Les méthodes probabilistes en phylogénie moléculaire : (2) L'approche bayésienne

Frédéric Delsuc, Emmanuel J.P. Douzery

LES MÉTHODES PROBABILISTES EN PHYLOGÉNIE MOLÉCULAIRE

(2) L'approche bayésienne

Frédéric DELSUC et Emmanuel J. P. DOUZERY

Laboratoire de Paléontologie, Phylogénie et Paléobiologie,
Institut des Sciences de l'Évolution de Montpellier (ISEM), UMR 5554-CNRS,
Université Montpellier II, Montpellier, France
E-mail : douzery@isem.univ-montp2.fr

Abstract. — Although being of relatively recent introduction into the field, the Bayesian approach of phylogenetic reconstruction has soon imposed itself as a leading method. This probabilistic method is based on the calculation of posterior probabilities of phylogenetic trees by the combination of a prior probability with the likelihood function. The use of Monte Carlo Markov chains (MCMC) to compute Bayesian posterior probabilities makes this approach particularly appealing. Indeed, these numerical methods allow the implementation of complex models of sequence evolution incorporating a large number of parameters in a reasonable computation time. The Bayesian approach thus appears as particularly promising for the future of phylogenetics. However, as a recently developed method, this approach has its still unresolved problems and the evaluation of its properties is still in its infancy. This latest aspect thus represents a major research axis for the next coming years.

Résumé. — Bien qu'étant d'introduction relativement récente en phylogénie, l'approche bayésienne de reconstruction d'arbres s'est très rapidement imposée comme une méthode incontournable. Cette méthode probabiliste est fondée sur le calcul de probabilités postérieures des arbres phylogénétiques par la combinaison d'une probabilité *a priori* avec la fonction de vraisemblance. L'utilisation de chaînes de Markov avec technique de Monte Carlo (MCMC) pour calculer les probabilités postérieures bayésiennes rend cette approche particulièrement attractive. En effet, ces techniques numériques permettent l'implémentation de modèles d'évolution des séquences complexes, incorporant un nombre élevé de paramètres en un temps de calcul raisonnable. L'approche bayésienne apparaît donc particulièrement prometteuse pour le futur de la reconstruction phylogénétique. Cependant, comme toute méthode récemment développée, cette approche comporte son lot de problèmes encore irrésolus et l'évaluation de ses propriétés n'en est qu'à ses prémices. Ce dernier aspect représente donc un axe de recherche privilégié pour les années à venir.

INTRODUCTION

Bien qu'étant l'une des méthodes d'inférence statistique parmi les plus anciennes, l'approche bayésienne n'a été que très récemment appliquée au problème de la reconstruction phylogénétique. Ainsi, en 1996, trois groupes ont indépendamment introduit l'approche bayésienne en phylogénie en formalisant le calcul des probabilités postérieures – c'est-à-dire calculées *a posteriori* – d'arbres phylogénétiques à partir de probabilités définies *a priori* (Li, 1996 ; Mau, 1996 ; Rannala & Yang, 1996). Cependant, en pratique, le calcul des probabilités postérieures des arbres phylogénétiques étant analytiquement impossible, il a été nécessaire d'implémenter des méthodes numériques connues en statistique sous le nom de chaînes de Markov avec technique de Monte Carlo (MCMC pour « Markov Chain Monte Carlo ») pour les estimer (Mau &

Biosystema 22, *Avenir et pertinence des méthodes d'analyse en phylogénie moléculaire*, 2004 : pages 75 à 86.

NEWTON, 1997 ; YANG & RANNALA, 1997 ; LARGET & SIMON, 1999). L'implémentation de ces algorithmes dans des programmes tels que BAMBE (SIMON & LARGET, 1998) et MrBAYES (HUELSENBECK & RONQUIST, 2001) permet l'analyse relativement rapide de jeux de données conséquents et confère à l'approche bayésienne un attrait indéniable ayant immédiatement séduit de nombreux utilisateurs.

LA MÉTHODE BAYÉSIENNE

Le théorème de Bayes

La notion de probabilité postérieure considérée ici est la probabilité de l'hypothèse *H* sachant les données *X* : *Pr* (*H* | *X*), ce qui diffère du maximum de vraisemblance où l'on cherche la probabilité d'observer les données *D* sous une hypothèse *X* : *Pr* (*D* | *X*). La probabilité postérieure d'une hypothèse pouvant être calculée par le théorème de Bayes, elle apparaît être fonction de la vraisemblance et de la probabilité *a priori* de cette hypothèse :

$$Pr(H \mid X) = \frac{Pr(X \mid H) \, . \, Pr(H)}{Pr(X)} \qquad (1)$$

où *Pr* (*X* | *H*) est la fonction de vraisemblance, *Pr* (*H*) la probabilité *a priori* de l'hypothèse *H* et *Pr* (*X*) la probabilité des données. La probabilité postérieure d'une hypothèse peut-être interprétée comme la probabilité que cette hypothèse soit vraie sachant les données.

L'inférence bayésienne de la phylogénie

L'inférence bayésienne de la phylogénie combine la probabilité *a priori Pr* (*T*) d'un arbre *T* avec la vraisemblance *Pr* (*X* | *T*) des données *X* sachant cet arbre *T* pour produire une distribution de probabilité postérieure *Pr* (*T* | *X*) sur les arbres en utilisant la formule de Bayes :

$$Pr(T \mid X) = \frac{Pr(X \mid T) \, . \, Pr(T)}{Pr(X)} \qquad (2)$$

La probabilité postérieure d'un arbre pouvant être interprétée comme la probabilité que cet arbre soit vrai sachant les données, les inférences sont réalisées à partir de la distribution de probabilité postérieure des différents arbres évalués au cours de l'analyse. De manière analogue à la méthode du maximum de vraisemblance, l'arbre de probabilité postérieure maximale peut ainsi être déterminé facilement. En reprenant les notations précédentes, la probabilité postérieure d'un arbre τ_i peut être formulée en utilisant le théorème de Bayes de la manière suivante :

$$f(\tau_i \mid X, \mathrm{v}, \theta) = \frac{f(X \mid \hat{\alpha}, \mathrm{v}, \theta) \, f(\hat{\alpha}_i)}{\sum_{j=1}^{B(s)} f(X \mid \tau_j, \mathrm{v}, \theta) \, f(\tau_j)} \qquad (3)$$

