# TP Phylogénie moléculaire

C. Brochier-Armanet, M. Gouy, V. Daubin et G. Perrière

2016-2018

## Table des matières

# 1 Recherche de séquences divergentes

1. Recherchez la séquence protéique de la globine $\alpha$ humaine (identifiant SwissProt P69905)
2. Combien d'homologues de cette séquence sont présents chez l'Humain ?
   *Précisez la stratégie de recherche que vous allez privilégier ? Justifiez vos choix.*
3. Il existe XX homologues de cette protéine chez l'Humain.
   *Pourquoi ne les avez-vous pas toutes détectées ?*
4. Refaites l'analyse réalisée en 1.2, mais en ajustant les paramètres du programme utilisé afin de prendre en compte la forte divergence des globines humaines.
   *Constatez-vous une différence par rapport à la première analyse ?*
5. Refaites l'analyse en utilisant le logiciel PSI-BLAST.
   *Constatez-vous une différence par rapport à la première analyse ?*
6. Téléchargez le fichier HSglobin_NA.fasta contenant les globines humaines.
7. Ouvrez le fichier avec SeaView (`File` → `Open`). Alignez les séquences avec Muscle (`Align` → `Alignment options` → `muscle` puis `Align` → `Align all`). Réouvrez le fichier dans une deuxième fenêtre (`File` → `New window` puis `File` → `Open`). Alignez les séquences avec ClustalΩ.
8. Comparez les alignements obtenus.
   *Constatez-vous des différences ? Si oui, à quelles positions ?*
9. Utilisez Gblocks pour éliminer les régions où l'alignement est ambigu (`Sites` → `Create set` → `Gblocks`).
   *Combien de positions sont gardées si on se base sur l'alignement obtenu avec Muscle ? Et si on se base sur l'alignement obtenu avec ClustalΩ ?*

# 2 Une bactérie de 250 millions d'années

## 2.1 Phylogénie en parcimonie

Dans un article publié dans *Nature*, Vreeland *et al.* (2000) ont annoncés qu'ils avaient isolés une bactérie âgée de 250 millions d'années à partir d'un cristal salin. La séquence de l'ARNr 16S de cette bactérie (nommée `unknown293`), alignée avec d'autres séquences provenant d'organismes actuels, est disponible dans le fichier permians.nxs.

Une chose importante à noter est que les séquences intitulées `BACSUCG.*` proviennent toutes de *Bacillus subtilis* 168 et qu'elles correspondent à différentes copies paralogues de l'ARNr 16S dans cette bactérie.

1. Sauvegardez ce fichier au format texte sur votre ordinateur.
2. Chargez-le dans SeaView (`File` → `Open`).
3. Alignez les séquences avec Muscle (`Align` → `Alignment options` → `muscle` puis `Align` → `Align all`) puis sauvez l'alignement dans un autre fichier (`File` → `Save as`) en conservant bien le format Nexus.
4. Faites la phylogénie en utilisant la parcimonie (`Trees` → `Parsimony`) avec les paramètres par défaut proposés par le programme.
   *Concernant les dits paramètres, en quoi l'option* `Randomize seq. order` *a-t-elle un interêt pour la parcimonie ? Quelles sont les informations importantes figurant dans la ligne de commentaire située en haut de la fenêtre contenant l'arbre ?*

5. Racinez l'arbre au moyen de la séquence de *L. casei* (`Re-root`), repassez en vue normale (`Full`) puis visualisez l'arbre sous forme de cladogramme, c'est-à-dire tel qu'il était présenté dans la publication d'origine.
6. Sauvez l'arbre construit dans le menu `Trees` (`File` → `Save to Trees menu`). Vous pouvez également le sauver dans un fichier (`File` → `Save (un)rooted tree`) pour une utilisation ultérieure.

## 2.2 Phylogénie en distances

1. Refaite la phylogénie sur les séquences alignées obtenues en 2.1.3 en utilisant le *Neighbor-Joining* (`Trees` → `Distance methods`). Utilisez les paramètres par défaut proposés par le programme.
2. Comparez l'arbre obtenu avec celui de parcimonie
   *Quelle est l'information importante apportée par les longueurs de branches dans le cas de l'analyse effectuée par Neighbor-Joining ? Que peut-on en conclure quant aux résultats de Vreeland et al. (2000) ?*

Vous pouvez consulter l'article de Graur et Pupko (2001) démontrant pourquoi cette bactérie est probablement d'origine beaucoup plus récente.

