TP Approche bayésienne / Détection de la sélection Institut Pasteur

Guy Perrière

24 novembre 2016

Table des matières

1 Une bactérie de 250 millions d'années ?

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 (notée `unknown293`), alignée avec d'autres séquences provenant d'organismes actuels est disponible dans le fichier permians.nxs.

1.1 Analyse en maximum de parcimonie

L'arbre publié dans l'article original ayant été construit en utilisant le maximum de parcimonie, nous allons tout d'abord refaire l'analyse en utilisant cette approche :

1. Sauvegardez le fichier de données au format texte sur votre ordinateur.
2. Chargez-le dans SeaView.
3. Alignez les séquences avec Muscle puis sauvez l'alignement dans un autre fichier (par exemple `permians_aln.nxs`) en conservant bien le format Nexus.
4. Refaites la phylogénie en utilisant la parcimonie avec les paramètres par défaut proposés par le programme. *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*.
6. Sauvegardez l'arbre construit, tout d'abord dans le menu `Trees`, puis dans un fichier pour une utilisation ultérieure.

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.

1.2 Analyse par approche bayésienne

Vous allez reprendre l'analyse précédemment effectuée avec la parcimonie en utilisant cette fois-ci une approche bayésienne. Le programme utilisé est MrBayes, dont le manuel d'utilisation est disponible ici.

1. Lancez MrBayes en tapant la commande `mb`. Pour voir la liste des options disponibles, tapez `help`.
2. 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 pour voir ce qu'ils signifient.
3. 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 ?
4. Lorsque l'analyse est terminée (ou quand vous l'aurez interrompue faute de temps par `Ctrl-C`), résumez les résultats (`sumt burnin=250` et `sump burnin=250`).
   Que signifie ce paramètre de burnin ?
5. 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 ?

6. 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 bayésienne ? 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 Détection de sélection dans les séquence de 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`

2.1 Estimation du ω 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 ω 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* κ *et* ω *?*

2.2 Estimation du ω spécifique aux hominoïdes

L'objectif est de calculer la valeur de ω 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 ω (`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* ω *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 ?*

2.3 Tests de significativité

Nous allons tout d'abord tester si le modèle utilisant deux valeurs différentes pour ω 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* Λ *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* χ^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* χ^2 *donnée en annexe du cours) ? Conclusion ?*

Enfin, nous allons tester si le valeur du ω 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.