où $B(s)$ est le nombre d'arbres possibles pour s taxons, v représente les longueurs de branches de l'arbre, et θ décrit les paramètres du modèle d'évolution des séquences. Une probabilité *a priori* non informative est généralement utilisée pour les topologies τ telle que $f(\tau_i) = 1/B(s)$, ce qui rend les différentes topologies possibles équiprobables. La vraisemblance des données pour les différentes topologies τ_i est calculée sous un modèle d'évolution donné. Cependant, dans l'expression précédente, la probabilité postérieure est conditionnée par les données mais aussi par des paramètres tels que les longueurs de branches (v) et les paramètres du modèle d'évolution (θ). Une approche possible pour prendre en compte ces paramètres est d'utiliser des valeurs prédéfinies de ces paramètres estimées par maximum de vraisemblance ou une autre méthode. Cette approche est appelée une analyse bayésienne empirique. Une approche alternative consiste à gérer l'incertitude sur ces paramètres en les incorporant dans l'analyse elle-même. Ce type d'approche, où les inférences sont conditionnées uniquement par les données observées, est appelé une analyse bayésienne hiérarchique. L'expression de la probabilité postérieure devient alors :

$$f(\tau_i \mid X) = \frac{f(X \mid \hat{\alpha}_i) \, f(\hat{\alpha}_i)}{\sum_{j=1}^{B(s)} f(X \mid \tau_j) \, f(\tau_j)} \qquad (4)$$

dont le calcul nécessite l'intégration de la fonction de vraisemblance sur toutes les combinaisons [$B(s)$] possibles de topologies (τ), longueurs de branches (ν) et paramètres du modèle d'évolution (θ) :

$$f(X \mid \tau_i) = \iint_{\upsilon_i \theta} f(X \mid \hat{\alpha}, \upsilon_i, \theta) f(\upsilon_i) f(\theta) d\upsilon_i d\theta \quad (5)$$

Une telle intégration étant analytiquement impossible, l'estimation des probabilités postérieures est réalisée numériquement par l'approche MCMC.

LES CHAÎNES DE MARKOV AVEC TECHNIQUE DE MONTE CARLO (MCMC)

L'idée sous-jacente aux MCMC est qu'une chaîne de Markov, prenant la forme d'une marche guidée à travers l'espace multidimensionnel des paramètres, peut être utilisée pour estimer une distribution de probabilité en échantillonnant les valeurs de ces paramètres de façon périodique. L'approximation de la distribution sera d'autant plus exacte que le nombre de pas effectués par la chaîne de Markov sera élevé (Lewis, 2001).

Un exemple de marche guidée

Le principe des chaînes de Markov peut-être visualisé par analogie avec la trajectoire d'un randonneur aveugle obéissant à des règles simples de déplacement sur un paysage (Lewis, 2001). Si le randonneur se déplace sur une surface plane délimitée en faisant des pas de longueur variable et en choisissant aléatoirement une direction à chaque pas, sa trajectoire sera du type de celle représentée sur la figure 1A. Ainsi, si le randonneur est autorisé à se déplacer pendant une durée suffisamment longue, la totalité de la surface va être visitée. Imaginons maintenant que la surface à parcourir comporte deux sommets – le mont Blanc (4 808 m) et les Grandes Jorasses (4 208 m) – représentés par des densités normales bivariées plus ou moins corrélées. Le randonneur suit alors les règles supplémentaires suivantes afin de se déplacer dans les trois dimensions du paysage : (1) si le randonneur peut faire un pas en montant, il le fait toujours ; (2) si le randonneur peut faire un pas en descendant, il ne le fait pas systématiquement. Il calcule d'abord le rapport R entre la hauteur qu'il atteindrait s'il descendait et celle de sa position précédente. Il choisit ensuite un nombre aléatoire entre 0 et 1. Si ce nombre est inférieur à R il descend, sinon il reste où il est. Par exemple, si le randonneur est au sommet du Mont Blanc, et qu'il peut descendre au refuge Vallot (4 362 m), R_1 vaut 0,9. S'il peut descendre à Chamonix (1 042 m), R_2 vaut alors 0,2. Malgré son aveuglement, il a donc R_1 = 90 % de chances de rester en haute altitude tout en descendant à Vallot, contre R_2 = 20 % de chances de revenir en vallée à Chamonix. En suivant ces règles élémentaires, la trajectoire adoptée par le randonneur doit l'amener à visiter les points du paysage de manière proportionnelle à leur altitude, les points les plus élevés étant les plus fréquemment visités. Un exemple de trajectoire suivie par le randonneur, dans ce cas-là, est représentée sur la figure 1B. Les mouvements du randonneur peuvent ainsi être utilisés pour estimer les caractéristiques du volume situé sous une portion spécifiée du paysage. En éliminant les premiers pas du randonneur, semblables à une marche d'approche, et appelés l'étape d'allumage (« burn-in »), une estimation du paysage peut être facilement obtenue. Cette estimation sera d'autant plus précise que le randonneur aura marché longtemps.

L'algorithme de Metropolis-Hastings

L'application des MCMC au problème de la reconstruction phylogénétique passe par la construction d'une chaîne de Markov dont chaque pas va impliquer une modification aléatoire de la topologie et des longueurs de branches de l'arbre ainsi que des paramètres du modèle d'évolution des séquences. L'algorithme généralement utilisé pour estimer la distribution de probabilité postérieure des arbres phylogénétiques est l'algorithme de Metropolis-Hastings (MH) basé sur les travaux de Metropolis *et al*. (1953) et Hastings (1970).

