisé par défaut (menu `Align`, option `Alignment options`), puis alignez les séquences (menu `Align`, option `Align all`). Une fois que le programme a fini de tourner, n'oubliez pas de cliquer sur `OK` pour charger l'alignement.

Repérez dans l'alignement les régions variables. Quelles sont les deux hypothèses permettant d'expliquer l'existence de ces régions si l'on se place du point de vue de la sélection naturelle ? **(3 points)**

Les régions hyper-variables peuvent s'expliquer par un relâchement de la sélection naturelle ou une forte sélection positive pour l'apparition de mutations non-synonymes. Dans le cas du HIV, il est connu que le virus mute très facilement afin de faciliter son pouvoir infectieux. En revanche, il n'est pas possible ici de directement trancher entre les deux hypothèses.

Avant de construire la phylogénie proprement dite, il est nécessaire de filtrer l'alignement.

Dans le menu `Sites`, sélectionnez l'option `Create set`, puis `Gblocks`. Différentes options du programme sont proposées. Sélectionnez les trois premières puis cliquez sur `OK`.

Sur l'alignement filtré, construisez tout d'abord un arbre en utilisant la méthode du maximum de vraisemblance. Dans le menu `Trees`, sélectionnez l'option `PhyML`. Laissez tous les paramètres par défaut puis cliquez sur `Run`. Le calcul doit prendre un certain temps. Si nécessaire, racinez l'arbre avec la séquence `FIV_OMA`. Une fois ceci fait, vous pouvez sauvegarder l'arbre dans le menu `Trees` de la fenêtre principale de `SeaView` (menu `File`, option `Save to Trees menu` de la fenêtre `PHYML_tree`). Afin de vous aider à reconnaitre les espèces hôtes des virus SIV, la relation ente le nom des séquences et l'espèce est donnée dans la Table 1. Les lettres et les chiffres figurant à la fin des noms de séquences (*e.g.*, `GAB1`) sont des indicateurs sur la provenance géographique des souches.

Séquence	Espèce
SIVcpz_Ptt	*Pan troglodytes troglodytes*
SIVcpz_Pts	*Pan troglodytes schweinfurthii*
SIVstm	*Stump-tailed Macaque*
SIVsyk	*Cercopithecus albogularis*
SIVagm	*African Green Monkeys*
SIVlhoest	*Cercopithecus lhoesti*
SIVsun	*Cercopithecus solatus*
SIVsm	*Cercocebus atys*

TABLE 1 – Concordance entre les noms des séquences SIV et les noms d'espèces.

Refaites maintenant une phylogénie en utilisant une méthode de distance. Dans le menu `Trees`, sélectionnez l'option `Distance methods`. Dans la boite de dialogue qui apparaît, sélectionnez `BioNJ`, la distance de `Kimura`, l'option permettant de calculer le *bootstrap* avec 1000 réplicats et déselectionnez l'option `ignore all gaps`. Une fois l'arbre obtenu, racinez-le avec la séquence `FIV_OMA`.

Comparez l'arbre de distance et celui du maximum de vraisemblance (topologie, soutien statistique des branches). Existe-t-il des différences significatives entre ces deux arbres ? Que pouvez-vous conclure par rapport au signal phylogénétique ? Par ailleurs, que pouvez-vous conclure quant à l'origine des virus HIV ? **(6 points)**

Les deux arbres sont quasiment identiques ce qui signifie que le signal phylogénétique semble être assez important pour pouvoir tirer des conclusions biologiques. HIV-1 et HIV-2 ont des origines bien différentes. Il y a eu de multiples transmissions parallèles (et non pas une seule) de primates à l'Homme au cours du passé. Le réservoir naturel de HIV-2 semble être proche de *Cercocebus* (SIVsm) ou des Macaques (SIVstm). Le réservoir naturel de HIV-1 n'est pas unique. Les gorilles semblent être le réservoir de HIV-1 groupe O, tandis que les séquences de chimpanzés sont phylogénétiquement plus proches des groupe M et N d'HIV-1.

2.2 Origine liée au vaccin contre la polio

Dans les années 1990, une hypothèse très controversée a été proposée pour expliquer l'origine de la transmission du HIV du Chimpanzé à l'Homme. Cette hypothèse était que le groupe M d'HIV-1, responsable de la majorité des cas déclarés du SIDA, est apparu chez l'Homme suite à la vaccination contre le virus de la polio de centaines de milliers de personnes

vivant au Congo de 1957 à 1960. En effet, lors de recherches dans un dispensaire de Stanleyville (aujourd'hui Kisangani) au Congo, ce vaccin aurait été cultivé à l'aide de cellules hépatiques de chimpanzés. Dans la région de Kisangani, c'est la sous-espèce *Pan troglodytes schweinfurthii* (Pts) qui est présente, contrairement à *Pan troglodytes troglodytes* (Ptt) qui se répartit plus à l'ouest du fleuve Congo (Figure 2).

FIGURE 2 – Carte de la République Démocratique du Congo montrant la répartition géographique des sous-espèces Ptt et Pts.

Lors d'une expédition dans le Kisangani en 2003, des échantillons de matières fécales de chimpanzés ont été collectés. Au sein de ces échantillons, une souche de SIV a été détectée, confirmant que le virus était bien présent dans cette région. Les séquences d'une portion du gène *env* et du gène *nef* ont été analysées phylogénétiquement.

L'objectif de cette partie est de vérifier si la souche de SIV circulant dans la région de Kisangani est phylogénétiquement proche des souches d'HIV-1. Pour cela, récupérez le fichier de séquences `HIVenv.fst` puis ouvrez-le avec `SeaView`. L'alignement est déjà réalisé et il est inutile de le filtrer pour réaliser l'analyse. Construisez la phylogénie de ces séquences par une méthode de distance (`BioNJ`, distance de `Kimura` et 1000 réplicats de *bootstraps*). Faire de même en maximum de vraisemblance en prenant toutes les options par défaut. Dans les deux cas, laissez le racinement au point moyen utilisé par défaut par le programme de visualisation d'arbre.

Comparer les arbres obtenus (topologie, soutien statistique des branches). Au vu des résultats, que concluez-vous relativement à l'hypothèse de l'origine de la transmission d'HIV-1 à l'Homme suite à la vaccination contre la polio ? **(6 points)**

On constate que la topologie des deux arbres est quasiment identique, mis à part une petite différence dans l'ordre de brachement des différentes sous-espèces Pts_TAN. Par ailleurs, la séquence de la région de Kisangani ne se groupe pas du tout avec les séquences d'HIV-1 connues, mais avec les séquences de SIV provenant de la sous-espèce Pts, ceci avec un fort soutien statistique (100% de *bootstrap* et $P = 1.0$ pour le test aLRT). Il ne semble donc pas que le virus circulant dans cette région soit à l'origine d'HIV-1, dont les séquences sont plus proches de l'autre sous-espèce Ptt se répartissant plus à l'Ouest du Congo.

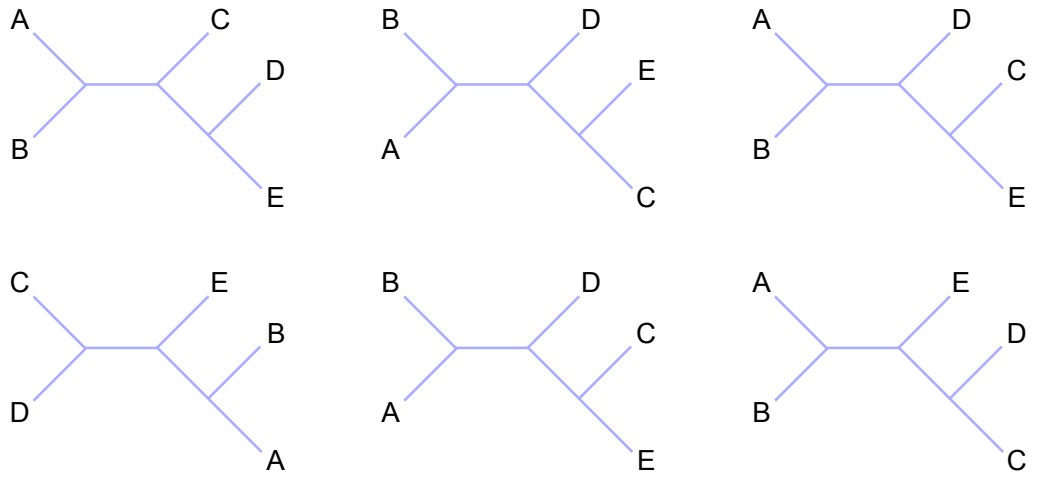
Phylogénie moléculaire
durée 2h

14 septembre 2015

Tous documents et moyens de calculs (calculatrice, ordinateur) autorisés
La partie pratique nécessite l'utilisation du programme SeaView

1 Question 1 (2 points) :

Construisez l'arbre du consensus majoritaire à 50% en utilisant les six arbres ci-dessous. Indiquez sur chacune des branches internes du consensus la fréquence des bipartitions.