## 2.3 Phylogénie en bayésien

Vous allez reprendre l'analyse précédemment effectuée avec la parcimonie et le *Neighbor-Joining* en utilisant cette fois-ci une approche bayésienne.

1. Ouvrez un terminal et placez-vous dans le dossier où se trouve le fichier contenant les séquences que vous avez alignées en 2.1.
2. Lancez MrBayes en tapant la commande `mb`. Pour voir la liste des options disponibles, tapez `help`.
3. Ouvrez le fichier de séquences alignées en tapant `exe nom_du_fichier`. En tapant `help lset`, vous pourrez observer les paramètres de l'analyse par défaut. Reportez-vous à la section correspondante du manuel de MrBayes pour voir ce qu'ils signifient.
4. Lancez une analyse en tapant `mcmc`. Observez l'évolution des différentes chaînes, et estimez le temps attendu pour l'analyse.
   *Combien de chaînes sont-elles lancées par défaut ?*
5. Lorsque l'analyse est terminée (ou quand vous l'aurez interrompue faute de temps par `Ctrl-C`), résumez les résultats (`sumt` et `sump`).
   *Que signifie ce paramètre de burnin ?*
6. Observez les arbres (par exemple avec la commande `showtree` ou en utilisant SeaView).
   *Que signifient les indices compris entre 0 et 1 pour chacune des branches internes ? Les résultats produits par l'analyse bayésienne corroborent-ils ceux obtenus avec le Neighbor-Joining ?*

# 3 Première utilisation du maximum de vraisemblance

1. Téléchargez le fichier c8alphapre.nxs qui contient plusieurs séquences alignées du précurseur de la chaine $\alpha$ du composant c8 du complément.

2. Ouvrez le fichier avec SeaView (`File` → `Open`).
3. Calculez un premier arbre en maximum de vraisemblance avec la version de PhyML implémentée dans SeaView (`Trees` → `PhyML`). Comme paramètres, utilisez le modèle WAG de substitution entre acides aminés et désactivez l'option permettant de prendre en compte la variation de la vitesse évolutive entre sites.
4. Conservez ce premier arbre au moyen de l'option `File` → `Save to Trees menu` de la fenêtre dans laquelle l'arbre s'affiche.
5. Calculez un second arbre avec PhyML, cette fois-ci en prenant en compte la variation de la vitesse évolutive entre sites.
6. Comparez les vraisemblances et les longueurs des branches des deux arbres. *Lesquelles sont supérieures ? Pourquoi ?*

# 4 Origine évolutive de la thésaurine du Xénope

Les oocytes prévitellogéniques contiennent deux types de complexes ribonucléoprotéiques, appelées thésaurisomes, qui sédimentent à 7S et 42S. La première (7S) est composée d'une molécule d'ARN 5S et du facteur de transcription IIIA, la seconde (42S) est constituée de quatre sous-unités, chacune d'elle étant composée de trois molécules d'ARNt, une molécule d'ARN 5S, de trois protéines : deux thésaurines a liant les ARNt et une thésaurine b liant l'ARN 5S. La fonction des thésaurisomes est le stockage à long terme des ARN 5S et ARNt chargés. Le thésaurisome 42S serait aussi impliqué dans la synthèse des protéines en fournissant des ARNt aux ribosomes. La question de l'origine évolutive des thésaurisomes est importante car elle renvoie à celle de la formation des réserves des oocytes chez le Xénope.

Des recherches basées sur la similarité de séquences dans les bases de données ont montré que la thésaurine a était homologue au facteur d'élongation EF-1a (appelé EF-Tu chez les bactéries).