Soit $\Psi = (\tau, \nu, \theta)$ un arbre, une combinaison de longueurs de branches et un ensemble de paramètres du modèle d'évolution donnés. L'algorithme MH construit une chaîne de Markov dont la distribution

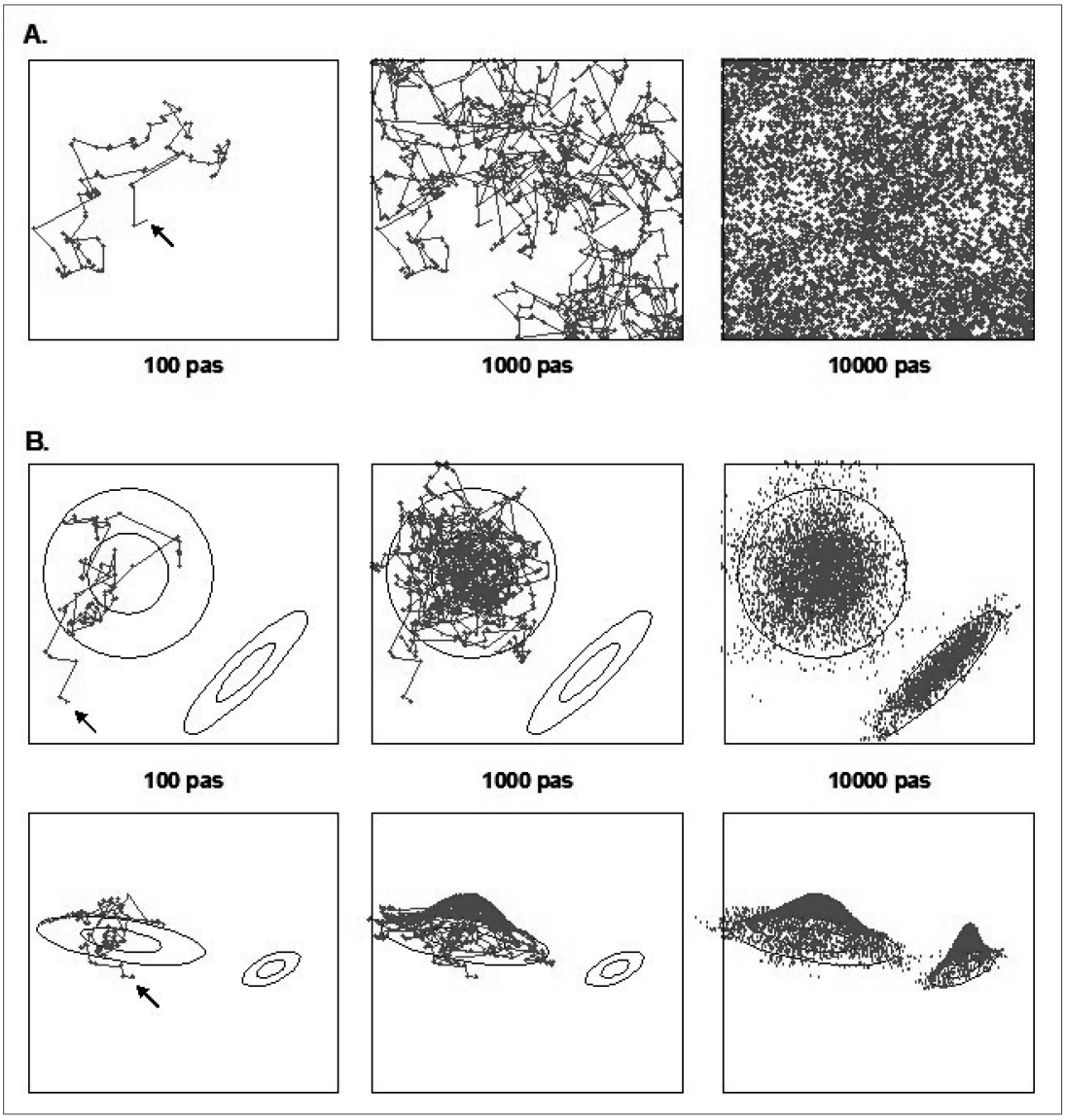

Figure 1. Illustration du principe des chaînes de Markov avec technique de Monte Carlo (MCMC).
A. Marche aléatoire sur une surface plane.
B. Marche guidée sur un paysage représenté en deux dimensions (en haut) et en trois dimensions (en bas).
Les points de départ sont indiqués par les flèches.
Cette figure a été réalisée avec le programme MCRobot 2.1 distribué par Paul O. Lewis
(http://lewis.eeb.uconn.edu/lewishome/software.html).

$$R = \min\left(1, \frac{f(\Psi'|X)}{f(\Psi|X)} \times \frac{f(\Psi|\Psi')}{f(\Psi'|\Psi)}\right)$$

$$= \min\left(1, \frac{f(X|\Psi')\, f(\Psi') / f(X)}{f(X|\Psi)\, f(\Psi) / f(X)} \times \frac{f(\Psi|\Psi')}{f(\Psi'|\Psi)}\right) \text{[Théorème de Bayes]}$$

$$= \min\left(1, \underbrace{\frac{f(X|\Psi')}{f(X|\Psi)}}_{\text{Ratio des vraisemblances}} \times \underbrace{\frac{f(\Psi')}{f(\Psi)}}_{\text{Ratio des probabilités } a \text{ priori}} \times \underbrace{\frac{f(\Psi|\Psi')}{f(\Psi'|\Psi)}}_{\text{Ratio de l'état proposé}}\right)$$

stationnaire est la probabilité postérieure (ici la probabilité conjointe de τ, ν, et θ). L'état Ψ correspondant à l'état actuel de la chaîne, un nouvel état Ψ' est proposé. La probabilité de proposer ce nouvel état Ψ' sachant l'état précédent Ψ est f (Ψ' | Ψ) et la probabilité de faire le mouvement inverse – ce qui n'est en fait jamais réalisé – est *f* (Ψ | Ψ'). Le nouvel état est accepté avec une probabilité *R* donnée ci-dessus.

Une variable aléatoire uniforme entre 0 et 1 est alors tirée. Si ce nombre est inférieur à *R*, alors l'état proposé est accepté et Ψ' = Ψ, sinon la chaîne reste dans son état originel Ψ. Ce processus est répété un très grand nombre de fois et la séquence des états visités forme une chaîne de Markov qui peut être échantillonnée de façon périodique. Les différents états par lesquels passent les MCMC sont souvent désignés sous le terme de « générations ». Les échantillons tirés de la chaîne de Markov représentent un échantillon de la distribution des probabilités postérieures. Décrits ainsi, les échantillons de la chaîne de Markov forment la densité de probabilité conjointe des topologies, des longueurs de branches, et des paramètres du modèle de substitution. La probabilité marginale des différents arbres, qui est simplement la fréquence avec laquelle ils ont été visités durant le parcours de la chaîne, donne une estimation de leur probabilité postérieure.