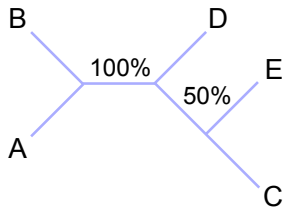


Réponse :

Pour établir l'arbre au consensus majoritaire à 50% il faut regarder les bipartitions induites par les deux branches internes présentes dans les six arbres. De gauche à droite et de haut en bas, les bipartitions en question sont :

— {AB|CDE} et {DE|ABC}, {AB|CDE} et {CE|ABD}, {AB|CDE} et {CE|ABD}
— {AB|CDE} et {CD|ABE}, {AB|CDE} et {CE|ABD}, {AB|CDE} et {CD|ABE}

La bipartition {AB|CDE} est retrouvée dans six arbres sur six (100%) et la bipartition {CE|ABD} est retrouvée dans trois arbres sur six. L'arbre du consensus majoritaire à 50% est donc :



2 Question 2 (6 points) :

Le modèle de substitution TIM3 est une version simplifiée du GTR dans lequel : i) les échangeabilités A $\leftrightarrow$ C et C $\leftrightarrow$ G sont égales entre elles ; et ii) les échangeabilités A $\leftrightarrow$ T et G $\leftrightarrow$ T sont égales entre elles.

Donnez l'expression de la matrice **Q** des taux de substitutions instantanés en utilisant l'ordre donné ci-dessous pour les lignes et les colonnes.

Réponse :

Sachant que, pour le modèle GTR, on a (Diapo. 89 du cours) :
— α = écheangeabilité A $\leftrightarrow$ C, β = écheangeabilité A $\leftrightarrow$ T,
— δ = écheangeabilité C $\leftrightarrow$ T, ϵ = écheangeabilité G $\leftrightarrow$ C,
— γ = écheangeabilité A $\leftrightarrow$ G, η = écheangeabilité G $\leftrightarrow$ T.
Il suffit de remplacer, dans la matrice de la Diapo. 87 du cours, les termes en ϵ par des termes en α et les termes en η par des termes en β. On obtient alors l'expression de **Q** pour le modèle TIM3 :

$$\mathbf{Q} = \begin{array}{c} \\ \mathrm{A} \\ \mathrm{C} \\ \mathrm{T} \\ \mathrm{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{A} & \mathrm{C} & \mathrm{T} & \mathrm{G} \end{array} \\ \begin{pmatrix} -\lambda_{\mathrm{A}} & \pi_{\mathrm{C}}\alpha & \pi_{\mathrm{T}}\beta & \pi_{\mathrm{G}}\gamma \\ \pi_{\mathrm{A}}\alpha & -\lambda_{\mathrm{C}} & \pi_{\mathrm{T}}\delta & \pi_{\mathrm{G}}\alpha \\ \pi_{\mathrm{A}}\beta & \pi_{\mathrm{C}}\delta & -\lambda_{\mathrm{T}} & \pi_{\mathrm{G}}\beta \\ \pi_{\mathrm{A}}\gamma & \pi_{\mathrm{C}}\alpha & \pi_{\mathrm{T}}\beta & -\lambda_{\mathrm{G}} \end{pmatrix} \end{array}$$

Quelle est l'expression de λ, le taux global de substitutions sous ce modèle ?

Réponse :

Sachant que $\lambda = \sum_i \pi_i \lambda_i$, on en déduit :

$$\begin{aligned} \lambda &= \pi_{\mathrm{A}}\lambda_{\mathrm{A}} + \pi_{\mathrm{C}}\lambda_{\mathrm{C}} + \pi_{\mathrm{T}}\lambda_{\mathrm{T}} + \pi_{\mathrm{G}}\lambda_{\mathrm{G}} \\ &= \pi_{\mathrm{A}}(\pi_{\mathrm{C}}\alpha + \pi_{\mathrm{T}}\beta + \pi_{\mathrm{G}}\gamma) + \pi_{\mathrm{C}}(\pi_{\mathrm{A}}\alpha + \pi_{\mathrm{T}}\delta + \pi_{\mathrm{G}}\alpha) \\ &+ \pi_{\mathrm{T}}(\pi_{\mathrm{A}}\beta + \pi_{\mathrm{C}}\delta + \pi_{\mathrm{G}}\beta) + \pi_{\mathrm{G}}(\pi_{\mathrm{A}}\gamma + \pi_{\mathrm{C}}\alpha + \pi_{\mathrm{T}}\beta) \end{aligned}$$

3 Question 3 (12 points) :

Chez les virus, la famille des Flaviviridiae comprend quatre genres : Hepacivirus, Flavivirus, Pegivirus et Pestivirus. Il s'agit de virus dont plusieurs représentants sont responsables

de maladies chez les mammifères. En particulier, le virus de l'Hépatite C (*Hepatitis C virus* ou HCV) est un Hepacivirus. D'après l'Organisation Mondiale de la Santé, 130 à 150 millions d'individus sont porteurs chroniques de l'Hépatite C et 350000 à 500000 personnes décèdent chaque année des suites de pathologies du foie provoquées par ce virus.

Les Flaviviridiae sont caractérisés par un génome ARN encodant une polyprotéine qui est clivée en peptides matures remplissant différentes fonctions. Le nombre de ces peptides est variable au sein des différents genres mais le NS5B est retrouvé chez tous les virus de la famille. En effet, ce peptide correspond à l'ARN polymérase ARN-dépendante nécessaire à la réplication du virus. L'objectif de l'étude que vous allez réaliser est de construire une phylogénie des Hepacivirus au moyen de séquences de NS5B afin d'essayer de déterminer l'origine évolutive de HCV.

Pour commencer, téléchargez le fichier contenant les séquences de NS5B à l'adresse :

`http://pbil.univ-lyon1.fr/members/perriere/files/NS5B.fst`

Lancez le programme `SeaView` puis chargez le fichier `NS5B.fst`. Les séquences de nom `HCV_*` correspondent aux sept génotypes connus pour HCV, l'hôte de ces virus est donc l'humain. Les autres séquences sont dénommées en utilisant un formalisme de type `hôte_virus_genre` dans lequel :

- `hôte` désigne la famille des hôtes du virus (`Bovine` = bovins, `Bat` = chauves-souris, `Monkey` = singes, `Canine` = chiens, `Equine` = chevaux, `Chimpanzee` = chimpanzé).
- `virus` désigne le genre du virus (`hep` = Hepacivirus, `pegA` = Pegivirus de type A, `pegB` = Pegivirus de type B)
- `genre` désigne soit le genre de l'espèce hôte (*e.g.*, `Aotus`) soit la souche du virus (*e.g.*, `GHC100`). Cette information est donnée à titre indicatif et n'est pas nécessaire pour interpréter les résultats.

Les données présentes dans ce fichier sont des séquences nucléiques. L'analyse phylogénétique va, quant à elle, utiliser les séquences protéiques correspondantes. Utilisez l'option de `SeaView` permettant de visualiser les séquences sous forme protéique puis alignez-les au moyen du programme `Muscle`. Filtrez l'alignement au moyen de `Gblocks` en utilisant les paramètres les moins stringents.