1. Téléchargez le fichier thesauORI.fasta qui contient un échantillon de séquences d'EF-1a et EF-Tu. Ouvrez le fichier avec SeaView (`File` → `Open`) puis alignez les séquences avec Muscle (`Align` → `Alignment options` → `muscle` puis `Align` → `Align all`). Sauvegardez l'alignement obtenu (`File` → `Save as`).
2. Eliminez les régions mal alignées au moyen de Gblocks (`Sites` → `Create set` → `Gblocks`) en utilisant le critère le plus stringent.
3. Construisez des arbres phylogénétiques par la méthode du *Neighbor-Joining* (modèle de Poisson, 500 réplicats de *bootstrap*) et par la méthode du maximum de parcimonie (100 réplicats de *bootstrap*).
   *Quelles informations vous apportent ces analyses quant à l'origine évolutive de la thésaurine a chez le Xénope ? Est-ce qu'une origine mitochondriale semble plausible ? Quelle hypothèse forte implique la position phylogénétique de la thésaurine a sur l'évolution des eucaryotes ?*
4. Vous allez maintenant réaliser l'analyse phylogénétique de votre jeu de données en utilisant la méthode du maximum de vraisemblance (modèle d'évolution LG, inclusion d'une loi Gamma pour tenir compte de l'hétérogénéité des vitesses d'évolution entre les sites, exploration de l'espace des arbres par NNI et SPR, tous les autres paramètres étant laissés par défaut).

*Quelles différences notables présente la topologie obtenue par maximum de vraisemblance comparativement aux topologies que vous avez obtenues avec le Neighbor-Joining et la parcimonie ? Ce résultat vous amène-t-il à réviser votre scénario ?*

5. Dans les cellules somatiques, l'insertion des ARNt chargés au niveau du site A des ribosomes requière l'action d'un facteur d'élongation spécifique, l'EF-1a.
*Quelle hypothèse fonctionnelle pourriez-vous avancer concernant l'apparition des thésaurisomes ?*

# 5 Endosymbiose mitochondriale

L'acquisition des mitochondries a été un événement clé dans l'histoire évolutive des eucaryotes. Tous les eucaryotes actuels connus possèdent une mitochondrie, à l'exception de quelques lignées, comme les Microsporidies, les Diplomonadines, les Trichomonadines, ou *Entamoeba*. Dans les années 1980, ces lignées dépourvues de mitochondries furent regroupées sous l'appellation *Archezoa*. Pour étudier cette acquisition, vous allez dans un premier temps reconstruire une phylogénie des eucaryotes basée sur l'ARNr 18S.

1. Téléchargez le le fichier euca18S.fasta contenant des séquences d'ARNr 18S eucaryotes représentatives de la diversité de ce domaine. Ouvrez le fichier avec SeaView (`File` → `Open`) puis alignez les séquences avec Muscle (`Align` → `Alignment options` → `muscle` puis `Align` → `Align all`). Eliminez les régions mal alignées avec Gblocks (`Sites` → `Create set` → `Gblocks`).
2. Réalisez l'analyse phylogénétique de ces séquences par la méthode du maximum de vraisemblance (`Trees` → `PhyML`) en utilisant les paramètres par défaut.
*Analysez la phylogénie obtenue, semble-t-elle cohérente ? En particulier, retrouvez-vous la monophylie de chacun des grands groupes eucaryotes ? Quelle est la position phylogénétique des Archezoa ? Forment-ils un groupe monophylétique ? Quelle hypothèse évolutive pouvez-vous formuler concernant l'absence de mitochondries chez Entamoeba ? Et pour les autres Archezoa ? Quelle hypothèse pouvez-vous faire concernant le moment où a eu lieu l'endosymbiose mitochondriale chez les Eucaryotes ?*

La formation des clusters [Fe/S] est une fonction primordiale pour toutes les cellules, qu'elles soient bactériennes, archéennes ou eucaryotes. En effet, de nombreuses activités cellulaires (*e.g.*, la photosynthèse, la réparation et la réplication de l'ADN, le contrôle de l'expression des gènes, etc.) dépendent de protéines porteuses de clusters [Fe/S]. Des dysfonctionnements au niveau des systèmes permettant de former ou de réparer les clusters [Fe/S] sont associés à de nombreuses maladies.

Vous allez maintenant vous intéresser à l'origine évolutive de processus chez les eucaryotes au travers de l'étude de la protéine IscS, une cystéine désulfurase qui utilise la L-Cystéine pour former de la L-Alanine et du soufre élémentaire. Ce dernier sera ensuite utilisé pour la formation ou la régénération de clusters [Fe/S].

1. Téléchargez le fichier IscS.fasta contenant un échantillon de séquences eucaryotes et procaryotes. Ouvrez le fichier avec SeaView (`File` → `Open`) puis alignez les séquences avec Muscle (`Align` → `Alignment options` → `muscle` puis `Align all`). Eliminez les régions mal alignées avec Gblocks (`Sites` → `Create set` → `Gblocks`).
2. Réalisez l'analyse phylogénétique de ces séquences avec PhyML (paramètres par défaut).