La convergence des chaînes de Markov

L'une des limitations potentielles de l'utilisation des MCMC pour estimer une surface est le problème de la convergence de ces chaînes. En effet, il faut s'assurer qu'un nombre suffisant de générations de MCMC a été réalisé afin d'avoir une estimation la plus exacte possible de la distribution des probabilités postérieures. Ce problème est illustré, dans la figure 1B, par le fait qu'en effectuant seulement 1 000 pas le randonneur n'a toujours pas visité le mont Blanc, second pic du paysage, car il est bloqué au sommet du premier, les Grandes Jorasses. Une façon de limiter ce problème d'optimum local, qui se traduit par un défaut de convergence de la chaîne, est de répéter l'analyse en utilisant des points de départ différents à chaque répétition de la randonnée (Haute-Savoie, Valais, Val d'Aoste…).

Les optimums locaux peuvent être potentiellement nombreux dans un paysage multidimensionnel aussi complexe que celui associé aux problèmes phylogénétiques. Ainsi, afin de gérer au mieux le problème potentiel de piégeage des MCMC dans ces optimums locaux, une variante de l'algorithme MH utilise un couplage de Metropolis des MCMC (MCMCMC ou MC^3 pour « Metropolis Coupling Markov Chain Monte Carlo »). Cet algorithme permet d'utiliser *n* MCMC simultanément dont *n* – 1 sont dites « chauffées » de manière graduelle. Ces chaînes « chaudes » – utilisant un pas plus large – permettant une exploration plus vaste de l'espace des paramètres – et sont utilisées pour guider la chaîne dite « froide » à partir de laquelle les inférences sont faites. Encore une fois l'exemple du randonneur permet de visualiser l'effet du « chauf-

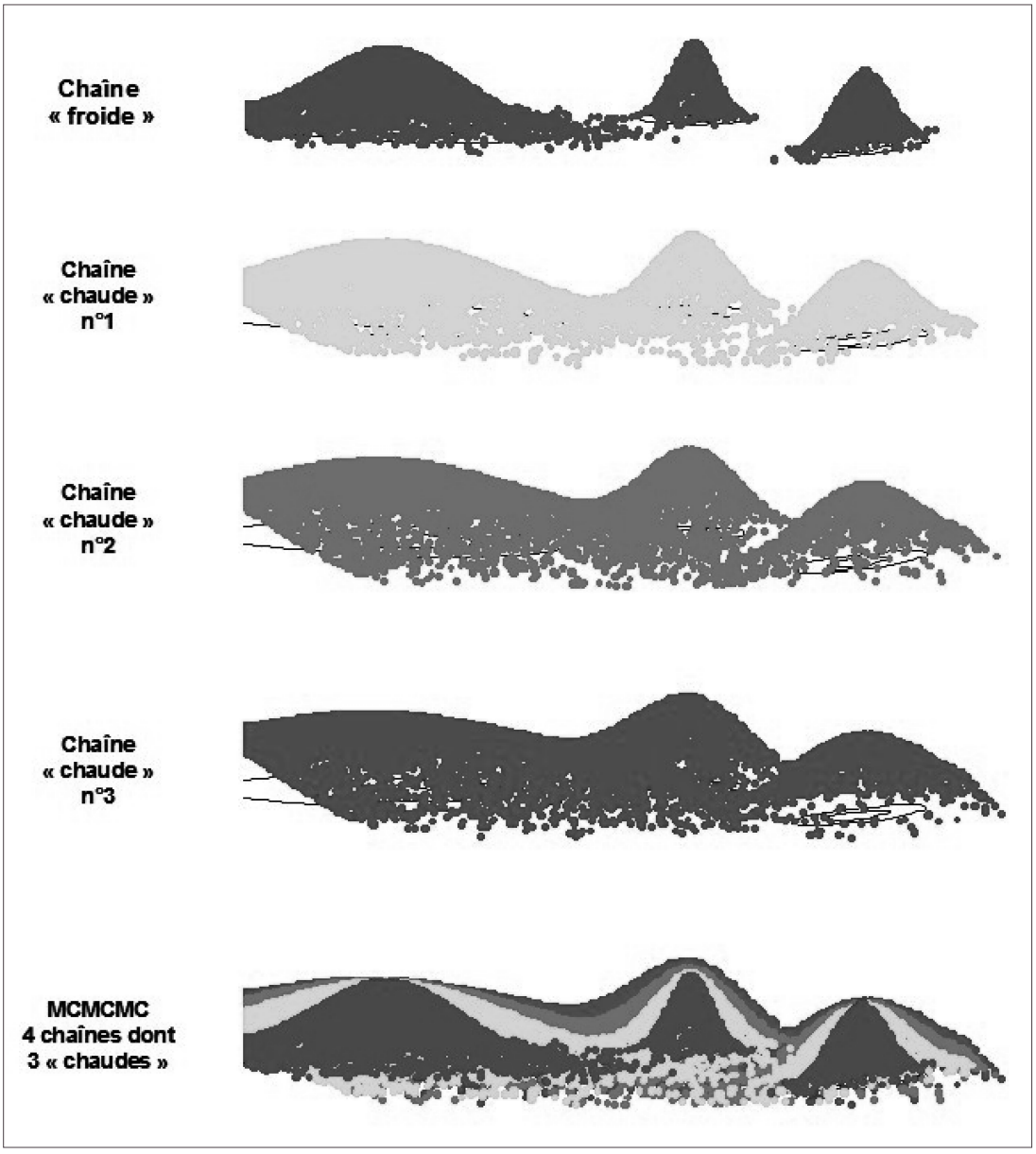

Figure 2. Illustration de l'effet du « chauffage » des chaînes de Markov sur le relief du paysage.
Les quatre chaînes ont effectué 100 000 générations chacune.
Le dernier graphique représente la superposition des quatre premiers illustrant
le parcours simultané des quatre chaînes.
Cette figure a également été réalisée avec le programme MCRobot 2.1.

fage » des chaînes sur l'exploration du paysage. L'effet recherché est l'atténuation du relief du paysage en comblant les vallées entre les collines (fig. 2). Les chaînes « chaudes » explorant le paysage plus largement que la chaîne « froide », des permutations entre les états des différentes chaînes « chaudes » et ceux de la chaîne froide sont tentées à chaque génération, ce qui peut permettre de sortir la chaîne « froide » d'un optimum local éventuel. L'utilisation de MC^3 contribue ainsi à la réduction des risques de défaut de convergence des MCMC (HUELSENBECK & RONQUIST, 2001 ; implémentation dans le logiciel MrBAYES).