Que pouvez-vous déduire de la qualité de l'alignement initial au regard du résultat du filtrage effectué par Gblocks ? **(1 point)**

Réponse :

Si vous avez bien utilisé les paramètres les moins stringents (*i.e.*, si vous avez coché les trois premières cases dans la boite de dialogue des options de Gblocks), vous devez constater que le filtrage n'enlève que deux régions de petite taille ce qui indique que l'alignement de départ était de bonne qualité.

Construisez les arbres au moyen de la méthode du maximum de vraisemblance sur l'alignement filtré en utilisant le modèle LG, le modèle WAG, puis le modèle JTT, tous les autres paramètres utilisés étant ceux par défaut. A chaque fois, racinez l'arbre en utilisant le groupe constitué par l'ensemble des séquences de Pegivirus.

Les arbres obtenus présentent-ils des différences du point de vue de la topologie obtenue ? Si oui, lesquelles ? Ces différences se situent-elles dans des régions de l'arbre soutenues du point de vue statistique ? Si vous deviez sélectionner l'un de trois ces arbres pour une publication éventuelle, lequel choisiriez-vous et quelle justification donneriez-vous à ce choix ? **(4 points)**

Réponses :

Les arbres que vous deviez obtenir si vous avez utilisé les bons paramètres sont donnés à la fin de ce document. Sur la base de ces arbres on constate que les différences topologiques ne portent que sur deux groupes :

— Le groupe constitué par les virus correspondants aux sept génotypes de HCV.

— Le groupe correspondant aux séquences d'Hepacivirus de bovins.

Dans les deux cas, ces différences se situent dans des régions à faible support statistique (aLRT < 0.95).

Concernant la sélection d'un arbre, étant donné que les trois modèles utilisés possèdent exactement le même nombre de paramètres, il est inutile d'effectuer un test de sélection de modèle et il suffit donc de prendre celui présentant la vraisemblance la plus élevée, c'est-à-dire l'arbre construit avec le modèle LG.

Sur la base de la topologie de l'arbre obtenu avec le modèle LG, construisez un scénario évolutif de l'apparition de HCV chez l'humain. Votre scénario est-il soutenu du point de vue statistique ? **(5 points)**

Réponses :

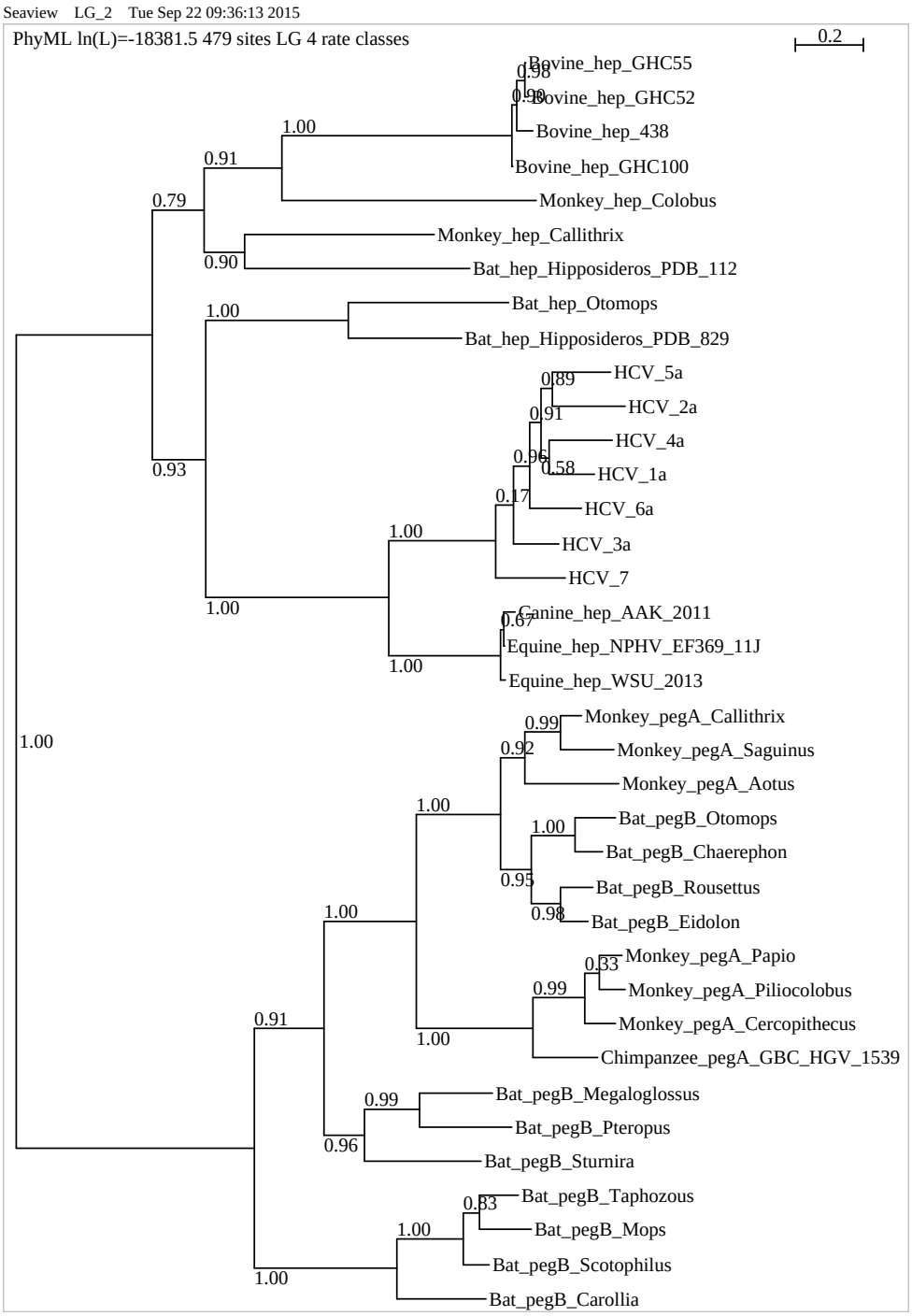
Lorsque l'on regarde le groupe le plus proche de HCV, on constate que celui-ci comprend des séquences provenant du cheval et du chien. On en déduit que HCV est probablement apparu chez l'humain suite à une contamination transmise par ces animaux domestiques. Cette première hypothèse est soutenue par les données puisque la valeur du test aLRT est égale à 1.0.