*Analysez la phylogénie obtenue. Que pouvez-vous dire de l'origine du gène codant pour la protéine IscS chez les Eucaryotes ? Observez attentivement la distribution taxonomique du gène IscS chez les Eucaryotes. Quelle information importante vous apporte-t-elle concernant l'origine des Archezoa et de l'endosymbiose mitochondriale ?*

# 6 Origine et évolution de la photosynthèse chez les eucaryotes

La capacité à réaliser la photosynthèse a été acquise par certains eucaryotes suite à l'endosymbiose chloroplastique. La photosynthèse Pour retracer l'histoire évolutive de la photosynthèse chez les eucaryotes vous allez réaliser l'analyse du gène *psbO* qui joue un rôle important au sein du photosystème II.

1. Téléchargez le fichier psbO.fst. Ouvrez le fichier avec SeaView puis alignez les séquences avec Muscle (`Align` → `Alignment options` → `Muscle` puis `Align` → `Align all`). Eliminez les régions mal alignées avec Gblocks (`Sites` → `Create set` → `Gblocks`).
2. Reconstruisez un arbre par la méthode du *Neighbor-Joining* (modèle de Poisson, 500 réplicats de *bootstrap*) et par la méthode du maximum de vraisemblance (exploration de l'espace des arbres par NNI et SPR, tous les autres paramètres par défaut).
   *Observez attentivement les groupes taxonomiques représentés dans l'arbre de psbO. N'y-a-t-il rien de surprenant ? Cherchez des renseignements sur les organismes incriminés. Quelles hypothèses (il y en a trois) pourriez-vous avancer pour expliquer votre observation ? Proposez un protocole expérimental permettant de les tester.*
3. Maintenant, comparez les phylogénies obtenues avec l'arbre de référence des Eucaryotes basée sur l'ARNr 18S que vous avez reconstruit précédemment.
   *Proposez un scénario permettant de réconcilier les deux arbres.*

# 7 Tests de comparaison de phylogénies

Rokas *et al.* (2003) ont publié une analyse phylogénomique d'un groupe de levures proches de *Saccharomyces cerevisiae*. Leur analyse portait sur un jeu de 106 gènes orthologues présent en une copie chez chacune des huit espèces considérées. Les alignements correspondant à cinq de ces 106 gènes sont disponibles à partir des liens suivants : YGL001Cnuc.fst, YOL097Cnuc.fst, YBL091Cnuc.fst, YER005Wnuc.fst et YNL155Wnuc.fst.

1. Reconstruisez les arbres de maximum de vraisemblance pour ces cinq gènes en utilisant le modèle GTR+$\Gamma_4$+I (`Trees` → `PhyML`).
   *Que pensez-vous de ces arbres ?*
2. Nous allons concaténer les cinq alignements. Pour ce faire, ouvrez le premier d'entre eux avec SeaView, puis en utilisant `File` → `Open`, ouvrez l'alignement suivant. Retournez ensuite à la première fenêtre et utilisez `File` → `Concatenate`, choisissez l'alignement que vous venez d'ouvrir et cliquez sur `OK`. Une fois tous les alignements concaténés ensemble, utilisez `File` → `Save as` pour enregistrer ce concaténat. Vérifiez que sa longueur est bien 3993 puis construisez l'arbre correspondant en utilisant

les mêmes paramètres qu'a l'étape 1.
*Comparez l'arbre du concaténat aux autres arbres obtenus, observez notamment le soutien de ses branches.*

3. Nous allons maintenant tester la significativité des différences observées entre les arbres, ceci au moyen de tests de comparaison de topologies. Comme il y a au total six tests à effectuer (un pour chaque alignement individuel plus un pour le concaténat), nous allons répartir la tâche entre les différentes personnes présentes.
   (a) Tout d'abord il faut que chaque groupe mette dans un fichier l'ensemble des six arbres qui ont été calculés, avec un arbre par ligne. Au cas où vous n'y arriveriez pas, le résultat de cette opération est disponible ici.
   (b) Allez sur le site web d'IQ-TREE. Pour effectuer le test, chaque groupe doit envoyer sur le site le fichier d'alignement qui lui a été assigné (`Input Data` → `Alignment file` → `Browse`). Concernant le modèle utilisé pour le calcul de l'arbre, nous allons bien sûr utiliser le même que celui employé précédemment, à savoir GTR+$\Gamma_4$+I (`Substitution Model Options` → `Substitution model` et `Rate heterogeneity`).
   (c) Faites ensuite dérouler le menu `Tree Topology Evaluation and Test`. A ce niveau, chaque groupe doit sélectionner au moyen du bouton `Browse` associé le fichier contenant les six arbres créé à l'étape 3.a. Laissez les options par défaut qui sont sélectionnées.
   (d) Soumettez votre analyse au moyen du bouton `SUBMIT JOB`.