UN EXEMPLE D'INFÉRENCE BAYÉSIENNE DE LA PHYLOGÉNIE

Afin d'illustrer l'approche bayésienne, un jeu de données d'ARNr 12S mitochondrial pour 10 espèces de mammifères xénarthres (tatous, fourmiliers et paresseux) a été analysé (DELSUC *et al*., 2003) en utilisant MrBAYES. Le modèle d'évolution employé utilise la matrice de substitutions HKY85 (paramètre κ décrivant le rapport transitions/transversions) et une loi gamma à huit catégories pour décrire l'hétérogénéité des taux de substitution entre sites (paramètre de forme α). La distribution des probabilités postérieures des arbres a été estimée par MC^3 en utilisant quatre chaînes simultanément (dont trois ont été « chauffées » de façon graduelle). 100 000 générations ont été réalisées pour chacune des chaînes en échantillonnant les différents paramètres toutes les 10 générations. Le degré de convergence des chaînes peut être vérifié en examinant l'évolution de la fonction de vraisemblance pendant le parcours de la chaîne « froide » afin de déterminer la période d'allumage (fig. 3A). Les générations réalisées pendant cette période sont éliminées des analyses et estimations subséquentes. Ici la vraisemblance a très vite convergé pour se stabiliser après seulement un millier de générations. De manière conservative, les 10 000 premières générations ont été éliminées (10 %) et les inférences sont alors réalisées sur les 90 000 générations suivantes. La distribution de probabilité postérieure des paramètres d'évolution du modèle – rapport transitions/tranversions (κ) et paramètre de la loi gamma (α) – peut être obtenue. À partir des 90 000/10 = 9 000 valeurs échantillonnées pour chaque descripteur du modèle, la moyenne, l'écart type, et l'intervalle de crédibilité à 95 % des paramètres peuvent alors être calculés (fig. 3B). Ainsi ce jeu de données possède un très fort rapport transitions/transversions (κ_{moyen} = 11,32), caractéristique des gènes mitochondriaux, et une très faible valeur de α (α_{moyen} = 0,17), soulignant la très forte hétérogénéité des taux de substitution entre sites dans cet ARN ribosomique. Le consensus majoritaire à 50 % des 9 000 arbres conservés après élimination des 1 000 premiers est présenté à la figure 3C. Les valeurs indiquées aux différents nœuds représentent les probabilités postérieures que le clade correspondant soit vrai.

LES PROBABILITÉS POSTÉRIEURES ET LA FIABILITÉ DES RECONSTRUCTIONS BAYÉSIENNES

La technique de re-échantillonnage du bootstrap non-paramétrique est la méthode la plus utilisée pour estimer le degré de confiance accordé aux nœuds d'un arbre phylogénétique obtenu par exemple par maximum de vraisemblance (FELSENSTEIN, 1985). Comme nous l'avons vu dans l'exemple précédent, la méthode bayésienne produit une collection d'arbres dont l'information phylogénétique peut être résumée en calculant le consensus majoritaire. La fréquence avec laquelle les différents nœuds apparaissent dans les arbres visités représente leur probabilité postérieure associée. Par rapport aux pourcentages de bootstrap, les probabilités postérieures bayésiennes présentent l'avantage d'être facilement interprétables statistiquement. Elles représentent en effet la probabilité qu'un clade donné soit vrai étant donné le modèle d'évolution, les probabilités *a priori*, et les données considérées (HUELSENBECK *et al*., 2001).

Cependant, l'approche bayésienne en phylogénie moléculaire comporte son lot de problèmes encore irrésolus (HUELSENBECK *et al*., 2002). De l'aveu même de ces auteurs, l'un des aspects les plus frustrants réside dans les différences parfois importantes observées pour un même nœud entre les