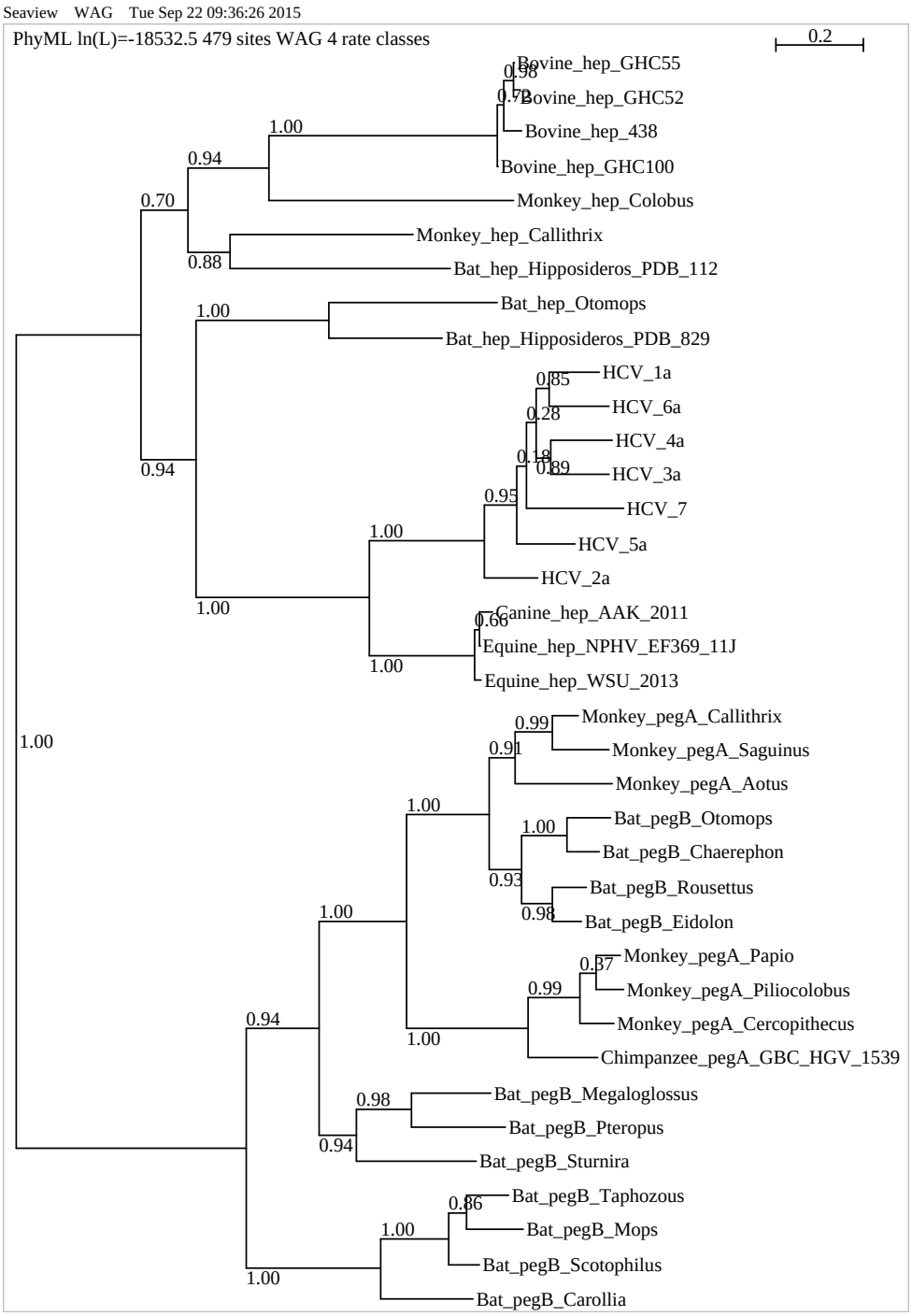
Par ailleurs, si on remonte plus profondément dans la phylogénie, on constate que le groupe frère de l'ensemble constitué par HCV + Hepacivirus canins et équins correspond à des virus de chauves-souris. Ce résultat permet d'émettre l'hypothèse supplémentaire d'une contamination interspécifique des chauves-souris au cheval et au chien. Cette deuxième hypothèse est par contre moins soutenue que la première du fait que la valeur du test aLRT est égale à 0.93 pour la branche correspondante.

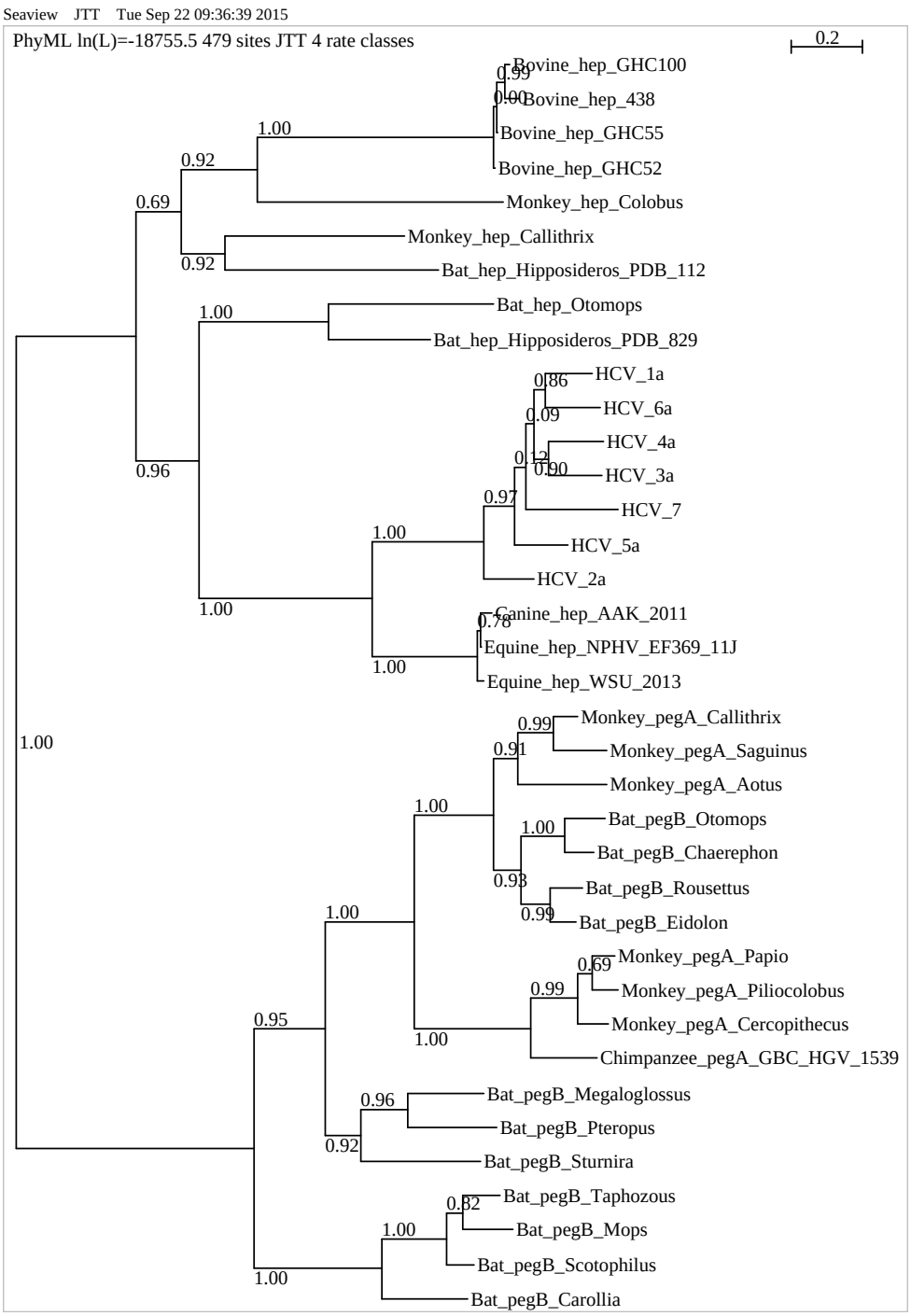
Enfin, cette phylogénie peut-elle être utilisée pour déterminer l'ordre d'apparition des différents génotypes de HCV au cours de l'histoire évolutive de ce virus ? Justifiez votre réponse. **(2 points)**

Réponses :

Du fait que, à l'exception d'une branche, les valeurs du test aLRT pour le groupe correspondant aux séquences d'HCV sont toutes inférieures à 0.95, cette étude ne permet pas déterminer l'ordre d'apparition des différents génotypes de ce virus.