   *Sur la page de résultats, notez quels sont les arbres qui ne rejettent pas l'hypothèse nulle au seuil $\alpha = 0.05$ avec le test ELW.*
   *Que pensez-vous de ces résultats ? Quelle est selon vous la source des incongruences observées par Rokas et al. ?*

# 8 Grippe de type C en Inde

Nous allons étudier l'origine évolutive de trois échantillons de virus de la grippe de type C isolés en Inde en 2011, 2012 et 2013 nommés respectivement : `C/India/P119564/2011`, `C/India/P121719/ 2012` et `C/India/P135047/2013`. Les données sont issues de l'article de Potdar *et al.* (2017). Tout d'abord, téléchargez le fichier influenza_india.fasta contenant 32 séquences du gène de l'Hémagglutinine Estérase (HE) du virus de la grippe C isolés en différents endroits de la planète entre 1947 et 2013.

1. Chargez le fichier sous SeaView et regardez les séquences protéiques correspondantes grâce à `Props` → `View as proteins`. Alignez-les en utilisant Muscle (`Align` → `Alignment options` → `Muscle` puis `Align` → `Align all`).
   *Que pensez-vous de la conservation du gène HE dans ces différentes souches ? Pensez-vous que reconstruire une phylogénie à l'aide des séquences protéiques soit une bonne idée ?*
2. Repassez en nucléotides et faites une sélection de sites conservés à l'aide de Gblocks en mode permissif (`Sites` → `Create set` → `Gblocks`). Enregistrez la sélection dans un nouveau fichier au format Nexus (`File` → `Save selection`).
3. Sélectionnez le modèle évolutif le plus adapté au jeu de donnée en utilisant le serveur d'IQTREE et le critère d'AIC corrigé (AICc).
   *Quel est le modèle sélectionné ?*

4. Nous allons maintenant utiliser la suite logicielle BEAST. Tout d'abord, lancez l'exécutable Beauti et ajoutez le fichier aligné et filtré obtenu en 2. (`File` → `Import data` → `Filter: all Files` ou en cliquant sur + en bas de la fenêtre).
5. Nous allons utiliser les dates fournies dans les noms des séquences. Pour cela, dans l'onglet `Tips`, choisissez d'utiliser les dates et cliquez sur `Guess date`. L'année étant la dernière donnée dans le nom, choisissez `Order: last`.
6. Regardez dans les différents onglets les paramètres par défaut (modèles et priors utilisés, etc). Changez le modèle utilisé si-besoin en fonction résultat obtenu en 3. *Retrouvez-vous dans BEAST tous les modèles testés par IQTREE ?*
7. Dans l'onglet `Clocks` choisissez l'horloge relaxée non-corrélée et une distribution LogNormale.
8. Enfin, fixez le nombre de pas de la chaîne à 100000, le nombre d'états intermédiaires à 100 et demandez un `log` tous les 100 pas. Générez le fichier BEAST correspondant (format `xml`).
9. Générez un deuxième fichier `xml` avec les mêmes paramètres qu'en 6. et 7. mais avec 2000000 de pas, un nombre d'états intermédiaires fixé à 2000 et un `log` tous les 2000 pas. Pensez bien à modifier le nom de sortie du fichier pour ne pas écraser celui créé en 8.
10. Lancez BEAST et chargez le fichier `.xml` créé en 8. Désélectionnez l'utilisation de la libraire BEAGLE et lancez l'analyse. En parallèle, faites de même avec le fichier `.xml` généré en 9. (*i.e.*, avec 2000000 pas). Cette analyse va prendre du temps, en attendant continuez le TP jusqu'à la question 12. avec les résultats obtenus avec 100000 pas.
    *Le fichier `.log` contient les valeurs des différents paramètres à la fréquence demandée en 8. Combien de lignes de valeurs de paramètres va-t-il contenir ? Combien d'arbres vont être générés ?*
11. Lancez le logiciel Tracer et chargez le fichier `.log` obtenu en 10. (`File` → `Import Trace file` ou en cliquant sur le +).
    *Qu'appelle-t-on la phase de burn-in ? A combien est-elle fixée par défaut dans Tracer ?*
12. Fixez le paramètre de *burn-in* à 25000 et regardez la trace.
    *Que pensez-vous des valeurs d'ESS (Effective Sample Size) ? Pensez-vous avoir atteint la convergence ? Comparez avec la trace obtenue avec 2000000 pas.*
13. Le fichier influenza_india_10_millions.log contient la trace de l'arbre obtenu avec le même jeu de données et les mêmes paramètres excepté que 10000000 de pas ont été faits.
    *Pensez-vous que la convergence a été atteinte ? A part augmenter le nombre de pas, que peut-on faire pour espérer atteindre la convergence ?*
14. Lancez maintenant le programme TreeAnnotator (distribué avec BEAST). Fixez le paramètre de *burn-in* à 25000 et la probabilité postérieure minimale à 0.5. Enregistrer l'arbre consensus dans votre répertoire de travail.
15. Ouvrez l'arbre obtenu avec le logiciel FigTree. Pour son racinement, aidez-vous de l'arbre phylogénétique influenza_C_D.pdf.
    *Selon vous, pourquoi est-il raciné par des séquences de virus de la grippe de type D ? Comment ont-elles été détectées ? Quelle(s) séquences de virus de type C allez-vous donc utiliser en groupe externe pour votre arbre ? Pourquoi ?*