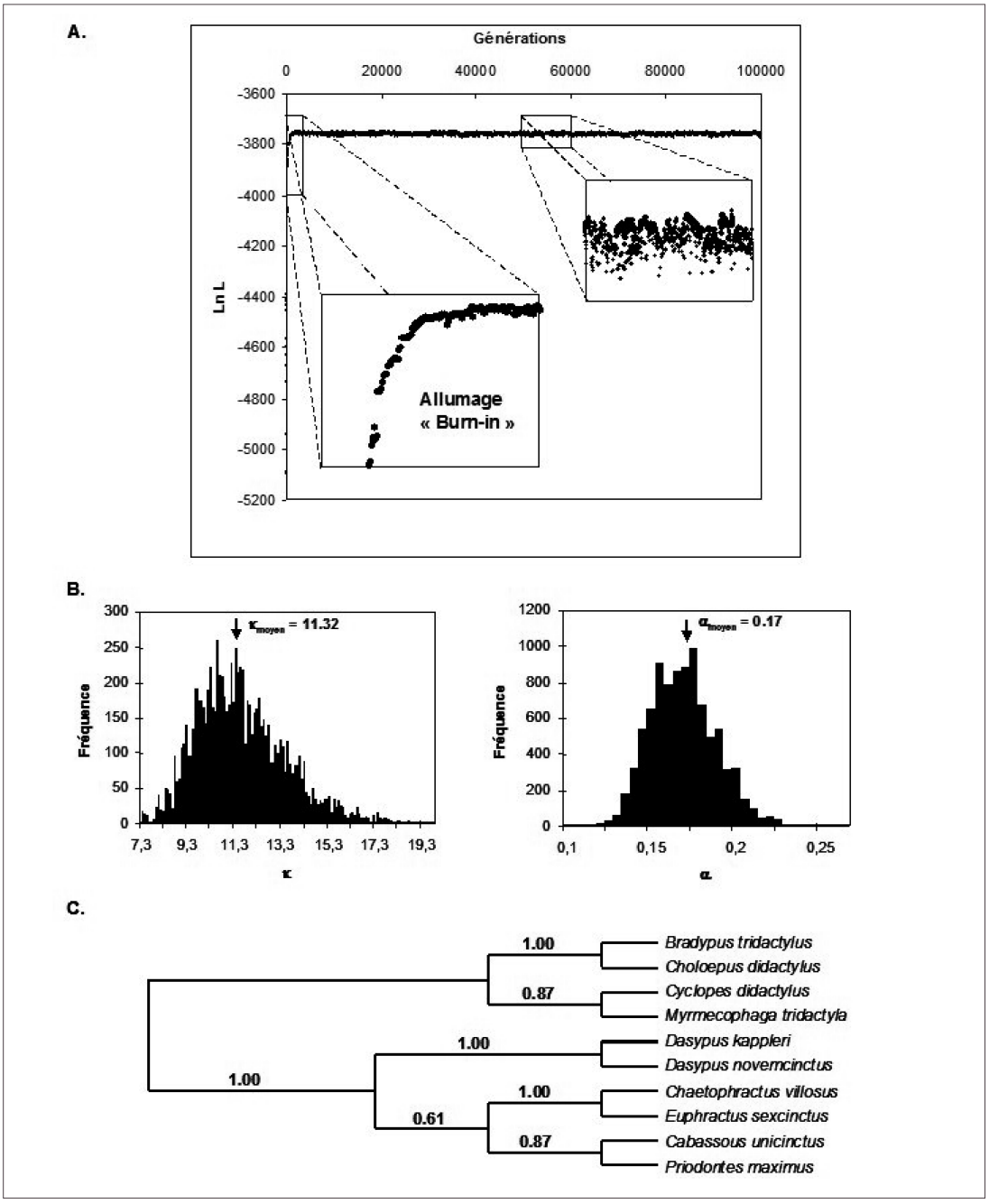

Figure 3. Exemple d'analyse bayésienne de l'ARNr 12S mitochondrial pour 10 espèces de Xénarthres utilisant un modèle HKY85 + Γ_8.

A. Convergence de la fonction de vraisemblance pour la chaîne de Markov dite « froide ».

B. Distributions des probabilités postérieures pour les deux paramètres du modèle d'évolution (κ et α).

C. Consensus majoritaire à 50 % des arbres échantillonnés après convergence des chaînes. Les valeurs aux nœuds indiquent les probabilités postérieures des clades correspondants. Données originales d'après Delsuc *et al.* (2003).

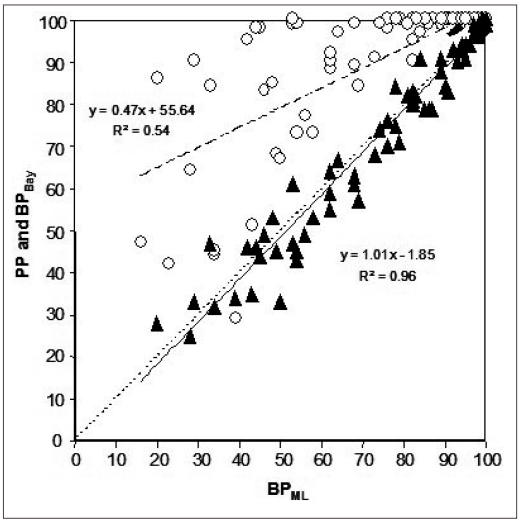

Figure 4. Corrélation linéaire entre valeurs de bootstrap non-paramétriques (BP_{ML}) et probabilités postérieures bayésiennes (PP ; valeurs multipliées par 100) d'une part [Cercles blancs], et probabilités postérieures bayésiennes obtenues après re-échantillonnage des données par bootstrap non-paramétrique (BP_{Bay}) [Triangles noirs] d'autre part. Les différents points représentent la compilation des valeurs obtenues pour les six jeux de données empiriques indépendants utilisés par Douady *et al.* (2003b). L'équation des droites de régression est donnée.

indices de confiance bayésiens – exprimés en terme de probabilités postérieures – et les valeurs de bootstrap non-paramétrique. En effet, des différences importantes entre les valeurs des deux indices ont rapidement été notées (Leache & Reeder, 2002 ; Whittingham *et al*., 2002 ; Wilcox *et al*., 2002) aboutissant parfois à des résultats conflictuels difficilement interprétables (Buckley *et al*., 2002 ; Douady *et al*., 2003a). Ainsi, la comparaison des valeurs obtenues pour divers jeux de données empiriques révèle une faible corrélation entre les deux indices et montre que les probabilités postérieures bayésiennes apparaissent quasi systématiquement plus élevées que les pourcentages de bootstrap correspondants (Douady *et al*., 2003b (fig. 4).

La compréhension de l'origine et de la nature des différences observées entre les deux types de valeurs a fait l'objet de nombreuses études récentes (Suzuki *et al*., 2002 ; Wilcox *et al*., 2002 ; Alfaro *et al*., 2003 ; Cummings *et al*., 2003 ; Erixon *et al*., 2003 ; Simmons *et al*., 2004) et des méthodes alternatives comme l'application du bootstrap non-paramétrique à la méthode bayésienne ont été proposées afin de rendre les deux indices comparables (Waddell *et al*. 2002 ; Douady *et al*., 2003b). Nous avons ainsi montré que l'application de la méthode de re-échantillonnage du bootstrap non-paramétrique à l'analyse bayésienne sur divers jeux de données empiriques aboutit à l'obtention de valeurs de soutien bayésiennes proches des valeurs de bootstrap obtenues en maximum de vraisemblance (fig. 4). Par conséquent, certains conflits topologiques révélés par la méthode bayésienne sont éliminés par l'application de la nouvelle approche proposée, en concluant plutôt à un manque de résolution (fig. 5).

Par ailleurs, les études de simulation montrent qu'au moins dans certaines circonstances, la méthode bayésienne tend à surestimer le degré de confiance accordé à un nœud donné (Suzuki *et al*., 2002 ; Alfaro *et al*., 2003 ; Cummings *et al*., 2003 ; Douady *et al*., 2003b ; Erixon *et al*., 2003). Une cause possible de ce comportement pourrait résider dans une sensibilité exacerbée de l'approche bayésienne à l'inadéquation du modèle utilisé pour décrire l'évolution des séquences (Lemmon & Moriarty, 2004). Ainsi, bien qu'étant une approche considérée comme conservative (Hillis & Bull, 1993 ; Wilcox *et al*., 2002 ; mais voir aussi Felsenstein & Kishino, 1993 ; Efron *et al*., 1996), l'utilisation des valeurs de bootstrap comme estimateurs du degré de confiance accordé aux nœuds d'un arbre phylogénétique semble être plus en phase avec ce que l'utilisateur attend d'une méthode, à savoir ne pas avoir tendance à soutenir fortement des hypothèses phylogénétiques alors qu'elles sont fausses (Douady *et al*., 2003b).