Seaview LG_2 Tue Sep 22 09:36:13 2015
PhyML ln(L)=-18381.5 479 sites LG 4 rate classes
0.2
Bovine_hep_GHC55
Bovine_hep_GHC52
Bovine_hep_438
Bovine_hep_GHC100
Monkey_hep_Colobus
Monkey_hep_Callithrix
Bat_hep_Hipposideros_PDB_112
Bat_hep_Otomops
Bat_hep_Hipposideros_PDB_829
HCV_5a
HCV_2a
HCV_4a
HCV_1a
HCV_6a
HCV_3a
HCV_7
Canine_hep_AAK_2011
Equine_hep_NPHV_EF369_11J
Equine_hep_WSU_2013
Monkey_pegA_Callithrix
Monkey_pegA_Saguinus
Monkey_pegA_Aotus
Bat_pegB_Otomops
Bat_pegB_Chaerephon
Bat_pegB_Rousettus
Bat_pegB_Eidolon
Monkey_pegA_Papio
Monkey_pegA_Piliocolobus
Monkey_pegA_Cercopithecus
Chimpanzee_pegA_GBC_HGV_1539
Bat_pegB_Megaloglossus
Bat_pegB_Pteropus
Bat_pegB_Sturnira
Bat_pegB_Taphozous
Bat_pegB_Mops
Bat_pegB_Scotophilus
Bat_pegB_Carollia


Seaview WAG Tue Sep 22 09:36:26 2015
PhyML ln(L)=-18532.5 479 sites WAG 4 rate classes
0.2
Bovine_hep_GHC55
Bovine_hep_GHC52
Bovine_hep_438
Bovine_hep_GHC100
Monkey_hep_Colobus
Monkey_hep_Callithrix
Bat_hep_Hipposideros_PDB_112
Bat_hep_Otomops
Bat_hep_Hipposideros_PDB_829
HCV_1a
HCV_6a
HCV_4a
HCV_3a
HCV_7
HCV_5a
HCV_2a
Canine_hep_AAK_2011
Equine_hep_NPHV_EF369_11J
Equine_hep_WSU_2013
Monkey_pegA_Callithrix
Monkey_pegA_Saguinus
Monkey_pegA_Aotus
Bat_pegB_Otomops
Bat_pegB_Chaerephon
Bat_pegB_Rousettus
Bat_pegB_Eidolon
Monkey_pegA_Papio
Monkey_pegA_Piliocolobus
Monkey_pegA_Cercopithecus
Chimpanzee_pegA_GBC_HGV_1539
Bat_pegB_Megaloglossus
Bat_pegB_Pteropus
Bat_pegB_Sturnira
Bat_pegB_Taphozous
Bat_pegB_Mops
Bat_pegB_Scotophilus
Bat_pegB_Carollia


Seaview JTT Tue Sep 22 09:36:39 2015
PhyML ln(L)=-18755.5 479 sites JTT 4 rate classes
0.2
Bovine_hep_GHC100
Bovine_hep_438
Bovine_hep_GHC55
Bovine_hep_GHC52
Monkey_hep_Colobus
Monkey_hep_Callithrix
Bat_hep_Hipposideros_PDB_112
Bat_hep_Otomops
Bat_hep_Hipposideros_PDB_829
HCV_1a
HCV_6a
HCV_4a
HCV_3a
HCV_7
HCV_5a
HCV_2a
Canine_hep_AAK_2011
Equine_hep_NPHV_EF369_11J
Equine_hep_WSU_2013
Monkey_pegA_Callithrix
Monkey_pegA_Saguinus
Monkey_pegA_Aotus
Bat_pegB_Otomops
Bat_pegB_Chaerephon
Bat_pegB_Rousettus
Bat_pegB_Eidolon
Monkey_pegA_Papio
Monkey_pegA_Piliocolobus
Monkey_pegA_Cercopithecus
Chimpanzee_pegA_GBC_HGV_1539
Bat_pegB_Megaloglossus
Bat_pegB_Pteropus
Bat_pegB_Sturnira
Bat_pegB_Taphozous
Bat_pegB_Mops
Bat_pegB_Scotophilus
Bat_pegB_Carollia

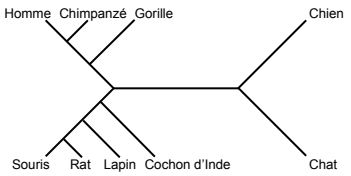
Phylogénie moléculaire
durée 2h

19 septembre 2016

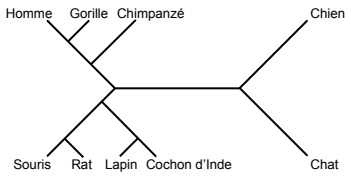
Tous documents et moyens de calculs (calculatrice, ordinateur) autorisés.
La question 3 nécessite l'utilisation du programme SeaView.

1 Question 1 (5 points) :

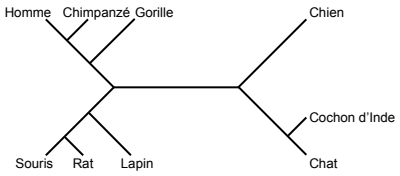
Calculez les trois distances de Robinson-Foulds *normalisées* (que l'on notera d_{AB}, d_{AC} et d_{BC}) entre les arbres A, B et C ci-dessous. Selon ces distances, quels sont les arbres les plus proches entre eux ? Le résultat obtenu vous paraît-il normal si l'on prend en compte la nature des espèces présentes dans ces trois arbres (argumentez votre réponse) ? Quelle est la limitation de la distance de Robinson-Foulds ainsi mise en évidence ?



A

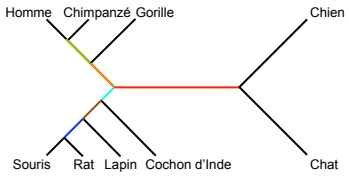


B

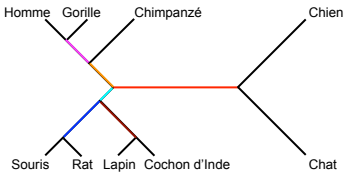


C

Réponse : Comme l'indique le calcul des distances ci-dessous, A est tout aussi proche de B que de C, bien que les différences évolutives apparaissent plus grandes (le Cochon d'Inde n'est certainement pas un carnivore, et son regroupement avec le chat est clairement artefactuel !) Il s'agit donc d'une limitation dans le sens que la méthode ne pondère pas en fonction de la distance phylogénétique *réelle* séparant les UTO.



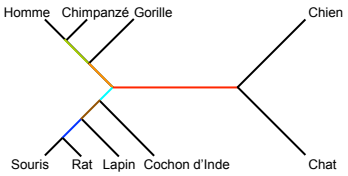
A



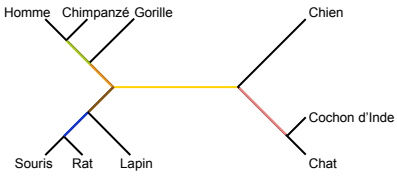
B

{Chien, Chat} vs. reste ⟷ {Chien, Chat} vs. reste
{Homme, Chimpanzé, Gorille} vs. reste ⟷ {Homme, Chimpanzé, Gorille} vs. reste
{Homme, Chimpanzé} vs. reste
{Homme, Gorille} vs. reste
{Souris, Rat, Lapin, Cochon d'Inde} vs. reste ⟷ {Souris, Rat, Lapin, Cochon d'Inde} vs. Reste
{Souris, Rat, Lapin} vs. reste
{Lapin, Cochon d'Inde} vs. reste
{Souris, Rat} vs. reste ⟷ {Souris, Rat} vs. reste

$$\left.\begin{matrix} b_t = 6 \\ b_c = 4 \end{matrix}\right\} \Rightarrow d_{AB} = 1 - 4/6 = 0{,}333$$



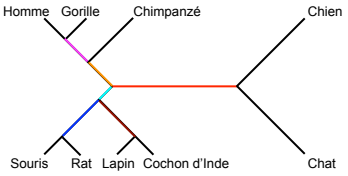
A



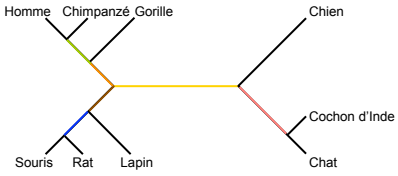
C

{Chien, Chat} vs. reste
{Chien, Chat, Cochon d'Inde} vs. Reste
{Homme, Chimpanzé, Gorille} vs. reste ⟷ {Homme, Chimpanzé, Gorille} vs. reste
{Homme, Chimpanzé} vs. reste ⟷ {Homme, Chimpanzé} vs. reste
{Souris, Rat, Lapin, Cochon d'Inde} vs. reste
{Chat, Cochon d'Inde} vs. reste
{Souris, Rat, Lapin} vs. reste ⟷ {Souris, Rat, Lapin} vs. reste
{Souris, Rat} vs. reste ⟷ {Souris, Rat} vs. reste

$$\left.\begin{matrix} b_t = 6 \\ b_c = 4 \end{matrix}\right\} \Rightarrow d_{AC} = 1 - 4/6 = 0{,}333$$