16. Nous allons modifier les paramètres de FigTree pour obtenir un arbre semblable à la figure 2.A de Potdar *et al.* Dans `Node labels` choisissez d'afficher les probabilités postérieures. Dans `Scale Axis` choisissez un axe inverse, fixez les paramètres `Label spacing` et `Tick spcaing` respectivement à 10 et à 1. Enfin, dans `Time Scale`, fixez le paramètre `Scale Factor` à $-1$.
*Analysez l'arbre obtenu. Les patients dont les échantillons ont été prélevés par Potdar et al. étaient-ils infectés de la même souche de grippe C ? Selon cet arbre, combien de souches différentes de virus de la grippe de type C circulaient au Japon en 2004 ?*
17. Dans leur article, Potdar *et al.* reconstruisent également les phylogénies de six autres gènes du virus de la grippe de type C.
*Pourquoi ? A quoi doit-on faire particulièrement attention avec les phylogénies de virus ?*

# 9 Sélection dans les lysosymes de primates

Pour cette partie pratique vous allez utiliser le logiciel PAML, dont le manuel d'utilisation est disponible ici. PAML est en fait une suite logicielle comprenant plusieurs exécutables et, dans le cadre de ce TP sur la détection de la sélection, nous allons utiliser uniquement le programme CodeML. Ce programme permet d'estimer le ratio $\omega = d_N/d_S$ via des analyses par sites, par branches ou les deux. Pour toute analyse, CodeML nécessite trois fichiers pour fonctionner :

— Un fichier contenant un alignment multiple au format Phylip.
— Un arbre phylogénétique au format parenthésé Newick.
— Un fichier de contrôle dans lequel les options du programme sont spécifiées.

A noter que, dans le cas de versions de PAML possédant une interface graphique, il est possible de générer ou de charger un fichier de contrôle au travers de l'interface.

Les trois fichiers nécessaires à CodeML pour effectuer l'analyse sont disponibles à partir des liens ci-dessous :

| | |
|---|---|
| Alignement | `lysosyme.nuc` |
| Arbre | `lysosyme.tree` |
| Contrôle | `lysosyme_M0.ctl` |

## 9.1 Estimation du $\omega$ global

1. Téléchargez les fichiers nécessaires à l'analyse sur votre ordinateur.
2. Visualisez l'arbre au moyen de SeaView. Jetez également un oeil sur le fichier de contrôle en utilisant un éditeur de texte.
3. L'estimation de $\omega$ sera faite au moyen du modèle M0 qui fait l'hypothèse que la valeur est identique pour l'ensemble des branches de l'arbre. Ce modèle est spécifié au moyen des options `model = 0` et `NSsites = 0` dans le fichier de contrôle. Des explications sur les différents modèles disponibles sont donnée à la p. 35 du manuel de PAML.