5. CONCLUSION

De par sa flexibilité permettant l'analyse de jeux de données de taille conséquente sous des modèles

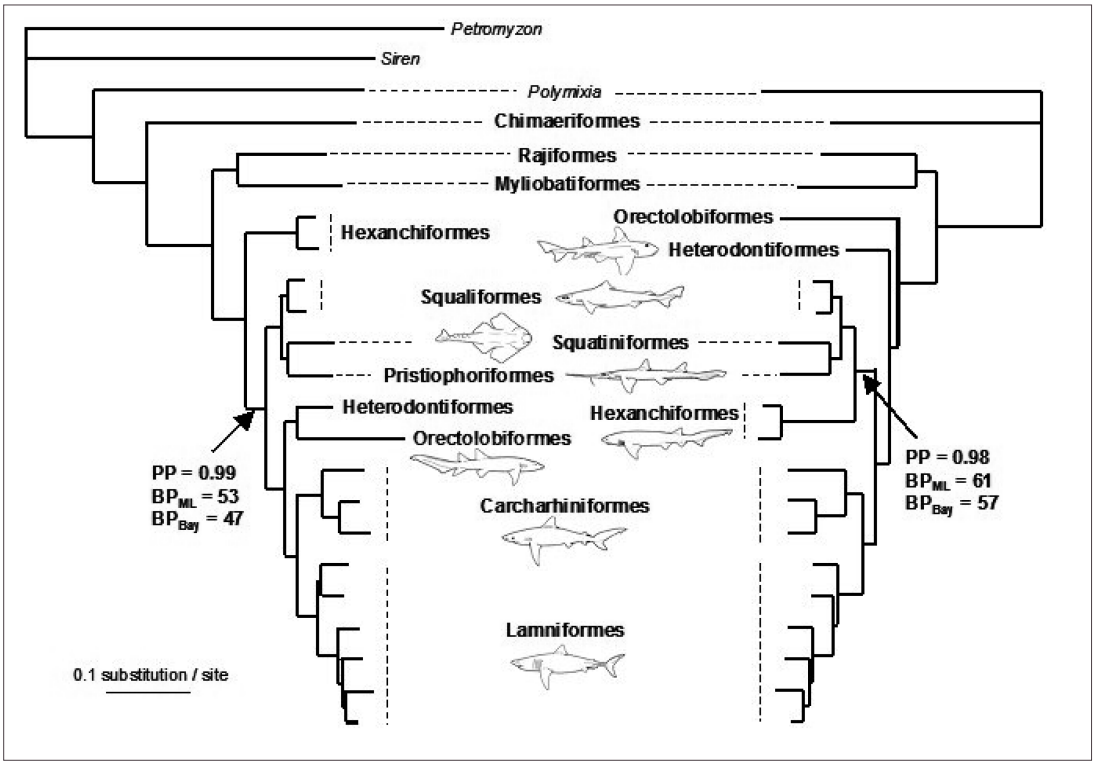

Figure 5. Illustration de l'effet du bootstrap appliqué à la méthode bayésienne lors d'un sévère conflit entre les arbres issus de l'analyse bayésienne des ARNr 12S et 16S combinés : le cas des Elasmobranches, avec des alignements de séquences différant uniquement par la composition du groupe externe. Données originales d'après Douady *et al.* (2003a). PP = probabilité postérieure bayésienne ; BP_{ML} = pourcentage de bootstrap obtenu après 100 réplications en maximum de vraisemblance (ML) ; BP_{Bay} = pourcentage de bootstrap obtenu après 100 réplications en inférence bayésienne.

d'évolution moléculaire de plus en plus complexes et incorporant un très grand nombre de paramètres (BOLLBACK, 2002 ; HUELSENBECK, 2002 ; RONQUIST & HUELSENBECK, 2003 ; HUELSENBECK *et al*., 2004 ; LARTILLOT & PHILIPPE, 2004), l'approche bayésienne apparaît comme particulièrement prometteuse. Néanmoins, certains problèmes comme la compréhension de la relation entre valeurs de bootstrap non-paramétriques et probabilités postérieures bayésiennes nécessitent clairement des études complémentaires. En particulier, l'évaluation de la sensibilité des méthodes bayésiennes à la distribution des paramètres définis *a priori* et au modèle d'évolution des séquences utilisé apparaît comme l'une des priorités. Ainsi, la comparaison des propriétés respectives des méthodes du maximum de vraisemblance et de l'approche bayésienne s'affiche comme l'un des axes de recherche privilégié des années à venir dans le domaine de la reconstruction phylogénétique à partir de données moléculaires (HOLDER & LEWIS 2003).

Remerciements. — Cet article est fondé sur une partie de la thèse de l'Université Montpellier II dirigée par EJPD, soutenue par FD le 17 décembre 2002, et intitulée : « Phylogénie moléculaire des Xénarthres (tatous,

fourmiliers et paresseux) : Application des méthodes probabilistes à la reconstruction de leur histoire évolutive au sein des Mammifères placentaires ». Les auteurs tiennent à remercier Nicolas Galtier et Hervé Philippe pour leurs commentaires éclairés en tant que membres du jury de cette thèse, ainsi qu'Alice Cibois et Jean-François Silvain pour leurs remarques pertinentes sur le manuscrit. Cet article représente la contribution ISEM 2004-029 de l'Institut des sciences de l'évolution de Montpellier (UMR 5554/CNRS).

RÉFÉRENCES BIBLIOGRAPHIQUES

ALFARO M.E., ZOLLER S. & LUTZONI F., 2003. Bayes or bootstrap ? A simulation study comparing the performance of bayesian markov chain monte carlo sampling and bootstrapping in assessing phylogenetic confidence. *Mol. Biol. Evol.*, 20 : 255-266.

BOLLBACK J.P., 2002. Bayesian model adequacy and choice in phylogenetics. *Mol. Biol. Evol.*, 19 : 1171-1180.