B



C

{Chien, Chat} vs. reste
{Chien, Chat, Cochon d'Inde} vs. Reste
{Homme, Chimpanzé, Gorille} vs. reste ⟷ {Homme, Chimpanzé, Gorille} vs. reste
{Homme, Gorille} vs. reste
{Homme, Chimpanzé} vs. reste
{Souris, Rat, Lapin, Cochon d'Inde} vs. Reste
{Chat, Cochon d'Inde} vs. reste
{Lapin, Cochon d'Inde} vs. reste
{Souris, Rat, Lapin} vs. reste
{Souris, Rat} vs. reste ⟷ {Souris, Rat} vs. reste

$$\left.\begin{matrix} b_t = 6 \\ b_c = 2 \end{matrix}\right\} \Rightarrow d_{BC} = 1 - 2/6 = 0{,}667$$

2 Question 2 (5 points) :

Dans le tableau ci-dessous figurent les comptages des conservations et des différentes substitutions observées entre deux séquences nucléotidiques A et B :

A/B	A	C	T	G
A	319	10	7	10
C	6	262	15	10
T	8	15	201	3
G	17	5	7	80

1. Calculez la distance de Jukes et Cantor (JC) entre ces deux séquences.
2. Calculez la distance de Felseinstein 1981 (F81) entre ces deux séquences.
3. Comparez les résultats. Comment expliquez-vous la différence observée relativement aux caractéristiques de ces deux modèles ?

Le détail des calculs devra être donné et toute réponse ne contenant que les résultats numériques sera considérée comme nulle et non avenue.

Réponse : Le nombre de sites dans l'alignement est égal au grand total du tableau, soit $\ell = 975$. Par ailleurs, le nombre de substitutions observées est égal à la somme des éléments non diagonaux, soit $n = 113$. On en déduit que la distance de Jukes et Cantor entre A et B est égale à :

$$d_{\mathrm{JC}} = -\frac{3}{4}\ln\left(1 - \frac{4}{3} \times \frac{113}{975}\right) \simeq 0.1259$$

Pour le calcul de la distance de Felsenstein 1981, il faut en plus connaître les valeurs des fréquences à l'équilibre des quatre nucléotides. Celles-ci peuvent être approximées par les fréquences dans le jeu de données considéré, soit :

$$\begin{aligned}
\pi_{\mathrm{A}} &= (319 + 10 + 7 + 10 + 319 + 6 + 8 + 17)/(2 \times 975) = 696/1950 \simeq 0.3569 \\
\pi_{\mathrm{C}} &= (6 + 262 + 15 + 10 + 10 + 262 + 15 + 5)/(2 \times 975) = 585/1950 = 0.3000 \\
\pi_{\mathrm{T}} &= (8 + 15 + 201 + 3 + 7 + 15 + 201 + 7)/(2 \times 975) = 457/1950 \simeq 0.2344 \\
\pi_{\mathrm{G}} &= 1 - \pi_{\mathrm{A}} - \pi_{\mathrm{C}} - \pi_{\mathrm{T}} = 1 - 0.3569 - 0.3000 - 0.2344 \simeq 0.1087
\end{aligned}$$

On pose $a = 1 - \pi_{\mathrm{A}}^2 - \pi_{\mathrm{C}}^2 - \pi_{\mathrm{T}}^2 - \pi_{\mathrm{G}}^2$, ce qui nous donne numériquement :

$$a = 1 - 0.3569^2 - 0.3000^2 - 0.2344^2 - 0.2344^2 \simeq 0.7159$$

On en déduit que la distance de Felsenstein 1981 entre A et B est telle que :

$$d_{\mathrm{F81}} = -0.7159 \ln\left(1 - \frac{113/975}{0.7159}\right) \simeq 0.1264$$

La (petite) différence entre les deux résultats est due à l'introduction des fréquences à l'équilibre dans le modèle. Dans le cas où les compositions en nucléotides auraient été encore plus divergentes, cette différence aurait été plus importante.

3 Question 3 (10 points) :

Les ADN polymérases γ permettent la réplication de l'ADN mitochondrial chez les Eucaryotes. Elles forment une sous-famille au sein des polymérases de type A. Quelle hypothèse pouvez-vous émettre *a priori* quant à l'origine évolutive des polymérases γ ?

Réponse (1 pt) : Du fait de leur nature mitochondriale, on peut supposer une origine bactérienne aux ADN polymérases γ.

Afin de tester cette hypothèse, téléchargez le fichier `DNAgam.fasta`. Chargez ce fichier sous SeaView puis alignez-le avec ClustalO. Filtrez l'alignement par Gblocks en utilisant tout d'abord les paramètres par défaut. Le nombre de sites sélectionnés vous semble-t-il suffisant pour réaliser une phylogénie ?

Réponse (1 pt) : Le nombre de sites sélectionnés est très faible, et donc clairement insuffisant pour reconstruire une phylogénie.

Supprimez la première selection faite par Gblocks puis réalisez une nouvelle filtration en utilisant cette fois-ci les paramètres les moins stringents de Gblocks. Combien de régions conservées reste-t-il après cette filtration ?

Réponse (1 pt) : Si vous avez bien utilisés les paramètres les moins stringents de Gblocks et que vous avez bien effectué l'alignement *avec ClustalO*, il doit vous rester 24 régions.

Sauvegardez cette nouvelle sélection de sites dans un fichier au format Fasta. Au moyen de la fonctionnalité *ad hoc* sur le site web d'`IQ-TREE` (onglet `Model Selection`) déterminez, en utilisant le critère BIC, quel est le modèle le plus approprié pour reconstruire la phylogénie de ce jeu de données ? Attention : du fait que les séquences sont des protéines, le serveur peut mettre quelques minutes avant de renvoyer ses résultats !

Réponse (2 pts) : Le modèle sélectionné est LG+I+G4 (soit LG avec les invariants et la correction par la loi Gamma). Un extrait de la sortie du serveur IQ-TREE montrant les modèles ordonnés par valeurs croissantes de BIC figure ci-dessous :

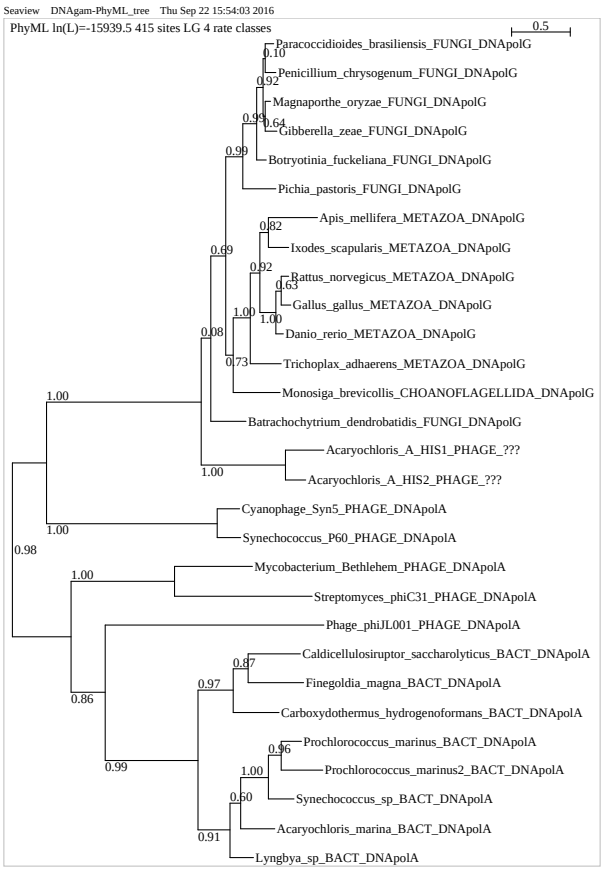
```
Best-fit model according to BIC: LG+I+G4

List of models sorted by BIC scores:

Model             LogL          AIC      w-AIC      AICc     w-AICc       BIC      w-BIC
LG+I+G4       -16365.2991  32844.5982 + 1.0000  32860.8839 + 1.0000  33080.5716 + 1.0000
LG+G4         -16381.1429  32874.2858 - 0.0000  32889.9713 - 0.0000  33106.1193 - 0.0000
Blosum62+I+G4 -16413.2037  32940.4073 - 0.0000  32956.6930 - 0.0000  33176.3807 - 0.0000
Blosum62+G4   -16431.2697  32974.5395 - 0.0000  32990.2250 - 0.0000  33206.3730 - 0.0000
LG+F+I+G4     -16371.5622  32895.1243 - 0.0000  32925.3672 - 0.0000  33209.7555 - 0.0000
...
```