4. Lancez le programme puis regardez le fichier de résultats dont le nom doit être `lysosyme_M0.mlc`.
   *Identifiez les différentes valeurs calculées. Quelle est la valeur du* $\ln L(\mathbf{t}, \kappa, \omega)$ *? Quelles sont les valeurs obtenues pour* $\kappa$ *et* $\omega$ *?*

## 9.2 Estimation du $\omega$ spécifique aux hominoïdes

L'objectif est de calculer la valeur de $\omega$ spécifique au groupe constitué par les hominoïdes et pour cela il est nécessaire de spécifier les branches de l'arbre concernées.

1. Dupliquez le fichier de contrôle `lysosyme_M0.ctl` et renommez le fichier dupliqué en `lysozyme_branch.ctl`. De la même façon, dupliquez le fichier `lysosyme.tree` et renommez le fichier dupliqué en `lysosyme_tagged.tree`.
2. Ouvrez le fichier `lysosyme_tagged.tree` avec un éditeur de texte puis ajoutez `#1` immédiatement après la première parenthèse qui suit `gibbon_Ggo` et la longueur de branche correspondante. Sauvegardez.
3. Pour vérifier que l'édition a bien marché, ouvrez `lysosyme_tagged.tree` avec SeaView et demandez à visualiser les valeurs de *bootstrap*. Le tag `#1` doit alors apparaître sur la branche conduisant au groupe des hominoïdes.
4. Ouvrez le fichier `lysozyme_branch.ctl` et éditez-le de façon à spécifier l'utilisation du modèle postulant l'existence de deux taux différents pour $\omega$ (`model = 2` et `NSsites = 0`). N'oubliez pas de modifier le nom du fichier d'arbre utilisé (`lysosyme_tagged.tree`) et de spécifier un autre nom pour le fichier de résultats (par exemple `lysosyme_branch.mlc`).
5. Lancez le programme.
6. Une fois que le programme a fini de tourner, ouvrez le fichier de résultats dans un éditeur de texte.
   *Quelle est la valeur du* $\ln L(\mathbf{t}, \kappa, \omega_0, \omega_1)$ *dans le cas de l'utilisation de ce modèle ? La valeur de* $\omega$ *spécifique de la lignée des hominoïdes est-elle différente de celle trouvée pour le reste de l'arbre ? Peut-on faire l'hypothèse que cette lignée est-elle soumise à une sélection purifiante, neutre ou positive ?*

## 9.3 Tests de significativité

Nous allons tout d'abord tester si le modèle utilisant deux valeurs différentes pour $\omega$ apporte un avantage significatif par rapport à celui avec une seule valeur. Pour cela il est nécessaire d'effectuer un test LRT.

1. *Calculez la valeur du rapport des vraisemblances* $\Lambda$ *en utilisant la formule de la Diapo. 18 du cours sur les modèles.*
2. *Sachant que le nombre de d.d.l. pour le test de* $\chi^2$ *est égal à la différence du nombre de paramètres entre les deux modèles, quel est ce nombre dans le cas présent ?*
3. *Au seuil* $\alpha = 5\%$*, le LRT est-il significatif (utilisez la table de* $\chi^2$ *donnée en annexe du cours) ? Conclusion ?*

Enfin, nous allons tester si le valeur du $\omega$ obtenue pour la lignées des hominoïdes est significativement différente de 1 (neutralité) :

1. Dupliquez le fichier `lysozyme_branch.ctl` et renommez le fichier dupliqué en `lysozyme_neutral.ctl`.

2. Ouvrez le fichier `lysozyme_neutral.ctl` et éditez-le de façon à spécifier le modèle postulant la neutralité (`fix_omega` = 1 et `omega = 1`). N'oubliez pas de spécifier un autre nom pour le fichier de résultats (par exemple `lysozyme_neutral.mlc`).
3. Lancez le programme.
4. Une fois que le programme a fini de tourner, ouvrez le fichier de résultats dans un éditeur de texte.
   *Récupérez la valeur du* $\ln L(\mathbf{t}, \kappa, \omega_0 = 1, \omega_1)$ *et effectuez le test LRT en comparant cette valeur avec celle obtenue précédemment. Conclusion ?*

*Selon vous, quelles analyses complémentaires serait-il possible de réaliser pour compléter cette étude ?*

Vous pouvez consulter l'article de Yang (1998) d'où ont été tirées les données de cet exercice.