BUCKLEY T.R., ARENSBURGER P., SIMON C. & CHAMBERS G.K., 2002. Combined data, Bayesian phylogenetics, and the origin of the New Zealand cicada genera. *Syst. Biol.*, 51 : 4-18.

CUMMINGS M.P., HANDLEY S. A., MYERS D. S., REED D. L., ROKAS A. & WINKA K., 2003. Comparing bootstrap and posterior probability values in the four-taxon case. *Syst. Biol.*, 52 : 477-487.

DELSUC F., STANHOPE M.J. & DOUZERY E.J.P., 2003. Molecular systematics of armadillos (Xenarthra ; Dasypodidae) : contribution of maximum likelihood and Bayesian analyses of mitochondrial and nuclear genes. *Mol. Phylogenet. Evol.*, 28 : 261-275.

DOUADY C.J., DOSAY M., SHIVJI M.S. & STANHOPE M.J., 2003a. Molecular phylogenetic evidence refuting the hypothesis of Batoidea (rays and skates) as derived sharks. *Mol. Phylogenet. Evol.*, 26 : 215-221.

DOUADY C.J., DELSUC F., BOUCHER Y., DOOLITTLE W.F. & DOUZERY E.J.P., 2003b. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Mol. Biol. Evol.* 20 : 248-254.

EFRON B., HALLORAN E. & HOLMES S., 1996. Bootstrap confidence levels for phylogenetic trees. *Proc. Natl. Acad. Sci. USA*, 93 : 13429-13434.

ERIXON P., SVENNBLAD B., BRITTON T. & OXELMAN B., 2003. Reliability of bayesian posterior probabilities and bootstrap frequencies in phylogenetics. *Syst. Biol.*, 52 : 665-673.

FELSENSTEIN J., 1985. Confidence limits on phylogenies : an approach using the bootstrap. *Evolution*, 39 : 783-791.

FELSENSTEIN J. & KISHINO H, 1993. Is there something wrong with the bootstrap on phylogenies ? A reply to Hillis and Bull. *Syst. Biol.* 42 : 193-200.

HASTINGS W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57 : 97-109.

HILLIS D.M. & BULL J.J., 1993. An empirical test of boostrapping as a method for assessing confidence in phylogenetic analysis. *Syst. Biol.*, 42 : 182-192.

HOLDER M. & LEWIS P.O., 2003. Phylogeny estimation : traditional and Bayesian approaches. *Nat. Rev. Genet.*, 4 : 275-284.

HUELSENBECK J.P., 2002. Testing a covariotide model of DNA substitution. *Mol. Biol. Evol.* 19 : 698-707.

HUELSENBECK J.P. & RONQUIST F., 2001. MRBAYES : Bayesian inference of phylogenetic trees. *Bioinformatics*, 17 : 754-755.

HUELSENBECK J.P., RONQUIST F., NIELSEN R. & BOLLBACK J.P., 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, 294 : 2310-2314.

HUELSENBECK J.P., LARGET B., MILLER R.E. & RONQUIST F., 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. *Syst. Biol.*, 51 : 673-688.

HUELSENBECK J.P., LARGET B. & ALFARO M.E., 2004. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. *Mol. Biol. Evol.*, 21 : 1123-1133.

LARGET B. & SIMON D.L., 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.*, 16 : 750-759.

LARTILLOT N. & PHILIPPE H., 2004. A Bayesian mixture model for across-site rate heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol*., 21 : 1095-1109.

LEACHÉ A.D. & REEDER T.W., 2002. Molecular systematics of the Eastern Fence Lizard (*Sceloporus undulatus*) : a comparison of parsimony, likelihood, and Bayesian approaches. *Syst. Biol*., 51 : 44-68.

LEMMON A.R. & MORIARTY E.C., 2004. The importance of proper model assumption in Bayesian phylogenetics. *Syst. Biol*., 53 : 265-277.

LEWIS P.O., 2001. Phylogenetic systematics turns over a new leaf. Trends. *Ecol. Evol*., 16 : 30-37.

LI S., 1996. Phylogenetic tree construction using Markov chain Monte Carlo. Ph. D. Dissertation, Ohio State University.

MAU B., 1996. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Ph. D. Dissertation, University of Wisconsin.

MAU B. & NEWTON M., 1997. Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo. *J. Comput. Graph. Stat*. 6 : 122-131.

METROPOLIS N., ROSENBLUTH A.W., ROSENBLUTH M.N., Teller A.H. & Teller E., 1953. Equations of state calculations by fast computing machines. *J. Chemical Physics*, 21 : 1087-1091.

RANNALA B. & YANG Z., 1996. Probability distribution of molecular evolutionary trees : a new method of phylogenetic inference. *J. Mol. Evol*., 43 : 304-311.

RONQUIST F. & HUELSENBECK J.P., 2003. MRBAYES 3 : Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19 : 1572-1574.

SIMMONS M.P., PICKETT K.M. & MIYA M., 2004. How meaningful are Bayesian support values ? *Mol. Biol. Evol*., 21 : 188-199.

SIMON D.L. & LARGET B., 1998. *Bayesian Analysis in Molecular Biology and Evolution (BAMBE)*. Department of Mathematics and Computer Science, Duquesne University, Pittsburgh, Pennsylvania.

SUZUKI Y., GLAZKO G.V. & NEI M., 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. *Proc. Natl. Acad. Sci. USA*, 99 : 16168-16143.

WADDELL P.J., KISHINO H. & OTA R., 2002. Very fast algorithms for evaluating the stability of ML and Bayesian phylogenetic trees from sequences data. *Genome Informatics*, 13 : 82-92.

WHITTINGHAM L.A., SLIKAS B., WINKLER D.W. & SHELDON F.H., 2002. Phylogeny of the tree swallow genus, *TachycinetaMol. Phylogenet. Evol.* 22 : 430-441.

WILCOX T.P., ZWICKL D.J., HEATH T.A. & HILLIS D.M., 2002. Phylogenetic relationships of the dwarf boas and a comparison of Bayesian and bootstrap measures of phylogenetic support. *Mol. Phylogenet. Evol.* 25 : 361-371.

YANG Z. & RANNALA B., 1997. Bayesian phylogenetic inference using DNA sequences : a Markov chain Monte Carlo method. *Mol. Biol. Evol.* 14 : 717-724.