Sous SeaView et au moyen de l'approche au Maximum de Vraisemblance (PhyML), reconstruisez la la phylogénie des ADN polymérases γ en utilisant le modèle déterminé à l'étape précédente. Les autres paramètres peuvent être conservés par défaut. Racinez (si nécessaire) l'arbre en utilisant les séquences d'ADN polymérases de type A (notées `DNApolA` dans le fichier). Sur la base de l'arbre obtenu, quelles conclusions pouvez-vous émettre concernant l'origine évolutive des séquences d'ADN polymérase γ ?

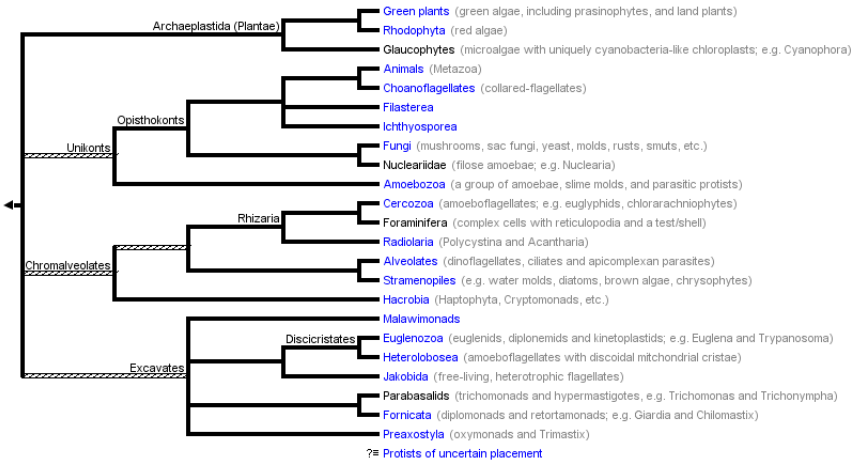
Réponse (2 pts) : L'arbre que vous devez obtenir si vous avez utilisé les bons paramètres est le suivant :



A la vue de cet arbre, une conclusion s'impose, les ADN polymérases γ sont probablement d'origine virale, plus particulièrement de bactériophages. La topologie obtenue, soutenue par de bonnes valeurs d'aLRT, suggère que ces phages, véhiculés par leurs bactéries hôtes, ont transféré le gène en question aux mitochondries des organismes dans lesquels ces bactéries ont pu pénétrer.

Une phylogénie des grands groupes Eucaryotes est disponible sur le site du `Tree of life project`. Au regard de l'arbre présenté sur cette page web, de quels grands groupes Eucaryotes l'ADN polymérase γ est-elle absente ? A 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 d'Eucaryotes ?

Réponses (3 pts) : La phylogénie des grands groupes Eucaryotes tel qu'elle figure sur le site est donnée ci-dessous. Il apparaît que l'ADN polymérase γ semble absente de quasiment tous les grands groupes, exceptés les Champignons (Fungi), les Animaux (Metazoa) et les Choanoflagellés (Choanoflagellates). Toujours au regard de cet arbre, l'apparition de ces enzymes chez les Eucaryotes se situerait donc au niveau de la branche conduisant aux Opisthochontes. Tous les autres groupes doivent donc posséder un mécanisme alternatif de réplication de l'ADN pour leurs mitochondries.


Archaeplastida (Plantae)
Green plants (green algae, including prasinophytes, and land plants)
Rhodophyta (red algae)
Glaucophytes (microalgae with uniquely cyanobacteria-like chloroplasts; e.g. Cyanophora)
Opisthokonts
Animals (Metazoa)
Choanoflagellates (collared-flagellates)
Filasterea
Ichthyosporea
Unikonts
Fungi (mushrooms, sac fungi, yeast, molds, rusts, smuts, etc.)
Nucleariidae (filose amoebae; e.g. Nuclearia)
Amoebozoa (a group of amoebae, slime molds, and parasitic protists)
Rhizaria
Cercozoa (amoeboflagellates; e.g. euglyphids, chlorarachniophytes)
Foraminifera (complex cells with reticulopodia and a test/shell)
Radiolaria (Polycystina and Acantharia)
Chromalveolates
Alveolates (dinoflagellates, ciliates and apicomplexan parasites)
Stramenopiles (e.g. water molds, diatoms, brown algae, chrysophytes)
Hacrobia (Haptophyta, Cryptomonads, etc.)
Malawimonads
Discicristates
Euglenozoa (euglenids, diplonemids and kinetoplastids; e.g. Euglena and Trypanosoma)
Heterolobosea (amoeboflagellates with discoidal mitchondrial cristae)
Excavates
Jakobida (free-living, heterotrophic flagellates)
Parabasalids (trichomonads and hypermastigotes, e.g. Trichomonas and Trichonympha)
Fornicata (diplomonads and retortamonads; e.g. Giardia and Chilomastix)
Preaxostyla (oxymonads and Trimastix)
?≡ Protists of uncertain placement

Phylogénie moléculaire
durée 2h

20 juin 2017

Tous documents et moyens de calculs (calculatrice, ordinateur) autorisés.

1 Question 1 (5 points) :

Le modèle de substitution pour les séquences nucléotidiques F84 (Felsenstein 1984) est une version améliorée du K2P (Kimura à deux Paramètres) dans lequel les fréquences à l'équilibre ont été introduites.

Donnez l'expression de la matrice **Q** des taux de substitutions instantanés en utilisant l'ordre donné ci-dessous pour les lignes et les colonnes :

$$\mathbf{Q} = \begin{array}{c} \\ \mathrm{A} \\ \mathrm{C} \\ \mathrm{T} \\ \mathrm{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{A} & \mathrm{C} & \mathrm{T} & \mathrm{G} \end{array} \\ \begin{pmatrix} -\lambda_{\mathrm{A}} & \pi_{\mathrm{C}}\beta & \pi_{\mathrm{T}}\beta & \pi_{\mathrm{G}}\alpha \\ \pi_{\mathrm{A}}\beta & -\lambda_{\mathrm{C}} & \pi_{\mathrm{T}}\alpha & \pi_{\mathrm{G}}\beta \\ \pi_{\mathrm{A}}\beta & \pi_{\mathrm{C}}\alpha & -\lambda_{\mathrm{T}} & \pi_{\mathrm{G}}\beta \\ \pi_{\mathrm{A}}\alpha & \pi_{\mathrm{C}}\beta & \pi_{\mathrm{T}}\beta & -\lambda_{\mathrm{G}} \end{pmatrix} \end{array}$$

Quelle est l'expression de λ, le taux global de substitutions sous ce modèle ?

$$\begin{aligned} \lambda = \sum_i \pi_i \lambda_i &= \pi_{\mathrm{A}}\lambda_{\mathrm{A}} + \pi_{\mathrm{C}}\lambda_{\mathrm{C}} + \pi_{\mathrm{T}}\lambda_{\mathrm{T}} + \pi_{\mathrm{G}}\lambda_{\mathrm{G}} \\ &= \pi_{\mathrm{A}}(\pi_{\mathrm{C}}\beta + \pi_{\mathrm{T}}\beta + \pi_{\mathrm{G}}\alpha) + \pi_{\mathrm{C}}(\pi_{\mathrm{A}}\beta + \pi_{\mathrm{T}}\alpha + \pi_{\mathrm{G}}\beta) \\ &\quad + \pi_{\mathrm{T}}(\pi_{\mathrm{A}}\beta + \pi_{\mathrm{C}}\alpha + \pi_{\mathrm{G}}\beta) + \pi_{\mathrm{G}}(\pi_{\mathrm{A}}\alpha + \pi_{\mathrm{C}}\beta + \pi_{\mathrm{T}}\beta) \\ &= 2\beta(\pi_{\mathrm{A}}\pi_{\mathrm{C}} + \pi_{\mathrm{A}}\pi_{\mathrm{T}} + \pi_{\mathrm{C}}\pi_{\mathrm{G}} + \pi_{\mathrm{G}}\pi_{\mathrm{T}}) + 2\alpha(\pi_{\mathrm{A}}\pi_{\mathrm{G}} + \pi_{\mathrm{C}}\pi_{\mathrm{T}}) \\ &= 2\beta\pi_{\mathrm{R}}\pi_{\mathrm{Y}} + 2\alpha(\pi_{\mathrm{A}}\pi_{\mathrm{G}} + \pi_{\mathrm{C}}\pi_{\mathrm{T}}) \end{aligned}$$

2 Question 2 (4 points) :

Soit l'alignement de cinq séquences nucléotidiques `A`, `B`, `C`, `D` et `E` telles que :

```
A   G G C C C G G C G G C T G C T G A C G C
B   T A A G C T T G G A A G C T T T G T A C
C   C C T G T A C C T G A T C C C G G C C C
D   G G A T C G C T T T C G T T A C T T T A
E   G G A T C G C T T T C G C T A C T T C A
```

```
I          *       * * * * *   * * * * *   *
```

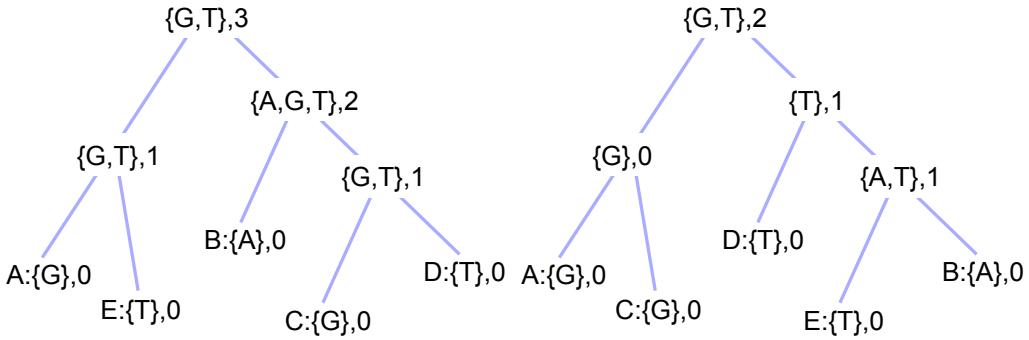
Indiquez par des * sur la ligne `I` de la figure ci-dessus quels sont les sites de cet alignement qui sont informatifs pour la parcimonie.

3 Question 3 (6 points) :

Soit le dixième site de l'alignement précédent (`GAGTT`) et les deux topologies non racinées au format Newick `((A,E),(C,D),B)` et `((A,C),D,(E,B))`.

Calculez pour chacune de ces deux topologies le nombre de substitutions correspondant à ce site au moyen de l'algorithme de Fitch. Quelle est la topologie pour laquelle ce nombre est minimal ?

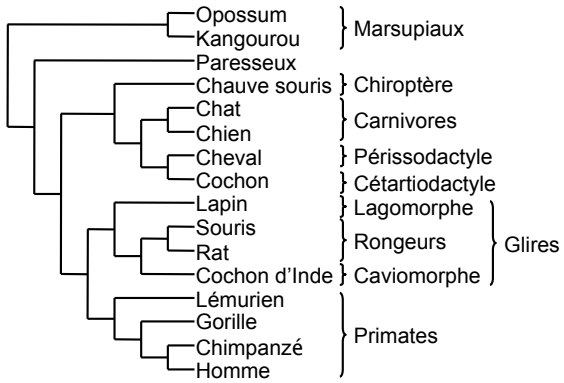
Pour calculer le nombre de substitutions, il faut d'abord raciner les arbres, la position de la racine n'ayant aucune influence sur le résultat du calcul. Dans la figure ci-dessous, le premier arbre est raciné avec le groupe {A, E} et le deuxième avec le groupe {A, C}. Le nombre de substitutions à chaque noeud inféré à partir de l'algorithme de Fitch est indiqué sur les arbres :



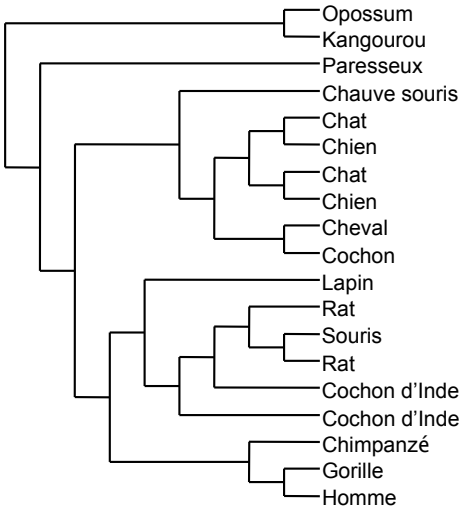
La topologie présentant le nombre minimum de substitutions pour le site considéré est celle située à droite (2 substitutions au lieu de 3).

4 Question 4 (5 points) :

En 2001, Murphy *et al.* (*Nature*, **409**:614-618) ont proposé une phylogénie pour les mammifères placentaires, l'arbre ayant été raciné avec les marsupiaux :

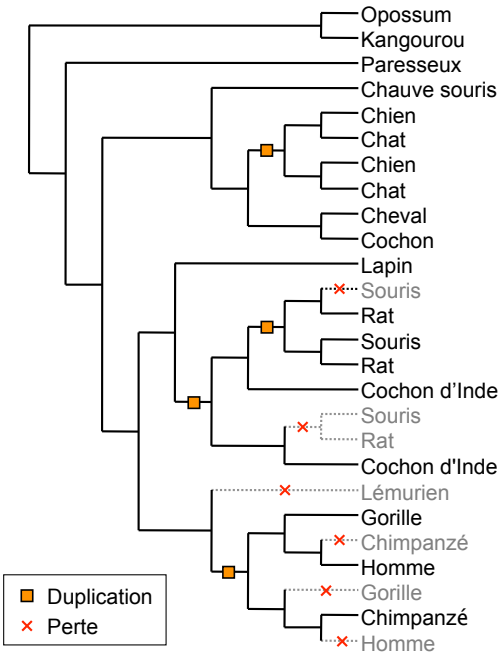


En prenant cet arbre comme référence pour les organismes considérés, interprétez en termes de duplications et/ou de pertes/absences de gènes la phylogénie suivante, réalisée sur une famille contenant plusieurs paralogues :



Pour ce faire, représentez sur un arbre l'ensemble des évènements en question. Par ailleurs, les deux gènes présents chez le rat sont-ils orthologues ou paralogues des deux gènes chez le cochon d'Inde ? Expliquez votre réponse.

L'arbre de la famille peut s'expliquer par un ensemble de quatre évènements de duplications, suivis de six pertes sélectives, comme montré ci-dessous :



Concernant les deux gènes chez le rat, ceux-ci sont **orthologues** de la copie chez le Cochon d'Inde qui est la plus proche d'eux sur l'arbre. En effet, si on remonte dans l'arbre, on voit que l'événement ancestral où se situe leur origine commune est un événement de spéciation. De la même

façon, ces deux gènes sont **paralogues** de la copie chez le Cochon d'Inde la plus éloignée (événement de duplication lorsque l'on remonte à l'origine commune).